www.ijarp.com ISSN 2456-9992 Page: 01-09

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 07

A REVIEW ON HEALTHCARE DISEASE PREDICTION USING CLUSTERING AND CLASSIFICATION ALGORITHMS

***[1]Prof. K. M. Poornima, [2]Dr. J. Joselin**

[1]Assistant Professor, Department of Software Systems & AIML Sri Krishna Arts and Science College Coimbatore, Tamil Nadu, India.

[2]Associate Professor, Department of Computer Applications Sri Krishna Arts and Science College Coimbatore, Tamil Nadu, India.

Article Received: 20 June 2026, Article Revised: 10 July 2026, Published on: 30 July 2026

***Corresponding Author: Prof. K. M. Poornima**

Assistant Professor, Department of Software Systems & AIML Sri Krishna Arts and Science College Coimbatore, Tamil Nadu, India.

Doi: https://doi-doi.org/101555/ijarp.4476

ABSTRACT

The rapid growth of healthcare data from Electronic Health Records (EHRs), medical imaging, laboratory reports, and wearable devices has increased the need for accurate disease prediction. Machine learning and data mining techniques, particularly clustering and classification algorithms, help identify hidden patterns in medical data and improve disease diagnosis. Clustering groups patients with similar characteristics, while classification predicts disease outcomes using labelled data. Hybrid approaches that combine these techniques enhance prediction accuracy and support early diagnosis and personalized treatment. This review discusses the applications, advantages, challenges, and future scope of clustering and classification algorithms in healthcare disease prediction.

KEYWORDS: Healthcare Disease Prediction, Machine Learning, Data Mining, Clustering, Classification, Electronic Health Records (EHR), Predictive Analytics, Artificial Intelligence.

1. INTRODUCTION

Healthcare systems worldwide are rapidly adopting digital technologies that generate vast amounts of structured and unstructured medical data. Hospitals collect information from patient records, laboratory tests, radiology reports, wearable sensors, pharmacy systems, and clinical notes. This abundance of healthcare data provides an opportunity to improve disease diagnosis through predictive analytics.[1]

Disease prediction aims to identify patients who are likely to develop a disease before clinical symptoms become severe. Early prediction helps physicians recommend preventive measures, personalize treatment, and reduce healthcare costs.[2][3]

Traditional statistical methods often struggle with large, heterogeneous datasets. Machine learning (ML) overcomes many of these limitations by learning complex patterns directly from data. Supervised learning algorithms are commonly used for disease classification, whereas unsupervised learning algorithms reveal hidden patient subgroups that may not be obvious to clinicians.[8][15]

1.1. Literature Review

Recent studies have shown that machine learning techniques significantly improve healthcare disease prediction by analysing large medical datasets. Classification algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Naïve Bayes are widely used for predicting diseases like diabetes, heart disease, cancer, and kidney disease. Clustering algorithms, including K-Means, Hierarchical Clustering, and DBSCAN, help group patients with similar characteristics, improving the accuracy of predictive models. Recent research suggests that hybrid approaches combining clustering and classification provide better prediction accuracy, efficient patient stratification, and enhanced clinical decision support.[11][5]

2. Problem Statement

Healthcare datasets are often large, complex, and contain missing or noisy information, making accurate disease prediction difficult. Traditional classification methods may not perform well due to data heterogeneity and class imbalance. Therefore, there is a need for a hybrid approach that combines clustering and classification algorithms to improve prediction accuracy, support early disease diagnosis, and assist healthcare professionals in making better clinical decisions.[2][6]

3. Research Objectives

- To study clustering and classification algorithms used in healthcare disease prediction.
- To analyse the performance of different machine learning algorithms for disease prediction.
- To develop a hybrid clustering-classification framework for improved prediction accuracy.

- To compare various algorithms using standard performance metrics such as accuracy, precision, recall, and F1-score.
- To identify challenges and future research directions in healthcare disease prediction.

4. Clustering Algorithms

Clustering is an **unsupervised machine learning** technique that groups patients with similar medical characteristics without using predefined labels. It helps identify hidden patterns and patient subgroups, improving disease analysis and treatment strategies.[13]

4.1. Common Clustering Algorithms

- **K-Means:** Groups data into K clusters based on similarity.
- **Hierarchical Clustering:** Creates a tree-like structure of clusters.
- **DBSCAN:** Identifies clusters based on data density and detects outliers.
- **Fuzzy C-Means:** Allows a patient to belong to multiple clusters with different membership values. [15]

5. Classification Algorithms

Classification is a **supervised machine learning** technique that predicts disease outcomes using labelled patient data. It is widely used for diagnosing diseases and predicting patient risk.[6]

