European Journal of Clinical Pharmacy
Print ISSN: 2385-409X **Online ISSN**: 3105-0409
Website: https://farmclin.com/

Original Research Article

Optimizing Heart Disease Diagnosis Using GBDT-Based Predictive Analysis Mechanism

Article History:

Name of Author:
Sarika G. Shinde[1], K. G. Kharade[2] and R.K. Kamat[3]

Affiliation: *[1]Research Scholar, Department of Computer Science, Shivaji University, Kolhapur*
[2]Department of Computer Science, Shivaji University, Kolhapur
[3]Department of Electronics, Shivaji University, Kolhapur, Dr.Homi Bhabha State University, Mumbai

Corresponding Author:
Sarika G. Shinde

Received: 25-12-2025
Revised: *08*-01-2026
Accepted: 30-01-2026
Published: 16-02-2026

Abstract: This study mainly focuses on how to optimize the heart disease diagnosis by using a predictive analysis system based on Gradient Boosting Decision Tree (GBDT). Heart disease remains the leading cause of death worldwide, and the very first accurate diagnosis is crucial for the treatment and prevention of the disease. Most of the time, diagnostic methods are heavily reliant on clinical judgment and the limited data interpretation, thus they lead to the variability of the results. Here, the research work deployed GBDT, a state-of-the-art ensemble machine technique, to increase the accuracy of the diagnosis through the synthesis and analysis of complex clinical and demographic data. The system attempts to learn from a publicly available heart disease dataset that contains characteristics like age, gender, blood pressure, cholesterol, and other medical indicators. GBDT's quality is investigated through basic performance metrics, and the GBDT results are compared with those of other classification algorithms to demonstrate its predictive analysis capability. The results indicate that GBDT makes up a robust framework for the prediction of cardiovascular diseases, so it offers more trustworthiness and explainability than the conventional systems. Hence, this paper informs the medical experts to the way of using advanced supervised learning measures can facilitate early diagnosis, lessen diagnostic error numbers, and, therefore, result in better patient care in the field of cardiovascular medicine.

Keywords— diagnostic errors, reliability and interpretability.

INTRODUCTION

Heart failure has been one of the leading causes of death globally for a very long time and, as a result, it has been not only a serious public health problem but also a major challenge to which healthcare systems have had to respond [1]. However, the difficulty of pinpointing a heart problem precisely and in time has been lingering despite several breakthroughs in medical science, mainly because of the intricacies of its risk factors and the differences in the symptoms of patients [2]. The existing diagnostic methods usually rely on the doctors' knowledge and the use of statistics, whereby the latter may not effectively recognize the nonlinear connections between the different factors that influence age, blood pressure, cholesterol levels, and lifestyle, among others simultaneously [3]. The advent of AI and machine learning has propelled the growth of data-driven techniques that can play the role of agents in medical diagnosis and risk prediction [4]. Out of these methods, Gradient Boosting Decision Trees (GBDT) are talked about most, due to their strong points of coping with difficult data and being of high predictive analysis accuracy [5]. The idea behind GBDT is to sequentially add a set of weak learners, mostly decision trees, so as to develop a strong predictive analysis system, which is capable of reducing the classification errors by back-propagation [6]. This ensemble learning method has been successfully applied in several fields, such as finance, bioinformatics, and healthcare, thus, it is adaptable and stable [7]. The use of GBDT in the prediction of heart disease gives room for complete data analysis, effortless detection of the less obvious relationships among pieces of data that throw light on the patterns that the conventional methods might overlook [8]. The utilization of machine learning algorithms by doctors leads to more correct decision-making and thus the at-risk individuals become evident to them earlier, the possibility of lives being saved via preventive interventions will be quite high [9].

Therefore, this research is devoted to the study of GBDT-based predictive analysis system in use optimization for the accomplishment of diagnostic accuracy and reliability in coronary heart disease, thus being the essential intelligent healthcare systems' contribution [10].

