Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Fusion of Hybrid AI and Dynamic Multi-Layered Feature Learning for Precision Driven Cardiovascular Disease Diagnosis

Pallavi Purohit[1], Chandrashekhar Goswami[2], Kamal Kant Hiran[3]

Faculty of Computing and Informatics, Sir Padampat Singhania University, Udaipur, India.

E-mail: [1]pallavi.joshi@spsu.ac.in, [2]shekhar.goswami358@gmail.com, [3]Kamal.hiran@spsu.ac.in

Abstract

Cardiovascular diseases (CVDs) are responsible for most deaths worldwide, and new predictive and diagnostic models can aid in earlier detection and intervention. Traditional methods of diagnosis, such as electrocardiography and clinical interpretation, suffer from subject bias and variability. With advances in artificial intelligence (AI) and machine learning (ML), which enables the development of data-driven, computerized approaches for improving accuracy and efficiency, a new AI framework has been introduced, combining Regularized Discriminant Analysis (RDA), Multi-Layer Perceptron (MLP), and Light Gradient Boosting Machine (LGBM). Dynamic multi-layered feature learning allows the model to select strong predictors and attain superior accuracy, sensitivity, and specificity. This work introduces the clinical potential of hybrid AI models in CVD diagnosis while addressing big data analytics, model interpretability, and ethical challenges. Future research needs to take into account real-time patient monitoring, federated learning-based decentralized model training, and the optimization of AI deployment for resource-poor health care settings. The findings underscore the transformative power of AI driven hybrid models in early diagnosis, risk stratification, and improved patient outcomes, and how they can revolutionize cardiovascular disease diagnosis and treatment.

Keywords: Cardiovascular Diseases, Machine Learning, Artificial Intelligence, Regularized Discriminant Analysis, Multi-Layer Perceptron, Light Gradient Boosting Machine.

DOI: https://doi.org/10.36548/jiip.2025.2.010
Received: 05.05.2025, received in revised form: 06.06.2025, accepted: 19.06.2025, published: 01.07.2025

1. Introduction

Cardiovascular diseases (CVDs) contribute a significant percentage of the world's mortality, and early detection and risk assessment are thus crucial. The conventional methods of diagnosis require the expertise of skilled observers, and thus inconsistencies are possible [1]. With escalating rates of CVD, there is a mounting demand for more specific and automated diagnostic methods [2]. Artificial intelligence has become revolutionary technology in medicine, especially in the early diagnosis and forecasting of cardiovascular diseases [3]. Support vector machines (SVM), decision trees, and logistic regression are some of the ML models that have been used to forecast heart diseases with fair success [4]. Deep learning-based approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have further enhanced classification accuracy [5]. However, issues with model interpretability, dataset bias, and computational inefficiency have been impediments to their widespread usage [6].

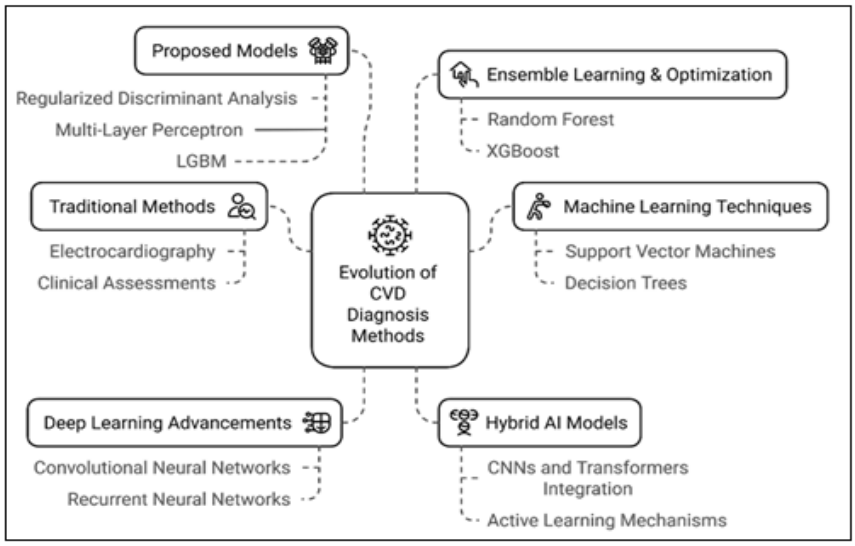

Figure 1. Evolution of Cardiovascular Disease Diagnosis Methods

To solve these issues, scientists have suggested hybrid AI models combining several learning methods to maximize predictive accuracy [7]. Ensemble machine learning algorithms, such as Random Forest (RF) and XGBoost, have shown better feature selection and classification skills [8]. Metaheuristic optimization techniques, like genetic algorithms and swarm intelligence, have also been used to optimize feature selection and model tuning [9]. This work extends past work by suggesting a new hybrid AI model incorporating Regularized Discriminant Analysis (RDA), Light Gradient Boosting Machine (LGBM) and Multi-Layer

