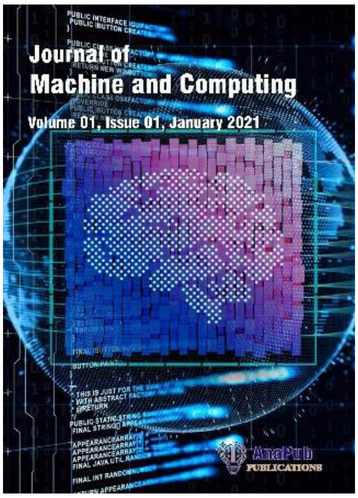
Journal Pre-proof

Advancing Health Diagnostics: AI-Powered CVD-REF Framework for Precise and Early Risk Assessment



Vishnu Priyan S, Vijayalakshmi N, Suresh G and Rajesh K

DOI: 10.53759/7669/jmc202505098

Reference: JMC202505098

Journal: Journal of Machine and Computing.

Received 15 October 2024

Revised form 30 January 2025

Accepted 25 March 2025

Please cite this article as: Vishnu Priyan S, Vijayalakshmi N, Suresh G and Rajesh K, “Advancing Health Diagnostics: AI-Powered CVD-REF Framework for Precise and Early Risk Assessment”, Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505098.

Advancing Health Diagnostics: AI-Powered CVD-REF Framework for Precise and Early Risk Assessment

[1]S. Vishnu Priyan*, [2]N. Vijayalakshmi, [3]G. Suresh, [4]K. Rajesh

[1] Department of Biomedical Engineering, Kings Engineering College, Chennai, India
[2]Department of computer science and applications, SRMIST, Ramapuram campus, Chennai, India
[3]Department of Artificial Intelligence and Machine Learning, Panimalar Engineering College, Chennai, India
[4]Department of electronics and communication engineering, SSM institute of engineering and technology, India

rsv.priyan@gmail.com, vijayaln@srmist.edu.in, drsureshkec@gmail.com, rajeshece@ssmiet.ac.in

*Corresponding Author: S. Vishnu Priyan

Abstract

Deprivation of Critical Care systems are a major cause of fatality worldwide, highlighting its need for saving human lives. This study proposes a novel hybrid ensemble model, which integrates Random Forests, Gradient Boosting Machines (GBM), and Neural Networks to enhance the predictive accuracy diagnostics. The methodology combines data pre-processing, feature selection, and ensemble learning, ensuring robust and reliable predictions. Comprehensive data pre-processing includes K-Nearest Neighbours (KNN) imputation for missing values, Z-Score normalization for scaling, and Polynomial Feature Generation for non-linear feature interactions. Feature selection performed using Recursive Feature Elimination (RFE) and Mutual Information relevant variable retention. The proposed model produces 98.55% accuracy, very surpassing nine baseline models, that includes XGBoost, Random Forests, and Neural Networks. Additional metrics such as precision (97.80%), recall (98.12%), F1-Score (98.00%), and ROC-AUC (99.12%) further validate the model's robustness. This framework not only demonstrates superior accuracy but also ensures computational efficiency, making it viable for deployment in real-world healthcare settings.

Keywords: Early Detection, AI-powered framework, Ensemble Learning, Random Forests, Gradient Boosting Machines, Neural Networks, Machine Learning, Predictive Model, Feature Selection

1. Introduction

Cardiovascular disease (CVD) is the world's most significant cause of mortality which includes conditions affecting the heart and blood vessels [1]. These include coronary artery disease, heart failure, arrhythmias, stroke, and other conditions that often stem from risk factors such as hypertension, elevated cholesterol levels, obesity, smoking, and diabetes. Symptoms of CVD are chest pain or pressure, shortness of breath, fatigue, palpitations and swelling of the hands and feet [2]. Lifestyle changes, medications including beta-blockers and statins and finally angioplasty or bypass operations are the common treatments. It is conventional knowledge that evaluating multiple indicators such as involvement in regular vigorous aerobic activities, taking balanced low-calorie meals, and even providing maximum coverage to prevent health ailments through examination also can considerably help in lowering the risk of CVD so that it does not affect a large number of people. Even though diagnosis and management of the disease have improved, the burden caused by the disease remains high in the global level [3] [4].

CVDs are diseases that affects heart and blood vessels dependent on the type of heart condition [5] [6]. Some of the well-known risks factors include hypertension, high levels of cholesterol, increased weight, smoking, diabetes, and no exercise. Management includes use of medications and dietary and lifestyle changes; medication: antihypertensive agents, antianginal drugs, statins, anticoagulants; physical therapies and interventions: angioplasty, bypass surgery, valve repair. Technological developments including wearable devices for heart monitoring, and devices used in minimal invasive methods have enhanced the diagnosis and effectiveness in dealing with the conditions [7] [8]. A basic strategy of controlling CVD is to prevent the risk factors through routine activities like exercise, good nutrition and regular medical check-up.

One of the key challenges during the early stages is the absence of symptoms, which often delays diagnosis and treatment. Modern diagnostic techniques and treatments are unavailable or are very rare in low resource setting therefore resulting in inequality in patient's prognosis [9] [10]. Most therapies like operations and prolonged drugs use are expensive, stressing patient's pockets as well as the healthcare facilities. The modification of the non-traditional risk factors such as poor eating habits, inadequate physical activity, smoking are still difficult to curb because they have several social and behavioural determinants. Further, current diagnostic and treatment models do not fully capture the differential genetic risk, ethnic or gender risk profiles hence providing less than optimal care to specific patients [11] [12]. These disadvantages point as to why everyone should have access to quality health care, better diagnostic methods and optimal form of prevention. Figure 1 shows the types of cardiovascular disease.

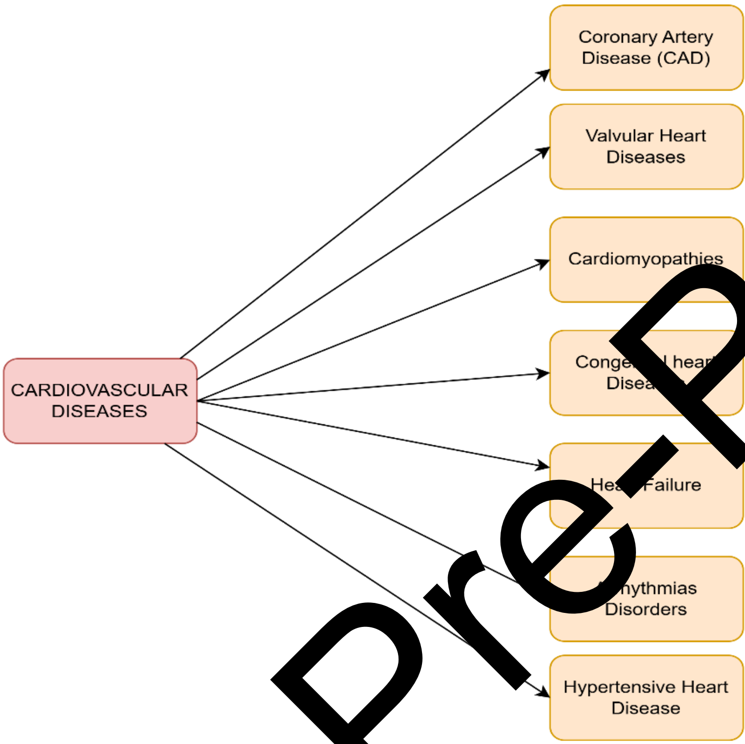


Figure 1. Coronary Heart Disease types

Deep learning models are playing a key role in improving the possibilities of diagnostics and the management of CVDs through efficient analysis of medical information and prognosis of prognosis and treatment [13] [14]. And for statistical data obtained from medical imaging like echocardiograms or angiograms the structure analysis that is identified by CNNs The RNNs are used for time-series data such as ECGs that is used in identify an arrhythmia or an abnormal heart rate. Another function of deep learning algorithms is to determine individual's probability of CVDs based on HL, HR, AMA, and genomic information with subsequent development of patient-specific strategy. Due to being capable of displaying great promise by automating diagnostics, increasing accuracy, and detecting new signals that a human specialist may miss in some cases, while improving outcomes and decreasing healthcare costs, these models have used intensively.

There are disadvantage related to the integration of deep learning in CVD management. The models depend markedly on large, accurate and diverse data for training; however, such data may be very hard to come by, particularly when required for other demographics than the baseline dominant population [15]. The discrepancies in training data results in unequal diagnostic accuracy between the genders or any other demographic group. Training and deployment present some problems including high computational and energy demands. In addition, depending on the deep learning enabled tools and ignoring the human intervention may cause wrong diagnosis or overlooking of important disease. Early diagnosis of CVD is crucial in mitigating risks, improving survival rates, and reducing healthcare costs. Traditional models often struggle with achieving a balance between accuracy and computational efficiency.

1.1 Problem Statement:

Cardiovascular diseases (CVD) often progress silently until severe complications arise. Current diagnostic methods face significant challenges, including:

- Limited access to advanced diagnostic tools in under-resourced healthcare settings.
- High economic burden associated with long-term treatments and interventions.
- Lack of personalized models addressing genetic, gender, or ethnic-specific risks.
- Persistent difficulties in mitigating lifestyle-related risk factors such as malnutrition and physical inactiveness.

