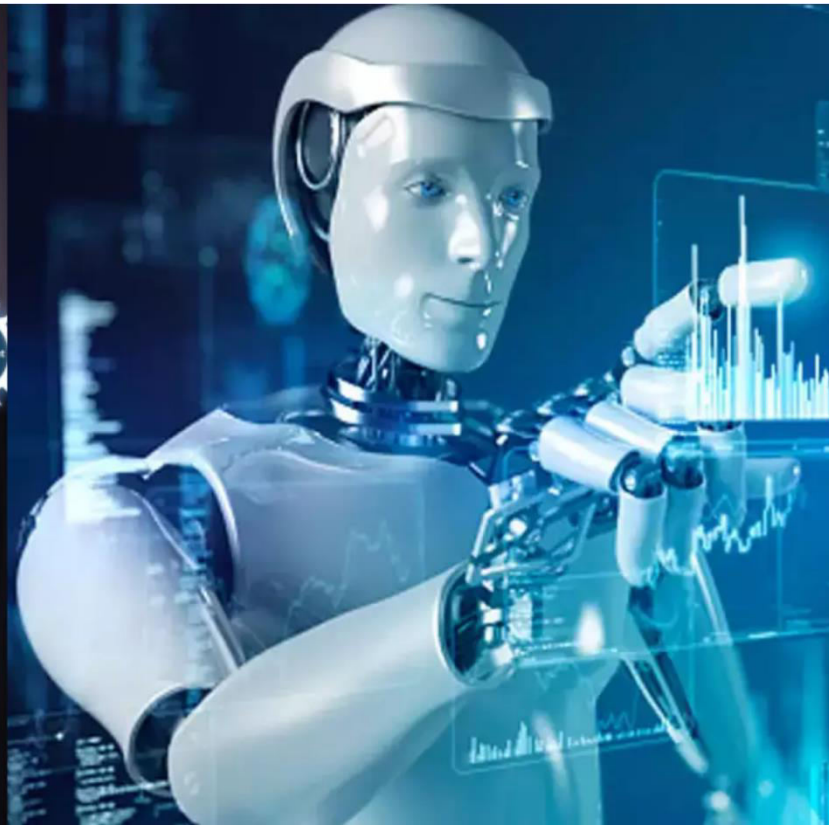
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502102|

An AI-Based Early Disease Prediction System Using Machine Learning Techniques

Nishant N. Zanzane[1], Rutuja R. Pagale[2], Shivani D. Dhumal3, Dr. Gajanan R. Joshi[4]

P.G. Research Student, Department of Computer Science, Vidya Pratishthan's Arts Science and Commerce College, Baramati(MH), Pune, India[1,2,3]

Assistant Professor, Department of Computer Science, Vidya Pratishthan's Arts Science and Commerce College, Baramati(MH), Pune, India[4]

ABSTRACT: Early and accurate disease prediction plays a vital role in improving healthcare outcomes and reducing medical costs. With the rapid growth of healthcare data, traditional diagnostic methods face limitations in analyzing large-scale and complex datasets. This research proposes an AI-based early disease prediction framework using machine learning techniques to assist clinicians in timely diagnosis. The system employs supervised learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Clinical data is preprocessed using normalization, missing value handling, and feature selection to enhance model performance. Experimental analysis demonstrates that ensemble-based models achieve superior accuracy, precision, recall, and F1-score compared to conventional classifiers. The proposed system can serve as an effective clinical decision support tool for early disease detection.

KEYWORDS: Artificial Intelligence, Disease Prediction, Machine Learning, Healthcare Analytics, Classification.

I. INTRODUCTION

The rapid advancement of healthcare technologies has led to the generation of large volumes of clinical data through electronic health records, diagnostic tests, medical imaging, and wearable devices. Despite the availability of such data, early disease detection remains a major challenge in clinical practice due to delayed diagnosis, complex symptom patterns, and limitations of traditional diagnostic methods. Early identification of diseases is critical, as timely intervention can significantly improve patient outcomes, reduce mortality rates, and lower overall healthcare costs. Despite significant progress in AI and machine learning–based disease detection, most existing systems are designed to predict only a single disease [1].