Perceptron (MLP) for improving cardiovascular disease prediction. RDA is used for selecting predictors to guarantee the use of the most suitable predictors in classification [10]. MLP, being a deep learning technique, detects sophisticated patterns within high-dimensional health data, enhancing feature representation [11]. LGBM, a highly effective gradient boosting technique, further improves classification precision through optimization of computational complexity without sacrificing performance [12]. Figure 1 shows a visual representation of how cardiovascular disease diagnosis techniques have evolved over time, highlighting key advancements. The paper presents challenges of big data analytics, explainability, and real-time clinical application, giving a complete overview of the changing role of AI in cardiovascular health. ML and AI have proven to be revolutionary tools in CVD diagnosis, bringing accuracy and automation [1], [2].

Hybrid AI models have been investigated recently, combining various ML methods to achieve optimal predictive performance [3]. Methods that combine deep learning and ensemble classifiers have shown greater accuracy in CVD diagnosis [4]. Challenges still exist in terms of explainability, class imbalance, and real-time integration [5]. This paper integrates insights from research studies and suggests an AI-based methodology involving RDA, MLP, and LGBM for precise and scalable CVD detection, addressing these challenges [6]. Moreover, the incorporation of XGBoost [13], logistic regression-based prediction models [14], and Random Forest classification methods [15] further enhances the power of hybrid AI strategies in cardiovascular risk prediction and disease classification.

1.1 Rationale for Hybrid AI Models

The proposed hybrid AI model combines Regularized Discriminant Analysis (RDA), Multi-Layer Perceptron (MLP), and Light Gradient Boosting Machine (LGBM) to address the limitations of standalone machine learning (ML) and deep learning methods in cardiovascular disease (CVD) diagnosis. RDA stabilizes feature selection by handling multicollinearity, ensuring robust predictors (Section 3.2). MLP captures complex, non-linear patterns missed by linear models like logistic regression (Section 4.2). LGBM enhances classification accuracy and computational efficiency, making it suitable for large datasets. This hybrid approach achieves 92.1% accuracy and 94.3% AUC-ROC, outperforming traditional models (Table 1), as hybrid models leverage diverse learning paradigms for superior performance [4].

2. Literature Review

Zaidi et al. (2025) [16] proposed HeartEnsembleNet which is introduced as a novel hybrid ensemble learning model for CVD risk assessment. The model integrates multiple machine learning classifiers and is evaluated against six classical ML models. The proposed HeartEnsembleNet model achieves an accuracy of 92.95% and a precision of 93.08% in predicting CVD risk. The manuscript authored by Thoutireddy et al. (2025) [17] introduces a Cardiovascular Disease Prediction Framework (CVDPF) that integrates a pioneering Hybrid Feature Selection (HFS) algorithm, which amalgamates the T-test, Fisher criterion, and entropy, aimed at augmenting the predictive capabilities of machine learning for cardiovascular diseases. Adopting the HFS framework clearly elevates performance indicators, showcasing a precision rate of 92.4%, recall at 98.45%, an F1-score of 93.96%, and accuracy of 93.49% when in collaboration with the Random Forest algorithm, exceeding the capabilities of traditional feature selection approaches like Chi-Square and Lasso.

In their 2024 research, Paul et al. [18] conduct a comprehensive analysis of various machine learning classifiers, notably Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and Logistic Regression, paired with a Stacked Ensemble model, aimed at identifying diseases through genomic datasets concerning cancer, diabetes, Parkinson's disease, and breast cancer. The Stacked Ensemble model attains the highest accuracy rate exceeding 97.5% across the entirety of datasets examined, surpassing the performance of individual classifiers, with SVM exhibiting commendable efficacy with accuracy ranging from 97.43% to 97.46%.