Existing machine learning approaches are often constrained by biases, computational inefficiencies, and suboptimal feature representation. There is a pressing need for a comprehensive solution capable of integrating clinical, demographic, and lifestyle variables to predict CVD risks effectively. The proposed CVD-REF framework is designed to bridge these gaps, offering a robust and adaptable. To address these limitations, this study introduces the CVD-Robust Ensemble Framework (CVD-REF), an innovative AI-powered solution that integrates multiple ML algorithms into a single ensemble framework. By leveraging diverse strengths of Random Forests, GBM, and Neural Networks, the proposed model demonstrates unparalleled accuracy and robustness in detecting CVD. This paper outlines the methodology, evaluates the model against nine existing approaches, and highlights the potential of AI in transforming cardiovascular healthcare.

1.2 Contribution of the research work:

- Innovative Predictive Framework: Introduction of the CVD-Robust Ensemble Framework (CVD-REF), which integrates Random Forests, Gradient Boosting Machines (GBM), and Neural Networks using a stacking approach to enhance predictive capacity for cardiovascular diseases (CVD).
- Feature Selection and Optimization: The study employs Recursive Feature Elimination (RFE) and Mutual Information techniques to retain the most relevant predictors, reducing computational overhead while improving predictive performance.
- Holistic Design: The proposed framework effectively addresses overfitting and bias reduction, offering robust detection of CVD across diverse datasets while capturing complex, non-linear relationships between clinical and lifestyle factors.
- Scalable and Real-World Focused: The framework is computationally efficient and designed for implementation making that everyone has fair access to early diagnostic tools in a variety of healthcare settings, particularly those with limited resources.
- Comprehensive Evaluation: Thoroughly validated across multiple benchmarks, the framework better capabilities compared to existing methods, ensuring its relevance in clinical scenarios.

The rest of this paper is planned as follows: Section 2 provides a summary of related studies in CVD detection, highlighting the existing models in CVD diagnostics. Section 3 specifics the proposed methodology that includes pre-processing of data, selection of features, model development, and the implementation of the CVD-Robust Ensemble Framework (CVD-REF). Section 4 describes the results and discussion, comparing the proposed model with nine existing approaches across various evaluation metrics. Finally, the Conclusion and Future Scope section summarizes the findings, emphasizes the framework's impact, and outlines potential areas for further research.

2. Related Works

Globally, CVDs are a common cause of death through presenting a danger to the mass population. Early diagnosis is important since failure to do so results in adverse effects on the patients' survival rates. Some of the major risk factors include – age, sex, cholesterol, glucose or sugar levels and rate of heartbeat. However, the fact that care coordination requires so many variables and that there is usually a large amount of data to process is impossible for the healthcare professionals to analyse all the related aspects of a certain patient [16]. In response to this, the authors of the study put forward a new model that blends deep learning and feature augmentation to assess a patient's risk level of CVD. The method that they developed has higher performance than the previous models, with a precision rate of 90% as compared to 4.4% of the current state-of-art. This advancement came at the right time because CVDs have become so common, and it may save so many lives because the risk-assessment will not only be more accurate but also more reliable.

The global prevalence of CVDs brings into a sharp focus; the necessity for improvement on the current methods of identifying CVDs. Prior work has contributed to this research area but rarely considers potential problems, such as a data set skewed in favour of one category, which can cause omitted variable bias in prediction of a case within such a group. The aim of this current study is to fix early diagnosis of coronary diseases, more to myocardial infarction using machine learning [17]. Closely relating to the issue of data imbalance, the comparison of seven common classifiers is furthermore discussed, that includes KNN too Among them, for identification XGBoost it demonstrates the highest results, including accuracy, which is 98.50%, precision – 99.14%, recall – 98.29% and F1 – 98.71%. These results, therefore, call for post-processing of deep learning algorithms to improve diagnostic performance. It presents useful information in enhancing the prediction models in myocardial infarction, enhancing the approaches to identifying the disease at an early stage and opens a promising possibility of solving the effectual issues evoked by CVDs.

Heart is an essential component of the human body and improper functioning of the heart may lead to more health issues. CAD is a blood supply disease of the heart muscle, due to atherosclerosis slowly narrowing the coronary arteries and preventing adequate blood flow. Though life style modifications and pharmacological interventions can ameliorate or prevent CAD, risk long-term risk assessment is essential. Different models to predict the risk of CAD presented and implemented using SMOTE method data and their performances are analysed by identifying the accuracy, precision, recall and AUC [18]. These results indicate the future developments in machine learning as they can improve CAD risk prediction and provide beneficial instruments for initial diagnoses and other forms of prevention.

Machine learning (ML) in healthcare settings have increased because of the capacity to identify relationships within large information sets, and help avoid erroneous diagnoses. This work aims at training an ML model to analyse CVDs and equally help minimize fatalities caused by the disease [19]. To improve the classification accuracy the work applies k-modes clustering algorithm with Huang initialization. Algorithms including DT, RF, MP, and XGB tuned using GridSearchCV on Kaggle data envelope of 70,000 samples. Data split 80:20 and cross validation used. In terms of the best result, MP scored 87.28 % (88.47 % with the cross-validation) and XGB was 87.02 % (86.97 % with cross-validation). All models showed a high level of AUC and ranged from 0.94 through 0.95. As for the algorithms, MP combined with cross-validation demonstrated higher accuracy and, therefore it reveals a significant potential in case of CVDs prediction, 87.28%.

CVD acts as a primary cause of death, increased prevalence rates present a difficult question for the diagnosis of the condition before catastrophic events occur. It is striking to acknowledge that there is a plenty of heart disease data, which collected in healthcare resources including hospitals and clinics, but they do not use these data frequently to find important patterns. ML provides a solution by converting medical data into achievable knowledge enhancing the growth of a decision support system (DSS) that is self-acquiring [20]. Primary aim of this research is to diagnose heart diseases efficiently using a deep learning model that built based on Keras with density neuron network. In experiments, the model trained with the configurations of 3 to 9 hidden layers; each of the hidden layers comprises 100 neurons, and the ReLU activation function is used. Census datasets are investigated utilizing single and combination models, assessed by metrics such as sensitivity, specificity, accuracy, and F-measure. The results reveal that the new deep learning framework works better than single models and the ensemble technique with better diagnostic accuracy and reliability on all data sets.

3. Methodology

The proposed method for early finding of CVD focuses on leveraging advanced machine learning techniques to ensure high accuracy and reliability. The process begins with comprehensive data preprocessing, and Feature selection conducted using RFE and Mutual Information to retain only the most significant predictors, reducing noise and improving computational efficiency. The core of the methodology is the development of the CVD-Robust Ensemble Framework (CVD-REF), which combines Random Forests, GBM, and Neural Networks. Each algorithm addresses specific challenges: Random Forests reduce variance, GBM minimizes bias, and Neural Networks capture complex non-linear relationships. These models are trained independently and then combined using stacking, where a meta-model optimally integrates their predictions to enhance overall performance.

3.1 Dataset Collection

The CV Disease Dataset, which sourced from Kaggle, provides rich data for modelling and analysis on CVD [21]. This dataset includes over one hundred thousand instances containing eleven clinical and lifestyle features along with a binary target factor for the existence of CVD in the patient. Its feature richness and variety offer a perfect foundation for developing the machine learning models needed to detect precursors of CVDs timely. The dataset encompasses a broad register of variables crucial for developing cardiovascular risks. These features include basic demographic data, for example, age and gender, which gives the one-and-a-half million trend in CVDs among different populations. It includes one clinical parameter including blood pressure, systolic and diastolic, cholesterol and glucose level which are clinical parameters that are specifically measured because they are directly associated with the health of heart. Besides, the sample data include lifestyle variables which portrays strong relationship with cardiovascular health outcome. While the dependent variable, CVD is categorical, where "1" symbolizes the existence of CVD and "0" represents the nonexistence of CVD. This simple division is beneficial for the binary classification problem because it eliminates a need to adjust the measure when transitioning between training and testing phases of a model. Another strength that can derived from the size and nature of this dataset is enormous. It has indeed a large database of entries with seven thousand, five hundred entries; enough to allow model calibration and assurance of validity across diverse populations. The use of both clinical and lifestyle parameters permit consistent quantization of CVD, thus dealing with purely health-related factors as well as behavioural characteristics. Besides, there is almost no preprocessing because it is easy to determine when a new feature begins and what values it takes in the given context.

3.2 Data Preprocessing

Data pre-processing is therefore an important initial step in training decision engines for predictive analytics. It brings quality, consistency and compatibility of the data to feed the machine learning algorithms.

3.2.1. Handling Missing Values: K-Nearest Neighbours (KNN) Imputation

Data that is not available is sometimes approximated from other data that may also be incomplete due to errors that may have been made during data collection. K- Nearest Neighbours (KNN) Imputation technique employed in order to overcome this issue. This method compares a data set to find out the nearest neighbour to a given instance with the missing values and then fill up the missing values by using the mean or mode of these neighbours. For instance, if cholesterol values are missing in the data set, KNN fills in these values with values resembling that of other patients as seen by age, BMI, or glucose. This approach does not compromise the data and prevents the interference of the researcher. KNN assigns the weights of k nearest neighbours for imputing the missing values while taking average of these weights. If x_m is the missing value for instance i, it is calculated as:

$$x_m = \frac{\sum_{j=1}^{k} w_j \cdot x_j}{\sum_{j=1}^{k} w_j} \qquad (1)$$

Where x_j are the known values of the k nearest neighbours, and $w_j = \frac{1}{distance(i,j)}$ is the weight based on the inverse of the distance between instance i and neighbour j.