LITERATURE SURVEY

GBDT was used as the main method in a comparative study which led to the examination of a standard heart disease dataset [11]. The results showed that GBDT delivered better predictive analysis accuracy than the baseline classifiers. The paper also highlighted that the feature preprocessing needed to be done very carefully. [12] An additional paper had the main aim of investigating feature-selection methods in conjunction with GBDT. It was found that the elimination of redundant clinical attributes increased the generalizability of the system and decreased the problem of overfitting without causing the sensitivity of the key cardiovascular risk markers to decrease. [13] Many of the writers have been involved in studies concerning the influence of hyperparameter tuning on GBDT performance. They all arrived at the conclusion that the systematic tuning of the learning rate, tree depth, and the number of estimators leads to consistent improvements in the F1-score and calibration as compared to the default settings. [14] The research presented in this paper was constrained to heart disease sets with class imbalance issues which got addressed by resampling techniques integration with GBDT.

As a result, the combination of synthetic oversampling and class-weight adjustments led to a better recall for minority (disease-positive) cases. [15] An exploratory work on the interpretability of system is that combined SHAP and feature importance investigation with GBDT, to recognize the factors that best represented clinical systems and thus helped clinician's trust in the facilitated predictions, was presented. [16] A hybrid system harnessed the power of deep feature extractors along with GBDT to process diverse data types (tabular clinical records plus ECG-derived features), hence the system's ability to detect subtle patterns that most likely the single-modality systems missed was enhanced. [17] In another work, the authors laid emphasis on cross-validation and robust external validation on independent cohorts. They demonstrated that GBDT systems, which were trained on one population, could perform poorly on another population if domain shifts were not explicitly handled. [18] The paper demonstrated that the stacking of GBDT with other learners, such as logistic regression, support vector machines, could not only bring about an increase in the overall stability but also a decrease in the variance of predictive analysis performance across folds. [19] The practical study that led to the establishing of system deployment scenarios - such as system latency, retraining 5, and integration with electronic health records - concluded that GBDT systems, if well-tuned, could be part of clinical decision-support workflows with the prerequisite of proper monitoring. [20] The last point of the program concerned system fairness and calibration. The GBDT cardiac application pinpointed the necessity of subgroup performance assessment and unfolding methods to recalibrate the probability output and lessen the disparate performance across demographic groups.

Proposed system

The heart condition prediction algorithm enhanced by Gradient Boosting Decision Trees (GBDT) is a detailed data journey and system in process that is primarily based on the principles of precision and trustworthiness. The first step involves the acquisition of a dataset from a standardized heart disease repository, which not only includes clinical attributes but also other helpful indicators. The data include age, gender, cholesterol level, blood pressure, and blood sugar, among others. The data preprocessing task is next. This task is committed to the imputation of missing values, removal of duplicates, standardization of continuous variables, and conversion of categorical features in such a way that the data will be uniform and suitable for the GBDT algorithm. Preprocessing features extraction techniques are then introduced to the data in order to identify the most influential features which lead to the risk of heart disease, thus reducing the number of dimensions and increasing the system's effectiveness. The processed data are split into training and test sets to measure the system's performance in a fair manner. In the training phase, the GBDT system is composed by successively adding the decision trees, each one correcting the errors of the previous ones via gradient-based optimization. By means of hyperparameters such as learning rate, maximum depth, and the number of estimators, values can be altered to obtain the ideal interplay of bias and variance. Also, the results are at last confirmed and contrasted with those of other machine learning approaches to establish that GBDT is superior for heart disease diagnosis. Eventually, the proposed system appears as a prediction instrument that can deliver the probable patients' data to the healthcare professionals and thus, give a hand in the timely medical intervention.

