Cardiac Risk Prediction Using Extra Trees-Based Classifier

Yash Soni[1]; Akhilesh A. Waoo[2]

[1,2]Department of C.S.E, AKS University, Satna (M.P.) India

Publication Date: 2026/02/07

Abstract: Heart Disease, also known as CVD, represents one of the major health issues worldwide. There is an urgent need for the availability of risk prediction systems that are reliable yet non-invasive in nature to facilitate timely clinical interventions. This study aims to explore how well the Extra Trees Classifier performs in predicting the risk of heart disease. The Extra tree-based method represents an advanced ensemble ML approaches that was trained on a comprehensive dataset containing 20 key clinical and lifestyle attributes of patients. In addition, this approach was carefully tuned and thoroughly evaluated to ensure reliable performance. Feature analysis plays the main role in this paper by ranking the most influential predictors of CVD risk according to their importance. This allows drawing data-driven conclusions that can inform clinically oriented risk assessment analyses. Based on the results described above, the Extra Trees Classifier is effective and reliable for predictive cardiology and thus serves as a good starting point for improved clinical decision-making.

Keywords: *Cardiac Risk Assessment, Machine Learning, Extra Trees-Based Approach, Ensemble Learning, Clinical Parameters, Predictive Modeling, Feature Importance, Kaggle Dataset.*

How to Cite: Yash Soni; Akhilesh A. Waoo (2026) Cardiac Risk Prediction Using Extra Trees-Based Classifier. *International Journal of Innovative Science and Research Technology*, 11(1), 3210-3216. https://doi.org/10.38124/ijisrt/26jan1615

I. INTRODUCTION

Cardiovascular disease is a significant health issue affecting individuals globally. It presents a considerable rate of illness and death. The diagnosis must be made as early as possible to reduce mortality rates and increase the success rate of treatment. It also enables the physicians to take informed and timely decisions. The conventional diagnostic procedures fully depend on the physician's skills, manual interpretation of symptoms, and invasive tests. These result in delayed diagnosis and variable results. Increasing the use of technology in healthcare and availability of large clinical datasets, ML is emerged as an effective tool for cardiac risk assessment through data analysis. This technology can unmask hidden patterns, evaluate complex clinical variables, and make fast and reliable predictions, thus enabling physicians to diagnose early. A wide range of conventional and advanced techniques, including tree-based methods, regression models, ensemble approaches, kernel-based classifiers, and neural architectures, are commonly applied in medical data analysis to categorize patients based on their risk levels. Among these, ensemble learning approaches typically lead to better performance due to their stability and ability to manage noise in the data. Various steps in preprocessing, such as label encoding, imputation, feature selection, and normalization, enhance the quality of data and hence improve the accuracy of these models. Recently, more attention has been paid to making models interpretable and user-friendly using web-based platforms such as Streamlit, along with the integration of datasets from multiple clinical sources, To help make systems for cardiac risk assessment practical and accessible [10].

This work proposes the use of an extra trees-based ensemble approach, which is know for its high accuracy and strong resistance to overfitting. The proposed system uses a publicly available clinical dataset sourced from kaggle and achieves an accuracy 96%. The chapter now presents a brief overview of related research conducted by various scholars in this field. The survey of the following literature reviews contributions from different researchers and presents the advances in the machine learning-based prediction of diseases [6].

II. LITERATURE REVIEW

- Nicholas et al. proposed a Decision Tree-based prediction model using seven clinical parameters. Their system showed moderate results but faced overfitting problems due to limitations inherent in single-tree models.
- Imam Husni Al Amin et al. examined multiple ensemble-based classification techniques within their study. Their findings indicated that one of the tree-based ensemble methods produced strong predictive results; however, its limited level of interpretability reduced its suitability for decision-making.
- Xia conducted an analysis using logistic regression, random forest, and support vector machines, where as random forest showed good performance but offered less transparent explanations in high-dimensional settings.