5.1. Common Classification Algorithms

- **Decision Tree:** Simple and interpretable model for classification.
- **Random Forest:** Ensemble of decision trees with high prediction accuracy.
- **Support Vector Machine (SVM):** Separates data into different classes using an optimal hyperplane.
- **Naïve Bayes:** Probability-based classifier suitable for medical diagnosis.
- **K-Nearest Neighbours (KNN):** Classifies patients based on the nearest neighbours. [12]

6. Proposed Hybrid Framework

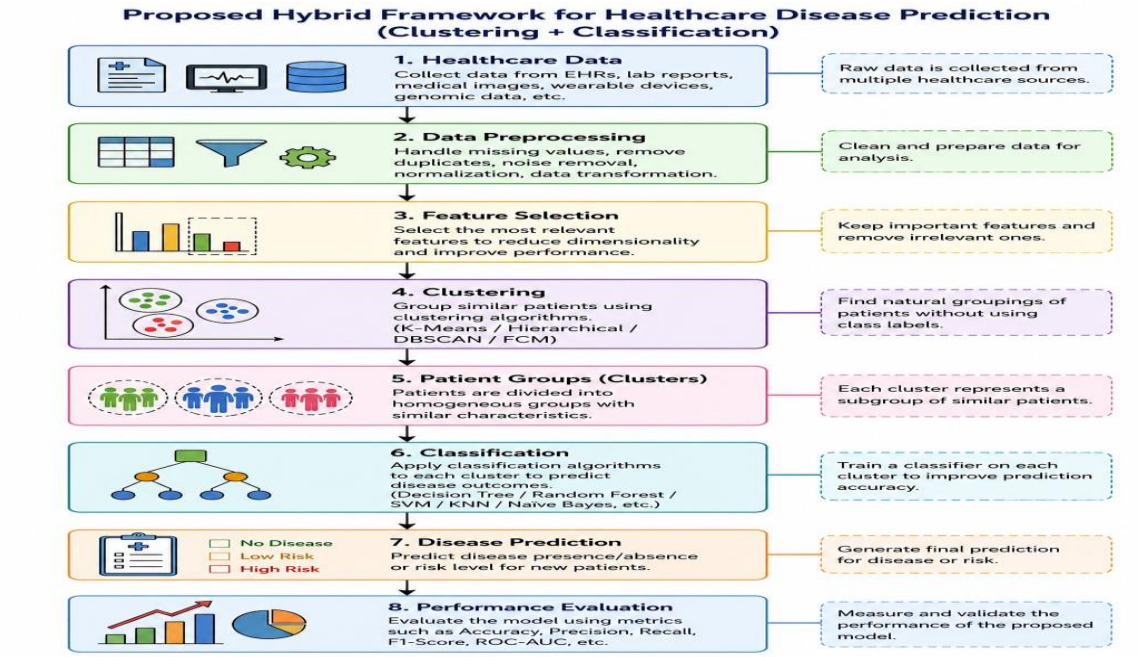

Fig 1. Framework for Healthcare Disease Prediction.

The proposed hybrid framework combines **clustering** and **classification** to improve healthcare disease prediction. First, patient data is cleaned and pre-processed. Next, clustering algorithms group similar patients based on medical characteristics. Finally, classification algorithms predict disease outcomes within each cluster, leading to improved accuracy and better clinical decision support.[9]

7. Data Description

The healthcare dataset consists of patient medical records collected from hospitals, Electronic Health Records (EHRs), or publicly available repositories such as the UCI Machine Learning Repository. It contains demographic information, clinical measurements, laboratory test results, and disease labels used for prediction.[4]

7.1. Sample Attributes

- Age
- Gender
- Blood Pressure
- Blood Glucose Level
- Cholesterol

- Body Mass Index (BMI)
- Heart Rate
- Medical History
- Disease Status (Healthy/Diseased)

8. Data Preprocessing

Data preprocessing improves the quality of the dataset before applying machine learning algorithms. Data preprocessing is a crucial step in healthcare disease prediction that improves the quality and reliability of medical data before applying clustering and classification algorithms. It removes errors, handles inconsistencies, and prepares the dataset for accurate analysis. [7]

8.1. Benefits of Data Preprocessing

- Improves prediction accuracy.
- Reduces computational complexity.
- Enhances model reliability and robustness.
- Minimizes the impact of missing and noisy data.
- Increases the efficiency of clustering and classification algorithms.[13]

