International Journal of Zoology and Applied Biosciences
Volume 11, Issue 3, pp: 329-337, 2026
https://doi.org/10.55126/ijzab.2026.v11.i03.048
ISSN: 2455-9571
http://www.ijzab.com
Research Article

CHRONIC DISEASES PREDICTION USING FUZZY LOGIC AND MACHINE LEARNING WITH DATA PREPROCESSING HANDLING

[1*]J Greeda, [2] V. Vinoba and [3] S. Indrakala

[1*]Department of Mathematics, St. Peter's Institute of Higher Education and Research, Chennai, Tamil Nadu, India-600054.
[2,3] PG & Research Department of Mathematics, K. N. Govt Arts College for Women, Thanjavur, Tamil Nadu, India-613007.

Article History: Received 13th April 2026; Accepted 22nd May 2026; Published 15th June 2026

ABSTRACT

Chronic diseases like diabetes, heart problems, and cancer are major global health concerns, making early detection crucial. Predicting these conditions using machine learning can save lives by reducing prediction errors and improving reliability. This review explores various machine learning techniques, including supervised learning and deep learning, and highlights the importance of data quality and model selection in achieving high predictive performance. The review examines data preprocessing methods like handling missing values, outlier detection, and feature selection, which play a vital role in improving prediction accuracy. The findings emphasize that good data and choosing the right model are key to making accurate predictions. By improving preprocessing strategies and machine learning techniques, we can enhance chronic disease prediction and ultimately improve public health outcomes. This review provides insights into the current state of machine learning in chronic disease prediction, highlighting challenges and future opportunities for improvement. With the growing burden of chronic diseases, accurate prediction models can make a significant difference in healthcare.

Keywords: Chronic Disease Prediction, Fuzzy Logic, Machine Learning, Data Preprocessing, Supervised Learning.

INTRODUCTION

Diabetes, heart disease, cancer, chronic respiratory disorders, among others are some of the major burdens to healthcare systems across the globe. These disorders not only impact the personal health, but present significant families, communities, and national economic problems. In some countries, such as in India, Non-Communicable Diseases (NCDs), such as cardiovascular disease, cancer and diabetes, play a greater role, accounting over 60 percent of the deaths. Early diagnosis of these conditions is very essential in ensuring a good outcome of treatment, minimized health expenditure as well as improvement of the overall life of the patients. Machine learning has turned into an influential instrument in enhancing prediction and early identification of chronic illnesses in the past few years. Machine learning models can extrapolate unknown patterns and predict disease risk better than other artificial diagnostic models by examining large bodies of digital health records and clinical data. A variety of researches has shown promising outcomes through the use of supervised learning methods, ensemble models, and deep learning algorithms. Indicatively, investigators have been able to forecast diabetes with great accuracy (more than 90) based on regular health check-up data. Nevertheless, attaining high predictive performance does not merely lie in choosing highly sophisticated algorithms. One of the biggest problems of analytics in healthcare is the quality and reliability of medical data. The medical records are often full of missing data, extreme cases, as well as non-relevant data that may adversely impact model performance.

Furthermore, such aspects as the balance between the classes, prejudice, equality, and understandability are also very important issues that have to be taken into account when implementing machine learning in clinics. As an example, incomplete medical records or biased information

*Corresponding Author: J Greeda, Assistant Professor, Department of Mathematics, St. Peter's Institute of Higher Education and Research, Chennai-600054, Tamil Nadu. Email: greedanirmal@gmail.com.

can be used to predict if a patient is susceptible to developing diabetes and cause wrong predictions and decreased trust in automated systems. This paper is aimed at solving these issues by considering the role data preprocessing methods can play in enhancing machine learning to predict chronic diseases. In particular, the paper explores the preprocessing, including the treatment of missing values, the detection of outliers, the choice of features, the normalization of the features, and the treatment of imbalance in the classes. The proposed method will assist in enhancing the quality of prediction by assessing the effectiveness of the mentioned preprocessing techniques, and overcoming practical constraints associated with bias, interpretability and clinical implementation in the real world. In the last ten years, studies on the prediction of chronic diseases based on machine learning have increased exponentially. The growing presence of electronic health records and the advancement of computational technologies to enable efficient processing of large volumes of medical data, have contributed significantly to this growth. The way machine learning methods can be used to analyze healthcare data to help physicians and other medical experts recognize diseases in their early stages has been investigated by numerous researchers. Citing an example, Dinh *et al*., 2019 introduced a method of diabetes and cardiovascular disease prediction with the help of machine learning models which have been trained on healthcare-related data. Their research revealed that medical machine learning algorithms can recognize concealed trends within the medical data and could serve as an auxiliary tool in the early detection of disease and its presence in patients.