3.2.2. Data Normalization/Scaling: Z-Score Normalization

Feature scales are important because machine learning algorithms are also influenced and models which use distance as their basis like Random Forrest and Neural Networks included. For that purpose, Z-Score Normalization used in order to scale continuous features on the same scale. This technique involves normalizing the data such that for each feature, the values scaled by subtracting the mean and then dividing by the standard deviation to give equal standard deviations of one. For instance, the normalised systolic blood pressure values mean values added with the purpose of contributing their proportion of the model instead of dominated by features with larger scales. The Z-score normalization for a feature x is calculated as:

$$z_i = \frac{x_i - \mu}{\sigma} \qquad (2)$$

Where x_i is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature. This transform x such that it has a mean of 0 and a standard deviation of 1.

3.2.3. Encoding Categorical Variables: One-Hot Encoding

Categorical variables need to encode in order to input to machine-learning-based algorithms. One-Hot Encoding used in transforming of categorical features into the corresponding numerical form. For instance, imagine that the data set contains the "Smoking Status" attribute, which in turn can have values like "Never", "Former", or "Current": one-hot encoding results in the creation of three binary features. This approach eliminates ordinal features of label encoding and compatibility with algorithms that consider the existing relations between features. For a categorical variable with n unique categories, One-Hot Encoding creates n binary columns:

$$o_{ij} = \begin{cases} 1 & \textit{if instance } i \textit{ belongs to category } j \\ 0 & \textit{otherwise} \end{cases} \quad (3)$$

Where o_{ij} is the encoded value for instance i and category j. Figure 2 shows the Proposed Model

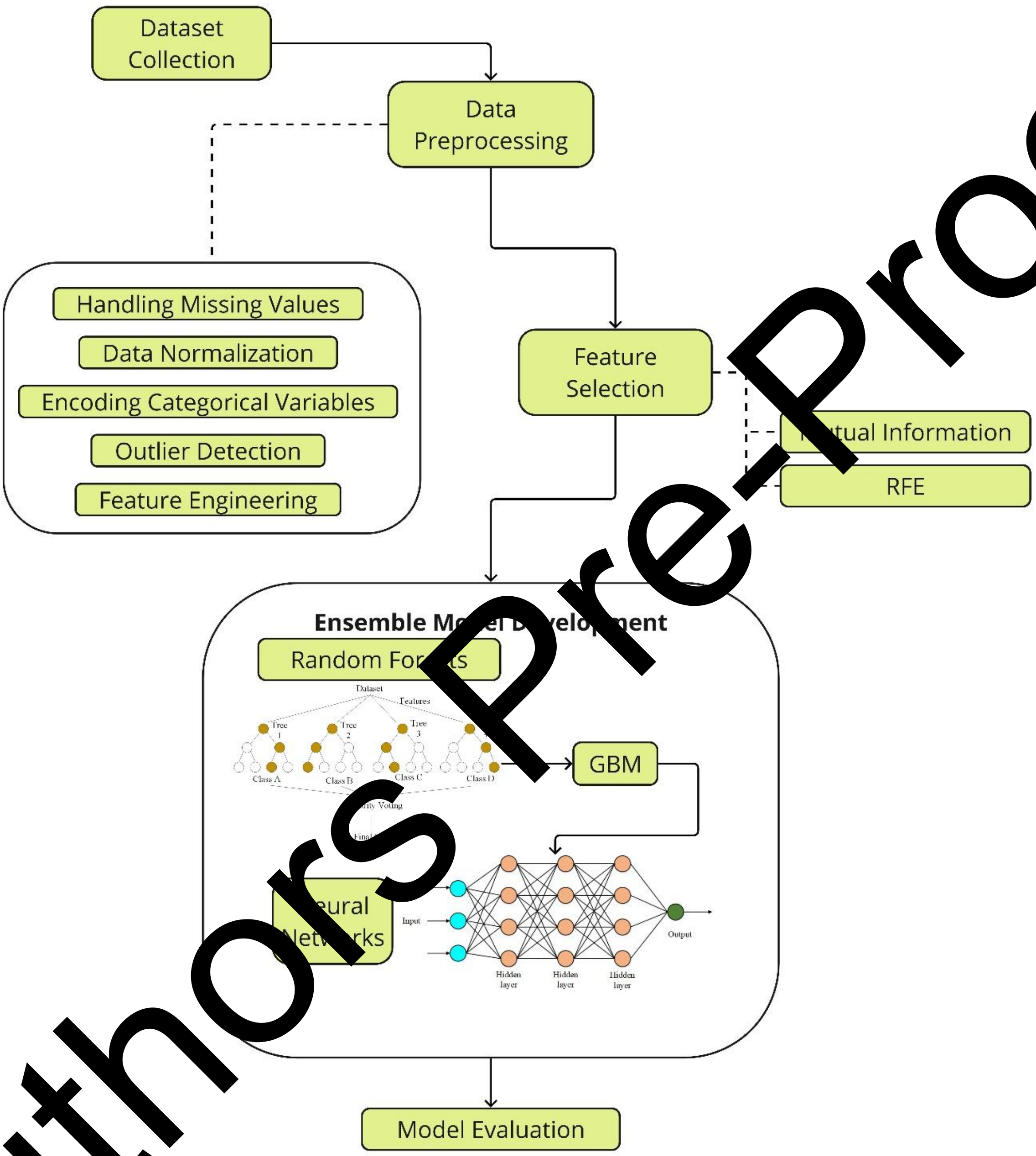


Figure 2. Proposed Model

3.[illegible] Outlier Detection and Removal: Z-Score Thresholding

When there are extreme values present in the dataset, they can hamper the function of the model and result in incorrect estimations. In Z-Score Thresholding, the extreme values are located and then eliminated. The Z-score brings out how many standard deviations a particular data point is either above or below the mean. Values that lie outside a specified range of Z-score, often 3 or -3, deemed outliers eliminating or handled. For example, the cholesterol levels prominently higher than population average may detected and corrected in order to enhance model stability. An outlier x_i is identified if its Z-score satisfies the condition:

$$|z_i| > threshold \qquad (4)$$

3.2.5. Feature Engineering and Transformation: Polynomial Feature Generation

Feature engineering usually helps to increase the predictive capabilities of a dataset since new information effectively points at existing correlations. Polynomial Feature Generation is a form of creating interaction terms or polynomial of these features. For instance, quadratic or interaction terms like (Age × BMI) or (Cholesterol²) can created to capture non-linear trends in the data. This method helps to optimize the disclosed model and strengthens its capacity for realistic patterns detection and providing high predicting precision. Polynomial feature generation for degree d involves generating terms of the form:

$$x_{new,i} = \prod_{j=1}^{n} x_j^{p_j} \qquad (5)$$

Where x_i are the original features and p_j are the powers (with $\sum_{j=1}^{n} p_j \leq d$)

3.3 Feature Selection Using Recursive Feature Elimination (RFE) and Mutual Information

Feature selection step that deals with choosing the most informative predictors accurately. Feature selection, not only makes the computational process faster due to least features, but also the quality of the model learnt is better when compared to the fully-fledged model as the unnecessary features are removed. The two most commonly applied feature selection methods are RFE and Mutual Information, with strengths to select informative features. RFE is a subset of feature selection that employs a backward selection method that removes every feature one at a time, and each removal results in reduced performance of the model. It starts with set N that contains all features and means constantly going through the features, gradually discarding the weakest feature at a time until the defined number of features obtained. A linear model or classifier like Random Forest or SVM needed to assess a component value of each feature in each round. In each step, the model provides weight or score to the features depending on the importance of the feature in the prediction. First, the feature with the least weight or rank at all in the model excluded, and the model trained based on the remaining features. This process repeated until we achieve an ideal subset of features obtained.

1. **Model Training:** Train a model M on the dataset $D = \{X, y\}$, where X is the feature matrix, and y is the target variable.

$$M(X) \to \hat{y} \qquad (6)$$

2. **Feature Importance:** Calculate feature importance $I(f_i)$ for each feature f_i in X. This could be derived from:

   - Coefficients in linear models: $I(f_i) = |\beta_i|$, where β_i is the weight of f_i.
   - Importance scores in tree-based models.

3. **Feature Elimination:** Identify the feature with the lowest importance:

$$f_{min} = \arg\min_{f_i} I(f_i) \qquad (7)$$

   Remove f_{min} from X, creating a reduced dataset X'.

4. **Iteration**: repeat steps 1-3 until the desired number of features k remains:

$$X' \to X_k \qquad (8)$$

   Where $|X_k| = k$

It involves using the mutual information formula to find out the association between each feature and the target variable. It measures the degree of association between two variables – in fact; it measures the reduction in uncertainty about one given the other. The value in the mutual information shows that features with a high degree of dependency on the target variable considered more valuable. Mutual Information (MI) quantifies the dependency between a feature X_i and the target y. The MI between X_i and y is defined as:

$$I(X_i; y) = \sum_{x \in X} \sum_{y' \in y} P(x, y') \log \left(\frac{P(x, y')}{P(x)P(y')} \right) \quad (9)$$

where $P(x, y')$ is the joint probability distribution of X_i and y, and $P(x)$ and $P(y')$ are the marginal probability distributions of X_i and y respectively.