- Jiang evaluated logistic regression, random forest, xgboost, and neural network models, nothing that each achieved accuracy level above 80%. The study highlighted that careful feature engineering played a key role in improving overall model performance.
- Meti and Lingraj proposed an XGBoost-based clinical prediction framework, which had 93% accuracy but required high computational power.
- Anusha et al. considered that Logistic Regression, even though interpretable and efficient, was unable to capture important and complex interactions between clinical features.
- Ayankoya et al. Compared logistic regression, random forest, and multilayer perceptron models, and found that MLP achieved the strongest results following parameter tuning.
- Shehzadi et al. explored hybrid statistical and ML methods over 1,025 clinical records and reported improved risk stratification, but lacked ensemble predictive models.
- Chaudhari et al. worked on optimization techniques and hyperparameter tuning, which led to improved accuracy on different ML models.
- Karna et al. Reviewed various advanced modeling methods and stressed the importance of good quality data and hybrid approaches.

III. METHODOLOGY

- *Research Design:*

This study adopts a quantitative approach that applies a supervised ML workflow for cardiac risk assessment using an extra trees-based approach. The main objective of this method is to transform raw patient case records into a practical and reliable predictive system capable of identifying whether an individual is likely to face potential cardiac-related risks. To achieve this, the workflow involves a detailed understanding of the data, careful preparation and cleaning of the dataset, systematic training and evaluation of the model, and final preparation of the system for deployment in a real-world setting. Every stage has been precisely drafted to ensure delicacy, conception, and practical connection [4].

- *Dataset Description:*

The system was trained using structured clinical data from Kaggle, that includes about 20 features related to patient demographics, lifestyle habits, medical history, and basic clinical measurements. Collectively, these describe the physical health of a case and possible threat factors for cardiovascular problems [2]

- *Key Data Categories:*

- *Physical and Demographic Character:*

This group includes basic physical details such as body weight, age, height, body mass index (BMI) and gender. These Features describe the general physical profile of an individual and are useful for understanding overall health conditions [5].

- *Lifestyle-Related Factors:*

These factors describe a person`s daily habits and lifestyle choices. Smoking, alcohol consumption, level of physical activity, eating habits, and stress level are some of these factors. Such variables are often associated with cardiovascular health outcomes [11].

- *Clinical Measurements:*

This category covers important medical reading, including systolic and diastolic blood pressure, heart rate, fasting blood sugar level, and total cholesterol. These signs reveal the patient’s physiological condition.

- *Medical History Indicators:*

It represents past or existing health conditions such as hypertension, diabetes, high cholesterol, family history of cardiac conditions, and previous cardiac events. These factors play an important role in assessing cardiovascular health across different populations [8].

- Target Attribute: Heart Disease is a binary variable, where:

1 means heart disease present & 0 means heart disease absent

- *Data Preprocessing Overview:*

Various preprocessing steps were implemented, among them were fixing missing values, changing categorical features into numerical form, and dividing the dataset into training and testing subsets. This is the kind of preprocessing that is very important for improving data quality and making sure that the prediction model is accurate and stable [6].

- *Manage Missing Values:*

Missing numerical values in the dataset were handled using medain imputation. This approach helps reduce the influence of extreme values and keeps the data distribution stable. Since the proposed prediction approach requires complete numerical inputs for each record, this step is necessary to ensure proper model functioning [8].

- *Encoding Categorical Features:*

The use of label encoding, categorical variables like gender, smoking status, and diet type were transformed into a numerical format suitable for exact understanding by the machine learning algorithm [2].

- Train–Test Split: The processed data was divided into:

- ✓ 80% Training set
- ✓ 20% Testing set

Stratified sampling was applied to maintain the same proportion of positive and negative cases in both the training and testing sets ensuring a fair evaluation [8].

- *Evaluation Metrics:*

The following metrics are used to evaluate how well the model performs:

- Accuracy – proportion of correct predictions
- Precision – how many predicted positives were correct
- Recall – ability to detect heart disease cases
- F1-score – balance between precision and recall
- ROC–AUC – overall separability between positive and negative classes [11].
- Confusion Matrix – detailed breakdown of prediction outcomes

IV. SYSTEM ARCHITECTURE AND WORKFLOW

➢ *System Architecture Overview:*

The project will have three layer-based architecture, which would make the structure more maintainable and scalable. This separation also means there can be independent development and upgrade of different layers.

- *Data Tier:*

The data layer acts as the base storage level of the system. It is used to store the main dataset, synthetic_heart_disease_dataset.csv. This layer also keeps the preprocessing files needed to handle the data, ensuring that cleaning and processing steps remain consistent and the system functions smoothly [2].

