International Journal of Innovative Research and Scientific Studies, 8(2) 2025, pages: 276-291

ISSN: 2617-6548

URL: www.ijirss.com

Diagnostic accuracy enhancement for cardiovascular disease prediction using dual optimized feature selection and fuzzy-based deep learning model

Lijetha C Jaffrin[1*], J. Visumathi[2]

[1,2]Department of Information Technology, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, India.

Corresponding author: Lijetha C Jaffrin (*Email: lijethacjaffrin@veltech.edu.in*)

Abstract

The most frequent reason for mortality in many nations is cardiovascular disease. Past experience and current clinical testing of diagnosing patients with comparable symptoms are frequently used by doctors to make the diagnosis of cardiovascular disease. Heart disease patients need to be diagnosed as soon as possible, treated as soon as possible, and closely monitored. Numerous data mining techniques have already been employed to diagnose and forecast heart conditions in order to meet their objectives. To help doctors forecast and detect cardiovascular disease, deep learning and machine learning may provide a stronger basis for prediction and decision-making from healthcare data sectors around the world. The aim of the research is to propose an accurate algorithm for the prior prediction of heart disease using dual feature selection methodologies. The features are selected by utilizing feature selection methods such as LASSO and MR-MR. The early prediction of cardiovascular disease (CVD) is performed using an improved fuzzy-based TabNet deep learning model with a fuzzy foundation. The dataset is considered from the Kaggle Heart Disease Repository. The area under the curve (AUC) for the recursive operating characteristic curve is estimated for the proposed algorithm. Additionally, error measures like mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are estimated for the model's predictions, and the magnitude of MSE is 0.038, MAE is 0.180, and RMSE is 0.195, respectively. The best approach for classifying and predicting CVD is the integration of the enhanced TabNet algorithm and fuzzy foundation. The suggested approach lowers costs and improves medical care for predicting heart illness. The strength of the suggested model is relatively satisfying, and it reveals good accuracy in predicting indications of heart disease in a specific individual when compared with previously implemented classifiers.

Keywords: Cardiovascular disease, Deep learning, Feature selection, Fuzzy decision-making, Prediction, TabNet.

DOI: 10.53894/ijirss.v8i2.5145
Funding: This study received no specific financial support.
History: Received: 16 January 2025 / **Revised:** 17 February 2025 / **Accepted:** 24 February 2025 / **Published:** 6 March 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

It was estimated that cardiovascular diseases (CVDs) were responsible for the leading causes of deaths worldwide. In a recent study by the Global Burden of Disease (GBD) and the World Health Organization (WHO), death rates due to cardiovascular disease were increasing worldwide every year [1, 2]. WHO data indicated that CVD was predicted to impact over 23.6 million individuals by 2030. Consequently, the administration of suitable treatments and early and precise diagnosis leads to the reduction of deaths from cardiovascular illnesses. It is crucial for people at high risk of heart disease to gain access to these services [3]. A major challenge for health organizations is to provide patients with affordable, high-quality clinical treatment. Accurate patient diagnosis and suitable therapy identification are both necessary for delivering high-quality care while avoiding inaccurate diagnoses [4]. Early detection of CVD also reduces mortality and expenses.

As a research curiosity for two decades, the use of Machine Learning (ML) has spawned a variety of commercial applications. It was a branch of artificial intelligence (AI) that fits models to data using statistical methods to recognize relevant patterns in huge, complicated, unstructured information [4]. It was a broad, interdisciplinary field with foundations in computer science, statistics, mathematics, cognitive analytics and among other fields [2]. To predict future data with high accuracy, historical data was used by Machine learning algorithms. Recently, ML techniques had a significant impact on the healthcare sector. By applying machine learning techniques to the healthcare sector, improvements such as accurate models of prediction, novel therapeutic strategies, Clinical Decision Support Systems (CDSS), drug discovery, and lowering medical care costs were possible [5, 6]. Recent developments in big data processing and the collection of everyday healthcare data had made it possible to apply machine learning (ML) to healthcare in practical applications. Several machine learning algorithms could be used on those datasets, which might be in an unstructured or structured format, to improve healthcare outcomes. Predictions of diseases were made using a variety of machine learning (ML) and Deep Learning techniques, including boosting algorithms, neural networks (ANN), K-nearest neighbor (KNN), decision trees (DT), linear regression, random forests (RF), and support vector machines (SVM) [7-9].

Deep learning has become a transformative technology in the medical field, especially when it identifies complicated conditions like cardiovascular disease (CVD). Deep learning automatically identifies significant patterns and correlations in massive datasets and provides medical diagnostics with supreme accuracy and efficiency. Utilizing the power of deep learning, a fuzzy-based TabNet model tailored for early and accurate CVD prediction was developed in this study [10]. Since deep learning models are excellent at identifying non-linear patterns and correlations in sizable datasets, they are well-suited for the complex nature of medical data. By adding domain-specific improvements like fuzzy logic, their diagnostic skills were further enhanced. This research sought to bridge the gap between machine learning's theoretical developments and their real-world applications in healthcare. The suggested method utilized deep learning, fuzzy logic, and improved feature selection, thus increasing diagnostic accuracy and making it possible to develop scalable and affordable early CVD detection tools. The aim of this research was to improve patient care and reduce the worldwide impact of cardiovascular illnesses [11].

This work presented a novel strategy that integrates fuzzy with enhanced TabNet deep learning model and dual optimal feature selection approaches for improving diagnostic accuracy for CVD prediction. Feature selection determined which features were the most pertinent and instructive, thus helping to improve model performance by lowering computational complexity and noise. In this, the model's predictive potential was greatly increased by deploying Maximum Relevance and Minimum Redundancy (MR-MR) and Least Absolute Shrinkage and Selection Operator (LASSO) approaches to guarantee robust feature selection. To handle ambiguity and imprecision in data, the proposed fuzzy-based TabNet model integrated a fuzzy logic framework, enabling more accurate decision-making in complex medical scenarios. In addition to improving prediction accuracy, this novel architecture also delivered higher specificity and lowers misclassification rates when compared to traditional classifiers [11] . Using data from the Kaggle Heart Disease Repository, the study had shown how well the model can classify and predict CVD. By addressing the limitations of existing classifiers and utilizing state-of-the-art feature selection techniques, ultimately, this work contributed to improving patient outcomes and medical quality by developing reliable and cost-effective solutions for CVD diagnosis.

The following are the main contributions and unique aspects of this work:

- To provide a sparse dual attribute selection mechanism such as Minimum Redundancy Maximum Relevance and Least Absolute Shrinkage and Selection Operator by penalizing less relevant features and reducing the feature set to include only the most predictive variables, improving model interpretability and efficiency.
- To deploy an enhanced deep learning model, the TabNet model, for classification.
- To integrate a fuzzy decision-making process after the model prediction to enhance classification certainty, to apply the fuzzy rules for decision refinement, to calculate "fuzziness score" for each prediction.

2. Literature Review

Researchers have developed a model for cardiovascular disease prediction using a variety of machine learning techniques. Given its potential to lower morbidity, mortality, and healthcare expenses, early diagnosis of cardiovascular disease (CVD) has drawn significant interest in medical research. Conventional diagnostic methods frequently fall short of the required standards of accuracy and dependability as the complexity of CVD data rises. To overcome these constraints, scientists are now investigating sophisticated computational techniques like Machine Learning (ML) and Deep Learning (DL).

The researchers El-Sofany, et al. [9] developed a mobile application that, given input symptoms, could rapidly diagnose heart illness using the best machine learning algorithm. Experimental findings showed that the XGBoost algorithm performed best on combined datasets and the SF-2 feature subset. They created a reasonable AI technique built on SHAP techniques to comprehend the system's ultimate prediction-making process. Additionally, the study showed that the suggested system might be modified by employing a domain adaptation strategy. This work has significantly advanced the field of machine learning-based heart disease prediction applications, offering novel insights and methods.

Haseeb, et al. [12] used two techniques to identify cardiac disease early. Artificial neural networks, ensemble learning, and conventional Machine Learning were used in the first technique. The second strategy used fusing approach to machine and ensemble learning methods in order to improve model performance.

The heart_statlog_cleveland_hungary_final dataset was divided using k-fold cross-validation, with k set to 10. Performance measures such as F1 score, sensitivity, accuracy, precision, and specificity were computed. The hybrid approach—which combines Bagging and Random Forest (RF)—was the best performer, with the highest average accuracy on the given data for the early detection of heart illness.