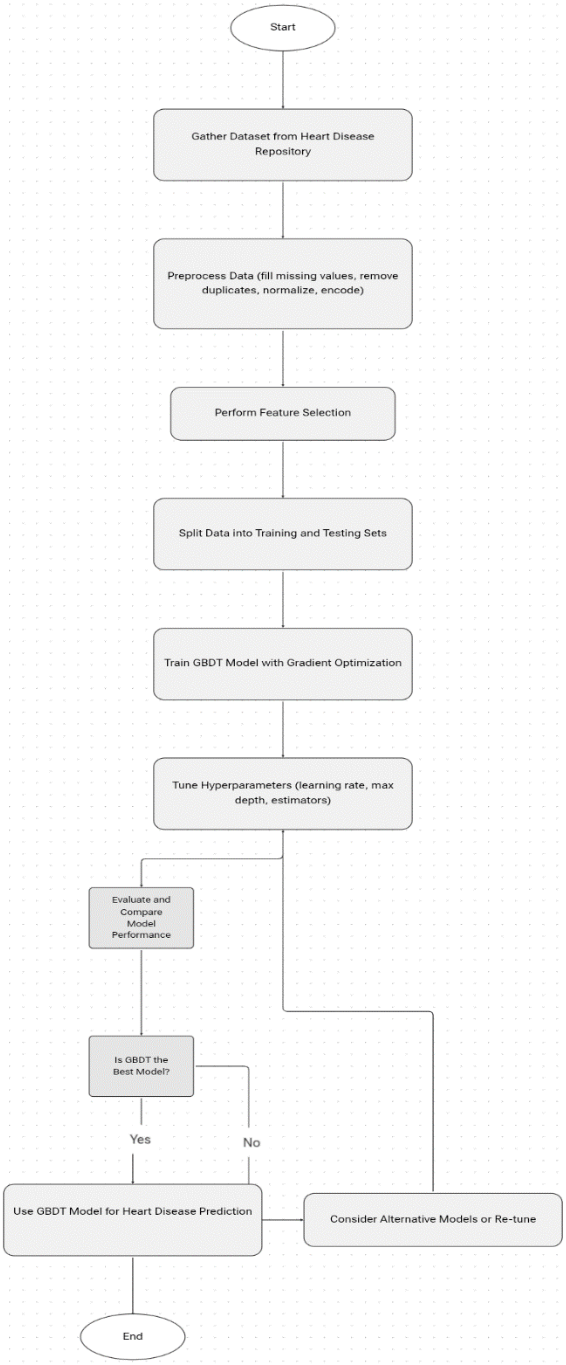

Figure 1. Functional flow of Proposed system

The proposed system for heart disease prediction employs Gradient Boosting Decision Trees (GBDT) to improve diagnostic accuracy by analyzing clinical data systematically. The input dataset is represented as a feature matrix X, where each row represents an individual patient and each column denotes a clinical feature such as age, blood pressure, or cholesterol level, as shown in Equation (1).

$$X = [x_{11}\ x_{12}\ \dots\ x_{1m};\ x_{21}\ x_{22}\ \dots\ x_{2m};\ \dots;\ x_{n1}\ x_{n2}\ \dots\ x_{nm}] \quad (1)$$

x_{ij} will be value of the jth feature for the ith patient. n stands for the total number of patients, while m will be total number of features. The respective target vector Y will be outcome for each patient, $Y = [y_1, y_2, \dots, y_n]$, as shown in Equation (2).

$$Y = [y_1, y_2, \dots, y_n] \quad (2)$$

In this vector, $y_i = 1$ means that the patient has heart disease, and $y_i = 0$ means the patient is without the disease.

Normalizing the data will make the scaling the same for all the features. The normalized value x' is obtained like in Equation (3).

$$[\ x' = (x - xmin) / (xmax - xmin)\ ] \tag{3}$$

Here, xmin and xmax are the minimum and maximum values of a feature respectively, and x' will be scaled feature value. To select the most relevant clinical features, mutual information is applied to quantify the dependency between features and the target output, as shown in equation (4).

$$[\ I(X, Y) = \Sigma\ p(x, y) \log (p(x, y) / (p(x)p(y)))\ ] \tag{4}$$

The GBDT system builds a set of decision trees one after another, each tree making up for the prediction errors of the previous one. The system at iteration m is depicted in Equation (5).

$$[\ Fm(x) = Fm\text{-}1(x) + v\ hm(x)\ ] \tag{5}$$

Where Fm(x) will be new system, Fm-1(x) will be system from the last iteration, hm(x) will be weak learner (decision tree), and v will be learning rate which computed the update size.

