iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.77181

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

Comparative Analysis for Heart Disease Prediction Using Extra Trees Classifier

Yash Soni[1], Akhilesh A Waoo[2]
Department of C.S.E, AKS University, Satna (M.P), India

Abstract: ***As one of the global outbreaks of modern innovation, the healthcare system poses a significant challenge to public health. For it kills people with all too great frequency. Without doubt, early and accurate prediction of heart disease is especially important. Focusing on patient records, machine learning methods have also been employed to predict whether a patient has the above-mentioned diseases or not. A comperhensive comparsion has been conducted in this paper to evaluate the prediction of cardiac related conditions using various learning-based models, such as decision tree, random forest, XGBoost, lightGBM and multilayer Percepton. First, finding related to the accuracy of various models are briefly discussed from existing literature. Then we experimentally evaluate our proposed Extra Trees Classifier using conventional classification metrics. Applications of Different Models. In this paper, we mainly use the synthetic_heart_disease_dataset, which contains clinical and demographic indicators widely used in cardiovascular risk assessment. Besides presenting the existing approaches, an ensemble-based Extra tress classifier is suggested to increase predicition accuracy by incorporating feature randomness and training strategies. Moreover, compared with conventional models, the proposed model has much lower variance and better generalization ability. Experiments show that our Extra Trees Classifier is significantly better than earlier methods such as Decision Tree, Random Forest, XGBoost, and Multilayer Perceptron. From the comparative analysis, it can be seen that ensemble learning methods achieve higher performance in predicting heart diseases. The route presented here could be an efficient tool for a clinical decision support system available for the earlier detection of heart disease.***

Keywords: ***Heart Disease Attitude, Group Learning, Kaggle Data Collection (cardiovascular disease patient network), State Competition Basis, Ensemble Studying Style, Extra Trees Classifier.***

I. INTRODUCTION

The human heart is one of the body's most essential organs, pumping blood and conveying essential nutrients such as oxygen to different portions of the human body.

If a disorder occurs in this cardiac section, it may lead to complications in various other organs. When the cardiovascular system malfunctions, it will have a bad effect on the normal function of different organs and can also pose a serious threat to life. Heart failure, a complete blockage, cardiac arrest, and other severe diseases are now causing severe fatalities all over the world. As a percentage of the total number of deaths globally last year, heart disease both closely appropriately significant contribution. [1]

With the increasing incidence of cardiovascular disease, early diagnosis and timely treatment have become two major concerns in today`s medical practice. Using older methods, heart disease is usually diagnosed through clinical tests combined with professional experience. However, these methods are time-consuming, expensive, and sometimes even subject to human error. Health Care data has been growing rapidly and this has created a strong need for intelligent machine-driven computer system to help doctors for making quick and accurate decision. In this setting, both data mining and machine-learning techniques have been globally used to locate patterns and extract useful information from clinical data for predication of illness. [1]

Using historical patient data, complex assumptions about future patient trajectories can be identified through predictive modeling approaches. The prediction of diseases is often considerably true and reliable. Decision trees, logistic regression, random forests, various classification models, and ensemble-based methods have been commonly used in cardiac risk assessments. Nevertheless, the accuracy of various approaches varies according to data, feature choice, and the process of applying different computational techniques. Therefore, comparing different ML approaches becomes necessary; by employing such techniques, one will get a clear idea of what approach is best for early-stage heart disease prediction. [12]

In many studies, the Kaggle dataset has been used as a standard benchmark. It includes various clinical and demographic covariates that are closely related to the progression of cardiovascular diseases. It includes various clinical and demographic covariates that are closely related to the development of cardiovascular diseases. By testing how different algorithms perform in systems developed using this data set,scientists can compare and measure their effectiveness under all similar conditions. A comparison needs to be

©IJRASET: All Rights are Reserved

made of all models. Ultimately, by doing this, the strengths and shortcomings of each model are shown in Fortune or otherwise. For real medical applications, this can make it easier to select an optimal algorithm with effective cost-benefit ratios. [4]