Halah, et al. [13] suggested an approach that selects pertinent features by combining Genetic Algorithm (GA) with Recursive Feature Elimination Method (RFEM), thus improving the resilience of model. They also employed Under-Sampling Clustering Oversampling Method (USCOM) strategies to resolve data imbalance, which enhanced predictive power of model. For classification task, additionally they used Multi-Layer Deep Convolutional Neural Network (MLDCNN) trained with adaptive elephant herd optimization method (AEHOM). Complete assessment revealed that the proposed Machine Learning-based Heart Disease Prediction Method (ML-HDPM) performed exceptionally well on important evaluation metrics.

The authors Israa Nadheer [14] used three feed-forward Neural Networks to successfully categorize the clustered groups. Additionally, they proposed a unified method that makes use of XGBoost ensemble classification, enhancing the overall classification of FNN model outputs by exploiting boosted ensemble learning. The Cleveland dataset was split into 70% training and 30% testing sets to produce sovereign datasets. The XGBoost model produced satisfactory testing results when MLP outputs were incorporated. This study offered a dependable and effective methodology for healthcare applications by demonstrating the effectiveness of combining feature engineering, data processing, and ensemble learning techniques for strong cardiovascular disease diagnosis.

The researchers Almazroi, et al. [15] proposed a method that used a Keras-based Deep Learning model, specifically a Dense Neural Network, to compute outcomes. The suggested model was tested using a range of hidden layer configurations in the dense neural network, from three to nine layers. With 100 neurons, each hidden layer utilized the ReLU activation function. Several datasets related to heart disease were used as benchmarks to conduct the study. Heart disease datasets were evaluated on both solo and ensemble models. The Dense Neural Network was also assessed with crucial metrics, including sensitivity, specificity, accuracy, and F-measure on all datasets. Because diverse attribute categories affect different datasets, various layer combinations perform differently. The findings of the suggested framework were examined through extensive experimentation. When compared with individual models and other ensemble approaches, the Deep Learning model proposed in this exploration [15] exhibited better accuracy, sensitivity, and specificity on all heart disease datasets.

The researchers Aarti, et al. [16] introduced a cutting-edge Hybrid Convolutional Neural Network (HCNN) for heart disease prediction on extensive dataset from UCI Machine Learning Repository. Attention mechanisms, leftover blocks, and convolutional layers were all included in the HCNN design. Even in the case of highly complicated patterns, these aided in extracting additional information from cardiovascular health data. By using these deep learning components, the HCNN displayed improved prediction abilities, achieving a fair F1 Score, strong categorization (AUC), and increased accuracy. The ability of the model to adapt to uncover intricate relationships within the data might help improve medical diagnosis. Its ability to quickly learn hierarchical patterns from unprocessed input made HCNN special. In order to effectively forecast cardiac disease, it was able to uncover hidden traits. To improve training efficiency for deep designs, the residue blocks prevented issues with fading gradients, while the convolutional layers assist the network in identifying local patterns. The network's remarkable capacity to distinguish between them was further enhanced by attention processes, which further concentrate it on key characteristics. This study made it possible for researchers, physicians, and data scientists to enhance cardiovascular health analytics by applying HCNN deep learning model. A detailed analysis of the model's architecture and performance measures was provided, which was a significant step toward more accurate and effective cardiac disease prognosis. This made it possible for further study and application of sophisticated neural networks in the area of medical diagnostic prediction.

The authors Bobburi, et al. [17] provided a systematic method that complies with the requirements for predicting patients' risk profiles using clinical data factors. The suggested appearance makes use of a Significant Neural Orchestrate to successfully handle the problems of underfitting and overfitting. This demonstrates outflanks on both planning and test data. Both Profound Neural Arrange (DNN) and Manufactured Neural Arrange (ANN) techniques were used to examine the model's efficacy, thus producing precise predictions on presence or absence of heart disease.

3. Materials and Methods

3.1. Data Collection

The Description of Kaggle heart disease dataset is in Table1 below. In this scientific study, dataset used was an unbalanced classification dataset with 1025 instances taken from Kaggle Heart Disease dataset [9]. This dataset offered one target variable and thirteen unique attributes to fully characterize the people under examination. There were two classes in the heart data set: class 0 and class 1, where 0 specifies that a person does not have heart disease and 1 shows that they have heart disease. Table 1. above showed 1025 occurrences and 14 attributes (columns). As a result of the dataset, women seemed to have lower risk of heart disease than men. To properly treat heart problems, precise diagnosis was necessary. Consequently, conventional methods of diagnosis and prognosis are unable to produce reliable outcomes.

3.2. Data Preprocessing

Preprocessing the data was obligatory to ensure that it could be used to train models. The data were first standardized using Standard Scaler and then tested and trained using train_test_split. This technique scaled each variable in dataset X to have a mean of 0 and a standard deviation of 1. Standard Scaler was a Python utility in the sklearn.preprocessing package that is used to standardize (normalize) features by scaling them to unit variance and removing the mean. Using calculated statistics like standard deviation and mean, the data was transformed. The final result of data preprocessing is a new X scaled dataset with a mean of 0 and a standard deviation of 1 for each feature. In addition, to prevent data leaks, the same scaler object was used to normalize the test and training data. The output of data preprocessing was shown in Table 2 below.

Table 1.
Description of features in Dataset.

Dataset Name	**Heart Disease Dataset**	
Instances	1025	
Classes	2- Class0- No Heart Disease in individual	Class1- Occurrence of Heart Disease in an individual
Total no of Features	14	
Input Features	13	
Output Features	1-Target variable	
Name of Feature	Description	Definition
age	Age	Indicates the person's age.
sex	Sex	Denotes a person's gender, with a "0" signifying a woman and a "1" signifying a man.
cp	Chest-pain type	Specifies the type of chest pain felt, and it can be classified as asymptotic (4), non-anginal pain (3), atypical angina (2), or typical angina (1).
trestbps	Resting Blood Pressure	refers to the person's resting blood pressure, which is expressed in millimeter-Hg.
chol	Serum Cholesterol	Indicates volume of cholesterol in blood in milligrams per deciliter.
fbs	Fasting Blood Sugar	This refers to comparing a person's fasting blood sugar level with cut-off of 120 mg/dL; value of "1" indicates true result (if fasting blood sugar is greater than 120 mg/dL) and value of "0" indicates false result.
restecg	Resting electrocardiogram	Having resting electrocardiogram findings that are classified as normal (0), aberrant ST-T waves (1), or left ventricular hypertrophy (2)
thalach	Max Heart Rate Achieved	Denoting highest ever attained heart rate
exang	Exercise-induced Angina	Denotes whether exercise-induced angina is present (1) or not (0).
oldpeak	ST Depression Persuaded by Exercise Relative to Rest	Showing a float or an integer.
slope	Peak Exercise ST Segment	Indicating whether the peak exercise ST segment is flat (2), downsloping (3), or upsloping (1).
ca	Number of Major Vessels Colored by Fluoroscopy	An integer or float that represents the number of major vessels (from 0 to 3) that have been colored by fluoroscopy.
Thal	Thalassemia	Classified as normal (3), fixed defect (6), or reversible defect (7).
target	Diagnosis of Heart Disease	Identifying whether a person has cardiac disease and classifying it as either present (1) or absent (0)

Table 2.
Normalized features and their Mean and Standard Deviation.

Standard Scaler Normalization			
S. No	**Feature**	**Mean**	**Standard Deviation**
1	age	-0.01197917	1.02130084
2	sex	0.0106728	0.99539542
3	cp	0.02050516	1.00830729
4	trestbps	0.00425462	0.99655698
5	chol	-0.03136069	0.93278486
6	fbs	-0.00598529	0.99407364
7	restecg	-0.01937919	0.99509783
8	thalach	0.02266935	0.99972692
9	exang	0.00138447	1.0004777
10	oldpeak	-0.02718553	0.97375129
11	slope	0.02117024	0.98499445
12	ca	0.01114114	1.01857959
13	Thal	0.00890558	0.97272186

3.3. System Design of Proposed Model

To develop a comprehensive framework for accurately predicting cardiovascular disease by utilizing a multimodal strategy that combines feature selection, dimensionality reduction, and sophisticated deep learning approaches was the notion of this research. By leveraging advanced deep learning algorithms, employing ensemble deep learning, and creative feature selection techniques, the model efficiently identified intricate patterns present in patient data. This comprehensive approach provided medical professionals with a useful tool to improve patient treatment outcomes and ensured the fast and accurate diagnosis of heart disease.

Additionally, data gathering, data pre-processing, and prediction were three crucial phases of the prediction system. To improve the predictive model, each of these phases was essential. A number of actions were carefully carried out during the rigorous pre-processing phase in order to maintain data integrity and model effectiveness. Furthermore, a traditional scalar method was utilized to recalculate the coefficients of every feature, bringing their mean and standard deviation to 0 and 1, respectively, thereby reducing the possibility of biases resulting from feature scale disparities and guaranteeing consistency in the influence of features.