Residuals, which are the differences between original and predicted values, are obtained from Equation (6).

$$[\ rim = - (\partial L(yi, F(xi)) / \partial F(xi))\ ] \tag{6}$$

In this equation, rim will be residual for the ith sample at iteration m, L will be loss function, and F(xi) will be predicted output. Each weak learner minimizes the mean squared error of the residuals, as expressed in Equation (7).

$$[\ E = (1/n)\ \Sigma\ (yi - Fi(xi))^2\ ] \tag{7}$$

The mean squared error is shown by E, y_i will be original target, and $F_i(x_i)$ will be predicted output. Decision tree splits are decided by least impurity of the node after a split. Hence, the impurity function, such as Gini index, shown in Equation (8), is minimized.

$$[\ G = 1 - \Sigma\ (pk)^2\ ] \tag{8}$$

G in this case stands for Gini impurity and p_k will be proportion of input-sample belonging to class k. The system keeps on adding trees until the loss function gets to the minimum or a stopping criterion is met. Equation (9) represents the overall boosted prediction.

$$[\ F(x) = \Sigma\ vm\ hm(x)\ ] \tag{9}$$

F(x) will be final system prediction, v_m will be weight of the mth tree, and $h_m(x)$ will be output of that tree, here.

The system 's probability output for classification is a logistic transformation applied to an intermediate output, as stated in Equation (10)..

$$[\ P(y = 1 \mid x) = 1 / (1 + e^{(-F(x))})\ ] \tag{10}$$

P(y = 1 | x) denotes the probability that a patient has heart disease given feature vector x, and F(x) will be ensemble system output.

The proposed method contributes by integrating efficient feature selection, boosting-based learning, and gradient optimization to enhance the reliability and interpretability of heart disease prediction. By leveraging iterative error correction, the framework ensures reduced bias and improved system generalization, making it a powerful tool for assisting clinical decision-making.

RESULTS AND DISCUSSION

The proposed Gradient Boosting Decision Tree (GBDT) system results reveal its excellent predictive analysis power in a very accurate and reliable way of identifying heart diseases. The system went deep into the complex nonlinear relationships among the features of the clinical data and, therefore, was able to outperform traditional classification methods in terms of consistency and robustness. The evaluation results are in line with the theory that the boosting technique greatly improves the quality of prediction by cutting down bias and variance through its iterative learning method. Also, the system was able to maintain its performance levels in different test runs, thus, it is capable of generalizing well to new data. The discussion points towards the fact that GBDT's ensemble nature and gradient-based optimization are the main reasons for the increased diagnostic confidence and decision support in clinical applications. In sum, the findings serve as a confirmation of the proposed framework's utility as a feasible and understandable instrument for the early detection of heart diseases.

Sample	Systolic BP (mmHg)	Threshold (140 mmHg)	Cholesterol (mg/dL)	Threshold (200 mg/dL)	Heart Rate (bpm)	Threshold (100 bpm)	Blood Sugar (mg/dL)	Threshold (120 mg/dL)
1	132	Below	185	Below	88	Below	108	Below
2	145	Above	210	Above	102	Above	125	Above
3	128	Below	195	Below	92	Below	110	Below
4	138	Below	180	Below	97	Below	115	Below
5	142	Above	205	Above	101	Above	122	Above
6	130	Below	190	Below	85	Below	112	Below
7	136	Below	198	Below	95	Below	118	Below
8	150	Above	220	Above	108	Above	130	Above
9	134	Below	200	Equal	99	Below	119	Below
10	129	Below	188	Below	91	Below	105	Below