- *Model Tier:*

The Model layer holds the main logic of the application and is responsible for using the trained ML models saved as extra_trees_model. pkl. This level covers only the execution of the prediction logic and all preparation feature processing occurring just before diagnosis generation.

- *Presentation Tier:*

This tier is the user interface of the system. It is realized as an application implemented by Streamlit (app.py). This layer is designed to collect patient inputs provided by the user and further show the results of the prediction in a clear and instantaneous way. It will provide direct interaction with the end-user.

➢ *System Components:*

- *Core Files:*

The core functionality of the system is implemented through a set of essential files. The primary data source used for this purpose is synthetic clinical dataset stored in CSV format. Preprocessing is managed by an imputer. pkl, which has the median-imputation pipeline that is needed. The trained classifier, saved as extra_trees_model.pkl, serves as the predictive core of the system. The whole Streamlit prediction system that glues it all together is described in app. py, and requirements. Txt is the list of all external libraries required to recreate the environment.

- *Preprocessing Component:*

The preprocessing component plays an important role in making sure that the raw input provided by user is converted into a form suitable for model processing. Its tasks include applying median imputation to patient data whenever some values are unavailable, label encoding of categorical columns, and, in general, transforming the input from a live user to match the exact formatting used during model training [4].

- *Predictive Model Device:*

The predictive model (Extra Tree) component is loaded in a serialized form, ensuring quick startup of the application, and the real-time prediction can be offered. The system process the input data and produces a simple yes/no decision related to the patient`s cardiac condition.

- *User Interface Component:*

The User Interface Component is a user-friendly Streamlit web app. It is structurally programmed by manual input of patient characteristics, and the predicted cardiac condition outcome is immediately displayed on screen. The UI is designed to be as simple and intuitive, navigation is easy, and the application is designed to be easy to use for both medical professionals and non-technical users,

➢ *System Workflow:*

Once the user interacts with the interface the system runs smoothly and efficiently. When patient details are entered into streamlit application the input information is collected and immediately sent to the preprocessing stage. During this step, missing values in the dataset are handled automatically, and additional values may be generated if required [10]. The cleaned data is then converted into same format used during model training so that it can be processed correctly. The processed data is passed to the prediction approach, and the result is displayed clearly. The user receives a cardiac condition: yes/no indication within a few seconds.

➢ *System Workflow Diagram:*

- *Explanation:*

The workflow begins with clinical dataset, which enters the preprocessing stage. During this satge:

- ✓ Label Encoding converts categorical entries into numerical form [5].
- ✓ Imputation replaces missing values using the median strategy [2].
- ✓ Train–Test Split divides the dataset for model training and performance evaluation [4].
- ✓ The cleaned data is then passed to the processed extra tree-based approach, where a trained prediction model is created. This model is subsequently used to produce cardiac Disease Predictions [1] for new or unseen patient inputs.