By distinguishing between two separate classes—class 0 and class 1, which stand for the absence and occurrence of cardiac illness, respectively—the database makes classification easier. In particular, 499 cases are classified as class 0, indicating that heart disease is not present, and 526 cases are classified as class 1, indicating that heart disease is present.

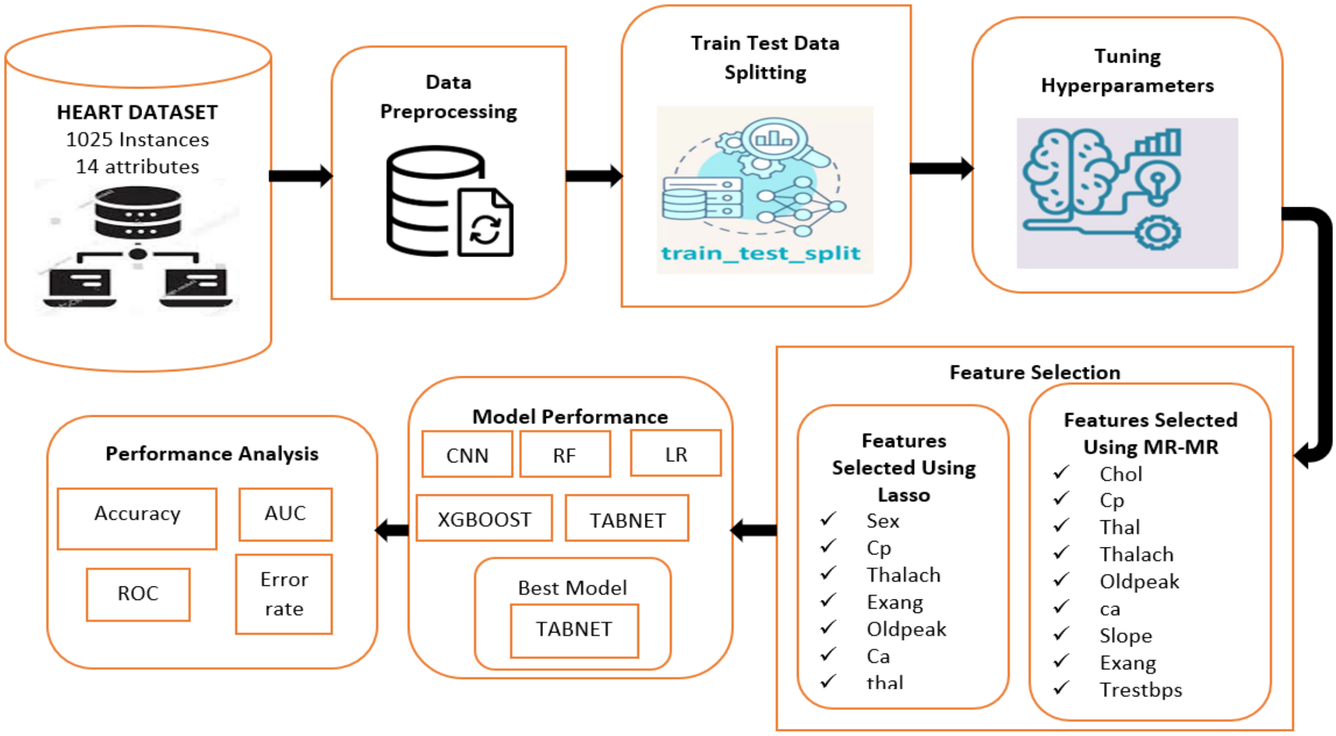

Figure 1.
Proposed Methodology for Cardiovascular Disease Prediction.

The Figure 1 for the dual-optimized feature selection fuzzy-based TabNet model was intended to show the methodical approach of the research methods, from data pre-processing to prediction. The workflow was visually represented by the diagram, which emphasized the combination of deep learning, fuzzy logic, and dual feature selection techniques. It offered a concise synopsis of the entire procedure, demonstrating how the creative fusion of different methods improved the diagnostic precision for the prognosis of cardiovascular illness. The main source of data was the Kaggle Heart Disease Repository. Cholesterol, blood pressure, age, and other clinical factors are the characteristics found in medical records. To make the dataset suitable for analysis, it was cleaned, normalized, and handled missing values. Two feature selection techniques were used: MR-MR (Maximum Relevance Minimum Redundancy), which picks features that are highly appropriate to the target and have little correlation with one another, and LASSO (Least Absolute Shrinkage and Selection Operator), which finds features with the strongest correlations to the target variable in order to reduce redundancy. For making predictions, this step produced the optimal subset of features. To handle uncertainty and ambiguity in medical data, a fuzzy inference system was incorporated. The translation of precise numerical inputs into fuzzy values was performed using linguistic terms (e.g., "low," "moderate," "high"), and the application of fuzzy rules was performed to enhance decision-making. The ambiguity and uncertainty in medical data were addressed by the incorporation of the fuzzy inference system. Fuzzy rules and linguistic phrases (such as "low," "moderate," and "high") were used in this module to convert exact numerical inputs into fuzzy values. TabNet was used as the fundamental deep learning architecture due to its attention-based feature learning and interpretability for organized medical data. The fuzzy logic module was incorporated into the TabNet framework to further increase accuracy and robustness. The fuzzy-based TabNet model was trained on an optimized dataset. Accuracy, recall, precision, F1-score, misclassification error, and specificity were the metrics used for evaluation. The ROC curve and confusion matrix were used for in-depth performance analysis. The model made predictions on patients' risk of CVD based on the processed input data. Clinicians were provided with practical advice for early intervention through interpretable outputs.

3.4. Proposed Model for Early Prediction of CVD

3.4.1. Feature Selection

Feature selection was a key component of creating a successful heart disease prediction model. It was vital to select features that are most relevant from the dataset to enhance the model's performance and reduce computational complexity. It also improved generalization to new data, decreased training time, and increased model accuracy by choosing the most informative features. By using feature selection for heart disease prediction, clinical, demographic, or lifestyle factors most strongly associated with heart disease risk could be identified. This procedure assisted healthcare professionals in making data-driven decisions by improving the predictive performance of cardiac models and offering insight into the fundamental causes of cardiac disease. The three main feature selection techniques were filtering, wrapping, and embedding; each had its own advantages based on the dataset and model being used.

3.4.2. LASSO

A useful method for selecting features and regularizing predictive models was Least Absolute Shrinkage and Selection Operator. This approach was successful especially when there were many features in the dataset, including some that were irrelevant or had poor correlation with the target variable. Using Lasso, it was possible to uncover the most substantial risk factors for heart disease, including lifestyle factors and clinical biomarkers. When the coefficients of the feature values were small, LASSO performed very well. Features with high coefficient values would be included in chosen feature subsets. Superfluous features could be identified by LASSO [18, 19]. Additionally, the most frequently found characteristics would eventually be considered the most important ones by continually carrying out the previously mentioned procedure, increasing the feature's dependability [20].

The formulation for the LASSO optimization problem was:

The minimization was

$$||y - X\beta||^2 + \lambda||\beta||_1 \tag{1}$$

In which, y represented the binary result (heart disease present or absent), β was a representation of the feature coefficients, X was the feature matrix and $||\beta||_1$ was the L1 norm of β and λ was the regularization parameter.

3.4.3. MR-MR

The feature selection method known as Minimum Redundancy Maximum Relevance (MRMR) reduced redundancy among the selected characteristics when choosing which features were most relevant for predicting job outcomes. Because datasets often contain strongly related features and irrelevant factors, this was particularly useful when predicting heart illness. Selecting features that were most relevant to the target variable while eliminating features that provide redundant information was the aim of this approach [21]. MRMR used mutual information to quantify the volume of information a feature contributes to the target variable's prediction. For instance, although having similar data, systolic and diastolic blood pressure might both be linked to heart disease. In order to ensure that only one was selected, it favored features that complement one another. It performed well for datasets with a high number of features and relatively few samples, such as those seen in clinical or genomics research. By eliminating unnecessary and superfluous features, it improved generalization on unknown data and reduced overfitting. Since it could handle both continuous and categorical data, it could be used to predict heart disease across a range of datasets. Because MRMR balances relevance and redundancy, it was a powerful choice for developing accurate and therapeutically meaningful prediction models [22]. By making it easier to identify important risk factors, MRMR enabled early diagnosis, customized treatment plans, and a better understanding of how various clinical and lifestyle factors affect the condition. The MRMR approach selected features using a scoring formula that balances relevance and redundancy:

Maximum Relevance was given by

$$\mathrm{Rel}_{\max} = (1/|S|) \sum I\,(fi;\ H) \qquad (2)$$

Where, S was the chosen feature set, fi was feature I, H was the result of heart disease and I was mutual information.