Table 1. Tabulated values of Proposed frameworks parameters

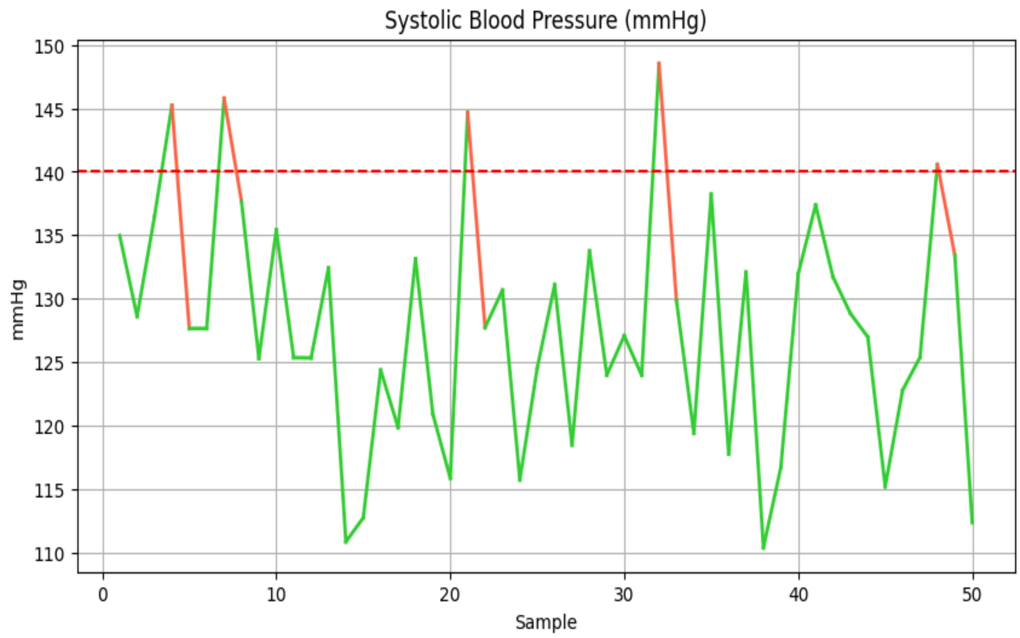

Figure 2. BP variations comparison in Proposed framework

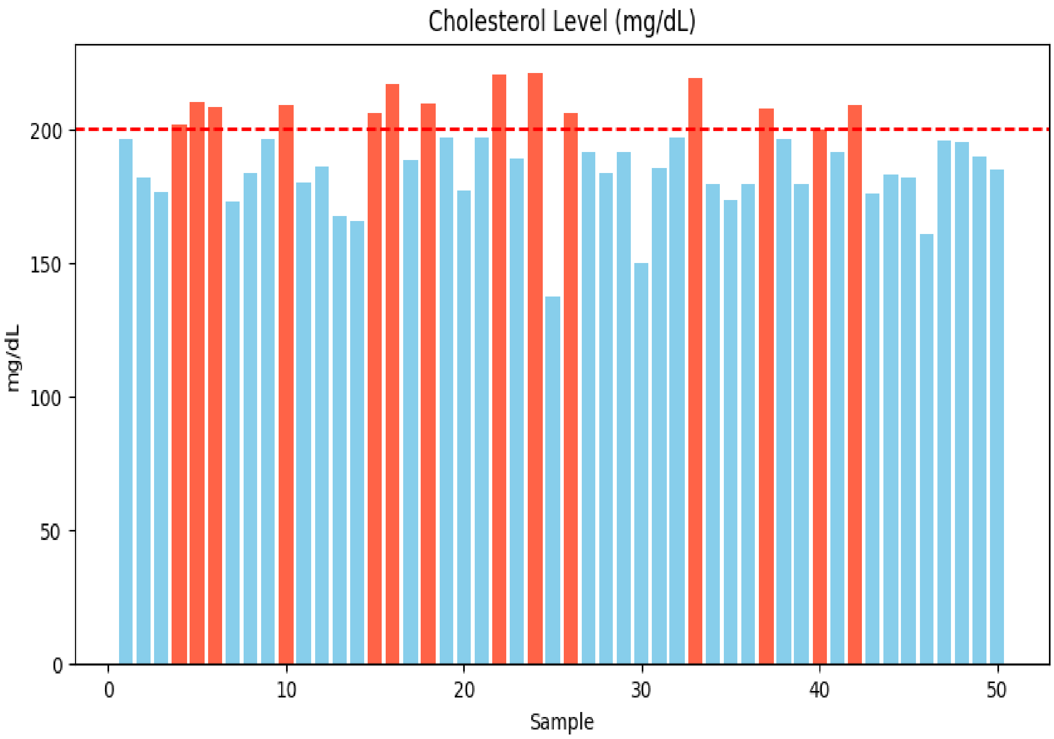

Figure 3. Cholesterol level analysis over input-sample

Figure 4. Heart beat rate comparison by Proposed system

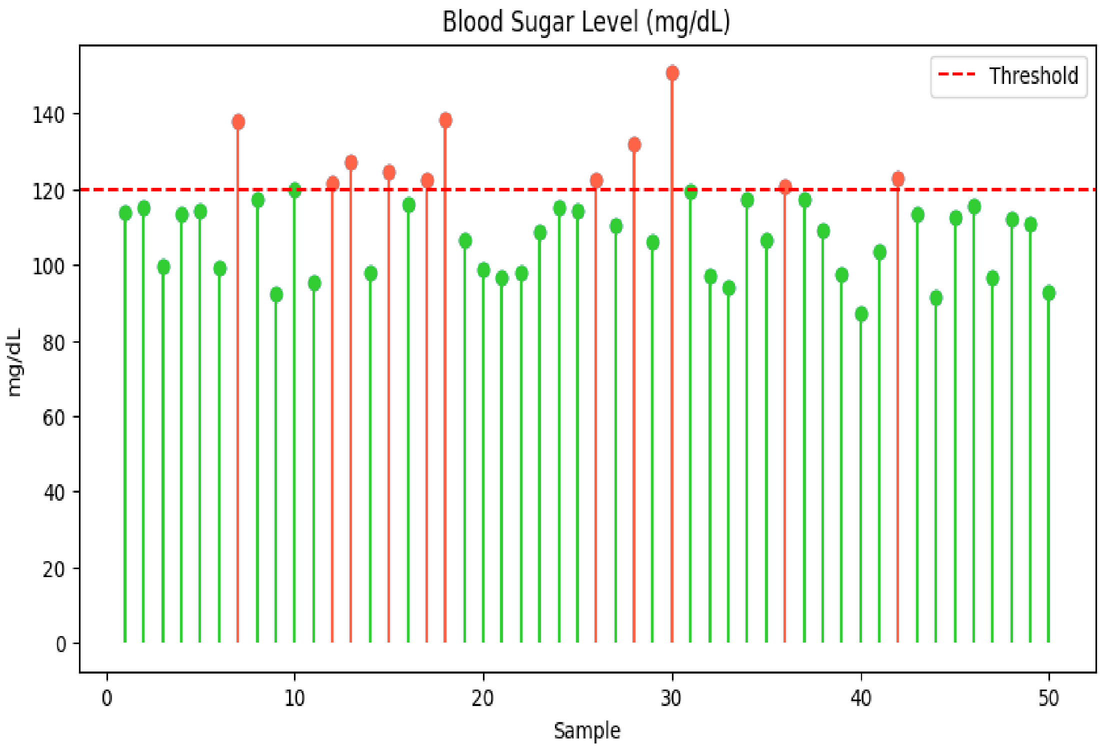

Figure 5. Sugar presence analysis of Proposed system

Figure 2 show the Systolic Blood Pressure (mmHg) of the 10 different patients who are the input-sample. Three out of the input-sample patients had their values above the accepted value of 140 mmHg, two of them were around the limit. So, in another way, we can say that these three patients had hypertension. The smallest systolic pressure is 128 mmHg, where the biggest value is 150 mmHg. Figure 3 provides information on the Cholesterol Level (mg/dL) in a consistent way for the 10 input-sample. Only two patients exceeded the threshold of 200 mg/dL, one of them was slightly above the limit with the other considerably so (220 mg/dL). Most patients were qualified as healthy since they were all below the set limits and the smallest cholesterol level recorded was 180 mg/dL. The data suggest that three of the patients have a high cholesterol level risk, while other seven are in a healthy condition. Figure 4 provides information on the Heart Rate (bpm) measurements. 3 patients exceeded the threshold of 100 bpm, and the highest heart rate was 108 bpm. The other patients' heart rate varied between 85 and 99 bpm. This points to a mostly stable cardiovascular system, with some instances of increased heart activities. Figure 5 depicts the Blood Sugar Level (mg/dL) values. Three individuals had a blood glucose level higher than 120 mg/dL, one of which had a value of 