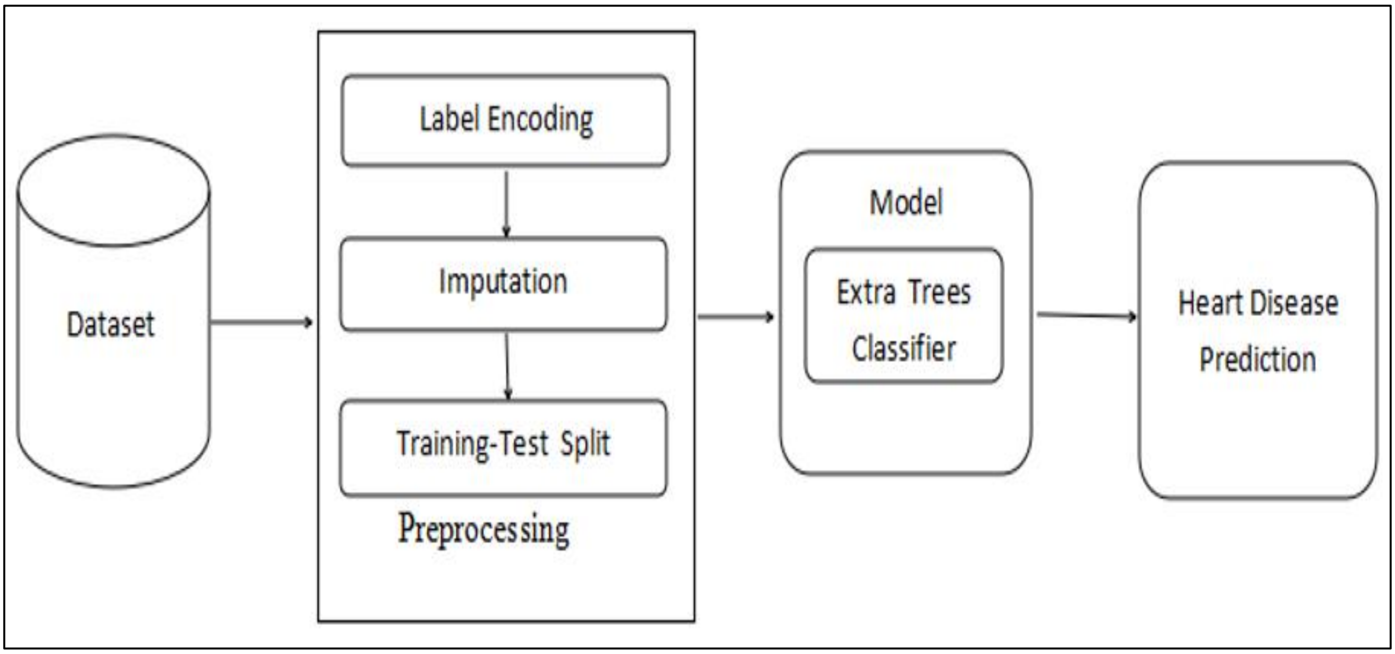

Fig 1 System Workflow Diagram of the Proposed Cardiac Assessment Approach

➢ *Development Tools & Environment:*

The following tools were used:

- Python version 3.10
- Scikit-learn (model training, preprocessing)
- Pandas & NumPy (data handling)
- Streamlit (web interface)
- Pickle (saving and loading trained model)
- Matplotlib/Seaborn (visualizations)

➢ *Deployment Using Streamlit:*

This application, developed in Streamlit and offers a simple, interactive and user-friendly interface for entering patient details. Once the information is provided the system quickly process the input and generates prediction results without noticeable delay. To ensure consistency application loads the same saved processing components and trained model that were used during development. This approach helps maintain uniform data handling and reliable predictions during actual use, preventing inconsistencies between the training and deployment stages [2].

V. OUTCOME

The results of the proposed method are discussed in this section. The model was tested using common evaluation measures such as accuracy, precision, recall, F1-score and ROC-AUC. These outcomes help explain how effectively the method separates individuals at risk from based on clinical data[4].

➢ *Model Performance Overview:*

The proposed method was trained using a kaggle clinical dataset after applying basic preprocessing steps such as label encoding and medium imputation. The dataset was divided into training and testing portion using an 80:20 split, and the model demonstrated strong predictive results. All evaluation metrics are shown below [9].

Table 1 Performance Metrics of the Proposed Approach

Metrics	Score (%)
Accuracy	96.17
Percision	97.17
Recall	94.52
F1 Score	95.83
AUC-ROC	99.60

This table shows that proposed (extra trees -based) method achieved high scores across all metrics, indicating strong generalization capability and reliable decision-making.

➢ *Confusion Matrix Analysis:*

The confusion matrix gives clear insight with regard to the strengths of the model in correctly classifying positive and negative cases. Extra Trees mis-classified only a very small number of instances, thus proving high sensitivity and specificity.