Minimum Redundancy was given by

$$\mathrm{Red}_{\min} = (1/|S|^2) \sum\sum I(fi;\ fj) \qquad (3)$$

where, fi, fj was feature pairs in S.

When selecting features, the MRMR algorithm optimizes two criteria: minimizing redundancy among chosen features and maximizing relevance to the target variable [23]. The relevance term (Relmax) ensures that the chosen attributes are relevant to the target. The redundancy term (Redmin) penalizes features that are highly correlated with previously selected features.

3.4.4. Prediction of CVD

A novel hybrid approach to cardiovascular disease prediction was presented in the proposed work, which blends improved tabular neural networks, or TabNets, with fuzzy logic. The goal was to enhance the accuracy and interpretability of prediction models for cardiovascular illness by combining fuzzy logic with an improved TabNet. To handle tabular data, a deep learning network called TabNet was enhanced using optimized feature selection and attention techniques. For medical datasets, these procedures were crucial. Fuzzy logic improved this by identifying the ambiguity and uncertainty found in clinical data, such as test results, patient histories, and symptoms. Together, they provided a robust framework for handling ambiguous data and spotting important trends, which improved the prediction of CVD risk [24].

The combination of fuzzy logic and augmented TabNet is used to forecast cardiovascular illness because it combines advanced feature representation with interpretable decision-making under ambiguity. In order to minimize noise and maximize the selection of clinically relevant variables, the enhanced attention approach in Enhanced TabNet employs sparse feature masking. At every decision stage, the feature importance is dynamically changed. The fuzzy logic system is mathematically integrated with this, using rule-based inference and membership functions to capture linguistic uncertainty and explain imprecise medical data. A new weighted aggregation approach is utilized to fuse both models. Through an adjustable parameter, α, the TabNet-predicted risk score—derived from high-dimensional latent space representations—is coupled with the interpretable risk output of the fuzzy system. Prediction accuracy is increased, and transparency is maintained by integrating complex machine learning models with the interpretability required for important healthcare decisions. This combination addresses feature complexity and data ambiguity at the same time, marking a significant advancement in predictive modeling [25].

The operation of the dual feature selection and fuzzy based modified TabNet algorithm1 for cardiovascular disease prediction is described below.

Algorithm 1 for Fuzzy based modified TabNet for Cardiovascular disease prediction.

Input: Dataset D = {X, y}, fuzzy membership functions F, fuzzy rules R

Output: Predicted cardiovascular disease stages

1. Normalize features X. Xscaled=X−μ/σ. Split dataset into training (D_train) and testing (D_test).
2. Apply LASSO to select features X_lasso. min‖y−Xβ‖22+λ‖β‖1 and Use MRMR to refine features to X_selected. Score(xi)=I(xi;y)−1/∣S∣ ∑xj∈SI(xi;xj)
3. For each feature x_i in X_selected: Define fuzzy sets (Low, Medium, High) using membership functions.
4. Generate Fuzzy Rules: Use domain knowledge to define rules R.
5. For each sample x in D_train: Compute membership values for each fuzzy set. Evaluate fuzzy rules R to calculate P_fuzzy(y|x).
6. Concatenate original features X_selected with fuzzy outputs P_fuzzy(y|x). Train TabNet model M with enhanced inputs and target labels y_train. Xenhanced = [Xselected, Pfuzzy(y|x)]
7. For each sample x in D_test: Compute fuzzy outputs P_fuzzy(y|x). Pass enhanced inputs (X_selected ∪ P_fuzzy(y|x)) through M. Output P_final(y|x).
8. Calculate accuracy, precision, F1-score, recall, MAE, MSE and Visualize feature importance and fuzzy contributions.

3.4.5. Sparse Feature Masking with Controlled Attention

TabNet selectively attends to the most pertinent features for every decision step using a learnable feature mask. In mathematics, this is represented as:

$$\mathrm{Mk} = \sigma\left(\mathrm{W}_{\mathrm{k}}^{\mathrm{T}}\mathrm{x} + \mathrm{b}_{\mathrm{k}}\right) \qquad (4)$$

Where Mk denotes step k's sparse mask vector, which establishes the significance of the features; Wk and bk are the weights and biases that can be trained for each step k. The vector x represents the input features, and Σ represents the Sigmoid activation to enforce sparsity. The innovative aspect is that feature importance is dynamically re-evaluated at each decision stage, which is crucial for medical data because patient-specific circumstances can affect feature relevance [26].

3.4.6. Sequential Aggregation of Latent Representations

Every decision step uses masked characteristics to alter the input x:

$$z_k = M_k \cdot f_k(x) \quad (5)$$

Where fk(x) denotes feature transformation function (e.g., layer of a neural network).

The following latent representations are aggregated in a sequential manner:

$$\text{Risk}_{\text{TabNet}} = g(z1, z2, \ldots, z_K) \quad (6)$$

Where g refers to an Aggregation function, usually an aggregate of weights. This combination of sequential attention and sparsity is innovative since it enables TabNet to adaptively focus on particular medical features without overfitting [27]. TabNet is enhanced with fuzzy logic, which adds a layer of clinical rule-based interpretable decision-making. Membership functions are used to map clinical variables (such as age and cholesterol) to fuzzy sets:

$$\mu_A(x) = \frac{1}{1+e^{-a(x-c)}} \quad (7)$$

Where μA(x) represents a membership value that indicates how much x belongs to the fuzzy set A (for example, "high cholesterol"). A variable regulates how steep the membership function is, and c denotes the fuzzy set's center.

3.4.7. Rule Activation and Aggregation

Using fuzzy inference rules, these membership values are combined:

$$R_j = \min\,(\mu_{A1}(x_1), \mu_{A2}(x_2), \ldots) \quad (8)$$

Where Rj represents the rule's activation level. Rule For instance: "IF age is high AND cholesterol is high, THEN CVD risk is high." All activated rules are weighted and aggregated to determine the output fuzzy risk:

$$R_j = \min\,(\mu_{A1}(x_1), \mu_{A2}(x_2), \ldots) \quad (9)$$

Where wj is Rule j's importance weight.

3.4.8. Integration of Enhanced TabNet and Fuzzy Logic

The result of the integration is a hybrid system that strikes a balance between fuzzy logic's interpretability and TabNet's accuracy.

3.4.8.1. Fusion of Weighted Risk Scores

The final risk score is a weighted combination of the two elements

$$\text{Final Risk Score} = \alpha \cdot \text{Risk}_{\text{TabNet}} + (1-\alpha) \cdot \text{Risk}_{\text{Fuzzy}} \quad (10)$$

Where α denotes a weighting parameter that is optimized during training to strike a balance between interpretability and prediction performance. This mathematical foundation ensures that the improved TabNet recognizes intricate patterns in clinical data, and fuzzy logic provides it with interpretability and resilience. This makes it especially appropriate for predicting cardiovascular illness in actual clinical situations [28].

4. Results and Discussion

Feature importance is the act of determining the relevance of each input feature in the decision-making process by computing the score of each feature in a machine learning model. Higher feature scores have a significant influence on the model's capability to predict the target variable. The majority of TabNet implementations, like the one included in the pytorch-TabNet library, include tools for determining feature importance. The selection of medical characteristics that could increase the prediction accuracy of heart disease was the main driving force behind this endeavor. The benefits of feature selection include improved data quality, reduced computing time for prediction models, enhanced predictive performance, and an effective data collection procedure [29]. The tabular representation of Feature Importance scores based on the TabNet model is shown in Table 3 below. The most crucial characteristics for predicting Cardiovascular Disease are "thalach," "oldpeak," "restecg," "chol," and "fbs," possessing appropriate scores related to the outcome. The TabNet model emphasizes the importance of concentrating on thalach, oldpeak, and restecg characteristics for improved diagnostic accuracy, as these are the most significant predictors of CVD. Critical characteristics for reliable predictions are identified by this feature importance analysis, which aids in the dual feature selection procedure.

Table 3.
Feature importance scores for each feature based on TabNet.

S. No	Feature	Feature Importance Score
1	age	0.0461
2	sex	0.0442
3	cp	0.0546
4	trestbps	0.0584
5	chol	0.0921
6	fbs	0.0835
7	restecg	0.1024
8	thalach	0.1868
9	Exang	0.0535
10	oldpeak	0.1454
11	slope	0.0508
12	ca	0.0196
13	Thal	0.0620

The following Figure 2. shows the different features in Kaggle heart dataset and their feature importance scores related with target variable. The feature "thalach" has high feature importance score of 0.1868, the feature "oldpeak" has next high feature importance score of 0.1454, the feature "restecg" has high feature importance score of 0.1024 and features "chol" and "fbs" has high feature importance score of 0.0921 and 0.0835 respectively.