This paper presents a comparative analysis of multiple data-driven prediction models used for cardiac risk assessment. In this work, various widely used learning approaches are reviewed. Further more, an ensemble-based tree classification approach is proposed to improve precision, accuracy, and overall robustness[11]. The main focus of this study is to compare the prediction performance across different approaches and demonstrate the practical value of the proposed method in supporting accurate early diagnosis. [3]

II. LITERATURE REVIEW

Prediction of heart disease through machine learning has been an area of intensive study because of the emergence of clinical datasets along with the need for early diagnosis. Several researchers have used different computational approaches to improve the accuracy and robustness of predictions. This section summaries independent studies by considering the models adopted, their performances and identified limitations.

Nicholas et al. Interested in such a system, we.CASCADE.et al developed the heart disease forecasting system by utilizing Decision Tree classier. The research identified clinical features to stratify patients with and without heart disease. Simple and interpretable decision rules were derived using the DT; however, with respect to prediction accuracy it was relatively low (about 78%). The finding on that single-tree models tend to be overfitting and unable to capture complex relations in medical data.

Imam Husni Al Amin et al. introduced a decision-tree based method to improve prediction accuracy through integrating multiple trees, using Random Forest. The generalization and overfitting was reduced by the ensemble model as compared to a single Decision Tree. In 6 cases, we have 88% accuracy in the proposed method and satisfactory performance is seen. Though it performed better, the model also consumed more computational resources and was not very interpretable.

Sakyi-Yeboah et al. used advanced boosting such as XGBoost and LightGBM to predict heart disease. They concentrated the work on how to optimize model parameters, increase accurate classification. Experimental results showed that the boosting-based methods achieved a performance level of approximately 93%, outperforming traditional machine learning methods. Although such approaches perform well, they were complex to train and needed careful hyperparameter tuning.

Madhushree Meti and Dr. Lingraj MD proposed a XGBoost-based approach for assessing cardiac risk using clinical data. The research indicated that the gradient boosting can work on nonlinear relationships and get better prediction performance, which about 93% accuracy. Nevertheless, the study mainly focused on performance optimization and lacked discussion regarding model interpretability and clinical explainability.

F. Y. Ayankoya et al. presented an MLP Neural network model for heart disease prediction. The model had the capability to learn complicated information from clinical data with a performance level of approximately 94%. Although the advanced capability of prediction was demonstrated in this neural network model, the black-box characteristic may decrease interpretability, which represents a major issue in clinical decision-making approaches. The results show the reviewed works confirm that ensembles and neural network-based models consistently present a superior performance than typical classifiers for heart disease prediction problems. However, most existing methods either need the compromising boosting algorithms or are not robust and interpretable. These limitations suggest the requirement for an effective ensemble model with improved accuracy, stability, and simplicity. the present work presents a novel classifier based on Extra Trees, referred as the Enhanced-Randomness-Superensemble (ERS), aiming at improving prediction performance and generalization by introducing more randomness in the ensemble [1] [11].

III. METHODOLOGY

The quantitative ML approach in the present work to compare of cardiac risk assessment models. The methodology will be carefully set up so the analysis is as reproducible and consistent as possible, so that it can be fairly evaluated against existing literature. It includes the detailed process with dataset selection, exploratory initial data exploration, data preparation steps, model implementation and model performance evaluation. Evaluate the proposed Extra Trees-based ensemble model using the same experimental set up as earlier models related work [11].