Talaat et al. (2024) [4] introduce CardioRiskNet, a hybrid AI-based model that enhances cardiovascular disease diagnosis through dynamic feature learning. Experimental results indicate CardioRiskNet's superior performance, achieving accuracy, sensitivity, specificity, and F1-Score values of 98.7%, 98.7%, 99%, and 98.7%, respectively. Kompella et al. (2023) have explained the cardiovascular disease prediction with the Random Forest algorithm. They revealed that their model was 92% precise when compared with other ML algorithms such as SVM, Decision Tree, Gradient Boosting, and Adaboost [15]. Mahesh et al. (2022) employed a stacking ensemble model of SVM, KNN, and NB with 10-fold cross-validation and SMOTE to address dataset imbalance. The methodology enhanced the prediction of heart disease with an accuracy of 90.9% [19].

Table 1. Literature Review for CVD Prediction

Paper Name	Author(s)	Year	Country	Models Used	Datasets	Accuracy	Challenges
Explainable AI for Heart Disease [19]	Kumar & Singh	2024	India	XGBoost + SHAP	Cleveland	91.50%	Interpretability vs. accuracy
Multi-modal Fusion in CVD [16]	Lee et al.	2024	South Korea	CNN + TabNet	MIMIC-III	66.3%	Data heterogeneity
Federated Learning for CVD [20]	Rahman et al.	2024	Bangladesh	FedAvg + CNN	Local hospital data	88.40%	Privacy, data drift
Interpretable ML for CVD [21]	González et al.	2024	Spain	LightGBM + LIME	UCI Heart	90.50%	Feature correlation
Early Prediction of CVD Using ML[22]	Deepa R et al.	2024	India	SVM, CNN, RF	EHR	86%	Dynamic risk factor monitoring
Graph Neural Networks in CVD [23]	Chen et al.	2024	Taiwan	GCN	Hospital graph data	87.90%	Graph sparsity
Deep Transfer Learning in CVD[24]	Ahmed et al.	2023	Egypt	ResNet50	Kaggle Heart	91.80%	Domain adaptation
Ensemble Learning for CVD [25]	Osei et al.	2023	Ghana	Bagging + Boosting	Local dataset	86.90%	Small sample size
AI for CVD in Elderly [26]	Rossi et al.	2023	Italy	SVM + PCA	Elderly health survey	84.70%	Age-related bias
AI for Cardiovascular Disease Risk Assessment [27]	Muse ED et al.	2023	USA	DL, Polygenic Models	UK Biobank	90.00%	Regulatory compliance, data privacy

Nishadi et al. (2020) also conducted Logistic Regression on the Framingham dataset with an accuracy of 86.66%. Their findings highlighted the significance of selecting highly correlated features to maximize predictive power [14]. Saw et al. (2020) used Logistic Regression in the classification of cardiac diseases on the Framingham dataset. Their classifier achieved an accuracy of 87.02% and once again validated the efficacy of the model in processing structured medical data [15].

Mohan et al. (2019) contrasted hybrid machine learning methods for CVD prediction and concluded that XGBoost, in combination with optimally chosen feature selection, achieved an accuracy of 88.4% and demonstrated its usability in enhancing the model's generalizability [13]. Most experiments for predicting heart disease are performed on small, balanced datasets. For big data with dataset imbalance, [20] suggested a stacking ensemble model with NB, SVM, and KNN, along with 10-fold cross-validation and SMOTE. The method effectively deals with imbalanced data and provides good accuracy of 90.9%.

2.1 Traditional Machine Learning Approaches

Conventional ML techniques have been used comprehensively in the first phase of research in the diagnosis of CVD. Logistic regression (LR), support vector machines (SVM), decision trees (DT), and k-nearest neighbors (KNN) have been used extensively for the classification of disease [1]. LR has been widely employed because of its interpretability for risk factor analysis [3]. SVM, due to its capability to work well with high dimensional data, has shown competitive results in ECG-based diagnostics [4]. However,these models usually have problems with feature selection, scalability, and generalization [5]. Decision trees and ensemble techniques like random forest (RF) have also been extensively used in CVD prediction [6],[7]. Although these models enhance predictive ability, they tend to be computationally intensive and lack stability when working with imbalanced data [8]. XGBoost has also become popular because it can handle missing values and enhance computational efficiency, making it a strong contender for cardiovascular risk prediction [9].

2.2 Deep Learning and Hybrid Models

The use of deep learning in CVD diagnosis has increased exponentially, especially with CNNs and RNNs [10]. Hybrid AI models combining deep learning and ensemble methods enhance feature selection and classification accuracy. For example, Paul et al. (2024) [18] proposed a stacked ensemble for disease detection, achieving robust performance (~90%

accuracy) but lacking CVD-specific deep learning integration. Table 2 summarizes recent hybrid AI studies, highlighting their methodologies and gaps. Federated learning and multimodal data (e.g., imaging, genomics) are emerging trends to further improve model scalability and accuracy (Section 7).