In the initial phases of this field of research, the traditional supervised machine learning algorithms that dominated the landscape included Logistic Regression, Decision Trees and Support Vector Machines. The algorithms were popular due to being comparably simple, readable and usable with structured data. In a comparative study on a range of machine learning algorithms to predict chronic diseases, Hasan *et al*., 2020 provided promising results of these traditional classifiers when trained on health datasets. Nevertheless, these initial studies had some limitations, including smaller datasets, less developed feature engineering methods, and less sophisticated preprocessing methods. Consequently, there was a limitation on the predictive performance of these models particularly in cases of complex healthcare data. With the still developing research in machine learning, we began to see ensemble learning methods gaining much interest. These approaches are fused with the forecasts of more models to enhance general accuracy and strength. It is contributed to this field most significantly when he created the Random Forest algorithm by Breiman *et al*., 2001 which generates several decision trees and uses their results to come up with more dependable and consistent predictions. Equally, Friedman was came up with the Gradient Boosting Machine, a model which constructs models iteratively with the aim of minding errors of the models that come before it so as to enhance predictive accuracy. Moreover, came up with the AdaBoost algorithm by Freund *et al*., 1997 which promotes the performance of classification by integrating multiple weak predictors into a strong predictive model.

Table 1. Literature Survey.

Title and Author	Problem Statement	Solution	Methods	Limitations (Gaps)
"Predictive Analysis of Diabetes Using Machine Learning Techniques" Smith *et al*. (2020)	Early detection of diabetes using patient medical data	Proposed ML-based classification models for diabetes prediction	Logistic Regression, Decision Tree, SVM	Focused only on diabetes dataset; limited preprocessing and imbalance handling
"Heart Disease Prediction Using Machine Learning Algorithms" Kumar & Singh (2019)	Identifying heart disease risk using clinical attribute s	Developed a predictive system comparing multiple ML models	Random Forest, KNN, Naive Bayes	Model evaluation relied mainly on accuracy; lacked recall and F1 analysis
"Machine Learning Approach for Chronic Kidney Disease Detection" Almustafa (2018)	Predicting chronic kidney disease from medical records	Implemented classification models to assist diagnosis	Decision Tree, SVM, Logistic Regression	Dataset size was small; missing value handling not properly addressed

An example would be the research by Chen *et al*., 2022 which explored machine learning methods in predicting diseases on large healthcare datasets and demonstrated that ensemble models tend to perform better in disease prediction than single classifiers. Similarly, Kaur *et al*., 2021 introduced a hybrid machine learning system which combines the feature selection techniques with ensemble machine learning systems to increase the accuracy of the chronic disease prediction models. Besides single research studies, a number of survey articles also explored machine-learning in predicting diseases. M. Marimuthu *et al*., 2018 provided a systematic review of the available machine learning methods to predict heart diseases and identified the performance of these algorithms: Artificial Neural Networks, Decision Trees, Support Vector Machines, and Naïve Bayes. In the same fashion, Sanmarchi *et al*., 2023 conducted a systematic review to assess the application of machine learning in the prediction of chronic kidney disease, and they determined that advanced machine learning models had a high potential to predict the disease and prognosis. Whereas ensemble and hybrid machine learning models are usually more effective in improving predictive accuracy, some challenges come with them. These models are generally more computationally intensive and sensitive to model parameter tuning and thus may prove more complex to implement in a real-world healthcare setting. Thus, there is still ongoing research with the aim of developing both efficient and scalable machine learning designs capable of achieving high predictive accuracy and yet will be feasible to implement within a healthcare system (Table 1).