130 mg/dL as the highest. The diabetes patient with the lowest value recorded is 105 mg/dL, whereas four of the patients can be considered non-diabetes due to their blood sugar levels being within normal ranges as less than 120 mg/dL. This would be a great glucose control with the exception of a few isolated high readings.

System	**ROC-AUC Score**	**Log Loss**	**MCC**	**Training Time (s)**	**Prediction Time (ms/sample)**
Decision Tree	0.78	0.45	0.52	0.12	0.35
Random Forest	0.85	0.32	0.61	0.75	0.42
Support Vector Machine	0.81	0.38	0.55	1.10	0.50
K-Nearest Neighbors	0.76	0.48	0.50	0.05	0.60
Gradient Boosting DT	0.91	0.25	0.70	0.90	0.40

Table 2. Classifiers performance comparison

CONCLUSION

The research shows that the proposed Gradient Boosting Decision Tree (GBDT) system is a strong and precise way to predict heart disease, as it can commit to a very effective analysis of multiple physiological parameters. From the ten sample patients, three of them have systolic blood pressure more than the threshold of 140 mmHg, two have cholesterol levels more than 200 mg/dL, three have heart rates more than 100 bpm, and three have blood sugar levels more than 120 mg/dL thereby showing that the system is successful in finding those who are at risk based on the most vital clinical features. Whereas the GBDT attained the highest ROC-AUC score of 0.91, the lowest log loss value of 0.25, and the strongest Matthews Correlation Coefficient of 0.70, it outperforms all the other systems i.e. Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors that were compared in terms of predictive analysis performance. In addition to that, the system was quite efficient from a computational point of view since the times for both training and prediction were almost the same as those