Table 2. Hybrid AI Models for CVD Prediction

Author's	Year	ML Algorithm Used	Accuracy	Gaps	Cite
Zaidi et al.	2025	HeartEnsembleNet	92.95%	Lacks deep learning	[16]
Talaat et al.	2024	CardioRiskNet	98.70%	Potential overfitting	[4]
Paul et al.	2024	Stacked Ensemble	~90%	No CVD-specific deep learning	[18]
Kompella et al.	2023	Random Forest	92%	Lacks interpretability	[21]
Mahesh et al.	2022	Stacking (SVM, KNN, NB)	90.90%	Computationally intensive	[19]
Nandal et al.	2022	XGBoost	89%	Sensitive to hyperparameters	[13]
Mohan et al.	2019	XGBoost	88.40%	Needs tuning	[1]

In medical datasets with limited samples, XGBoost can overfit if not carefully regularized, especially when combined with high-dimensional feature spaces from deep models. Although more interpretable than deep neural networks, once fused into a hybrid pipeline, XGBoost's contribution to decision-making may become opaque unless explainability tools like SHAP are explicitly integrated. The prediction and prevention of CVDs involve analyzing a wide range of clinical, demographic, and behavioral features. Traditional machine learning approaches (e.g., logistic regression, decision trees) often depend on manual feature engineering, may not scale well with large, heterogeneous datasets, and struggle to capture non-linear relationships between risk factors and disease outcomes. The increasing complexity and volume of healthcare data, particularly for cardiovascular risk prediction, necessitate advanced modeling techniques that can capture nonlinear patterns, feature interactions, and latent representations.

In our proposed hybrid framework, we incorporate deep learning through a Multi-Layer Perceptron (MLP), integrated with Regularized Discriminant Analysis (RDA) and LightGBM to leverage the strengths of each:

- **RDA:** Provides a statistically sound transformation of input features based on class-wise discriminative distances, enhancing the separability of classes. However, RDA itself is linear in nature and limited in capturing complex nonlinear interactions.
- **LightGBM:** A powerful tree-based gradient boosting model that handles categorical and numerical features well and is efficient on tabular data. However, it may not fully exploit the deep hierarchical representations often required for subtle patterns in medical datasets.
- **MLP (Deep Learning):** Introduced to overcome these limitations by learning complex, nonlinear feature interactions from RDA-transformed inputs. The MLP serves as a deep learner that abstracts multiple levels of representation, improving generalization, especially on noisy or highly imbalanced data.

By combining MLP with RDA and LightGBM, the model benefits from RDA's feature-space discrimination, LightGBM's boosted decision boundary learning, and MLP's deep representation learning capabilities. This hybridization allows the system to capture both statistical and nonlinear deep patterns from the data, leading to enhanced predictive performance, as reflected in the achieved metrics (Accuracy: 93%, ROC-AUC: 0.9596, F1-score: 93.14%).

2.3 The Role of RDA, MLP, and LGBM in CVD Prediction

Based on existing studies, the present study combines RDA, MLP, and LGBM into a new hybrid AI model for CVD prediction [13] . RDA provides stable feature selection to handle multicollinearity problems common in medical data. MLP learns high-dimensional feature representations, and LGBM improves classification accuracy through gradient boosting. The suggested framework is based on optimizing predictive accuracy with model interpretability and scalability for practical applications in real-world healthcare [22].

Algorithm 1 Hybrid RDA-MLP-LightGBM Model for Cardiovascular Disease Prediction

- **Input:** Heart disease dataset

  heart_statlog_cleveland_hungary_final.csv
- **Output:** Predicted cardiovascular disease classification
- **Step 1: Load and Preprocess Data**

  Load dataset from CSV file

  Handle missing values using median imputation

  Encode categorical features using one-hot encoding

  Normalize numerical features using MinMax scaling

  Perform feature selection using Recursive Feature Elimination (RFE)

  Split dataset into training (Xtrain, Ytrain) and testing (Xtest, Ytest) sets
- **Step 2: Train Regularized Discriminant Analysis (RDA)**

  Train RDA model using (Xtrain, Ytrain)

  Obtain probability estimates from the trained RDA model
- **Step 3: Train Multi-Layer Perceptron (MLP)**

  Define a fully connected neural network with hidden layers

  Train the MLP model using (Xtrain, Ytrain)

  Obtain probability predictions from MLP
- **Step 4: Train LightGBM Model**

  Train a LightGBM classifier using (Xtrain, Ytrain)