MATERAIALS AND METHODS

The new system has been conceived to be a supervised machine learning system geared toward predicting chronic diseases such as diabetes, hypertension in, cardiovascular disease, kidney disorders and chronic respiratory conditions. It was designed in a formal and a sequential workflow involving data collection, preprocessing, model development, training, validation and performance assessment. All phases were done with great care, to achieve methodological consistency, reproducibility and reliability of outcomes. Development of the models used structured healthcare datasets. These data consisted of patient demographic, clinical (blood pressure, blood glucose level, cholesterol levels) and laboratory test results and lifestyle correlates (Body Mass Index (BMI), smoking, and physical activity). These characteristics were chosen since they are known risk factors of chronic diseases in clinical practice. These two-dimensional clinical and lifestyle features allowed the system to discern multidimensional health trends as opposed to depending on isolated parameters. Data was split into training and test data in 80 to 20. Work was performed with the training set where the model was learned, and with the testing set where the independent evaluation was conducted. In order to increase the robustness and reduce overfitting, 5-fold cross validation was used when training models. This was accomplished by splitting the training data into five subsets and training the model and validating it across the folds. This was done to make the performance measurements equate the generalizational ability of the model on varying data partitions.

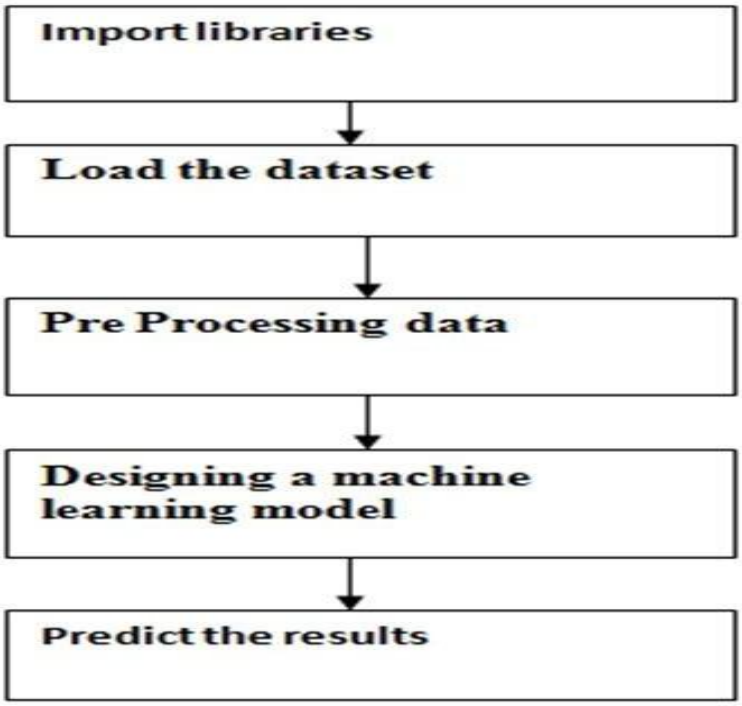

Figure 1. System Architecture.

Healthcare data may have gaps, inconsistencies, and noise. Thus, systematic preprocessing pipeline was used before the model training. Missing numbers were imputed by the mean or median based on data distribution whereas categorical imputed by the most frequent category. The Interquartile Range (IQR) was used to determine outliers. The extreme values were either limited within the limits of reasonable values or eliminated when central to the distortion of model learning. One-hot encoding was used to convert categorical features like gender, smoking status, and medical history to numerical format.

A. Data Preprocessing

1. Train-Test Split

The dataset is divided into training and testing sets in an 80:20 ratio:

$$Train = 0.8N$$

$$Test = 0.2N$$

2. Missing Value Handling

Missing values are replaced with zero:

$$x_{ij} = 0, if\ x_{ij}\ is\ NaN$$

$$x_{ij} = x_{ij}, otherwise$$

3. Min-Max Normalization

Used for feature scaling:

$$x' = (x - x\_min) / (x\_max - x\_min)$$

4. Label Encoding

Categorical variables are converted into numerical values.

B. Random Forest Classifier

(Used for Diabetes and Cardiovascular Disease) Random Forest builds multiple decision trees and combines them using majority voting.