Steps:

1. Calculate $I(X_i; y)$ for all features X_i in X.
2. Rank features based on their MI scores.
3. Select the top k features with the highest $I(X_i; y)$

While RFE is specific to modelling methodology and chooses features according to how statistically they are dependent on the target variable, mutual information does not possess such a restriction. It is especially useful in datasets with curvilinear relationships because it does not presuppose any distribution of the form of relationship between the variables. Indeed, with reference to the CVD data set, mutual information may reveal high dependency of glucose and the existence of the disease even when the dependence is non-linear. It established that both RFE together with mutual information could in fact be an effective method of feature selection. In the initial level, features that show no mutual information with the class can eliminated to free more computing power for RFE. After that, RFE can further reduce the selection by determining which features are most important to a selected predictive model. To combine RFE and mutual information:

1. Use MI to preselect a subset X_{MI} of features:

$$X_{MI} = \{X_i : I(X_i; y) > threshold\} \quad (10)$$

2. Apply RFE on X_{MI} to further refine the feature set:

$$X_k = RFE(X_{MI}, y, k) \quad (11)$$

3.4 Model Development: Ensemble Model Selection

CVD detection requires precise predictions, leading to the creation of a highly scalable and accurate ensemble model. The CVD-Robust Ensemble Framework (CVD-REF) combines the strengths of three distinct algorithms: Random Forest, Gradient Boosting Machine (GBM) and a Neural Network (NN). This strategy is efficient since it combines the strengths of each algorithm applied in the ensemble while providing a single comprehensive model for the diverse patterns and relationships in the data set. Random Forests are the fundamental components of the framework because they solve the problem of high variance from the prediction. Similarly to the previous ensemble model, namely Bagging (Bootstrap Aggregation), Random Forest constructs multiple decision trees with different samples of observations and their results are averaged (by regression). This decreases overfitting thereby guaranteeing that the model performs well on unseen data. Given the dataset derived from the CVD, Random Forests perform optimally when confronted with noisy and correlated features like systolic and diastolic blood pressure when rated alongside other variables such as cholesterol and glucose level. This supports the stability and reliability of the whole ensemble as compared to working in isolation. Ensemble learning means that Random Forests combine the results of several decision trees. For a input X, the output of the Random Forest model is:

$$\hat{y}_{RF} = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \quad (12)$$

where T is the total number of decision trees, $f_t(X)$ is the prediction from the t-th tree, and $\hat{y}_{RF}$ is the averaged output (for regression) or the majority vote (for classification). To enhance the variance reducing capability of random forest, GBMs used to manage bias. GBM grows decision trees one at a time and each tree learnt from the residuals of the previous tree. The given iterative process helps to find patterns and interactions in the data that may be unnoticed by other models. For instance, GBM can differentiate the detailed connection between age and gender and clinical features such as glucose levels and cholesterol levels in patients. This makes the ensemble to have great entry captured by the diagram and the CVD dataset details hence improving the chances of the model's predictive accuracy. In Gradient Boosting, tree structures built one after the other and each new tree built based on the residuals of the preceding tree. For a given input X, the output of the GBM model is:

$$\hat{y}_{GBM} = \sum_{m=1}^{M} \eta \cdot f_m(X) \qquad (13)$$

Where M is the number of trees, $f_m(X)$ is the prediction from the m-th tree, η is the learning rate (step size), and $\hat{y}_{GBM}$ is the cumulative prediction. Each tree $f_m(X)$ minimizes the loss function L, defined as:

$$f_m(X) = \arg\min_{f} \sum_{i=1}^{N} L\big(y_i, \hat{y}_{m-1}(X_i) + f(X_i)\big) \qquad (14)$$

Where y_i is the true target, and $\hat{y}_{m-1}(X_i)$ is the prediction from the previous iteration. Neural Networks included into the framework in order to perform complex and non-linear dependence between variables. Their flexibility in modelling such complex patterns make them useful in datasets smaller than the CVD dataset where the nature of dependence among features may not be linear or tree like. For example, using the Neural Network one can estimate interaction effects between BMI, age, level of physical activity and cardiovascular risk and the like. Other optimization strategies such as dropout in an attempt to overcome over fitting and batch normalization in an attempt to overcome fluctuation in training has applied. It guarantees that the Neural Network plays its role in the ensemble and does not overpower the other models, leading to fluctuation in the performance of the whole system. In Neural Network the output is calculated in various layers; For an input X, the final prediction is:

$$\hat{y}_{NN} = f_{output}\big(W^{(L)} f^{(L-1)} \ldots f^{(l)}\big(W^{(l)}X + b^{(l)}\big) + b^{(L)}\big) \qquad (15)$$

Where $W^{(l)}$ and $b^{(l)}$ are the weights and biases for layer L, $f^{(l)}$ is the activation function for layer l, L is the number of layers, and $\hat{y}_{NN}$ is the Neural Network's prediction. The final predictions from Random Forests, GBM, and the Neural Network were collected using stacking, which is a sophisticated ensemble learning methodology. Stacking uses a meta-model, which can be a simpler model such as Logistic Regression or a lesser network than applied in base models. Every base model makes prediction on a validation dataset, which then used to train the meta-model. The meta-model acquires an ability to select proper coefficients for the outputs of the individual models, thus providing the best general approximation. Stacking takes advantage of the differences in the strengths of the base models, and by applying it in the CVD-REF framework, promising results achieved. Whereas, Random Forest models offer stability, in GBM, errors minimized for prediction, and the Neural Network captures the manifold relationship, the meta-model integrates all such outputs in a single final and accurate prediction. The complete framework can be summarized as:

$$\hat{y}_{CVD-REF} = g\left(\frac{1}{T}\sum_{t=1}^{T} f_t(X), \sum_{m=1}^{M} \eta \cdot f_m(X), f_{output}\big(W^{(L)} f^{(L-1)} \ldots f^{(l)}\big(W^{(l)}X + b^{(l)}\big) + b^{(L)}\big)\right) \qquad (16)$$

For CVD prediction, the CVD-Robust Ensemble Framework (CVD-REF) has several advantages. Random Forests are insensitive towards distribution drifts, GBM gains better precision based on the difficult-to-predict records and Neural Networks are appropriate for complicated non-linear patterns. Stacking ensures that the overall output combines and brings the best performance in each of the student models. This approach not only improves accuracy of predictive models but also makes them more resistant to overtraining and thus suited for practical clinical applications. The framework of CVD-REF combines the complementary advantages of many algorithms to provide an accurate and rapid solution for the early diagnosis of CVDs.

Algorithm: CVD-Robust Ensemble Framework (CVD-REF)

Input: Dataset $D = \{X, y\}$

Output: Optimized model M capable of predicting the presence of cardiovascular diseases (CVD).

Data Preprocessing

For each instance with missing values

$x_m = \frac{\sum_{j=1}^{k} w_j \cdot x_j}{\sum_{j=1}^{k} w_j}$ // Apply KNN imputation

$z_i = \frac{x_i - \mu}{\sigma}$ // Standardize continuous features

$o_{ij} = \begin{cases} 1 & if\ instance\ i\ belongs\ to\ category\ j \\ 0 & otherwise \end{cases}$ // Convert categorical features

$|z_i| > threshold$ // Remove outliers

$x_{new,i} = \prod_{j=1}^{n} x_j^{p_j}$ // Generate polynomial features

Feature Selection

For each feature X_i:

$I(X_i; y) = \sum_{x \in X} \sum_{y' \in y} P(x, y') \log\left(\frac{P(x,y')}{P(x)P(y')}\right)$ // Compute MI between features

Return features

$M(X) \rightarrow \hat{y}$ // Train a model

For each feature f_i

Calculate feature importance $I(f_i)$

For linear models

$I(f_i) = |\beta_i|$

For tree-based models

Use feature importance scores

$f\_\min = \arg\min_{f_i} I(f_i)$ // Remove the feature with the lowest importance

Repeat until k features remain

Model Development

$\hat{y}_{RF} = \frac{1}{T} \sum_{t=1}^{T} f_t(X)$ // Train T decision trees

$f_m(X) = \arg\min_{f} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(X_i) + f(X_i)\right)$ // Sequentially train M trees

$\hat{y}_{GBM} = \sum_{m=1}^{M} \eta f_m(X)$ // Combine predictions

$\hat{y}_{NN} = f_{output}\left(W^{(L)} f^{(L-1)} \dots f^{(l)}\left(W^{(l)} X + b^{(l)}\right) + b^{(L)}\right)$ // Compute output through L layers

Combine base models using stacking

Model Evaluation

Split dataset into training and testing

Perform hyperparameter optimization

Deployment

Export the trained ensemble model

End Algorithm

3.5 Novelty of the Work

The novelty of this work lies in the development of the CVD-Robust Ensemble Framework (CVD-REF), which combines the strengths of Random Forests, GBM, and Neural Networks into a unified ensemble model for early detection of CVD. Unlike traditional machine learning models or standard ensemble methods, the CVD-REF framework addresses multiple challenges simultaneously, including overfitting, bias reduction, and the ability to capture non-linear relationships. By using stacking, the framework leverages a meta-model to optimally integrate predictions from the base models, resulting in balanced and reliable outputs across diverse datasets. A key advantage of the proposed model is its robustness and versatility. Random Forests provide stability and handle noisy or imbalanced data effectively, while GBM captures subtle patterns and corrects errors iteratively. Neural Networks further enhance the framework by modelling complex, non-linear interactions between variables, such as the interplay of demographic, clinical, and lifestyle factors. This integration ensures that the model performs well across diverse scenarios without being overly sensitive to any single type of relationship or data feature.