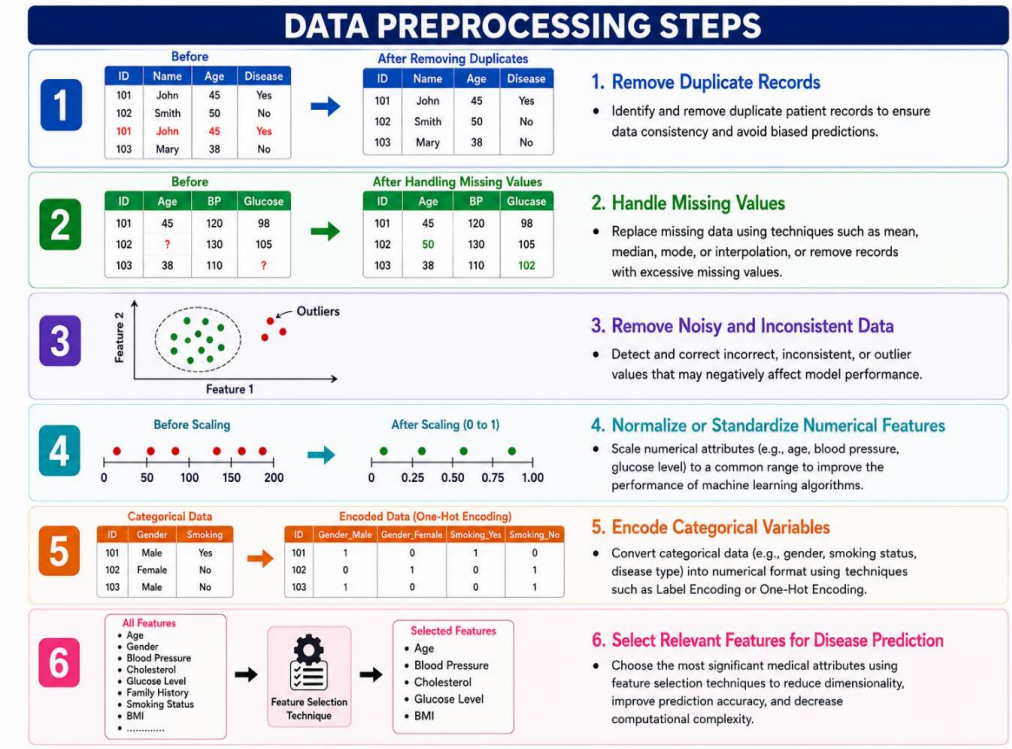

Fig 2. Steps involved in Data Preprocessing.

9. Experimental Methodology

The proposed methodology combines clustering and classification algorithms for disease prediction.[1]

9.1. Methodology Steps

- Collect healthcare dataset.
- Perform data preprocessing.
- Select important features.
- Apply a clustering algorithm (e.g., K-Means) to group similar patients.
- Train a classification algorithm (e.g., Random Forest or SVM) on each cluster.
- Predict disease outcomes.
- Evaluate model performance using standard metrics.[14]

10. Performance Evaluation

The performance of the proposed model is evaluated using the following metrics.

Metric	Formula	Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Percentage of correctly predicted cases among all predictions.
Precision	TP / (TP + FP)	Measures the proportion of correctly predicted positive cases among all predicted positive cases.
Recall (Sensitivity)	TP / (TP + FN)	Measures the ability of the model to correctly identify actual disease cases.
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean of Precision and Recall, providing a balance between them.
ROC-AUC	Area Under the ROC Curve	Evaluates the model's ability to distinguish between diseased and healthy patients across different thresholds.

11. Comparative Analysis

The comparative analysis indicates that the Hybrid (Clustering + Classification) approach achieves the highest prediction accuracy (95%) by effectively grouping similar patients before classification. Random Forest (90%) and Support Vector Machine (88%) for healthcare datasets. K-Nearest Neighbors (78%) and Decision Tree (75%) are simpler to implement but offer comparatively lower accuracy and may face issues such as slow prediction or overfitting. [3][6]