A. Dataset Description

This paper uses an artificially created dataset of heart disease obtained from Kaggle, which is designed to be as close as possible to the real thing. This dataset contains organized data for patients in concerning of several reason associated with cardiac risk: demographic factors (age, gender), clinical variables (blood pressure, Cholesterol level, Heart rate), factors associated with lifestyle (Smoking status, Physical activity), along with indicators associated with past medical conditions (Diabetes, Hypertension,Family history of heart disease).[1][4]

The target variable in the dataset shows whether a patient is influenced by heart disease or not, and is known as a binary class attribute. Since the dataset clearly separates positive and negative cases, this dataset is ideal for applying a classification task in a machine learning model. Moreover, the variety and relevance of attributes present within the dataset allow meaningful learning tasks to be carried out for assessing cardiac risk.[1]

Attribute Category	Attributes Included
Demographic Factors & Physical Measures	Age, Gender, Height, Weight, Body Mass Index (BMI)
Lifestyle Factors	Smoking status, physical activity, exercise-induced angina
Clinical Measurements	Systolic BP, Diastolic BP, Heart Rate, Cholesterol, Increased level of sugar in Blood
Medical History	Hypertension, Diabetes, Hyperlipidemia, Family History, Previous Heart Attack
Target Outcome	Cardiac Risk is represented as either present (1) or absent (0)

Table 1. Dataset Attributes

B. Data Preprocessing

In order to prepare the data for machine learning analysis, various data preprocessing tasks have been considered. Missing numerical values are dealt with using the median imputation method. This ensures that the data is not altered by outliers. Additionally, the categorical variables are transformed using label encoding. This enables the machine learning algorithms to adequately handle the data. Finally, the data is split using an 80:20 stratified sampling ratio. This ensures that the proportion of classes is maintained. Various data preprocessing tasks are considered to make the data suitable before implementing machine learning analysis.

C. Feature Distribution Analysis (Histogram Analysis)

Histograms were applied to check for distribution characteristics of physical variables, including age, Weight, Height, and BMI, as well as binary clinical features, such as Hypertension, Diabetes, and so on. Inspection of these charts indicates that physical variables are generally evenly distributed, but the distributions of all medical history variables are bimodal, skewed, and heavily imbalanced (peaked around binary values), as is common in most medical datasets.

The observation that distributions are non-normal and there are discrete points within the data gives a clear empirical justification for the use of Extra Trees tree models, as these models can handle skewed distributions without being affected by issues caused by normalization. [11]

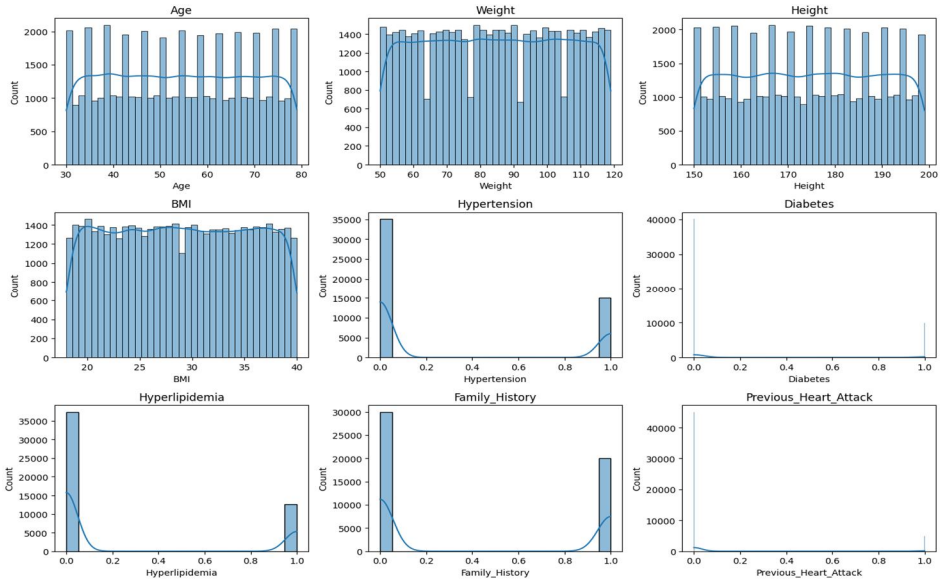


Fig 1: Histogram of Selected Clinical Features

D. Feature Relationship Analysis (Correlation Heatmap)

The correlation heatmap will help in identifying how different input variables are related along with whether there could be multicollinearity may exist among them. The heatmap indicates that there appears to be some relationship between some of the clinical variables, for instance, blood pressure,together with cholesterol, shows some relationship, while most of the remaining variables appear to be weakly correlated. It is reasonable to state that there does not appear to be any multicollinearity within the dataset; thus, this make it possible to apply tree- based models without removing any variables.