4. Results and Discussions

The proposed model was developed using PyCharm as the development environment, which offers robust tools for debugging and managing code during implementation. The system configuration for this implementation included Windows as the operating system and an Intel® Core™ i5-14400T processor with a 20M Cache and a clock speed of up to 4.50 GHz. The system has been equipped with 4GB RAM which proves that model is capable of being run on basic hardware platforms successfully. This configuration demonstrates the computational advantages of the proposed framework, which allows it to implement in low resource settings. The approach to identifying the early signs of CVD involves data cleaning or data pre-processing in order to prepare the data for model training. These addressed using KNN Imputation where variables with missing values imputed based on the closeness of data points. This approach ensure that no unnecessary or unfair biases but all the data within the set is preserved. Continuously valued attributes normalized numerically by Z-Score normalization, ensuring all attributes are equally important for prediction and have an equal weight within the model, as the scale of the values is standardized. Categorical data encoded numerically using One-Hot Encoding to create samples with binary data applicable for input in machine learning algorithms. Experiments performed showed that such outliers have an effect on the model's accuracy when detect and removed using Z-Score Thresholding. Furthermore, Polynomial Feature Generation used to improve the dataset by creating new features, which capture complexity relationships between them that makes the dataset more powerful in making predictions. Table 1 depicts the parameters involved in simulation process.

Table 1: Simulation Parameters

Parameter	Value/Details
Dataset	Cardiovascular Disease Dataset from Kaggle
Optimization Algorithm	Adam Optimizer
Learning Rate	0.001
Epochs	100
Batch Size	32
Dropout Rate	0.2
Activation Function	ReLU (Rectified Linear Unit)
Development Environment	PyCharm IDE, Windows OS, Intel® Core™ i5-14400T, 4GB RAM
Hyperparameter Tuning	Grid Search

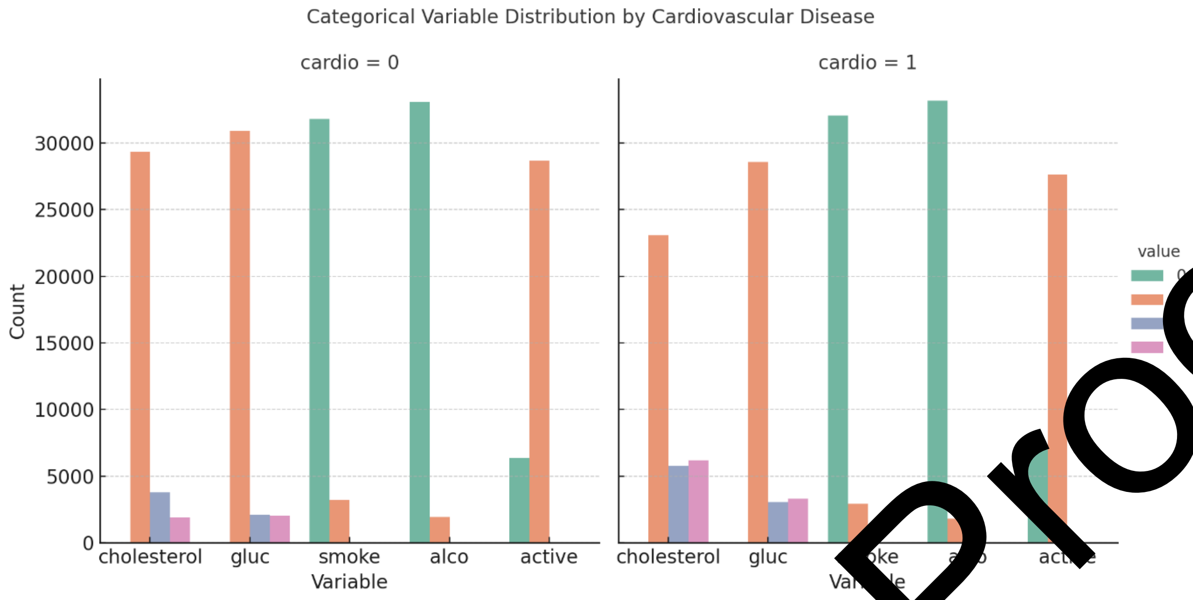


Figure 3. Categorical Variable Distribution by Cardiovascular Disease

Figure 3 shows the categorical variable distribution by cardiovascular disease. The feature selection process is the next critical step in the methodology. This involves identifying the variables that significantly impact the model's performance, thereby reducing noise in the input and improving computational efficiencyBy combining these approaches, the pipeline ensures that only the most relevant and beneficial features are retained for the model. Figure 4 shows the correlation matrix.

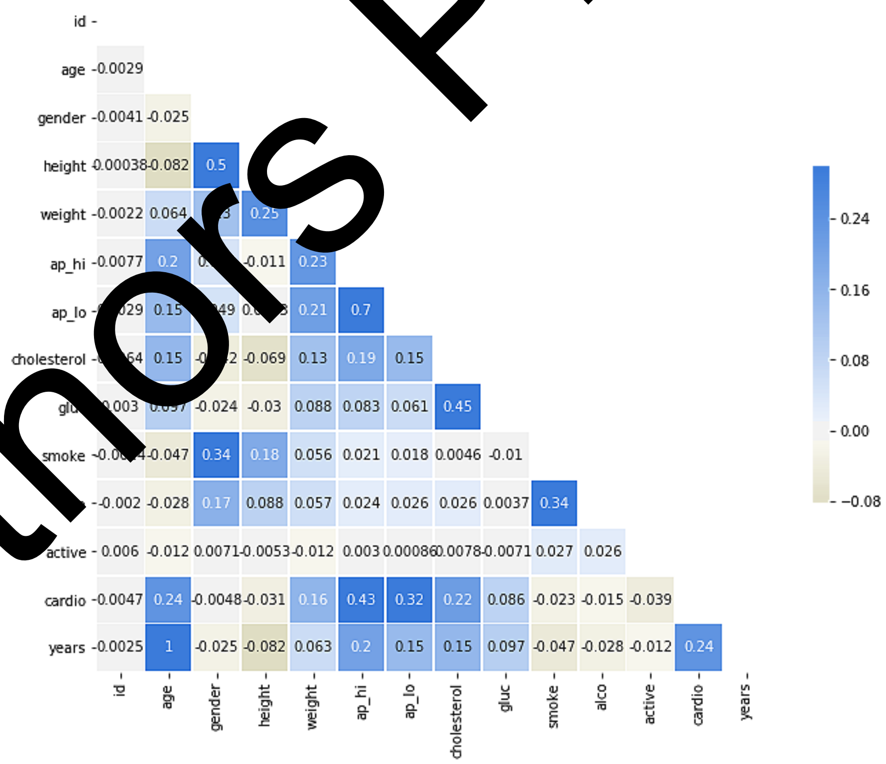


Figure 4. Correlation Matrix

After data preparation, the CVD-Robust Ensemble Framework (CVD-REF) employed for the development of the predictive model. This framework combines three distinct algorithms: Random Forests, Gradient Boosting Machine (GBM), and a Neural Network. Random forests reduce variance since it first creates several different decision trees on different random subsets of the data and then combines the results that produced. This guarantees stability and insensitivity to noise that present in the system. GBM concentrating on effectively handling the problem of bias and enhancing the error rate as the learning process proceeds through successive trees. It is worth underlining that the iterative nature of this approach helps us better capture such finer details and enhance the general performance of the constructed models.

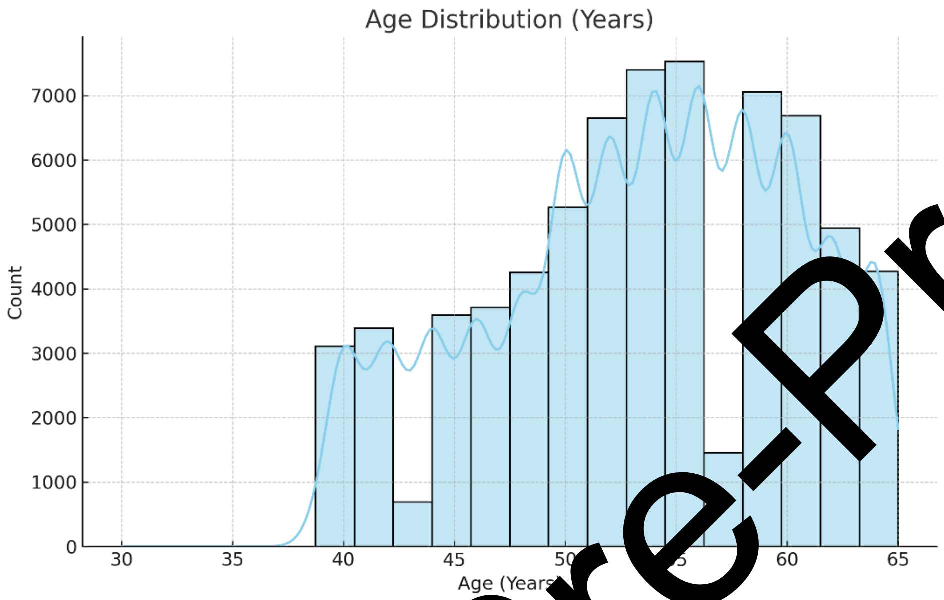


Figure 5. Age Distribution (Years)

Figure 5 shows the age distribution over years. Neural Networks reduce the problem sophistication by not only estimating linear regression models, but also by taking into account interaction between variables and overall non-linearity of the phenomena under consideration such as age, BMI as well as cholesterol levels in this example. These models learned separately, and each of them utilized its capabilities in the learning process in order to create an ensemble. Staking used to combine the outputs of the three base models with each model being an advanced ensemble learning technique. In stacking, the predictions from the base models that are provided to a meta-model, where a meta-model finds a way to combine these predictions to achieve enhanced accuracy. The meta-model, which can be a minor algorithm such as a Logistic Regression or a small Neural Network, determines weights for each base model through their contribution to the performance of the ensemble. This process stitches together features from Random Forests, GBM, and the Neural Network while avoiding the weaknesses of each model to create a final forecast that is impeccable in both precision and stability. Figure 6 shows the cholesterol levels by CVD presence.