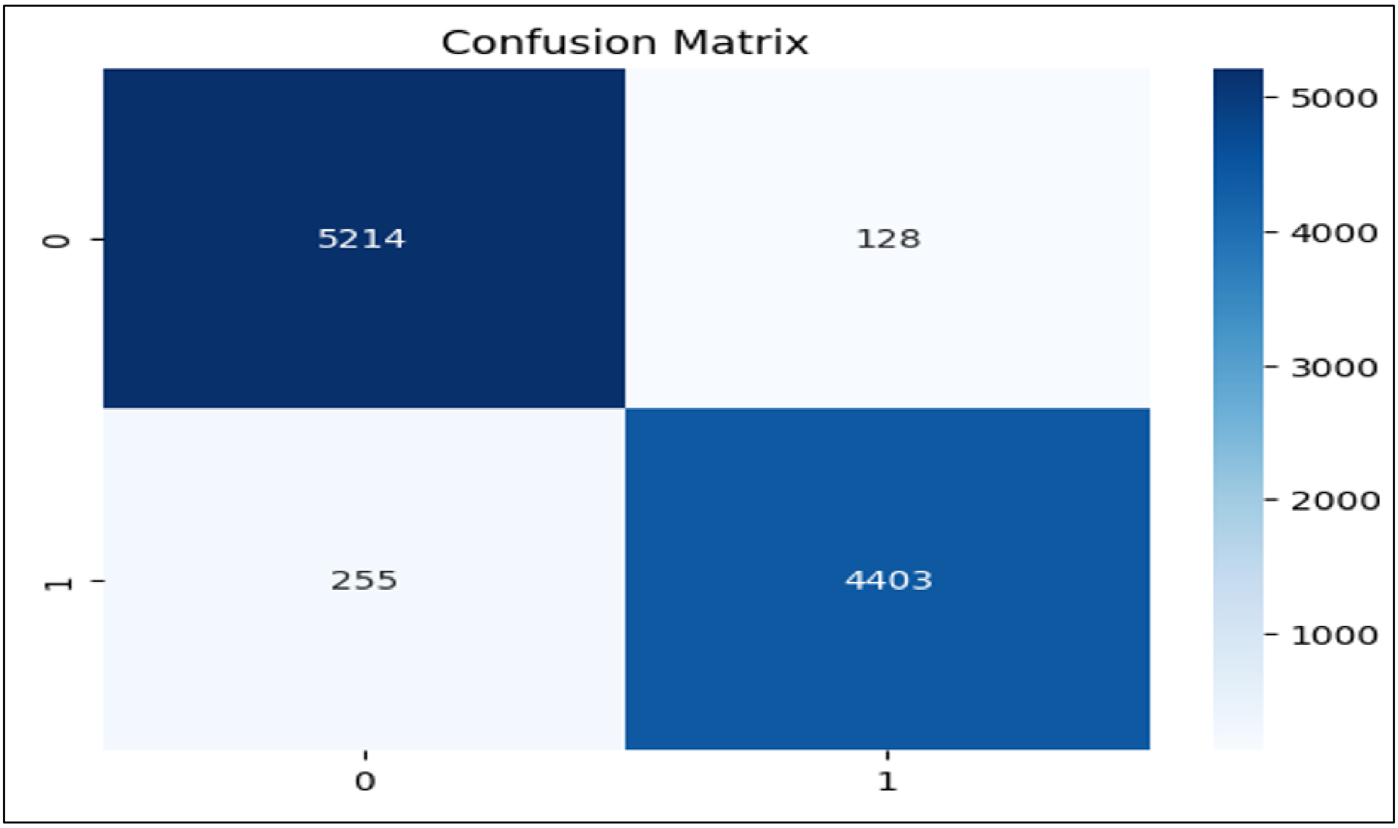

Fig 2 Confusion Matrix of Extra Trees Classifier

The matrix indeed confirms that most heart disease cases are identified by the model while keeping the rate of false negatives low, which is the primary requirement for any diagnostic medical application, because overlooking a positive case may lead to delayed treatment [10].

- *ROC Curve:*

This graph explains the model`s performance at different threshold values of the data and it reaches the top-left corner of the plot [1].