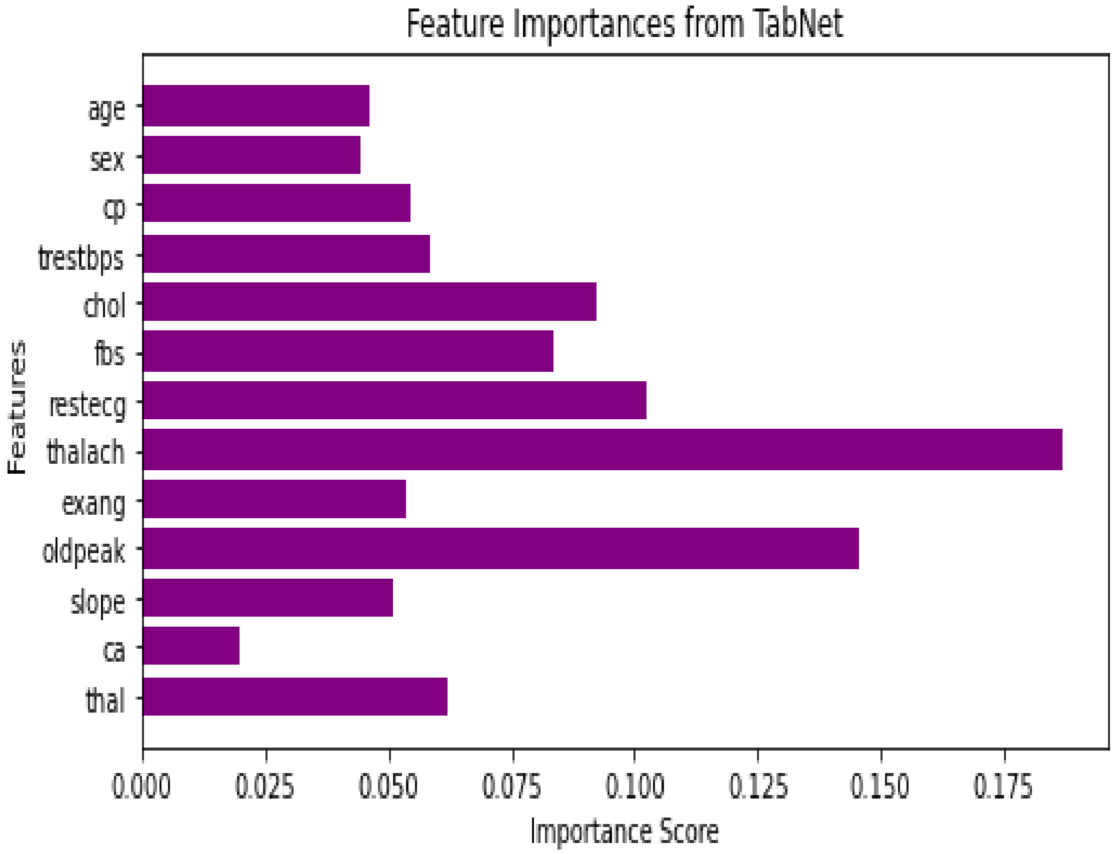

Figure 2.
Feature Importance from TabNet.

The following Figure 3 is the graphical analysis representing the number of individuals affected by heart disease and its percentage in different age groups and based on gender. This analysis examines how heart disease is distributed among various age and gender categories using data. In particular, the following are examined: In which age groups does heart disease most commonly occur? In each age range, are males or females more likely to be impacted? How many people in each demographic group have heart disease? Age is divided into groups (20–29, 30–39, etc.) using pd.cut() to create the age_group column. The percentage of people afflicted by the disease is determined for each age group and gender. The stacked bar chart below shows the number of males and females with the disease (1) in each age group. The light blue color in the stacked bar chart represents the count of females affected by heart disease in each age group, and the salmon color represents the count of males affected by heart disease in each age group [30].

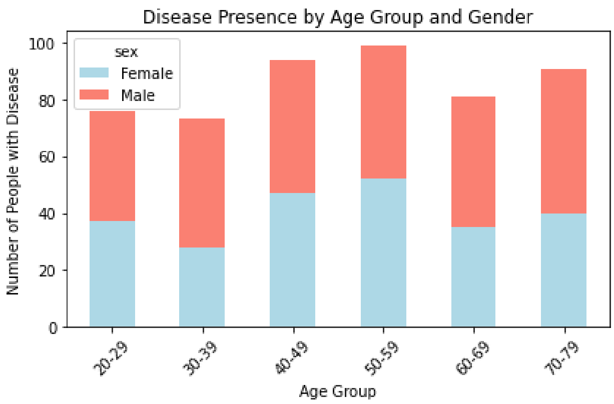

Figure 3.
Counts and Percentages of Disease Presence by Age Group and Gender.

Table 4.
Gender-wise count affected by heart disease in each age group.

	Disease-affected count based on gender		Disease affected Percentage	
Age Group	**Male**	**Female**	**Male**	**Female**
20-29	39	37	46.98	46.83
30-39	45	28	54.87	41.79
40-49	47	47	53.40	49.47
50-59	47	52	52.80	54.73
60-69	46	35	52.27	49.29
70-79	51	40	50.49	45.97

Table 4 represents the analysis that provides a thorough viewpoint on this important health issue by using data-driven insights to display and quantify disease prevalence. This analysis determines whether males or females are more affected in particular age groups and identifies the age group with the largest number of sickness cases. The probability of contracting the disease for each gender within each age range is displayed in the percentage column. The above table reveals that females in the age group between 50-59 are highly affected by heart disease, with the percentage of affected females being 54.73%. Males in the age groups between 30-39 and 40-49 are also highly affected by heart disease, with the percentage of affected males in the 30-39 age group being 54.87% and the percentage of affected males in the 40-49 age group being 53.40%.

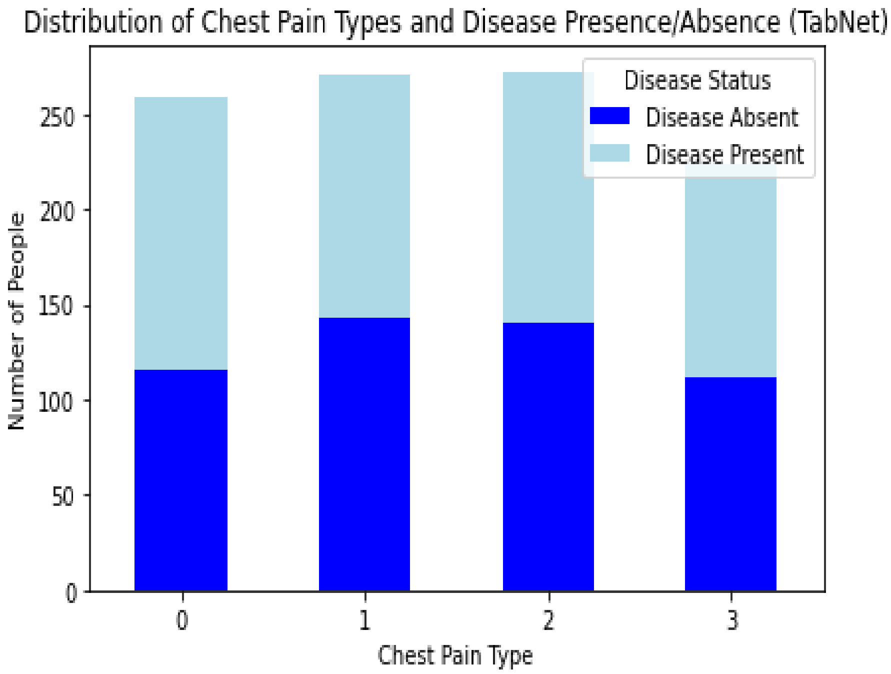

Figure 4.
Distribution of chest pain types in disease-affected individuals.

The above Figure 4 exposes the graph that shows the distribution of Chest Pain Types in disease-affected individuals. A stacked bar plot is created, showing the distribution of disease presence or absence for each chest pain type. Disease presence is shown in salmon color and absence in light blue color. In the above stacked bar plot, the x-axis represents the chest pain forms (0, 1, 2, 3) and the y-axis shows the number of people. Each form of chest pain is represented by stacked bars indicating the number of patients with and without the disease [31]. The dataset is divided into groups based on disease state and chest pain type, and then size () is used to determine how many people are in each group.

Table 5.
Count of individuals with Chest pain types.