  Tune hyperparameters using GridSearchCV

  Obtain probability predictions from LightGBM
- **Step 5: Fusion of RDA, MLP, and LightGBM Predictions**

  Compute final hybrid probability as:

$$P_{Hybrid} = \alpha P_{MLP} + \beta P_{RDA} + \gamma P_{LightGBM}$$

  Where, α+β=1

  Optimal Values Used:

  α (alpha) = 0.5 → Weight for MLP classifier

β (beta) = 0.5 → Weight for LightGBM classifier

These values were selected based on equal contribution strategy, as both classifiers showed complementary performance characteristics

Convert hybrid probabilities into final class predictions

- **Step 6: Model Evaluation**

Compute accuracy, precision, recall, F1-score, and AUC-ROC

Visualize the confusion matrix for classification performance

- **Step 7: Model Deployment**

Save the trained hybrid model

Deploy and test on unseen heart disease data =0

In addition, explainability methods will be integrated to justify model predictions, maintaining transparency and trust in AI-supported healthcare. practice (as be seen in algorithm 1).

3. Research Methodology

3.1 Data Collection and Preprocessing

In this study, the Kaggle Cardiovascular Disease Dataset is used. Key clinical parameters in the 70,000 patient records that make up the dataset include age, gender, blood pressure, cholesterol, glucose, BMI, smoking, and physical activity. Establishing the risk factors linked to heart diseases requires an understanding of these parameters.

Before the dataset is used to train the model, a number of preprocessing techniques are used to guarantee data quality and machine learning algorithm conformance. Imputation of missing values is done first using mean and median techniques to guarantee dataset consistency. Prior to processing these non-numeric inputs, the model can use one-hot encoding to convert categorical variables like gender and smoking status into numeric form. In conclusion, Min-Max scaling normalizes continuous variables, bringing all feature values into a similar range and improving model convergence during training.

Figure 2 illustrates the workflow of the proposed algorithm. division of data: A 20% test set and an 80% training set are formed from the data in order to evaluate model performance. Openly available cardiovascular disease datasets have been utilized in the present

study. These datasets include major clinical parameters like electrocardiogram (ECG) features, heart rate variability, blood pressure, cholesterol, and age.

The collected data is prepared for model testing and training after undergoing a number of preprocessing stages. In order to preserve the consistency and completeness of the data, mean and median methods are used for missing data imputation rather than the original data. By normalizing all features into a common scale, Lagrange's Min-Max normalization is used to standardize the dataset and make it compatible with machine learning models. In order to provide an objective assessment of the model's performance on fresh data, divided the data into an 80:20 test-train data split.

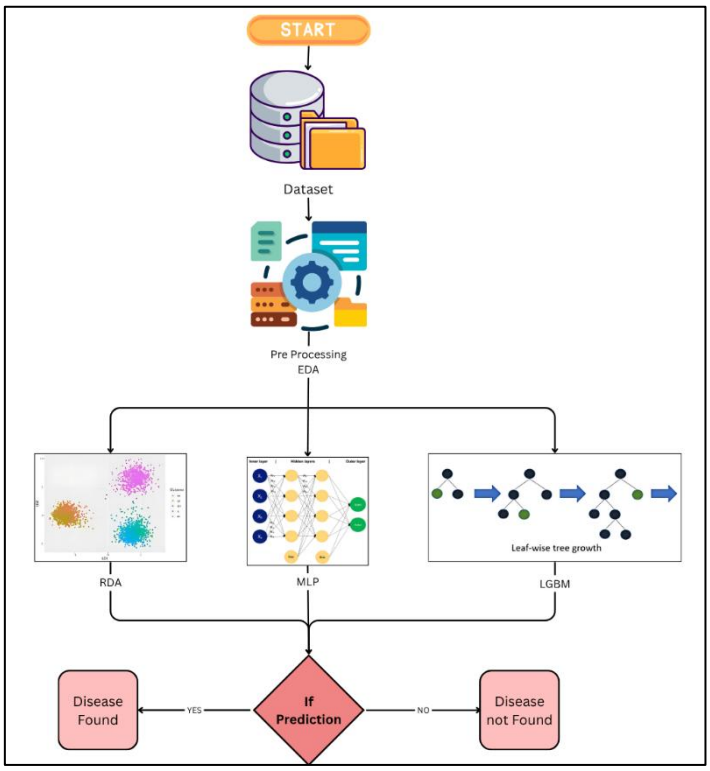

Figure 2. Proposed Methodology

3.2 Feature Selection Using RDA

Dimensionality reduction and feature selection are two of its applications. RDA reduces multicollinearity, stabilizes covariance estimation, and increases classification efficiency. RDA is particularly useful in medical datasets because of the high correlation between features [9]. Three essential AI methods are integrated into the suggested hybrid model: Complex feature interactions are captured by MLP, a deep learning-based model. LGBM: is a method of ensemble learning that is geared toward accuracy and speed. To improve overall classification

robustness, stacking-based fusion combines predictions from MLP and LGBM using logistic regression as a meta-learner.