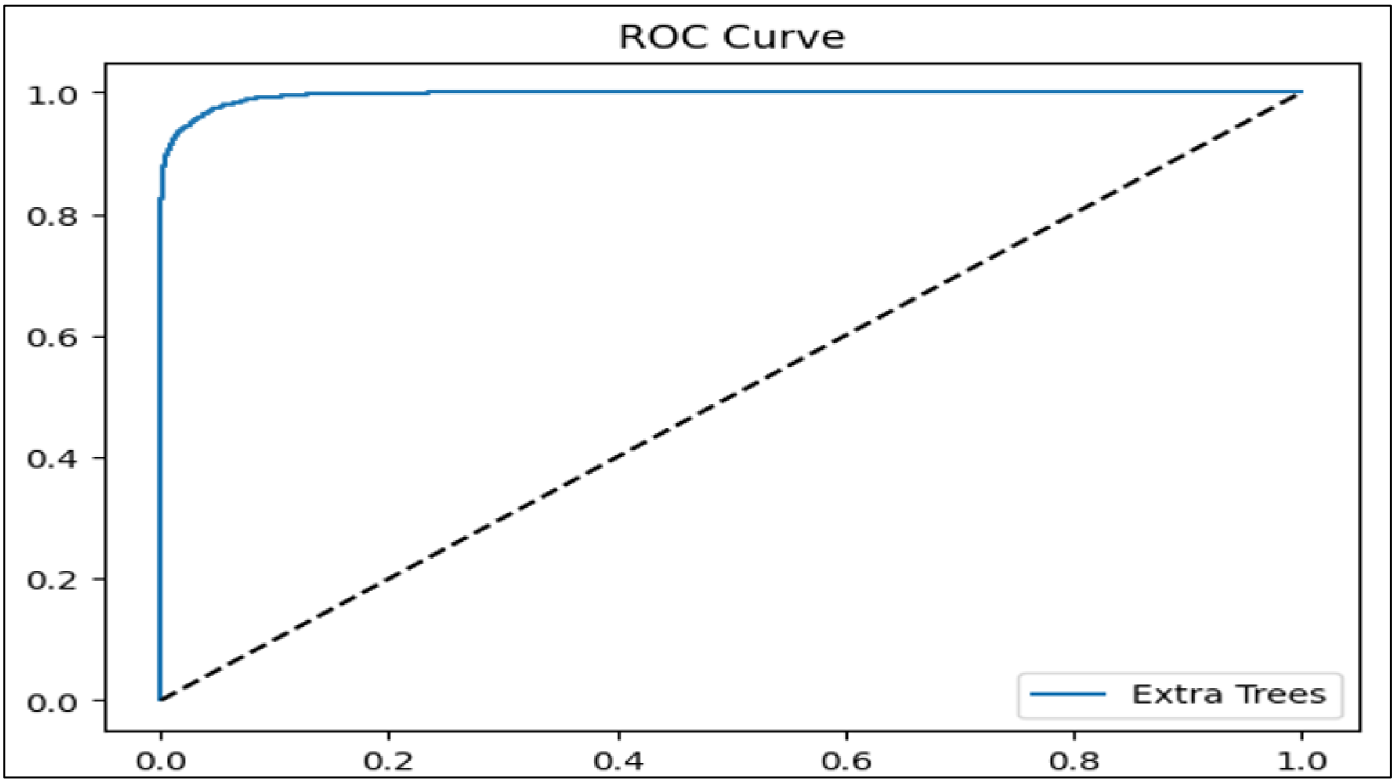

Fig 3 ROC Graph Showing the Model`s Performance

➢ *AUC Score:*

An AUC value of 0.996 for the proposed approach shows that it can clearly separate the classes and performs better many commonly used ML approaches.

➢ *Precision–Recall Curve:*

The percision-recall curve is useful for medical data because correctly identifying positive cases is important. The model shows a high curve which indicates that it performs well in both precision and recall [6].