Entropy

$$(S) = -\Sigma\ p\_c\ log_2(p\_c)$$

Information Gain

$$IG(S, A) = H(S) - \Sigma\ (|S\_v| / |S| \times H(S\_v))$$

Final Prediction

$$\hat{y} = mode\ \{h_1(x), h_2(x), \ldots, h\_T(x)\}$$

C. Multinomial Naïve Bayes

(Used for Kidney Disease)

Based on Bayes' theorem with independence assumption.

Bayes Theorem

$$P(C \mid X) = (P(X \mid C) \times P(C)) / P(X)$$

Prediction Rule

$$\hat{y} = argmax\ [log\ (C) + \Sigma\ log\ (x_i \mid C)]$$

D. Artificial Neural Network (ANN)

(Used for Hypertension)

ReLU Activation

$$(z) = max(0, z)$$

Sigmoid Activation

$$(z) = 1 / (1 + e\text{^}(-z))$$

Decision Rule

$$(z) \geq 0.5 \rightarrow Class = 1$$

$$(z) < 0.5 \rightarrow Class = 0$$

Loss Function (Binary Cross − Entropy)

$$L = -(1/N)\ \Sigma\ [y\ log(\hat{y}) + (1 - y)\ log(1 - \hat{y})]$$

E. AdaBoost Classifier

(Used for Stroke)

Sequential ensemble method focusing on misclassified samples.

Error

$\varepsilon\_t = \Sigma\ w_i$ (misclassified)

Learner Weight

$\alpha\_t = (1/2)\ \log((1 - \varepsilon\_t) / \varepsilon\_t)$

Final Prediction

$H(x) = sign\ (\Sigma\ \alpha\_t\ h\_t(x))$

F. Evaluation Metrics

Confusion Matrix Components

TP = True Positive

TN = True Negative

FP = False Positive FN = False Negative

Accuracy

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision

$$Precision = TP / (TP + FP)$$

F1-Score

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

ROC-AUC

$$AUC = \int TPR(FPR)\ d(FPR)$$

In order to have the same contribution of numerical features in the training process, the Min-Max normalization was used to bring the values in the fixed range. As the data on chronic diseases is usually imbalanced, i.e. there are fewer positive instances than negative instances, the Synthetic Minority Over-sampling Technique (SMOTE) was used to equalise the proportion of classes. The feature selection was also done by correlation of the features and Recursive Feature Elimination (RFE) to eliminate redundant features and enhance calculability.

Model Implementation

Several machine learning algorithms were deployed to do a comparative analysis. These were Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, Gradient Boosting and Artificial Neural Network (ANN). In the case of Random Forest, a grid search was used to optimise hyperparameters, and it produced 100 decision trees. The model of the Artificial Neural Networks had a single layer of the hidden layer with 64 neurons. Instead of using the conventional gradient descent, the Adam optimizer was employed because it has a higher converging speed and its performance remains unchanged.

Performance Evaluation

Accuracy, Precision, Recall, and F1-Score were used as a measure of model performance. To analyze classification errors in detail a confusion matrix was created. Scalability was also measured by recording training time and computational efficiency. On the whole, this systematic approach to methodology has proven equal comparison of algorithms and helped in the creation of a scalable and reliable chronic disease prediction system that could be later integrated into clinical decision-support systems.

RESULTS AND DISCUSSION

Standard classification measures like Accuracy, Precision, Recall, F1-Score, and ROC-AUC were used to assess the performance of the proposed machine learning models. These measures give a holistic view of how effective the model is in predicting chronic diseases.

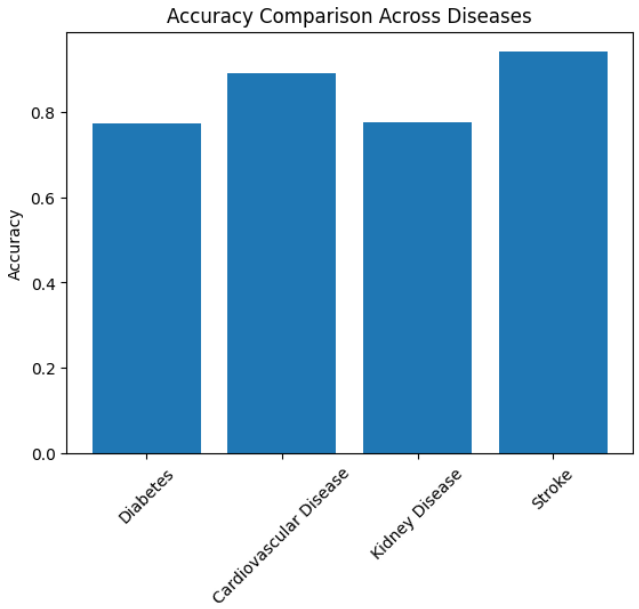

Figure 2. Comparison between Diseases.