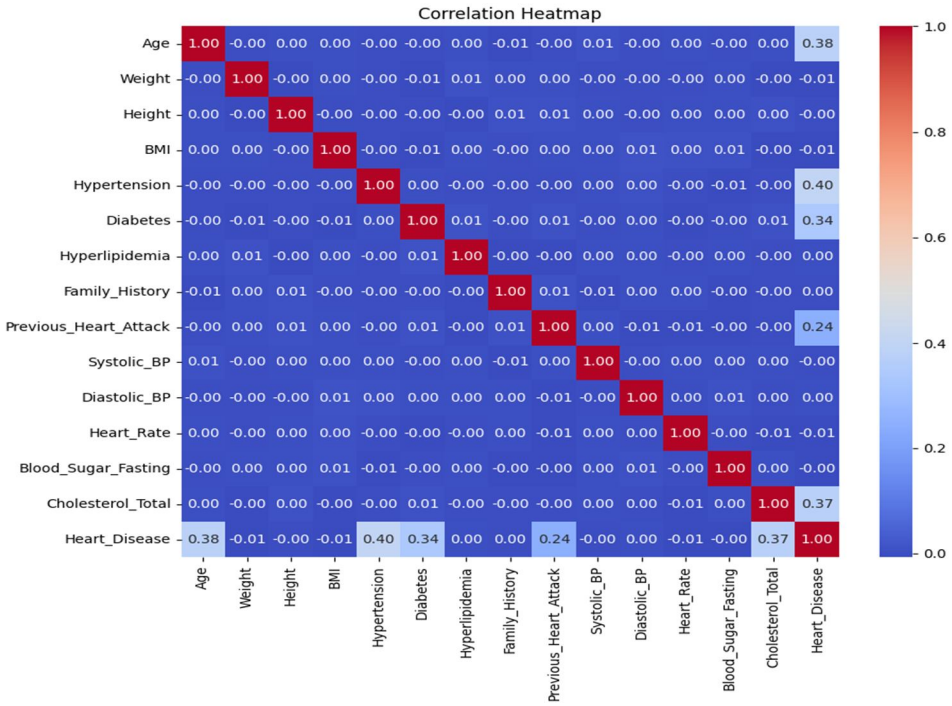


Fig 2*:* Correlation Heat map of Dataset Features

E. Data Preprocessing and Transformation

To make this dataset suitable for ML approaches, various pre-processing techniques were applied. For missing numerical values, median imputation was used, which reduces the impact of outliers and helps keep the original data pattern intact. For handling categorical variables, encoding was done, and these variables were represented in numerical form, enabling machine learning algorithms to handle them. The dataset was divided into separate training and testing groups through 80:20 stratified sampling, which ensures that a balanced proportion of cardiac risk and non- risk remains in both sets. [1]

F. Comparative *Analysis*

In this part of the study, a comparative analysis is carried out on ML- based approaches used to evaluate accuracy reported in existing literature in comparison with the experimental results achieved by the Extra Trees-based model. [11]

Table 2: Comparative Performance with Existing Studies

Author(s)	Model	Accuracy (%)
Nicholas et al.	Decision Tree	78
Imam Husni Al Amin et al.	Random Forest	88
Sakyi-Yeboah et al.	XGBoost / LightGBM	93
Madhushree Meti & Dr. Lingraj	XGBoost	93
F. Y. Ayankoya et al.	Multilayer Perceptron	94
Proposed Model	Extra Trees Classifier	96.17

G. Model Implementation

The performance of the proposed method will be evaluated by comparing it`s with results reported for various established ML algo approaches, including Decision Tree, Random Forest, XGBoost, LightGBM, and Multilayer Perceptron, as documented in earlier studies. These are used as benchmark approaches. [8]