Chest Pain Type/ Target	Target			
	No Heart Disease (0)	**Heart Disease (1)**	**Total**	**Percentage Affected**
0	116	143	259	55.21%
1	143	127	270	47.03%
2	140	132	272	48.52%
3	112	112	224	50%

Table 5 above shows the distribution of chest pain types in individuals and whether they have the disease or not. Chest pain type 0 represents Typical Angina, chest pain type 1 represents Atypical Angina, chest pain type 2 represents Non-Anginal Pain, and chest pain type 3 represents Asymptomatic. It is estimated that 55.21% of individuals with heart disease have chest pain type Typical Angina. There are 47.03% of individuals suffering from heart disease who experience Atypical Angina. There is an estimated 48.52% of individuals suffering from heart disease who have chest pain of the Non-Anginal type. The percentage of individuals affected by heart disease with chest pain type Asymptomatic is 50%. Based on the analysis, Typical Angina has the highest percentage of heart disease patients, while Non-Anginal Pain is more common in heart disease patients.

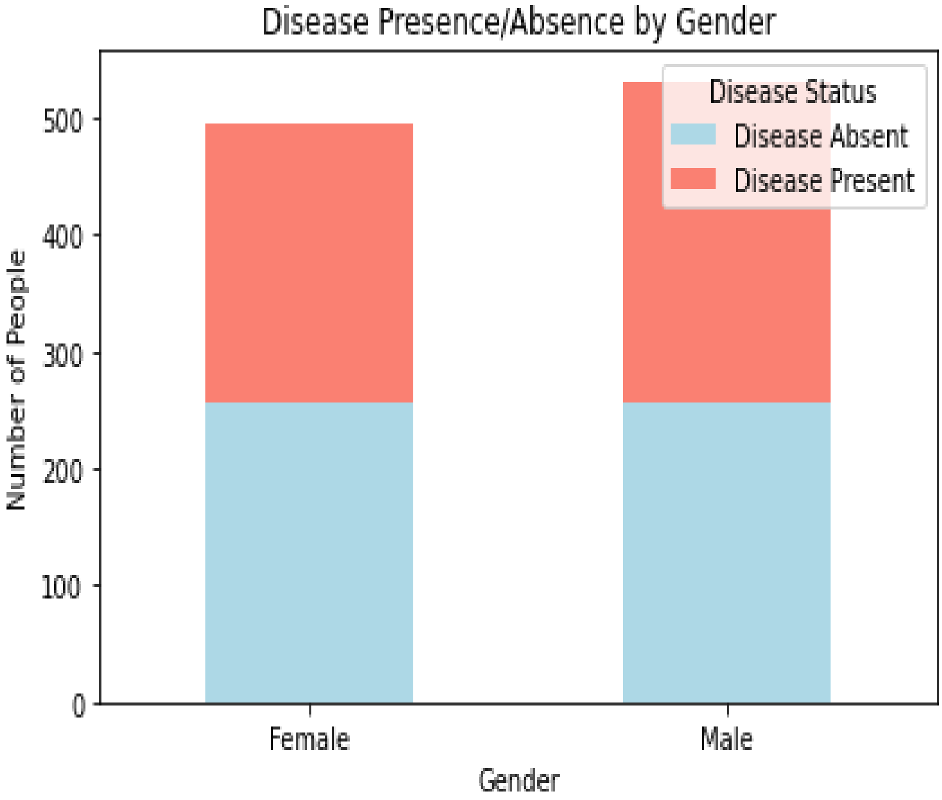

Figure 5.
Distribution of disease presence across genders.

It is crucial to determine the existence of cardiac disease in men and women to comprehend these differences and adjusting treatment and prevention plans appropriately. The study was conducted to observe number of men and women with heart disease, as shown in Figure 4. Through examining the distribution of disease presence across genders, this study seeks to determine which gender is more likely to suffer from heart disease, comprehend the percentage of both men and women affected, and offer insights for directing gender-specific health interventions. This study highlights the gender-based burden of heart disease through data summaries and visualizations, which will be helpful to researchers and medical practitioners who are trying to reduce gender disparities in cardiovascular health outcomes.

Table 6.
Percentage of disease affected males and females.

Sex	Disease Presence /Absence		Percentage Affected
	No Heart Disease (0)	Heart Disease (1)	
Female	255	239	48.380
Male	256	279	51.789

The number of males and females with heart disease is analyzed in Table 6. The objective of the analysis is to determine the percentage of each gender affected by heart disease by comparing the prevalence of the ailment in men and women. Finding disparities in the prevalence of heart disease between the sexes, calculating the count of men and women with diagnoses, and providing data-driven insights for developing gender-specific prevention and treatment strategies are the primary objectives. With a 51.789% impacted rate, heart disease is most common among men. With an impacted percentage of 48.380%, heart disease affects women less frequently than it does men.

4.1. Performance Metrics

Assessing the efficacy of predictive models is crucial when using machine learning and deep learning to predict cardiovascular disease (CVD). Accuracy of model's classification and outcome prediction is evaluated using performance metrics, which are measurable benchmarks. These measures ensure model's reliability in real-world scenarios exposing model's strengths and weaknesses. Accuracy, recall, precision and F1-score are the main performance metrics used in this study.

For early intervention and better CVD treatment, it is crucial to identify individuals who are at risk using accurate prediction models. The prediction performance of various machine learning methods for CVD is evaluated in this study. Models like CNN, TabNet, Random Forest, XGBoost, and logistic regression are evaluated using key performance indicators, including accuracy, precision, F1-score, recall, and AUC-ROC. Based on the dataset and its properties, the goal is to identify the model that produces the most accurate predictions. In order to determine which model is optimal for CVD prediction tasks, this study will evaluate the benefits and shortcomings of each model and clarify how machine learning models might enhance clinical decision-making by analyzing these metrics and displaying the model outputs.

The prediction of disease was examined with several machine learning models for the available input Kaggle heart dataset. The forecast was made using five distinct machine learning models. Three of the five models had an accuracy of 90% or above. Figure 6a illustrates that the TabNet model achieves the highest accuracy of 99.61% among all the models. The accuracy is high because TabNet employs a sequential attention mechanism, which aids in the model's concentration on the most relevant characteristics for outcome prediction, and it also learns feature representations directly from raw tabular data without the need for considerable pre-processing or feature engineering. With an accuracy of 78.98%, the Logistic Regression model was the least accurate of the models that were examined. The accuracy rate for the Convolutional Neural Network was 86.77%. XGBoost demonstrated a 96.83% accuracy rate. Using Random Forest, the accuracy was 95.83%.

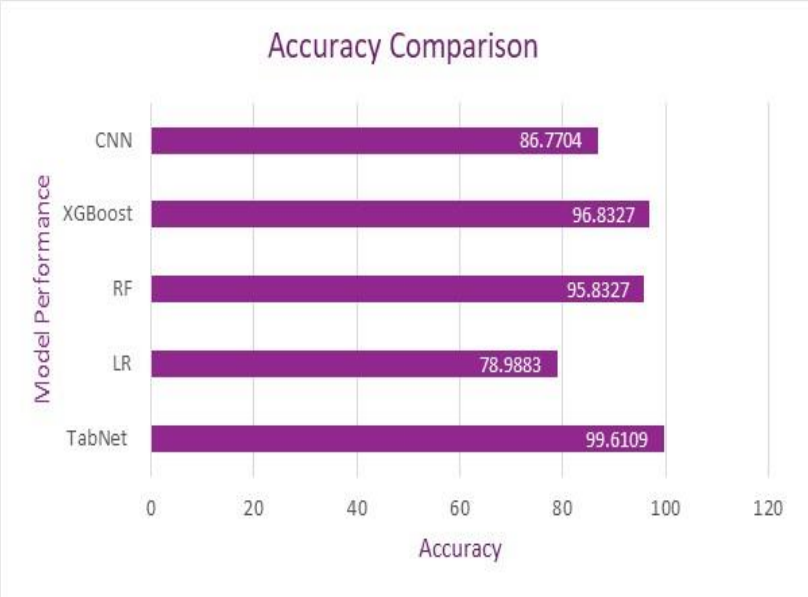

Figure 6a.
Models' performance analysis using accuracy.

Figure 6b.
Models' performance analysis using precision.

The Figure 6b depicts the model performance analysis using the performance metric precision. The TabNet model possesses the highest precision of 99%. The high precision of predicting cardiovascular disease using TabNet suggests that the model is good at reducing false positives, or correctly identifying those who actually have the condition. In order to prevent the model from being misled by irrelevant or weakly correlated features, TabNet's sequential attention mechanism concentrates on the most pertinent features for prediction. This effectively separates people with true cardiovascular disease symptoms, lowering the possibility of false positives. The precision of XGBoost, Random Forest, and CNN are 96.3%, 95.1%, and 85.1% respectively. The Logistic Regression model possesses the least precision of 79.8%.