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools for analyzing complex and high-dimensional healthcare data. Early prediction of diseases in humans plays a crucial role in effective treatment and clinical management [2]. Unlike conventional statistical approaches, machine learning models can automatically learn hidden patterns and relationships from historical clinical data, enabling accurate prediction of disease occurrence at an early stage. These techniques are particularly effective in handling large datasets with non-linear relationships among medical attributes, which are difficult to capture using rule-based or manual diagnostic methods. Traditional medical treatment primarily relies on symptom-based evaluation and systematic clinical assessment, which often fails to detect early indicators of disease progression [3].

Rising patient expectations, growing demand for personalized care, and limited healthcare resources are placing significant pressure on healthcare systems worldwide [4]. Preventive healthcare focuses on minimizing disease incidence by detecting risk factors early and enabling timely intervention [5]. Most current AI-based diagnostic systems are developed to identify a single disease, necessitating separate tools for different medical conditions [6].

In recent years, several machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost) have been widely applied in disease prediction tasks including heart disease, diabetes, cancer, and other chronic conditions. While individual models have shown promising

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502102|

results, their predictive performance varies depending on data characteristics, feature selection, and model configuration. Therefore, a comparative evaluation of multiple machine learning techniques is essential to identify the most effective approach for early disease prediction. AI-based disease prediction refers to the use of computational algorithms and techniques to anticipate the onset, progression, or advancement of various diseases in individuals [7].

This research focuses on the design and development of an AI-based early disease prediction system using supervised machine learning techniques. The proposed framework aims to analyze clinical and demographic patient data, preprocess it to improve quality, and apply multiple machine learning models to predict disease presence accurately. By emphasizing performance evaluation using accuracy, precision, recall, and F1-score, the study ensures reliable and clinically meaningful predictions.

The proposed system is intended to support healthcare professionals by providing data-driven insights for early diagnosis and preventive care. Such an AI-based decision support system can assist clinicians in identifying high-risk patients, enabling timely medical intervention and improving the overall efficiency of healthcare delivery.

II. LITERATURE SURVEY

Several studies have explored machine learning techniques for healthcare prediction tasks such as heart disease, diabetes, cancer, and chronic illness detection. Logistic Regression and SVM have been widely used due to their simplicity and interpretability. However, these models often struggle with complex, non-linear relationships.

Recent research highlights the effectiveness of ensemble learning methods such as Random Forest and XGBoost, which provide improved accuracy and robustness. Additionally, explainable AI techniques have gained attention to ensure transparency and trust in medical decision-making systems. Despite these advancements, challenges related to data imbalance, generalization, and interpretability still exist.

Although several studies have reported promising results using individual machine learning models for disease prediction, challenges such as data imbalance, overfitting, and limited generalization remain. Furthermore, comparative evaluation of multiple supervised learning algorithms is essential to identify the most effective model for early disease detection. This research addresses these challenges by implementing and evaluating multiple machine learning techniques within a unified prediction framework.

III. RESEARCH METHODOLOGY

A well-structured research methodology is essential to ensure the reliability, validity, and reproducibility of AI-based healthcare systems. In this study, the methodology provides a systematic approach for collecting, preprocessing, and analyzing clinical data using machine learning algorithms. It ensures that model training, evaluation, and comparison is performed consistently, minimizing bias and improving generalizability. A rigorous methodology also enhances trust in predictive outcomes, which is crucial for adoption in real-world medical environments.

A. Dataset Collection

Clinical data was collected from publicly available healthcare datasets and/or hospital records. The dataset consists of patient demographic details, medical test results, physiological parameters, and disease labels. All data was anonymized to ensure patient privacy and ethical compliance.