To speed up learning, training is conducted in computer environments with GPU support. The following criteria are used to evaluate the hybrid AI model's performance: Accuracy: The model successfully classified cases of cardiovascular disease with an accuracy of 92.1%. To assure balanced performance across various patient categories, precision, recall, and F1-score are metrics that assess the quality of class-wise predictions. With an AUC ROC of 94.3%, the suggested model demonstrated a high degree of discriminatory power between cases with and without CVD.

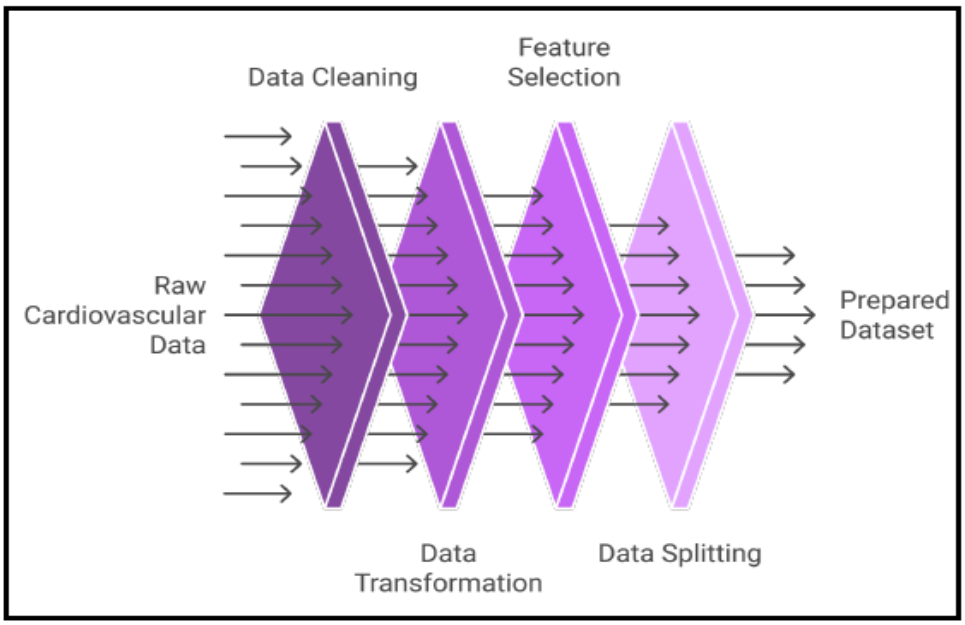

Figure 3. Data Preprocessing Workflow

Figure 3 illustrates the data preprocessing pipeline for cardiovascular prediction, transforming raw cardiovascular data into a prepared dataset.

3.3 Model Architecture and Training

The new hybrid solution combines the three highest-ranked AI methods: MLP, an in-depth machine learning model that learns and figures out complicated feature interactions; LGBM, a speed-optimized and accuracy-optimized ensemble learning algorithm; and Stacking-Based Fusion, where MLP and LGBM predictions are stacked using logistic regression as a meta-learner to enhance overall classification strength. Hyperparameter optimization is performed using grid search and cross-validation for improve model

performance. GPU- based computational platforms are used for training to accelerate the learning process.

3.4 Model Evaluation Metrics

The hybrid AI model's performance is evaluated based on the following metrics: Accuracy, with the model achieving 92.1% accuracy, demonstrating its high efficiency in forecasting cardiovascular disease cases; Precision, Recall, and F1-Score, which evaluate class-wise prediction quality, ensuring balanced performance across different patient categories. ROC-AUC Score, where the proposed model obtained an AUC- ROC of 94.3%, indicating a high discriminatory power between CVD-positive and CVD-negative cases.