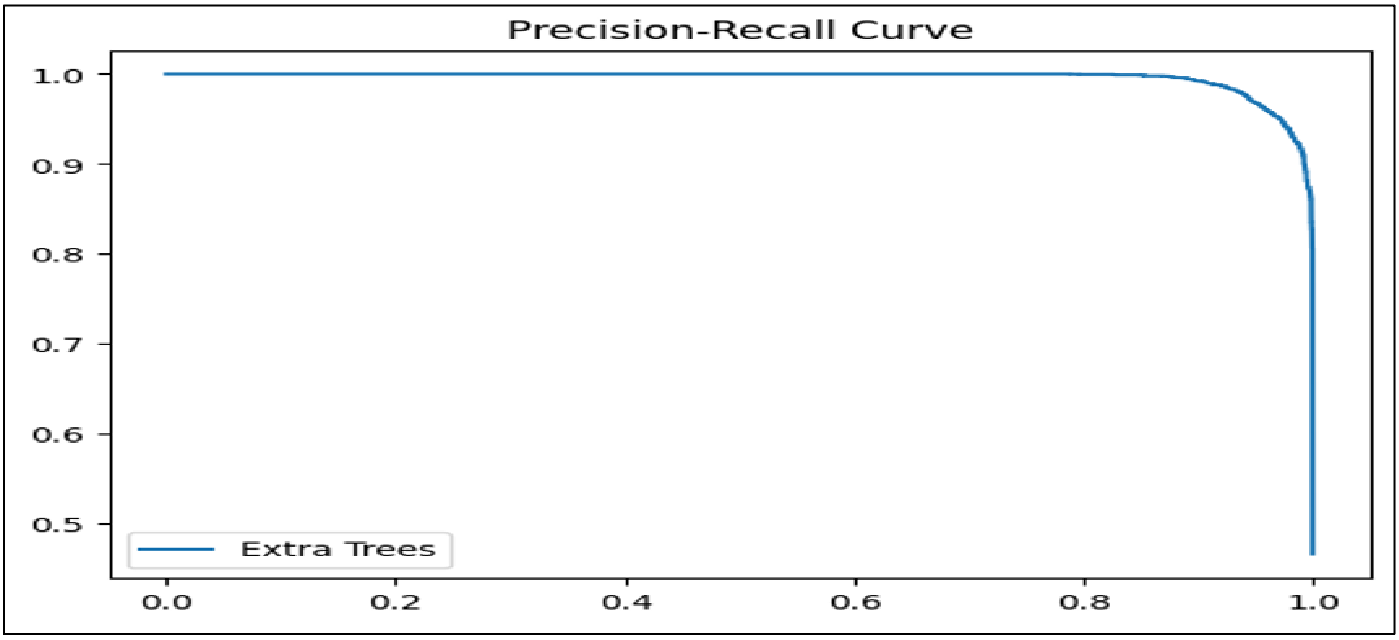

Fig 4 Precision–Recall Curve

When the threshold value is changed the model`s performance remains almost the same. This indicates that the model is stable and produces reliable predictions instead of fluctuating results.

➢ *Comparison with Existing Studies:*

A detailed comparison with results reported in earlier studies shows that the proposed extra tree-based method achieves performance that is equal or better than many commonly adopted ML approaches. These approaches include decision tree models, random forests, logistic regression, multilayer perceptron networks, and gradient boosting techniques, which are frequently used in related in related research. Overall, the finding demonstrates the strong predictive capability of the proposed method when compared with established ML approaches [8].

Table 2 Comparison with Existing Studies

Researcher	**Proposed Algorithm**	**Accuracy (%)**
Nicholas	Decision Tree	78
Imam Husni Al Amin	Random Forest	88
Sakyi-Yeboah	XGBoost / LightGBM	93
Madhushree Meti1 & Dr. Lingraj	XGBoost	93
F.Y. Ayankoya	Multilayer Perceptron	94
Proposed Model	Extra Trees-based Method	96.17

➢ *The Good Performance of this Method is Mainly Due to the Following Factors:*

- Using several random trees.
- Reduced overfitting
- Accessibility of clinical interactions with the patient.
- This makes Extra Trees a good model for predicting cardiology applications.

VI. CONCLUSION

Assessing cardiac risk has become an important focus in modern healthcare, as early detection can support timely and lifesaving treatment decisions. In this study, an effective cardiac risk assessment system based on extra tree-based approach was developed using a clinical dataset obtained from kaggle, along with thorough data preprocessing steps such as label encoding and median imputation [5].

➢ *The Model Obtained:*

- 96% accuracy
- High precision, recall, and F1-score
- Exceptional AUC-ROC of 0.99

These finding demonstrate that the extra trees-based method is highly effective in identifying Potential cases of

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1615

cardiac risk. Its ensemble structure helps reduce overfitting and allows better handling of complex interactions among features that are difficult for traditional methods to capture. In addition, a user-friendly web interface developed using Streamlit supports practical useby medical professionals as well as end users [6].

This study also introduces a comprehensive and practical end to end ML workflow, covering all stage from data preparation to system deployment, and provides a dependable solution for early assessment of cardiovascular risk [2].

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.