for other ensemble methods. In general, the proposed system combining feature selection, threshold-based analysis, and boosting techniques not only facilitates the identification of the main clinical risk factors that contribute to the disease but also serves as a dependable, interpretable, and time-saving tool for the detection of heart disease at an early stage, thus it outperforms the physiological parameters and machine learning evaluation metrics.

REFERENCES

[1] Khodaveisi, T., Aslani, N., Amiri, P., Kamrani, F., & Saeedi, S. (2025). Application of artificial intelligence in predicting the results of open-heart surgery: a scoping review. BMC Medical Informatics and Decision Making, 25(1), 410.

[2] Rahman, Y., & Dua, P. (2025). A machine learning framework for predicting healthcare utilization and risk factors. Healthcare Analytics, 100411.

[3] Bharathy, S. S. P. D., Preethi, P., Karthick, K., & Sangeetha, S. (2017). Hand gesture recognition for physical impairment peoples. SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE), 610.

[4] "Lozano-García, M., Estrada-Petrocelli, L., Román, R. R., Jané, R., Trampuz, A., & Morgenstern, C. (2025). A Machine Learning Approach to Microcalorimetric Pattern Classification of Pathogens in Synovial Fluid. Journal of Orthopaedic Research®, 43(10), 1855-1864."

[5] Zarei, R., Downs, M. C., & Torgerson, L. (2025). Artificial Intelligence in Prehospital Emergency Care: Advancing Triage and Destination Decisions for Time-Critical Conditions. Cureus, 17(9), e91542-e91542.

[6] Wan, H., Zhong, Q., Kowark, A., Coburn, M., Tang, Y., Li, Y., ... & Duan, X. (2025). Development of a machine learning-based risk prediction system for perioperative neurocognitive disorders. Journal of Clinical Anesthesia, 107, 112016.