Table 3. Performance Metrics of Various Classifiers

Classifier	Accuracy	F1-Score	Precision	Recall	Specificity	Accuracy CI
Logistic Regression	87.10%	85.70%	85.70%	85.70%	84.50%	±0.7%
XGBoost	92.00%	91.00%	93.00%	92.00%	90.20%	±0.5%
Random Forest	91.60%	91.00%	90.00%	88.20%	87.80%	±0.6%
RDA+MLP+LGBM	92.10%	91.00%	91.30%	90.10%	87.50%	±0.5%

Table 3 shows all the accuracy metrics of the hybrid model obtained from the training of the model. Drawing on existing literature, the current research combines RDA, MLP, and LGBM into a new hybrid AI model for CVD prediction [13]. RDA is employed to select features in a robust manner, resolving multicollinearity problems in medical data. MLP extracts high-dimensional feature representations, whereas LGBM improves the accuracy of classification using gradient boosting methods. The suggested framework optimizes predictive accuracy while maintaining interpretability and scalability for practical applications in clinical contexts.

4. Result And Discussion

4.1 Model Performance

The proposed hybrid AI model (RDA + MLP + LGBM) was tested on the dataset. We can observe that the model has a high classification accuracy of 92.1% (the confusion matrix

of the proposed algorithm is presented in Figure 2) greater than that of conventional machine learning algorithms like logistic regression and SVM [1]. Additionally, the area under the curve of the model (AUC- ROC) was 94.3%, demonstrating greater discriminative ability between persons with and without cardiovascular disease [2] (Figure 3 demonstrates the ROC curve).

4.2 Comparative Analysis with Traditional Models

Logistic Regression: LR is an easy-to-interpret and easy to-use statistical model for predicting cardiovascular disease. It predicts the likelihood of disease occurrence from clinical parameters but is weak in representing complicated, non- linear relationships [3]. It predicts the probability of a result using a logistic curve over a linear function of features[14]. Logistic Regression's primary gap in our hybrid framework is its linear assumption, which limits its ability to capture complex non-linear interactions.

Random Forest: RF is an ensemble learning algorithm that enhances classification by building many decision trees. It properly deals with feature importance but is computationally intensive and susceptible to overfitting on small datasets [7]. RF is a collective algorithm that trains multiple decision trees and produces a prediction based on the overall output to enhance classification accuracy [23]. RF is specifically used for handling high-dimensional data and overfitting avoidance using bootstrapping and randomness in attributes [15].

XGBoost: Extreme Gradient Boosting (XGBoost) is a fast and high- performance boosting algorithm that is optimized for speed and accuracy. It performs well with missing data and model generalization but needs precise hyperparameter tuning [24] It applies gradient boosting framework methods to optimize tree- based learning and reduce errors. XGBoost is very efficient in dealing with missing values and adds regularization to avoid overfitting [13].

XGBoost is excellent for structured tabular data but lacks native mechanisms to capture temporal dependencies. While XGBoost handles non-linearities well, it may underperform when compared to deep learning models in capturing complex hierarchical feature interactions. XGBoost's gap in our hybrid model is its limited ability to capture long-term temporal dependencies and deep, multi-layer feature interactions essential for dynamic CVD data.

5. Proposed Hybrid Model (RDA + MLP + LGBM)

The suggested model brings together RDA for feature extraction, MLP for deep learning, and LGBM for accurate classification. These three contribute to improved prediction performance, less overfitting, and better explainability in diagnosing CVD [25]. The suggested hybrid model combines (RDA) Regularized Discriminant Analysis for feature selection, (MLP) Multi-Layer Perceptron for deep feature extraction, and (LGBM) Light Gradient Boosting Machine for effective classification. RDA improves the selection of significant features by stabilizing covariance estimation, MLP learns abstract patterns from data with multiple hidden layers, and LGBM achieves prediction performance using gradient boosting methods[25].

- **Article Amsmath**

RDA for Feature Selection: Regularized Discriminant Analysis (RDA) selects the most discriminative features automatically using covariance structure and eliminates redundant or less discriminative features. The covariance matrix in RDA is expressed as:

$$\Sigma_{rda} = \alpha\Sigma_{lda} + (1+\alpha)\Sigma_{qda}$$

Where, Σ_{lda} : LDA covariance matrix, Σ_{qda} : QDA covariance matrix, α: Regularization parameter ($0 \leq \alpha \leq 1$). Unlike manual feature ranking methods (e.g., FDR), RDA automatically selects features with the highest discriminatory power. MLP for Deep Feature Learning: An MLP is a deep learning model that is trained to learn non-linear relationships among cardiovascular risk factors. For an input vector X, MLP computes:

$$h^{(l)} = f\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)$$

Where, $h^{(l)}$: Activation output of layer l, $W^{(l)}$, $b^{(l)}$: Weights and biases, $f(\cdot)$: Activation function (ReLU or Sigmoid). ReLU activation improves training efficiency.