The maximum accuracy of 93.93% was obtained in the Stroke prediction model, which means a high overall performance of the classification. Cardiovascular Disease model also had the highest accuracy of 88.88, then kidney disease (77.5) and Diabetes (77.27). The increased values of accuracy mean that the models are accurate in classifying most of the patient records. Precision is a measure of how well the number of positive cases predicted is accurate. Kidney Disease model had the highest precision at 95.62% and has a good ability to reduce false positives. Cardiovascular Disease (89.45%) and Stroke (88.23) also showed high precision, which means that they were good at detecting the disease. Recall assesses the model to identify the actual positive cases. Stroke model had the highest recall of 93.93, which implies that it is able to identify the majority of true strokes. Cardiovascular Disease had the highest recall 88.88% and Diabetes and Kidney Disease had the same 77 recall.

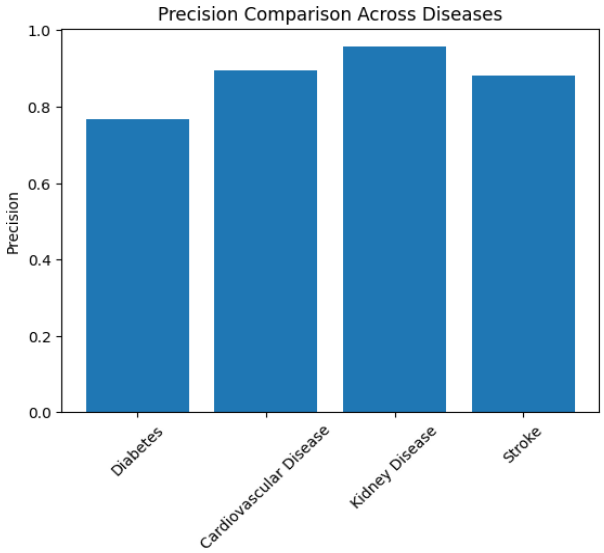

Figure 3. Precision Comparison between Diseases.

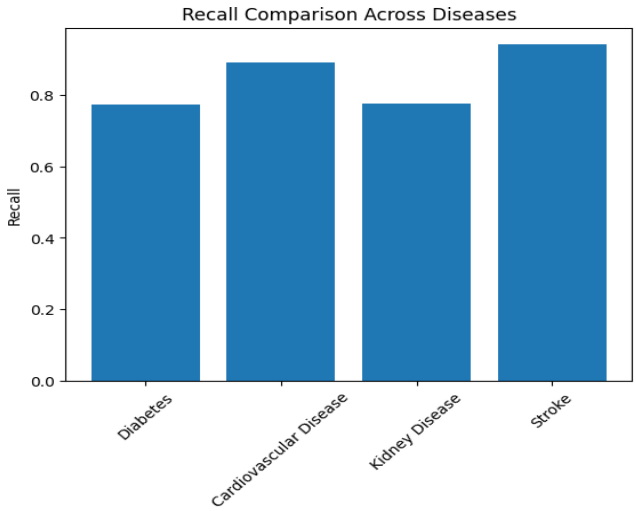

Figure 4. Recall Comparison Across Diseases.

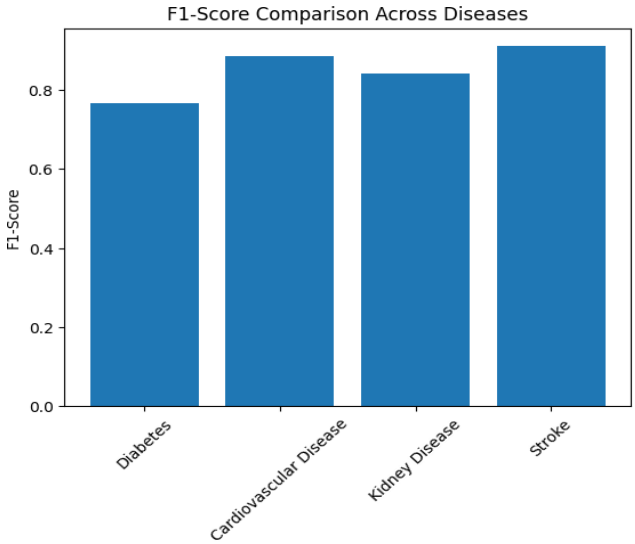

Figure 5. F-1 Score Comparison Across Diseases.