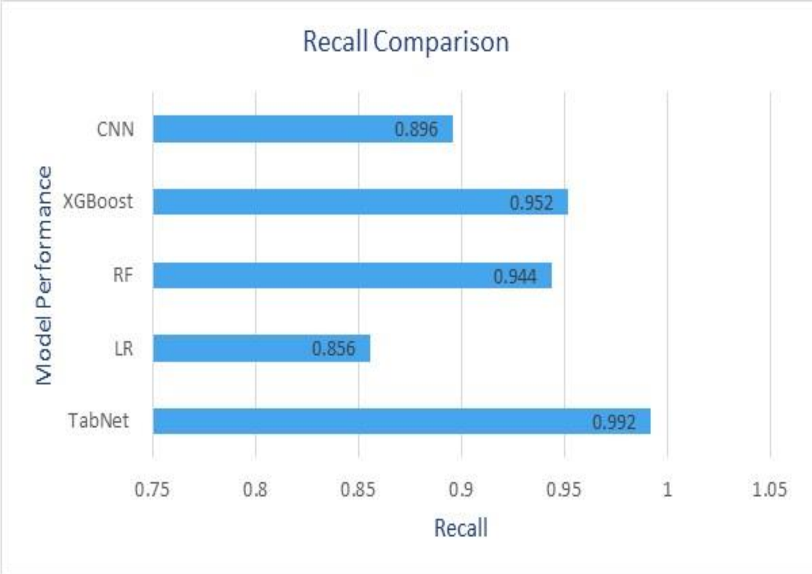

Figure 6c.
Models' performance analysis using recall.

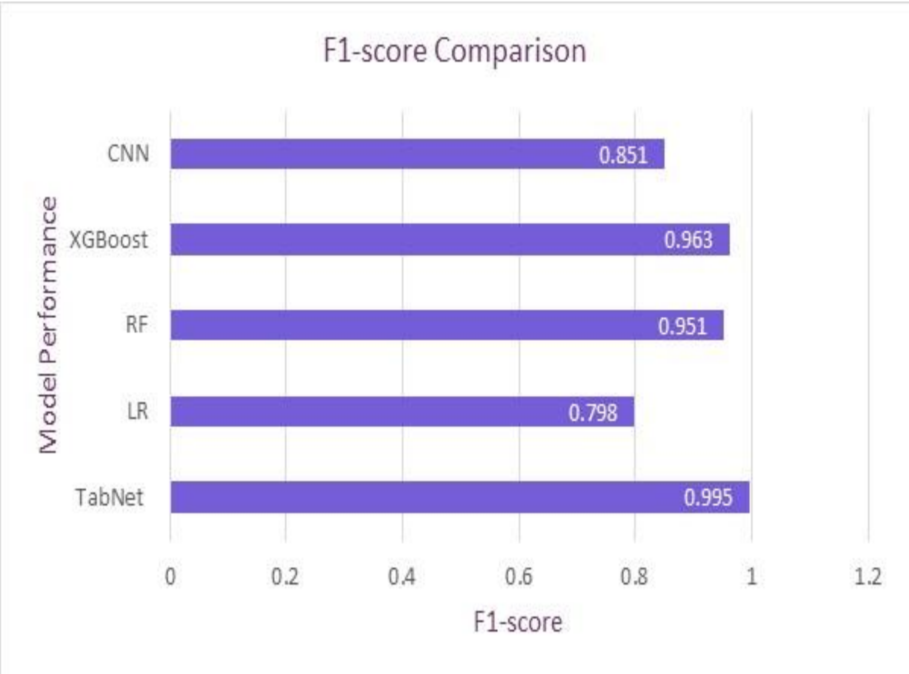

Figure 6d.
Models' performance analysis using F1-score.

The above Figure 6c expresses the model performance analysis using Recall performance metric. The traditional Machine Learning models frequently have trouble striking a balance between recall and precision, particularly when dealing with unbalanced datasets or substantial non-linear feature interactions. The recall measure of TabNet for cardiovascular disease prediction was higher than other models which was 99.2%, because even in situations when other models could overlook them, the model picks up on minute patterns and interactions in the data that are essential for identifying actual occurrences of cardiovascular disease. Other models such as XGBoost, Random Forest, CNN possessed recall of 95.2%, 94.4%, 89.6% for predicting the disease. The Logistic Regression had least recall measure of 85.6% than other models.

The above Figure 6d expresses the model performance analysis using F1score performance metric. The F1score of TabNet model was high which was 99.5%. This indicated that the model was successful in detecting the majority of positive cases and reliable in distinguishing real positives. The F1score of remaining models XGBoost, Random Forest, CNN and Logistic Regression were 96.3%, 95.1%, 85.1% and 79.8% respectively.

Table 7.
Model Performance Analysis for Cardiovascular Disease Prediction.

Performance Metrics **Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
CNN	86.77%	85.1%	89.6%	85.1%
XGBoost	96.83%	96.3%	95.2%	96.3%
RF	95.83%	95.1%	94.4%	95.1%
LR	78.98%	79.8%	85.6%	79.8%
TabNet	99.61%	99%	99.2%	99.5%

The above Table 7 summarizes the performance analysis of various models such as XGBoost, CNN, Random Forest, Logistic Regression, and TabNet for cardiovascular disease prediction using various performance metrics like accuracy, recall, precision, and F1-score. With an outstanding accuracy of 99.61% and an F1-score of 99.5%, TabNet attains the top results across all measures. In this investigation, this model is the most successful at predicting CVD since it has nearly flawless prediction abilities. XGBoost is a dependable option for prediction tasks needing high precision since it delivers strong performance with 96.83% accuracy and 96.3% precision. Random Forest (RF) performs well, with measures such as an accuracy of 95.83% and an F1-score of 95.1% that are somewhat lower than XGBoost. CNN shows that it can accurately identify affirmative cases, as seen by its respectable recall (89.6%). However, in terms of overall performance, it falls short of RF, XGBoost, and TabNet. With an accuracy of 78.98% and precision of 79.8%, Logistic Regression (LR) performs the least out of all the models, underscoring its shortcomings in challenging predicting tasks.

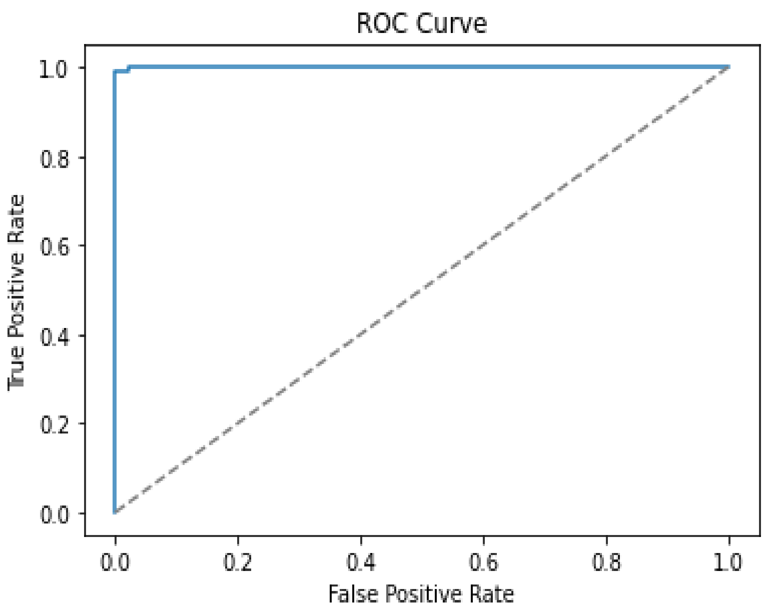

Figure 7.
ROC curve for CVD prediction.

The ROC curve in Figure 7 demonstrates how a binary classification model performs in making predictions. The performance of a random classifier is shown by the dashed diagonal line. In the event that the curve falls along this line, the model's predictions are as accurate as a wild guess. The performance of the model at various threshold levels is shown by the blue line. Starting at (0, 0), the curve should ideally move toward the top-left corner (high TPR, low FPR), which denotes improved model performance. Given that the ROC curve is in the upper-left corner, the model is performing well and has good discrimination between the two classes. Near-perfect performance is indicated by the ROC curve, which shows that the AUC value is very near to 1 (probably >0.99). The curve illustrates how well the model works at various thresholds for decisions. For example, FPR may rise if a threshold favoring high TPR is established. A criterion that favors low FPR could make TPR lower. This ROC curve illustrates how well the model performs in differentiating between the two classes (e.g., disease presence vs. absence). Models that obtain virtually flawless metrics, like TabNet, are consistent with this level of performance.