The proposed model, an Extra Trees Classifier, on the Kaggle synthetic dataset for training. In the Extra Tress-based method, randomness is intentionally added to feature selection and split criteria, leading to the creation of many tree-based structures. This method can effectively reduce the variance and prevent overfitting to improve generalization performance. [11] [4]

H. System Workflow

As soon as the user engages with the interface, it will begin to run without any awkward lag or artificiality involved. Once the user interacts with the interface, the system begins to run smoothly without noticeable delay. After the user enters the required patient details, the data is captured and sent for preprocessing . At this stage, an imputation steps is applied to handle any missing values in the dataset, ensuring that the input data remains consistent. The cleaned and processed data is then prepared in the required format and passed to the Extra Tree-based classification model. Based on this processed input, the system generates a prediction, and the user receives a clear Yes or No result for cardiac risk within a seconds. [11]

I. System Workflow Description

The user-provided data is first preprocessed using imputation and encoding so that it is aligned with the training structure. After this, the processed data is passed to the Extra trees-based classification approach, which produces a binary result indicating whether cardiac risk is present or not. [1] [11]

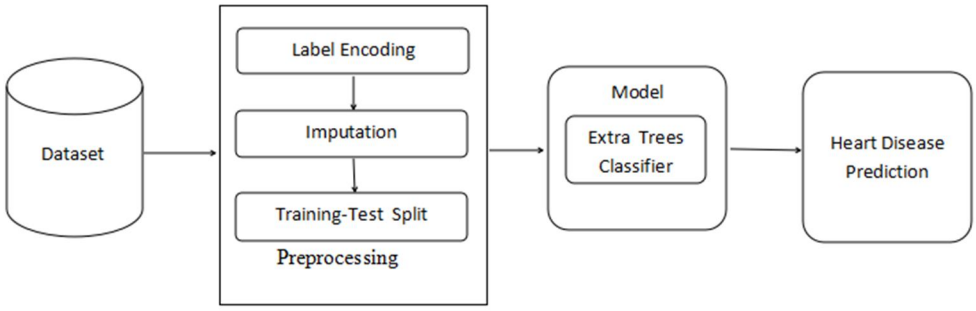


Fig 3: System Workflow Diagram

IV. RESULTS AND DISCUSSION

The experimental evaluation shows that ensemble-based learning methods are highly effective for assessing cardiac risk using structured clinical datasets. In this work, an Extra Trees-based approach with improved performance is introduced, achieving an accuracy of 96.17% on this datasets. [12]

The strong performance of the proposed approach can be attributed to the ensemble strategy, which increases randomness during feature selection and the choice of split points for the trees in the proposed approach. This added randomness helps the method better capture non-linear relationships among patient attributes related to demographic and lifestyle factors. [11] [12]

From a clinical perspective, reducing incorrect predictions is essential, particularly in cases where false negatives may occur. The balance between sensitivity and precision, as reflected by the chosen evaluation metrics, confirms that the proposed system in a well-controlled and consistent manner. Consequently, the system can be considered dependable for deployment in real-time scenarios focused on early disease detection.[3] When comparing models discussed in earlier studies, it can be seen that traditional decission trees are limited in how much complexity they can capture. Random forestes improve generalization but come with higher computational cost. Boosting techniques such as XGBoost and LightGBM deliver strong performance, though they require careful tuning of parameters. Neural networks also perform well, but their lack of interpretability makes them harder to explain in clinical settings. In comparison, the Extra tree-based approach offers a balanced solution across these aspects. Overall, the experimental findings show that ensemble learning with randomized decision trees provides an effective and dependable method for cardiac risk assessment. [2]