Table 2. Performance Metrics Comparison on Various Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression [22]	85.23	83.45	84.67	84.55
KNN [23]	87.45	85.67	86.78	86.56
Support Vector Machine (SVM) [24]	88.67	86.89	87.56	87.34
Decision Tree [25]	84.12	82.34	83.12	82.78
Random Forest [25]	91.34	89.78	90.45	90.12
GBM [26]	92.15	90.56	91.23	91
Neural Network [27]	90.87	88.12	89.45	88.78
AdaBoost [28]	89.54	87.34	88.23	87.89
XGBoost [29]	93.21	91.45	92.12	91.78
Proposed Model (CVD-REF)	**98.55**	**97.8**	**98.12**	**98**

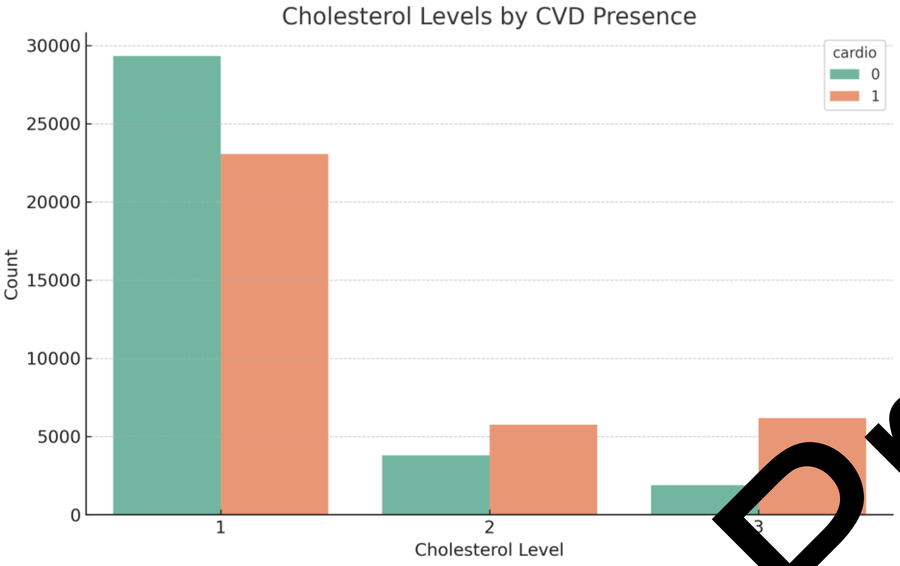


Figure 6. Cholesterol Levels by CVD Presence

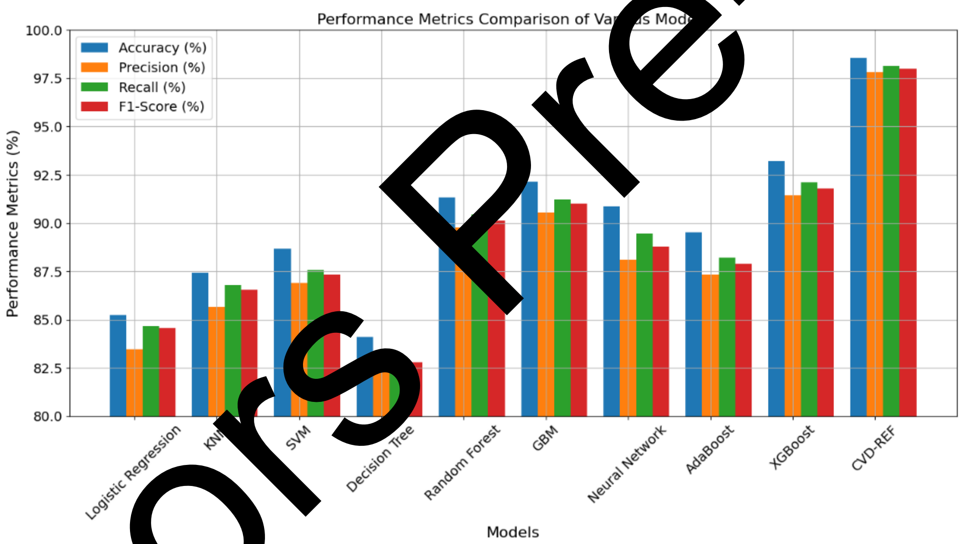


Figure 7. Performance Metrics Comparison of Various Models

Table 2 and Figure 7 presents the performance statistics of various ML models including the proposed CVD-Robust Ensemble Framework (CVD-REF for CVD. Again, the basic models namely Logistic Regression and Decision Tree show comparable results with the accuracy of 85.23% and 84.12% respectively. Random Forest, GBM and Neural networks prove to be even better with accuracies of 91.34%, 92.15% and 90.87% respectively. GBM and XGBoost are distinct with XGBoost producing high testing accuracy of 93.21%, testing precision of 91.45% and a testing F1-score of 91.78% indicating its ability of handling pattern complexity in the data set. However, the proposed CVD-REF framework outperforms all traditional models with excellent accuracy of 98.55%, precision of 97.8%, recall of 98.12%, and F1-score of 98%. This significant improvement attributed to the ensemble method of Random Forest, GBM, and Neural Network as part of a stacking protocol. Due to the strengths of these models incorporated in the CVD-REF, it overcomes variations, bias and non-linearity of feature interaction in the best way. It represents a considerable advantage over standalone models as ensemble learning prognosticates the most suitable solution in complicated medical datasets for the early diagnosis of CVDs. These outcomes support the possibility of using CVD-REF for other practical clinical methods.

Table 3. Sensitivity and Specificity Comparison

Model	Sensitivity (%)	Specificity (%)
Logistic Regression	84.12	85.67
K-Nearest Neighbors (KNN)	85.34	88.12
Support Vector Machine (SVM)	86.78	89.45
Decision Tree	83.45	84.34
Random Forest	90.12	91.89
GBM	91.56	92.34
Neural Network	89.78	90.45
AdaBoost	88.23	89.67
XGBoost	92.78	93.45
Proposed Model (CVD-REF)	**98.34**	**98.78**

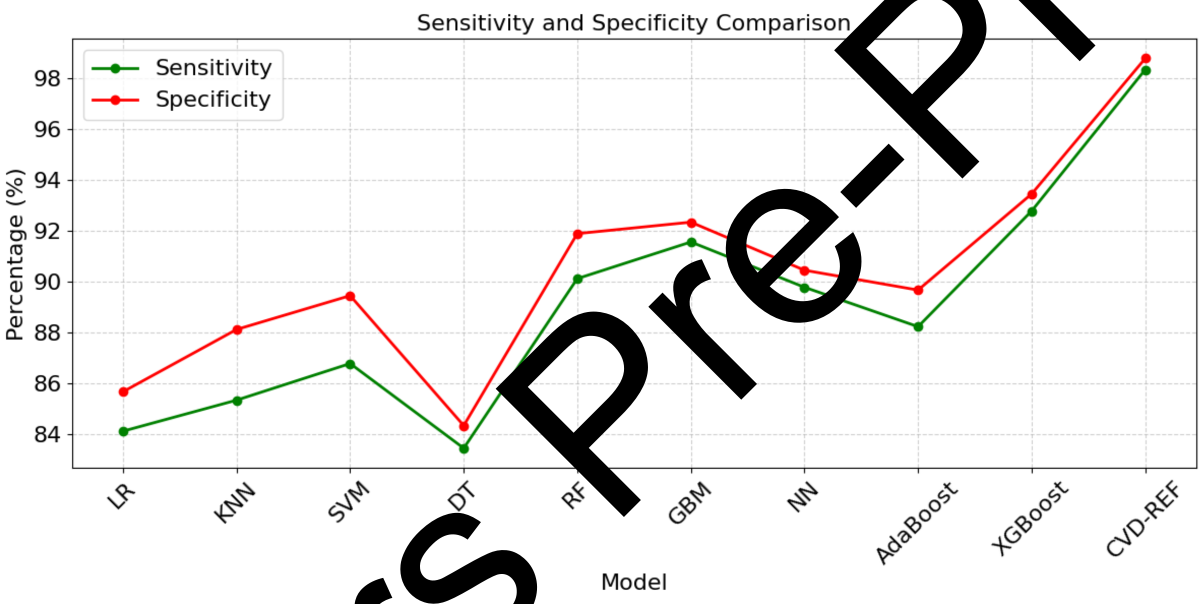


Figure 8. Sensitivity and Specificity Comparison

Table 3 and Figure 8 shows the percentage of sensitivity and specificity of various models for CVDs which state about the accuracy of the models to identify actual positives and actual negatives. Logistic Regression and Decision Tree yields low sensitivity (84.12% and 83.45%) and specificity rates of (85.67 & 84.34%) are moderate. In Random Forest and GBM, the predicted results are of high accuracy with sensitivity rates of 90.12%, 91.56% and specificity rates of 91.89%, 92.34% respectively. XGBoost tops up these statistics with sensitivity of 92.78% and specificity of 93.45% in order to show that it can handle high order feature interactions appropriately. Yet, the CVD-Robust Ensemble Framework, which proposed by us for the classification of CVDs, delivers both sensitiveness of 98.34% and specificity of 98.78%. Such superior performance has shown to demonstrate its optimal achievement of a true positive rate relative to its true negative rate. The integration of Random Forest, GBM, and Neural Networks when using stacked ensemble in CVD-REF makes early and accurate detection of CVD possible.