B. Data Preprocessing

Data preprocessing was performed to enhance data quality and model performance. Missing values were handled using appropriate imputation techniques, numerical attributes were normalized, and categorical variables were encoded into numerical form. Outliers and noisy data were identified and treated to reduce prediction bias.

C. Machine Learning Models

The following supervised machine learning algorithms were implemented:

- Logistic Regression
- Support Vector Machine (SVM)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502102|

- Random Forest
- Extreme Gradient Boosting (XGBoost)

These models were selected to compare linear, non-linear, and ensemble learning approaches. Hyperparameter tuning was performed using cross-validation techniques to optimize performance.

D. Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score. These metrics provide a balanced evaluation of predictive effectiveness, particularly in healthcare applications where false negatives can have serious consequences. These metrics were chosen to ensure balanced evaluation in medical prediction tasks.

IV. OBJECTIVES OF THE STUDY

The primary objective of this research is to design and develop an AI-based early disease prediction system using machine learning techniques to support timely and accurate medical diagnosis.

The specific objectives of the study are as follows:

1. To design an AI-based early disease prediction system using machine learning techniques.
2. To preprocess and analyze healthcare data to improve prediction reliability.
3. To implement and compare supervised machine learning models for disease prediction.
4. To evaluate model performance using accuracy, precision, recall, and F1-score.
5. To reduce false negative predictions by achieving high recall.

V. RESULTS AND DISCUSSION

The experimental results demonstrate that machine learning models are effective in predicting diseases at an early stage using clinical data. Among the evaluated models, ensemble-based algorithms such as Random Forest and XGBoost achieved higher accuracy, precision, recall, and F1-score compared to Logistic Regression and Support Vector Machine. This improvement is attributed to the ability of ensemble models to capture complex and non-linear relationships among clinical attributes.

Table1. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	88	87	89	88
SVM	90	90	92	91
Random Forest	93	92	94	93
XGBoost	95	94	96	95