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

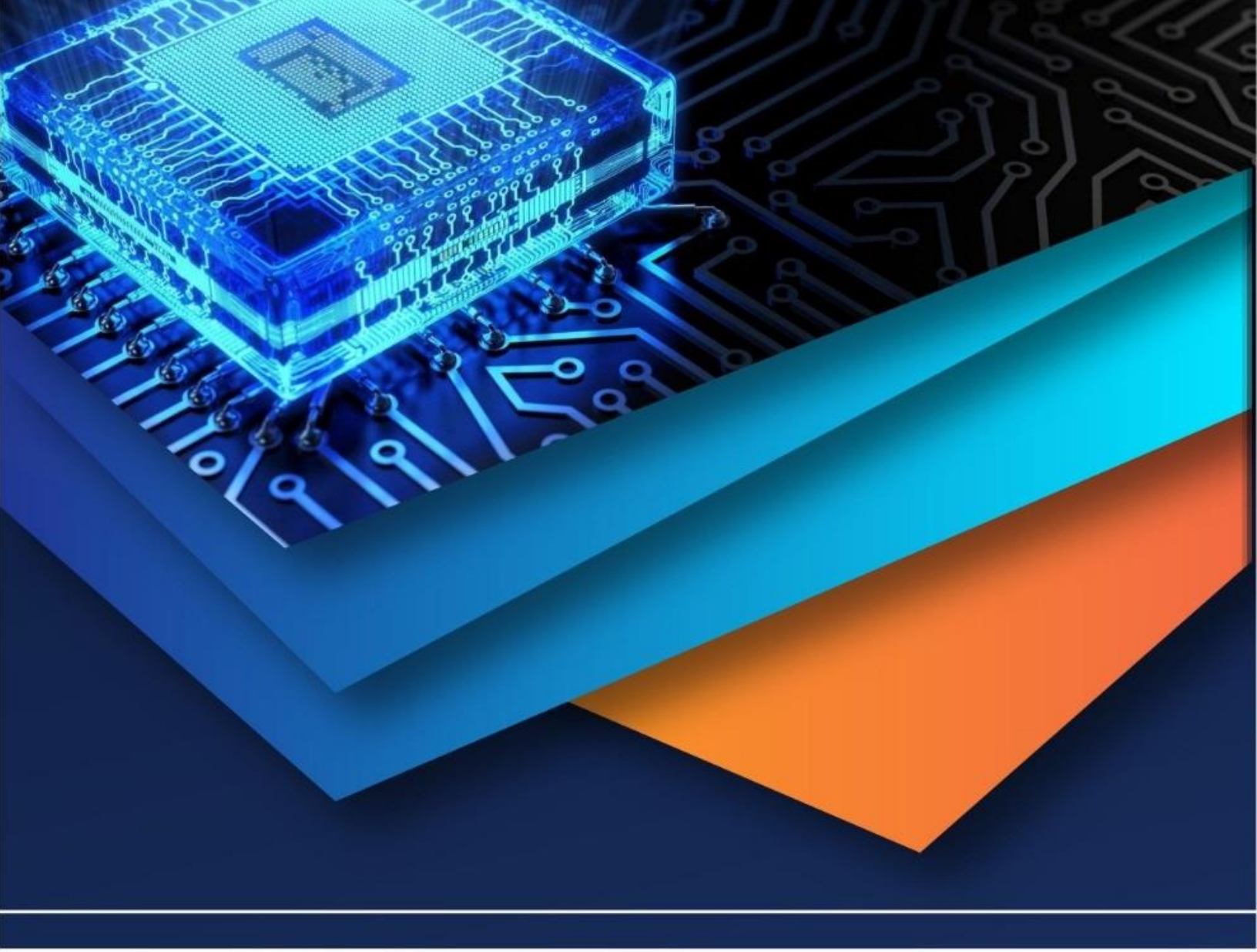
V. CONCLUSION

This paper presents a detailed comparison of how different ML-based approaches perform in assessing cardiac-risk. [4] With the help of a synthesized Kaggle heart disease dataset, this paper explores an ensemble learning-powered Extra Trees Classifier. [12]
The developed model achieved an accuracy of 96.17%, outperforming many perviously reported methods in the literature. This improvement can be largely attributed to the ensemble structure adopted in the Extra Trees based approaches, which introduces additional randomness during training and helps reduce overfitting.
By combining multiple randomized decision trees, the extra trees techniques is able to capture complex patterns in the data while maintaining good generalization performance. Overall, the results demonstrate that ensemble learning methods, particularly those implemented through the extra trees framework, are effective and reliable for supporting cardiac risk assessment in clinical decision-making. [11] [12]