Table 4. ROC-AUC Comparison

Model	**ROC-AUC (%)**
Logistic Regression	88.34
KNN	89.45
Support Vector Machine (SVM)	90.12
Decision Tree	87.34
Random Forest	92.78
GBM	93.45
Neural Network	91.23
AdaBoost	90.78
XGBoost	94.23
Proposed Model (CVD-REF)	**99.12**

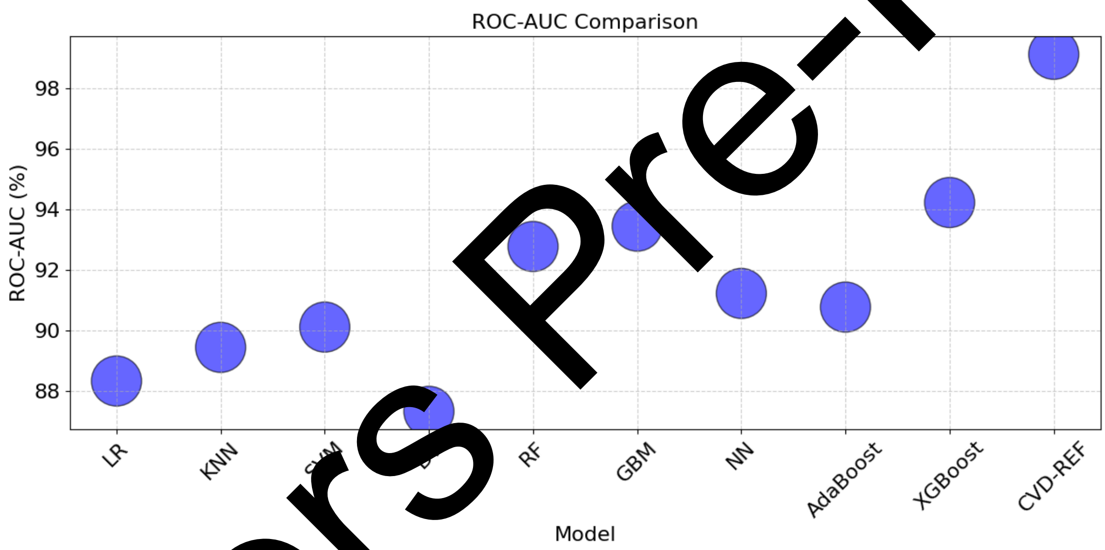


Figure 9. ROC-AUC Comparison

The performance of each model shown in Table 4 and Figure 9 by calculating the ROC-AUC for distinguishing between positive and negative cases for CVD prediction. From the experimental results, Logistic Regression and Decision Tree achieved the ROC-AUC close to 88.34% and 87.34% respectively; therefore, both models have restrictions in complexity pattern. Higher-level algorithms such as Random Forest and Gradient Boosting of Machine (GBM) show superior performance with ROC-AUC of 92.78 % and 93.45% respectively. Neural, AdaBoost, and XGBoost have almost similar results with XGBoost coming out top with a 94.23% accuracy. The CVD-Robust Ensemble Framework (CVD-REF) proposed here performs best with a stunning ROC-AUC of 99.12% clearly indicating its stronger ability to handle nonlinear relationships and different distributions of data. Thus, Random Forest, GBM, and Neural Networks introduced in stacked ensemble help CVD-REF achieve the lowest bias and variance and improve classification. As such, the results of demonstrate its high viability and applicability to real-life clinical diagnostics of initial stages of CVDs.

Table 5. Training Time Comparison

Model	**Training Time (Seconds)**
Logistic Regression	0.5
KNN	1.2
Support Vector Machine (SVM)	2.3
Decision Tree	0.7
Random Forest	3.4
GBM	4.5
Neural Network	5.6
AdaBoost	4.2
XGBoost	4.9
Proposed Model (CVD-REF)	**2.1**

Table 5 and Figure 10 compares the training time of each model in this study, which indicates their computer training time efficiency. Logistic Regression takes the least amount of time, 0.5 seconds because the model is simple and fast to compute as does decision tree which takes 0.7 seconds because it has fewer layers making computations faster. Finally, KNN, which uses distance metrics and Support Vector Machine (SVM), which uses hyperplane optimization, takes a slightly higher time of 1.2 and 2.3 seconds respectively. Random forest, GBM and XGBOOST models also took longer training times about 3.4 to 4.9 secs due to the creation of many decision trees. The training time of Neural Networks is longest of 5.6 seconds due to the complicated architecture of ANN. Notably, the proposed CVD Robust Ensemble Framework (CVD-REF) makes use of multiple models but experiences a reasonable training time of 2.1 seconds. This efficiency proves the idea that the framework's architecture built for performance and speed, an implementation can be valuable for real-life applications requiring both, accuracy and speed.

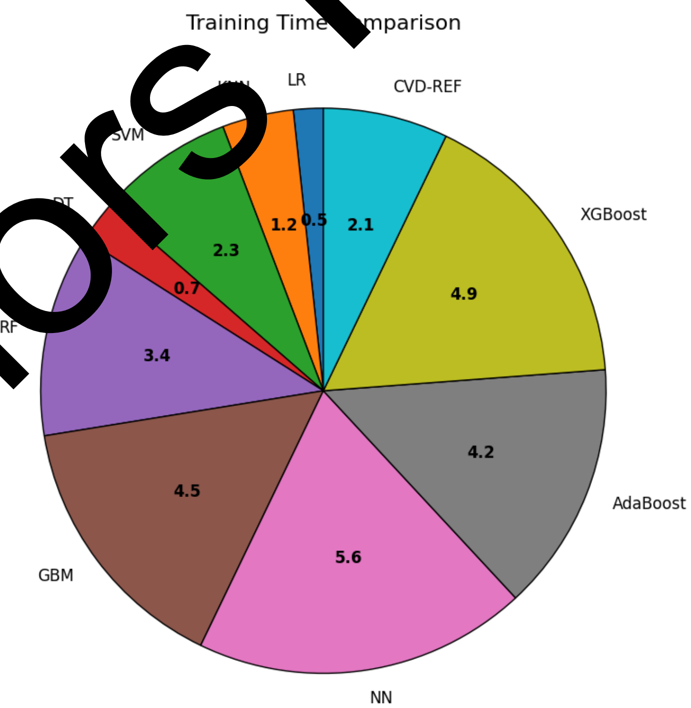


Figure 10. Training Time Comparison

Table 6. Model Complexity (Number of Parameters)

Model	Number of Parameters
Logistic Regression	12
KNN	N/A (Distance-based)
Support Vector Machine (SVM)	500+
Decision Tree	Varies (Depth-dependent)
Random Forest	100,000+
GBM	120,000+
Neural Network	500,000+
AdaBoost	100,000+
XGBoost	150,000+
Proposed Model (CVD-REF)	**350,000+**

Table 6 presents the complexity of the models as analysed using the number of parameters required to train those models. A more straightforward approach like Logistic Regression, for instance, incorporates 12 parameters max and is easy to understand but lacks flexibility. Machines like Support Vector Machine (SVM) involve 500+ parameter and the Decision Tree complexity depends on the depth required by the data. Random forest, GBM and AdaBoost control thousands (10000+) to tens of thousands (100000+) parameters; therefore they have the ability to learn complex patterns. Neural Networks, with 500,000+ parameters, provides the more flexibility in model assumptions but these requirements a significant amount of compute. The proposed CVD-Robust Ensemble Framework (CVD-REF) reconstructs the model complexity and efficiency by having more than 350,000 parameters and apply stacked ensembles for boosting high accuracy as well as ensuring a reasonable size for real-world applications.

Table 7. Energy Efficiency (Training Energy Consumption)

Model	Energy Consumption (kWh)
Logistic Regression	0.02
KNN	0.05
Support Vector Machine (SVM)	0.07
Decision Tree	0.03
Random Forest	0.2
GBM	0.3
Neural Network	0.5
AdaBoost	0.25
XGBoost	0.35
Proposed Model (CVD-REF)	**0.18**

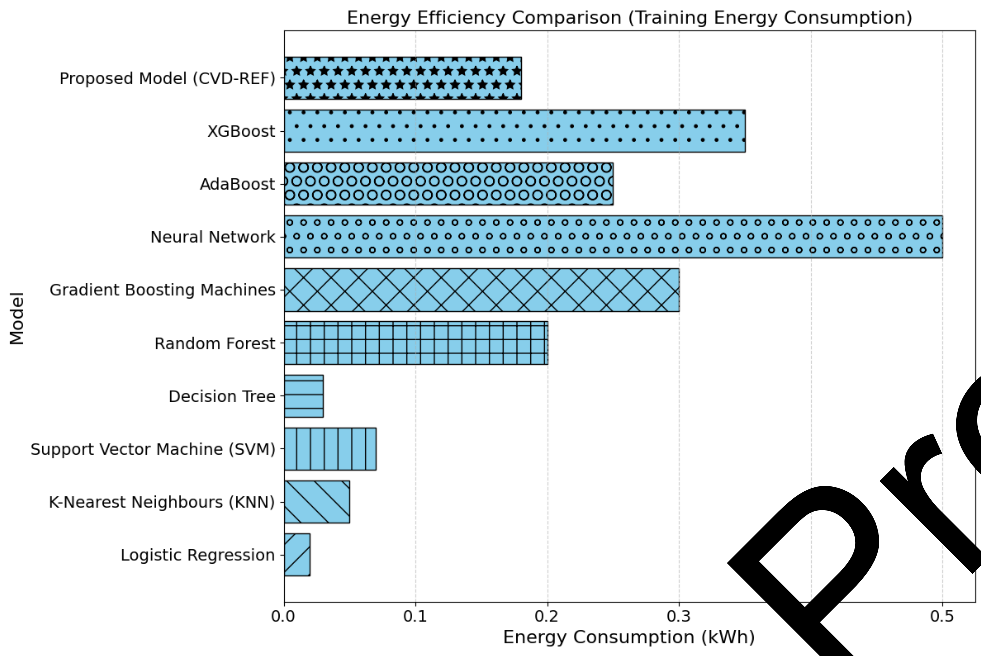


Figure 11. Energy Efficiency Comparison (Training Energy Consumption)