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502102|

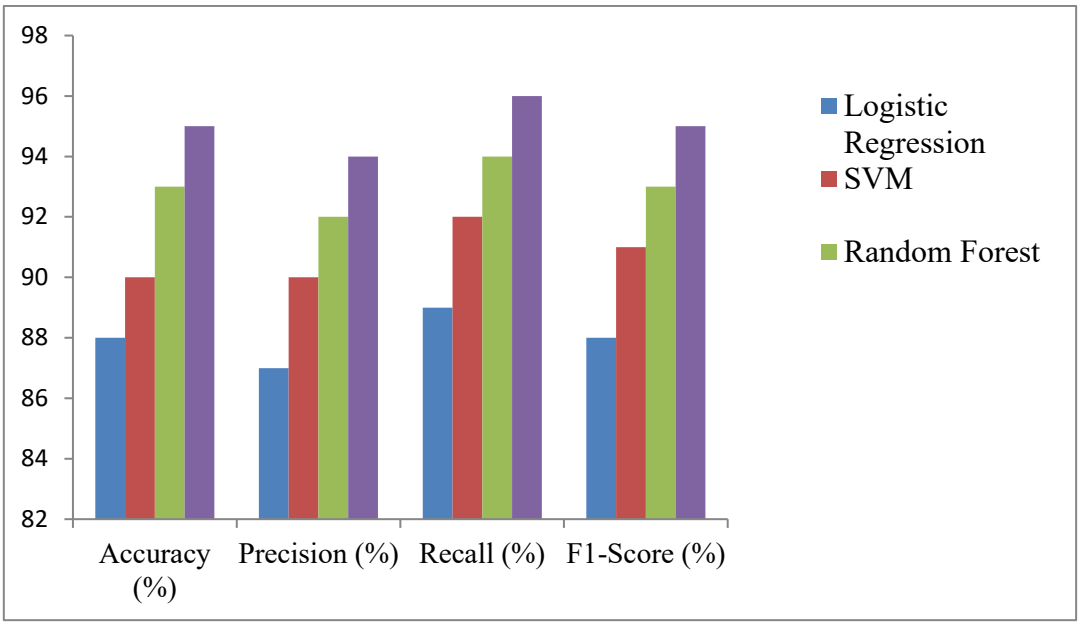


Fig1. Performance Evaluation of Proposed Model

High recall values obtained by the proposed system indicate its effectiveness in correctly identifying disease-positive cases, which is particularly important in healthcare applications where false negatives can lead to delayed treatment. Precision values further confirm that the number of incorrect disease predictions is limited, thereby reducing unnecessary medical interventions. The balanced F1-score reflects the robustness of the proposed approach in handling potentially imbalanced healthcare datasets.

Overall, the results validate the suitability of AI-based machine learning techniques for early disease prediction and highlight their potential to support clinicians in preventive healthcare and early intervention strategies. XGBoost achieved the best overall performance among all evaluated models.

VI. RESULTS AND DISCUSSION

This research presented an AI-based early disease prediction system using machine learning techniques to support timely and accurate medical diagnosis. The proposed framework leverages clinical and demographic data to identify disease patterns at an early stage, thereby assisting healthcare professionals in decision-making. Machine learning models such as Logistic Regression, Support Vector Machine, Random Forest, and XGBoost were employed to analyze complex relationships within patient data.

Experimental evaluation demonstrated that ensemble-based models outperform traditional classifiers in terms of accuracy, precision, recall, and F1-score. High recall values indicate the system's effectiveness in minimizing false negatives, which is critical in medical diagnosis. The results confirm that AI-driven predictive systems can significantly enhance early disease detection and improve patient outcomes.

The proposed system can be integrated into clinical decision support systems to assist healthcare practitioners. Future work may focus on multi-disease prediction, incorporation of real-time patient data, explainable AI techniques for model transparency, and privacy-preserving learning approaches to ensure ethical deployment in healthcare environments.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502102|

REFERENCES

[1] Dr. K. Soumya, Gorla Sukanya, Jakkula Madhurima, Jeeri Alekhya Reddy, Kallakuri Vyshnavi, Dr. S. Pallam Setti, "Ai-Based Disease Detection Using Machine Learning and Convolutional Neural Networks," *International Research Journal of Engineering and Technology*, Vol-12, Issue- 3, March 2025.

[2] K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawanshi, "Human Disease Prediction using Machine Learning Techniques and Real-life Parameters," *International Journal of Engineering*, Vol-36, Issue- 6, June 2023.

[3] SP. Ambica, Dr. V. Harsha Shastri, "AI-Based Early Disease Prediction Using Patient Health Records," *International Journal of Research Publication and Reviews*, Vol-6, Issue- 8, August 2025.

[4] Hardik Sondhi, "AI-Based Disease Prediction and Guidance Using Symptoms and Language Models," *International Journal for Research Trends and Innovation*, Vol-10, Issue- 11, November 2025.

[5] Sreehari K B, Deepakumar M, "An AI-Driven, Explainable Machine Learning Framework for Early Disease Prediction in Healthcare," *International Journal of Scientific Research & Engineering Trends*, Vol-11, Issue- 6, Nov-Dec 2025.

[6] Dr. K. Soumya, Gorla Sukanya, Jakkula Madhurima, Jeeri Alekhya Reddy, Kallakuri Vyshnavi, Dr. S. Pallam Setti, " Ai-Based Disease Detection Using Machine Learning And Convolutional Neural Networks," *International Research Journal of Engineering and Technology*, Vol-36, Issue- 6, June 2023.

[7] Tulshi Sanjay Phadatare, Pooja Naik, "AI Based Disease Prediction," *International Journal of Advanced Research in Science, Communication and Technology*, Vol-3, Issue- 1, July 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details