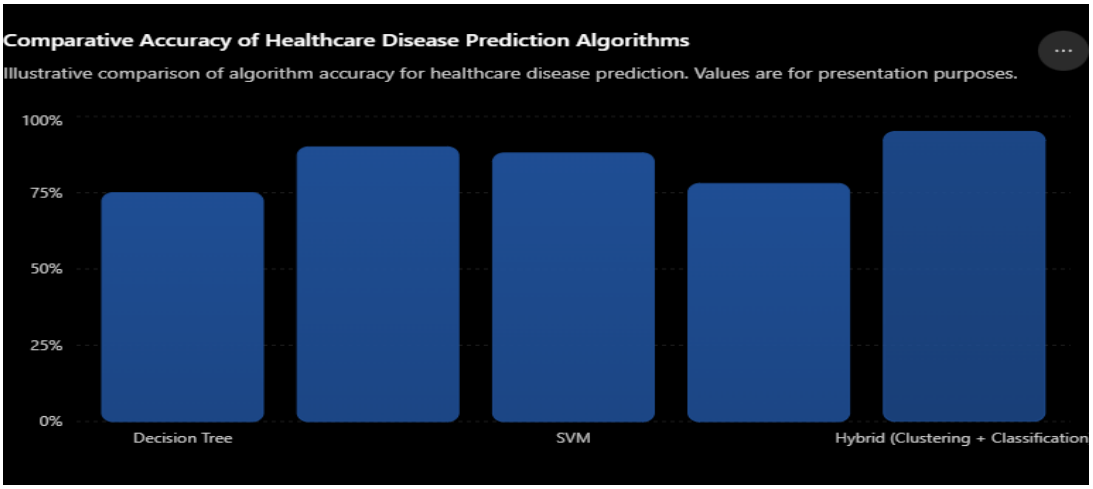

Fig 3. Comparative Analysis of various algorithms.

The review indicates that machine learning techniques have significantly improved healthcare disease prediction by analysing large and complex medical datasets. Classification algorithms such as Random Forest, SVM, and Decision Tree provide reliable prediction results, while clustering algorithms help identify hidden patient groups with similar clinical characteristics. Hybrid approaches that combine clustering and classification achieve better prediction accuracy, improved patient stratification, and enhanced clinical decision support compared to individual algorithms.[14]

12. CHALLENGES AND LIMITATIONS

Despite their effectiveness, healthcare disease prediction models face several challenges, including poor data quality, missing values, class imbalance, high-dimensional datasets, and patient privacy concerns. In addition, some machine learning models are difficult to interpret, making them less suitable for clinical decision-making. Hybrid models may also require higher computational resources and more complex implementation.[2][7]

13. CONCLUSION

Healthcare disease prediction using clustering and classification algorithms has become an important application of data mining and machine learning. Clustering helps identify similar patient groups, while classification accurately predicts disease outcomes. Hybrid approaches provide better prediction performance than individual algorithms by improving patient grouping and reducing data complexity. Although challenges such as data quality, interpretability, and privacy remain, continued advancements in artificial intelligence are expected to enable more accurate, efficient, and reliable disease prediction systems for intelligent healthcare.

REFERENCE.

1. Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The use of machine learning for analyzing real-world data in disease prediction and management: Systematic review. JMIR Medical Informatics.
2. Kumar, R., Garg, S., Kaur, R., Menon, S. V., et al. (2025). A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. Frontiers in Artificial Intelligence.
3. Ghule, S. S., & Patil, B. A. (2025). AI-driven prediction of health diseases: Applications, challenges, and future prospects. International Journal of Latest Technology in Engineering. Management & Applied Science.
4. Kothinti, R. R. (2024). Artificial intelligence in disease prediction: Transforming early diagnosis and preventive healthcare. International Journal of Novel Research and Development, 9(5), I330–I340.
5. Bansal, A., Sinha, A., Dhoundiyal, S., & Joshi, U. (2024). Heart disease prediction using machine learning. International Journal of Research Publication and Reviews, 5(2), 3013–3019.
6. Ramesh, B., Srinivas, G., Reddy, P. R. P., Rasool, M. D. H., Rawat, D., & Sundaray, M. (2023). Feasible prediction of multiple diseases using machine learning. E3S Web of Conferences, 430, 01051.
7. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2023). A guide to deep learning in healthcare. Nature Medicine, 29(1), 12–24.
8. Radha, R., & Gopalakrishnan, R. (2022). Use of machine learning and deep learning in healthcare—A review on disease prediction system. In P. Singh (Ed.), Fundamentals and Methods of Machine and Deep Learning. Wiley.
9. Ahsan, M. M., & Siddique, Z. (2021). Machine learning based disease diagnosis: A comprehensive review.
10. Ahsan, M. M., & Siddique, Z. (2021). Machine learning-based heart disease diagnosis: A systematic literature review.
11. Sadhasivam, J., Senthil, J., Ganesh, R. M., & Chellapan, N. (2021). Liver disease prediction using machine learning classification. Webology, 18(2), 441–452.
12. Shinde, S. A., & Rajeswari, P. (2018). Intelligent health risk prediction systems using machine learning: A review. International Journal of Engineering and Technology, 7(3), 1019–1023.

13. Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(1), 1–16.
14. Alanazi, H. O., Abdullah, A. H., & Qureshi, K. N. (2017). A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. Journal of Medical Systems, 41(4), Article 69.
15. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.