REFERENCES

[1] Nicholas, G.Hoendarto, and J.Tjen, "cardiac risk Prediction with Decision Tree," Social Science and Humanities Journal, vol. 9, no. 1, pp. 6451-6457, Jan. 2025, doi: 10.18535/sshj. v9i01.1444.
[2] H. Al Amin, S. Wibisono, E. Lestariningsih, and M. L. M.A, "Optimizing cardiac risk Prediction with Random Forest and Ensemble Methods," COGITO Smart Journal, vol. 11, no. 1, pp. 180-[Page Numbers], June 2025.
[3] Sakyi-Yeboah et al., "cardiac risk Prediction Using Ensemble Tree Algorithms: A Supervised Learning Perspective," Applied Computational Intelligence and Soft Computing, vol. 2025, Art. ID 1989813, 18 pages, 2025, doi: 10.1155/acis/1989813.
[4] Xia, "Influencing Factors and Prediction of Heart Disease," Highlights in Science, Engineering and Technology, vol. 123 (BFSPH 2024), pp. 586-592, 2024.
[5] Jiang, "cardiac risk Prediction Using Machine Learning Algorithms," Master's Thesis, University of California, Los Angeles, 2020.
[6] M. Meti and Dr. Lingraj, "Heart Boost: Clinical Data-Driven cardiac risk Prediction Using XGBoost," International Research Journal on Advanced Engineering Hub (IRJAEH), vol. 3, no. 9, pp. 3517-3525, Sep. 2025, doi: 10.47392/IRJAEH.2025.0517.
[7] A.T L, A. BK, and D. D, "cardiac risk Prediction Using Logistic Regression," Indian Journal of Computer Science and Technology, vol. 4, no. 2, pp. 356-359, May-Aug. 2025, doi: 10.59256/indjcst 20250402048.
[8] F. Y. Ayankoya et al., "cardiac risk prediction using machine learning model," Global Journal of Engineering and Technology Advances, vol. 24, no. 2, pp. 036-049, 2025, doi: 10.30574/gjeta 2025.24.2.0223.
[9] B. Shehzadi et al., "cardiac risk Prediction Statistical Analysis and Classification of cardiac riskUsing Clinical Parameters," Social Sciences & Humanity Research Review, Jan.-Mar. 2025, pp. [Page Numbers], ISSN: 3007-3162.
[10] Y. Chen, "Predicting cardiac risk Using Machine Learning: Analysis and New Insights," Dean&Francis [Journal Title Implicit], pp. [Page Numbers], ISSN: 2959-6157.
[11] S. Chaudhari, C. S. Gautam, and A. A. Waoo, "Optimizing cardiac risk Prediction Accuracy using Machine Learning Models," International Journal of All Research Education and Scientific Methods (IJARESM), vol. 12, no. 6, June 2024, pp. [Page Numbers], ISSN: 2455-6211.
[12] V. V. R. Karna et al., "A Comprehensive Review on cardiac risk Prediction using Machine Learning and Deep Learning Algorithms," Archives of Computational Methods in Engineering, pp. [Page Numbers], 2024, doi: 10.1007/s11831-024-10194-4.
[13] Anjali Regala, SD Ravikanti, and RG Franklin, "Design and implementation of cardiac risk prediction using naive Bayesian", International conference on trends in electronics and informatics (ICOEI), pp. 292-297.
[14] VV Ramalingam, A Dasapopath and MK. Raja, "cardiac risk prediction using machine learning techniques-a survey", International journal of Engineering & Technologies, Vol. 7, no. 5.8, pp. 684-7.
[15] E. I. Elsedimy, S. M. M. Abo Hashish, and E. Alzgara, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization", Multimedia Tools and Applications, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)