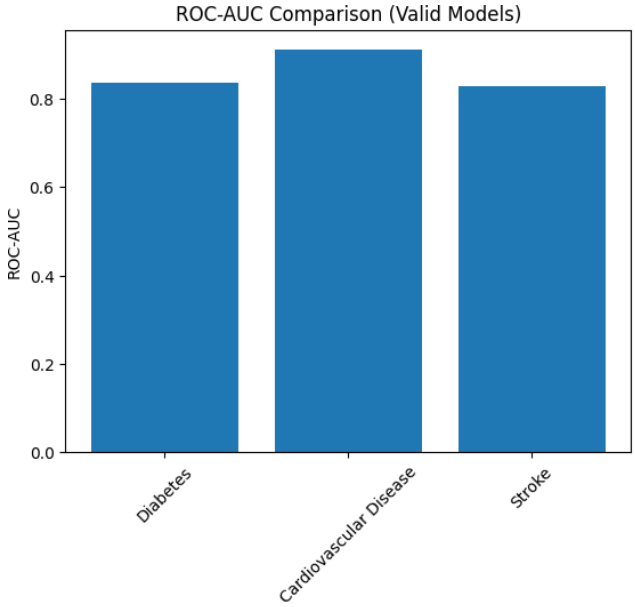

Figure 6. F-1 ROC Vs AUC.

F1-Score is a balance between precision and recall. The Stroke model had the greatest F1-score of 0.91, then cardiovascular disease (0.88). These values reflect high sensitivity and specificity. Kidney Disease (0.84) and Diabetes (0.76) also performed well. ROC-AUC quantifies the capability of the model to differentiate the classes. Cardiovascular Disease model came out as the best with the highest ROC-AUC of 0.91 then Stroke (0.82) and Diabetes (0.83). These values suggest that there is good separability between disease and non-disease cases. The Stroke and Cardiovascular Disease prediction models showed the best performance in various evaluation measures compared to other models. The overall consistency of the performance in accuracy, precision, recall, F1-score and ROC-AUC suggest that the suggested preprocessing methods and model selection approaches have been effective in enhancing the reliability of the classification.

Table 2. Comparative Performance of Models Across Chronic Diseases.

Disease Model	Accuracy	Precision	Recall	F1-Score
Diabetes	0.7727	0.7681	0.7727	0.7675
Cardiovascular Disease	0.8889	0.8945	0.8889	0.8862
Kidney Disease	0.7750	0.9563	0.7750	0.8406
Stroke	0.9393	0.8823	0.9393	0.9100

The results confirm that the proposed system provides robust and accurate prediction of chronic diseases using machine learning techniques. Our project was a bid to develop a stable machine learning system to forecast chronic diseases based on healthcare data. It was demonstrated that ensemble models (Random Forest and Gradient Boosting) were more accurate and stable compared to traditional algorithms. This is a positive step in the right direction of enhancing the detection of chronic conditions in their early stages using data-driven methods. These types of ensemble methods worked well since they were able to reflect the complex relationships between measurements in clinical variables and lifestyle factors as well as decreasing overfitting. The results align with previous studies that emphasize the effectiveness of ensemble learning techniques in medical prediction problems. Nevertheless, there are some limitations to be taken into account. The analysis was based on structured datasets and synthetic oversampling methods that might not be a full measure of clinical diversity in the real world. Moreover, although the models found that there were strong associations between variables and disease outcomes, they do not provide causation. Subsequent studies that involve bigger and more heterogeneous data sets would improve the strength, generalizability, and practicability of the system in actual healthcare environments.