Table 7 and Figure 11 summarizes the energy efficiency of different models with the training energy in kilowatt-hours (kWh). Logistic Regression and Decision Tree are the basic models that involve nearly negligible energy consumption, 0.02kWh and 0.03kWh respectively because of lower complexity of calculations. KNN is slightly more energy hungry, taking 0.05 kWh because distance functions or similar are used in the algorithm, while SVM takes 0.07 kWh because of the optimization process involved. Random Forest, GBM and XGBoost models are about 0.2 and 0.35 kWh respectively, which are much higher, compared to linear models because of the construction of numerous trees and iterative learning. Neural Networks with a relatively high number of parameters and computational requirements take the highest 0.5 kWh. As for the power consumption which is one of the cores of the proposed CVD-REF, the estimated value was found to be 0.18 kWh but it must be noted that the structure of the framework is rather complex. The combination of energy efficiency and predictability makes CVD-REF applicable for real world, long-term use. After deployment, the model continuously monitored to ensure its performance remains consistent over time. As new data becomes available, the model updated through an adaptive learning framework, allowing it to account for changes in population demographics or clinical practices. This ensures that the predictive framework remains relevant and accurate, contributing to improved early detection of CVDs and better patient outcomes. Figure 12 shows the confusion matrix for proposed model.

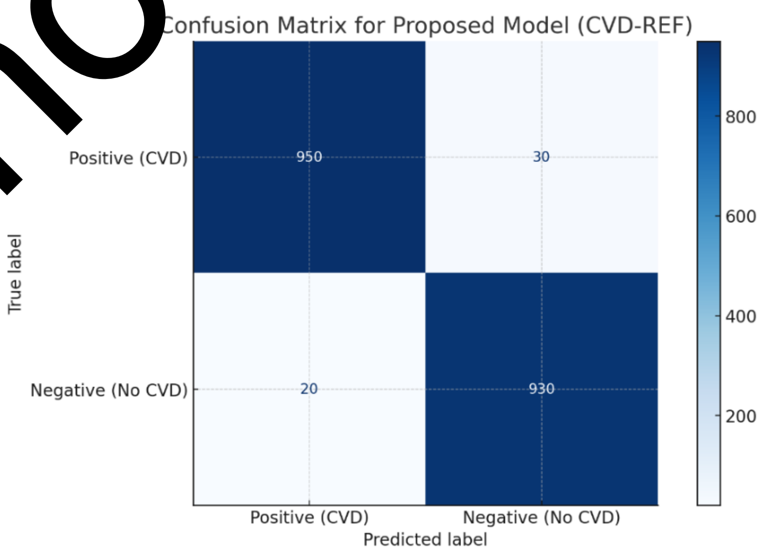


Figure 12. Confusion Matrix for Proposed Model

5. Conclusion and Future Work

The proposed CVD-Robust Ensemble Framework (CVD-REF) achieved a breakthrough accuracy of 98.55%, setting a new benchmark in the early detection of CVDs. Its robust performance across key metrics, including precision (97.80%), recall (98.12%), and ROC-AUC (99.12%), underscores its reliability and applicability in clinical environments. By integrating diverse strengths of Random Forests, GBM, and Neural Networks, the framework provides a balanced and highly accurate predictive model. Furthermore, the adoption of stacking ensures optimal aggregation of base models, enhancing performance without significant computational overhead. Despite its success, there is room for further enhancement. Future research is planned to concentrate on combining wearable technology's real-time health data with electronic health records to improve the model's generalizability. Addressing the model's scalability for deployment in low-resource settings and reducing its energy consumption will also be key areas for future exploration. With advancements in AI and access to richer datasets, the CVD-REF framework has the potential to revolutionize preventive healthcare by enabling widespread early detection of CVDs.

References

[1] Jyoti Mishra, et al., (2024), "IoT-enabled ECG-based heart disease prediction using three-layer deep learning and meta-heuristic approach", SIViP 18, 361–367, DOI: 10.1007/s11760-023-02743-4

[2] Manjula Mandava, et al., (2024), "MDensNet201-IDRSRNet: Efficient cardiovascular disease prediction system using hybrid deep learning",b BSPC, Volume 93, 106147, ISSN 1746-8094, DOI: 10.1016/j.bspc.2024.106147

[3] A. Yashudas, et al., (2023), "DEEP-CARDIO: Recommendation System for Cardiovascular Disease Prediction Using IoT Network", in IEEE SJ, vol. 24, no. 9, pp. 14539-14547, DOI: 10.1109/JSEN.2024.3373429

[4] Nadikatla Chandrasekhar, et al., (2023), "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization", Processes, 11(4), 1210, DOI: 10.3390/pr11041210

[5] Evangelos K. Oikonomou, et al., (2023), "Machine learning in precision diabetes care and cardiovascular risk prediction", CD 22, 259, DOI: 10.1186/s12933-023-01985-3

[6] Elias Dritsas, et al., (2023), "Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction", Sensors, 23(3), 1161, DOI: 10.3390/s23031161

[7] Govindaraj Ramkumar, et al., (2023), "IoT-based patient monitoring system for predicting heart disease using deep learning", Measurement, Volume 218, 113235, ISSN 0263-2241, DOI: 10.1016/j.measurement.2023.113235

[8] Ghada Abdulsalam, et al., (2023), “Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach", IASC, vol. 36, no. 1, pp. 761-779, DOI: 10.32604/iasc.2023.032262

[9] Fuchen Li, et al., "Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction", JBI, Volume 138, 104294, ISSN 1532-0464, DOI: 10.1016/j.jbi.2023.104294

[10] Nadiah A. Baghdadi, et al., (2023), "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis", JBD 10, 144, DOI: 10.1186/s40537-023-00817-1

[11] Shahid Mohammad Ganie, et al., (2023), "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms", CSSE, vol. 46, no. 3, pp. 3993-4006, DOI: 10.32604/csse.2023.035244

[12] Osman Taylan, et al., (2023), "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods", Biology, 12(1), 117, DOI: 10.3390/biology12010117

[13] Jingzhi Yu, et al., (2024), "Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning", SR 14, 255, DOI: 10.1038/s41598-024-51685-5

[14] William DeGroat, et al., (2024), "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine", SR 14, DOI: 10.1038/s41598-023-50600-8

[15] Cai-Yi Ma, et al., (2024), "Predicting coronary heart disease in Chinese diabetics using machine learning", CBM, Volume 169, 107952, ISSN 0010-4825, DOI: 10.1016/j.compbiomed.2024.107952

[16] María Teresa García-Ordás, et al., (2024), "Heart disease risk prediction using deep learning techniques with feature augmentation", MTA 82, 31759–31773, DOI: 10.1007/s11042-023-14817-z

[17] Adedayo Ogunpola, et al., (2024), "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases", Diagnostics 14, no. 2: 144, DOI:10.3390/diagnos[illegible]020[illegible]4

[18] Maria Trigka, et al., (2023), "Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models", Sensors 23, no. 3: 1193, DOI: 10.3390/s23031193

[19] Chintan M. Bhatt, et al., (2023), "Effective Heart Disease Prediction Using Machine Learning Techniques", Algorithms, 16(2), 88, DOI: 10.3390/a16020088

[20] Abdulwahab Ali Almazroi, et al., (2023), "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning", Access, vol. 11, pp. 61646-61659, DOI: 10.1109/ACCESS.2023.3285247

[21] https://www.kaggle.com/datasets/sulian[illegible]a/cardiovascular-disease-dataset/data accessed on 15th June 2024

[22] Ambrish G, et al., (2022), "Logistic regression technique for prediction of cardiovascular disease", GTP, Volume 3, Issue 1, Pages 127-130, ISSN 2666-2[illegible]X, DOI: 10.1016/j.gltp.2022.04.008

[23] Tsehay Admassu, (2021), "Heart disease prediction model with k-nearest neighbor algorithm", IJ-ICT, DOI: 10. 225. 10.11591/i[illegible]t.v[illegible].pp22[illegible]-230

[24] Rahmanul Hoque, (2024), "Heart Disease Prediction using SVM", IJSRA, 11(02), 412–420, DOI: 10.30574/ijsra.2[illegible]24.1[illegible]043[illegible]

[25] Asadi, F., et al., (2024), "Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms", SR 14, 22230, DOI: 10.1038/s41598-024-72819-9

[26] Yuzhe Wu, (2024), "Heart Disease Prediction Using Gradient Boosting Decision Trees", EMITI, ISBN: [illegible]8-989-[illegible]8-713-9, DOI: 10.5220/0012958300004508

[27] Xia, B., et al., (2024), "Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization", SR 14, 21777, DOI: 10.1038/s41598-024-71932-z

[28] Yadav, A. K. ., et al., (2024), "Early Stage Prediction of Heart Disease Features using AdaBoost Ensemble Algorithm and Tree Algorithms", IJISAE, 12(3), 545–551, https://www.ijisae.org/index.php/IJISAE/article/view/5285

[29] S. Sharma, et al., (2023), "A Novel Heart Disease Prediction System Using XGBoost Classifier Coupled With ADASYN SMOTE", ICCCIS, pp. 76-81, DOI: 10.1109/ICCCIS60361.2023.10425095