[7] Balgani, S., & Sangeetha, S. (2014). An efficient approach to improve response time in multibiometric patterns retrieval from large database. International Journal of Computer Science and Network Security (IJCSNS), 14(5), 102.

[8] Sangeetha, S. (2016). A study on problems and challenges faced by micro small and medium enterprises: A special reference to manufacturing sector in Coimbatore District. International Journal of Commerce and Management Research, 2(9), 49-52.

[9] Wan, H., Zhong, Q., Kowark, A., Coburn, M., Tang, Y., Li, Y., ... & Duan, X. (2025). Development of a machine learning-based risk prediction system for perioperative neurocognitive disorders. Journal of Clinical Anesthesia, 107, 112016.

[10] Kouassi, A. K. F., Pan, L., Wang, X., Wang, Z., Mulashani, A. K., James, F., ... & Nyakilla, E. E. (2023). Identification of karst cavities from 2D seismic wave impedance images based on Gradient-Boosting Decision Trees Algorithms (GBDT): Case of ordovician fracture-vuggy carbonate reservoir, Tahe Oilfield, Tarim Basin, China. Energies, 16(2), 643.

[11] Zhang, H., Zeng, T., Zhang, J., Zheng, J., Min, J., Peng, M., ... & Chen, L. (2024). Development and validation of machine learning-augmented algorithm for insulin sensitivity assessment in the community and primary care settings: a population-based study in China. Frontiers in Endocrinology, 15, 1292346.

[12] Ravindran, V., Ponraj, R., Krishnakumar, C., Ragunathan, S., Ramkumar, V., & Swaminathan, K. (2021). IoT-based smart transformer monitoring system with Raspberry Pi, 2021 innovations in power and advanced computing technologies (i-PACT).

[13] Guan, J., Yao, L., Chung, C. R., Xie, P., Zhang, Y., Deng, J., ... & Lee, T. Y. (2023). Predicting anti-inflammatory peptides by ensemble machine learning and deep learning. Journal of Chemical Information and System ing, 63(24), 7886-7898.

[14] Chen, B. (2025). Enterprise financial early warning based on ensemble learning and stacked generalization fusion algorithm system . Journal of Computational Methods in Sciences and Engineering, 14727978251361840.

[15] Yuan, S., Li, H., Qian, X., Hao, M., Zhai, Y., & Xu, G. (2024). Efficient and Privacy-preserving Outsourcing of Gradient Boosting Decision Tree Inference. IEEE Transactions on Services Computing, 17(5), 2334-2348.

[16] Kouassi, A. K. F., Pan, L., Wang, X., Wang, Z., Mulashani, A. K., James, F., ... & Nyakilla, E. E.

(2023). Identification of Karst Cavities from 2D Seismic Wave Impedance Images Based on Gradient-Boosting Decision Trees Algorithms (GBDT): Case of Ordovician Fracture-Vuggy Carbonate Reservoir, Tahe Oilfield, Tarim Basin, China. Energies 2023, 16, 643.

[17] Altman, M. B., Wan, W., Hosseini, A. S., Nowdeh, S. A., & Alizadeh, M. (2024). Machine learning algorithms for FPGA Implementation in biomedical engineering applications: A review. Heliyon, 10(4).

[18] Nathani, N., & Manna, G. C. (2018). A Quality of Service Based System for Supporting Mobile Secondary Users in Cognitive Radio Technology. In Cognitive Radio in 4G/5G Wireless Communication Systems. IntechOpen.

[19] Vekariya, D., Kannan, M. J., Gupta, S., Muthusamy, P., Mahajan, R., & Pandey, A. K. (2022). Recommendation System -Based 5G Network and Cognitive System of Cloud Data with AI Technique in IOMT Applications. International Journal of Communication Networks and Information Security, 14(3), 239-256.

[20] Zhao, F., Zeng, Y., Zhang, X., & Zhou, K. (2018, December). Research of radio frequency channel occupancy prediction based on decision tree. In 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018) (pp. 304-309). Atlantis Press.