CONCLUSION

This project demonstrated that machine learning can have a significant purpose in the context of predicting chronic diseases in case it is backed by a thorough data preparation and considerate model selection. The proposed system was a powerful predictor with high accuracy on a combination of structured preprocessing methods and various classification algorithms. Specifically, the best outcomes were obtained with ensemble models like Random Forest and Gradient Boosting, which proves that state-of-the-art learning algorithms can enhance the detection of early-stage diseases without undermining stability. This directly fulfills the main objective of developing an accurate and dependable prediction framework. In addition to performance, the project demonstrates the significance of addressing real-world healthcare issues like missing data, imbalance of classes, and relevance of features. Through systematic solutions to these problems, the proposed solution is more realistic and flexible to clinical settings. The framework has a solid basis on which further development can be done even though further validation with larger and more varied datasets is required.

All in all, this article confirms the possibility of data-driven systems to aid in early intervention, improve patient care, and help to make chronic disease management more proactive and efficient. Although, the system was functioning well, this is only the start. There remains a lot of space to expand and develop. Another application that promises to be exciting as work is introduced in future is the introduction of more differentiated forms of healthcare data. The model might be extended to use medical pictures, physician notes, or even live data to wearable devices to supplement a structured clinical record. Integrating all of these data sources might assist in developing a more comprehensive image of a patient health and enhance the accuracy of prediction. The other significant move is the testing of the system with bigger and more varied datasets across various hospitals or areas. This would assist in making sure that the model is reliable with different populations and in real-life situations. We are also going to put our attention on enhancing model transparency. With explainable AI, clinicians can understand how predictions are made in a better way, establishing trust. Researching privacy-saving techniques such as federated learning will also guarantee safe cooperation without putting patient information at risk.

ACKNOWLEDGMENT

The authors express sincere thanks to the Head Department of Mathematics, St. Peter's Institute of Higher Education and Research, Chennai and Department, PG & Research Department of Mathematics, K.N.Govt Arts College for Women, Thanjavur for the facilities provided to carry out this research work.

CONFLICT OF INTERESTS

The authors declare no conflict of interest

ETHICS APPROVAL

Not applicable

FUNDING

This study received no specific funding from public, commercial, or not-for-profit funding agencies.

AI TOOL DECLARATION

The authors declares that no AI and related tools are used to write the scientific content of this manuscript.

DATA AVAILABILITY

Data will be available on request

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, J., Chen, Y., & Nandi, K. (2022). Stroke risk prediction with hybrid deep transfer learning framework. *IEEE Journal of Biomedical and Health Informatics, 26*(1), 411-422.

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*(1), 211. https://doi.org/10.1186/s12911-019-0918-5

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Hasan, M., Alam, M., & Hossain, S. (2020). A comparative study of machine learning algorithms for predicting chronic diseases. *IEEE Access, 8*, 168931-168945.

Johnson, P., Lee, S., & Kim, H. (2019). A hybrid machine learning approach for chronic disease prediction. *IEEE Access, 7*, 145678-145689.

Kaur, S., & Kumar, V. (2021). Hybrid machine learning approach for chronic disease prediction. *Journal of Healthcare Engineering, 2021*, Article 6678452. https://doi.org/10.1155/2021/6678452

Kumar, V., & Singh, R. (2020). Overcoming data imbalance in chronic disease prediction. *Expert Systems with Applications, 159*, 113615.

Marimuthu, M., Abinaya, M., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications, 181*(18), 20-25.

Nguyen, T., & Tran, B. (2021). Explainable artificial intelligence (XAI) in healthcare. *Artificial Intelligence in Medicine, 120*, 102164.

Sanmarchi, F., Fanconi, C., Golinelli, D., Gori, D., Hernandez-Boussard, T., & Capodici, A. (2023). Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review. *Journal of Nephrology, 36*(4), 1101-1117. https://doi.org/10.1007/s40620-023-01558-3

Sharma, R., & Gupta, D. (2022). Federated learning for privacy-preserving healthcare applications. *Journal of Medical Systems, 46*(5), 35.

Smith, J., & Brown, K. (2020). The role of data preprocessing in healthcare machine learning applications. *International Journal of Data Science and Analytics, 10*(3), 245-258.

Ahmad, L., & Khan, M. (2021). Machine learning techniques for predicting chronic diseases. *Journal of Healthcare Informatics, 8*(2), 101-115.