Figure 8.
Error metrics for Cardiovascular disease Prediction.

The bar graph in Figure 8 shows graphically the MSE, MAE, and RMSE error metrics computed for the model's predictions in comparison to actual values in this bar graph. The three errors metrics are represented on the X-axis (Error Measure): RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and MSE (Mean Squared Error). The magnitude of the determined error values for each measure is displayed on the Y-axis (Error Value). The size of each error measure is shown by the heights of the bars: RMSE: 0.195, MAE: 0.180, and MSE: 0.038. To differentiate each error measure, a separate hue is used, with red denoting RMSE, green for MAE, and blue for MSE. Since square root scaling is not used, MSE, which measures squared differences, has the smallest value and is usually lower than RMSE. Since MAE assesses the average absolute difference between predictions and actual data, it is larger than MSE. Because it amplifies greater deviations by taking the square root of squared errors, RMSE has the highest value of the three. The model's error measures are clearly compared in the graph. The model appears to function well with low prediction errors, as indicated by the low values for all metrics. However, slightly larger variances are shown by MAE and RMSE, highlighting the need to reduce outliers for future progress.

5. Conclusion

According to findings, the most accurate and successful method for early cardiovascular disease (CVD) prediction is the enhanced fuzzy-based TabNet deep learning model. The application of LASSO and MR-MR feature selection approaches guarantees that the most pertinent characteristics are chosen for the prediction model, improving its accuracy and resilience. These are important findings that support this conclusion. With a high level of accuracy of 99.61% and dependability, the enhanced TabNet algorithm demonstrates remarkable performance. The suggested model shows better accuracy and prediction ability for early CVD detection when compared to conventional classifiers like CNN, XGBoost, and Random Forest. By facilitating early intervention, the model's accuracy in predicting the signs of heart disease lowers healthcare expenditures and enhances the standard of medical care. Its application of a fuzzy basis improves judgment in situations that are unclear or ambiguous, which is crucial for medical diagnoses. Because of its reliable performance and affordability, the enhanced fuzzy-based TabNet model is recommended for clinical applications in early CVD prediction. Adoption of it may result in reduced morbidity linked to late-stage cardiovascular illnesses, improved patient outcomes, and more efficient use of available resources. The model's performance is evaluated using metrics and error measures, which offers a thorough understanding of how well it predicts cardiovascular illness.

References

[1] A. Roth Gregory, "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study," *Lancet*, vol. 392, no. 10259, pp. 1736–88, 2018.

[2] W. H. Organization, *Global action plan for the prevention and control of noncommunicable diseases 2013-2020 Global action plan for the prevention and control of noncommunicable diseases 2013-2020*. Geneva, 2013.

[3] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease, In Information and communication technology for competitive strategies." Singapore: Springer, 2019, pp. 447–454.

[4] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94-98, 2019. https://doi.org/10.7861/futurehosp

[5] M. K. Devi *et al.*, "Design and implementation of advanced machine learning management and its impact on better healthcare services: a multiple regression analysis approach (MRAA)," *Computational and Mathematical Methods in Medicine,* vol. 2022, no. 1, p. 2489116, 2022. https://doi.org/10.1155/2022/2489116

[6] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," *Med,* vol. 2, no. 6, pp. 642-665, 2021. https://doi.org/10.1016/j.medj.2021.04.006

[7] V. Binson, S. Thomas, M. Subramoniam, J. Arun, S. Naveen, and S. Madhu, "A review of machine learning algorithms for biomedical applications," *Annals of Biomedical Engineering,* vol. 52, no. 5, pp. 1159-1183, 2024. https://doi.org/10.1007/s10439-024-03459-3

[8] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: Classification and comparison," *International Journal of Computer Trends and Technology,* vol. 48, no. 3, pp. 128-138, 2017.

[9] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports,* vol. 14, no. 1, p. 23277, 2024. https://doi.org/10.1038/s41598-024-74656-2

[10] C. N. Vanitha, K. Vanitha, C. Narmatha, S. A. Krishna, and R. Dhivakar, "Heart disease prediction using enhanced deep learning," presented at the In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 528-532), 2022.

[11] A. Jaiswal, M. Singh, and N. Sachdeva, "Empirical analysis of heart disease prediction using deep learning," presented at the 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, (2023) pp. 1-9. https://doi.org/10.1109/ACCAI58221.2023.10201235, 2023.

[12] K. B. Haseeb, Ahmad, M. A. Aslam, and H. Mustafa, "Heart disease detection: A comprehensive analysis of machine learning, ensemble learning, and deep learning algorithms," *Nano Biomedicine and Engineering,* vol. 16, no. 4, pp. 677-690, 2024. https://doi.org/10.26599/NBE

[13] A.-A. Halah, A *et al.*, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports,* vol. 14, no. 1, p. 7819, 2024. https://doi.org/10.1038/s41598-024-58489-7

[14] Israa Nadheer, "Heart disease prediction system using hybrid model of Multi-layer perception and XGBoost algorithms," presented at the BIO Web of Conferences, 97, 00047 (2024), https://doi.org/10.1051/bioconf/20249700047, 2024.

[15] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, "A clinical decision support system for heart disease prediction using deep learning," *IEEE Access,* vol. 11, pp. 61646-61659, 2023. https://doi.org/10.1109/ACCESS.2023.3285247

[16] K. Aarti, J. Komal, K. I. Piyush, S. Nilesh, A. F. Rupa, and N. Navnath, "Advanced heart disease prediction: Deep learning-enhanced convolutional neural network in the internet of medical things environment," *J.ElectricalSystems,* pp. 1-10, 2024. https://doi.org/10.52783/jes.748

[17] N. V. Bobburi, R. Basati, A. Gunti, S. L. Bulla, A. Chilaka, and K. Sudha, Mr. R., "Harnessing deep neural networks for heart disease prediction," *Journal of Science and Technology,* vol. 9, no. 4, 2024. https://doi.org/10.46243/jst.2024.v9.i4.pp1-9

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology,* vol. 58, no. 1, pp. 267-288, 1996.

[19] C. Zhou and A. Wieser, "Jaccard analysis and LASSO-based feature selection for location fingerprinting with limited computational complexity in progress in location based services 2018 14," Springer International Publishing, 2018, pp. 71-87.

[20] P. Ghosh, A. Karim, and S. Atik, "SA, M. Saifuzzaman," expert model of cancer disease using supervised algorithms with a lasso feature selection approach,"" *International Journal of Electrical and Computer Engineering,* vol. 11, no. 3, 2020.

[21] L. C. Jaffrin and J. Visumathi, "Impact of feature selection techniques on machine learning and deep learning techniques for cardiovascular disease prediction-an analysis," *Edelweiss Applied Science and Technology,* vol. 8, no. 5, pp. 1454-1471, 2024.

[22] J. Kanimozhi, N. Mohanasuganthi, L. C. Jaffrin, G. Preethi, and S. A. Ayshwariya, "Virtual medical assistant system for diseases detection using machine learning," presented at the International Conference on Smart Technologies and Systems for Next Generation Computing, 2023.

[23] L. C. Jaffrin, J. Visumathi, and G. U. Srikanth, "Role of exploratory analytics and visualization in heart disease prediction," *International Journal of Engineering Trends and Technology,* vol. 70, no. 1, pp. 24–34, 2022.

[24] C. Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, "Gender differences in brain-heart connection, in brain and heart dynamics." Cham, Switzerland: Springer, 2020, p. 937.

[25] M. S. Oh and M. H. Jeong, "Sex differences in cardiovascular disease risk factors among Korean adults," *The Korean Journal of Medicine,* vol. 95, no. 4, pp. 266-275, 2020.

[26] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications,* vol. 5, no. 3, p. 176_183, 2013.

[27] L. C. Jaffrin and G. U. Srikanth, "A KNN algorithm based predictive model for heart disease progression," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science),* vol. 16, no. 5, pp. 76-81, 2023.

[28] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," *Acta Sci. Nutritional Health,* vol. 3, no. 7, p. 147_151, 2019.

[29] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," presented at the In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207), 201.

[30] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," presented at the In The Impact of Digital Technologies on Public Health in Developed and Developing Countries: 18th International Conference, ICOST 2020, Hammamet, Tunisia, June 24–26, 2020, Proceedings 18 (pp. 299-306). Springer International Publishing, 2020.

[31] L. C. Jaffrin and J. Visumathi, "An intelligent model for early prognosis of heart illness," *AIP Conference Proceedings,* vol. 2802, no. 1, p. 120016, 2024. https://doi.org/10.1016/j.colsurfa.2019.123980