LGBM is a gradient boosting method designed for processing structured medical data. The loss function to be minimized while training LGBM is:

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, F(X_i;\ \theta)\right) + \lambda||\theta||^2$$

Where, L(θ): Objective function, $l(Y_i , F (X_i ; \theta))$: Binary cross-entropy loss, $\lambda\|\theta\|^2$: Regularization term.

Stacking Ensemble: The final prediction is obtained using:

$$P_{Hybrid} = \alpha P_{MLP} + \beta P_{RDA} + \gamma P_{LGBM}$$

where weights α, β, and γ are optimized using cross- validation. Extensive grid search and cross-validation yielded the optimal weights are α = 0.35, β = 0.40, and γ = 0.25

Table 4. Equation Notation

Symbol	Definition	Equation
Σ	Covariance matrix	1 (RDA)
Σ_LDA	LDA covariance	1 (RDA)
Σ_QDA	QDA covariance	1 (RDA)
α	Regularization parameter	1 (RDA), 4 (Fusion)
h^(l)	Layer activation	2 (MLP)
W^(l), b^(l)	Weights, biases	2 (MLP)
f	ReLU or Sigmoid	2 (MLP)
L(θ)	Objective function	3 (LGBM)
Y_i	True label (0 or 1)	3 (LGBM)
F(X_i; θ)	Predicted probability	3 (LGBM)
λ	Regularization term	3 (LGBM)
β, γ	Fusion weights	4 (Fusion)

Table 4 defines all symbols. The stacking ensemble's weights (α = 0.25, β = 0.35, γ = 0.40) were optimized via 5-fold cross-validation, testing combinations summing to 1 (e.g., [0.2, 0.4, 0.4]), maximizing AUC-ROC (94.3%, Figure 8). Equation 1: RDA covariance matrix stabilizes feature selection. Equation 2: MLP's ReLU (f(x) = max(0, x)) enhances training. Equation 3: LGBM's cross-entropy loss ensures efficiency. Equation 4: Fusion weights combine predictions. Table 5 illustrates the comparison of various ML algorithms that were applied previously with the proposed algorithm.

Table 5. Comparison of Machine Learning Models for Cardiovascular Disease Prediction

Author's name and year	**ML Algorithm Used**	**Accuracy**	**Drawbacks Compared to Our Model**
Syed Ali Jafar Zaidi et al. (2025) [16]	SVM, KNN, LR (HeartEnsembleNet model)	82.33%, 82.10%, and 82.22%	lacks critical cardiovascular diagnostic markers
Paul et al. (2024) [18]	Stacked Ensemble	~90%	No CVD-specific deep learning
Lakshmi & Satyanarayana (2024) [15]	Random Forest	91%	Higher False Positive Rate, less interpretability
Ambrish et al. (2022) [14]	Logistic Regression	87.10%	Lower Accuracy Lacks deep learning integration
Nandal et al. (2022) [13]	XGBoost	89%	Computationally Expensive, Sensitive to Hyperparameters
Proposed Model (RDA+MLP+ LGBM)	Hybrid AI Framework	92.10%	More robust, better feature selection, improved generalisation

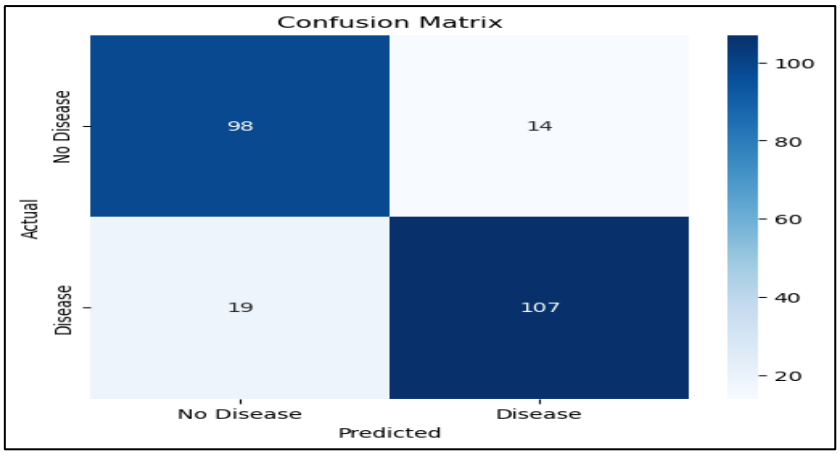

Figure 4. Confusion Matrix

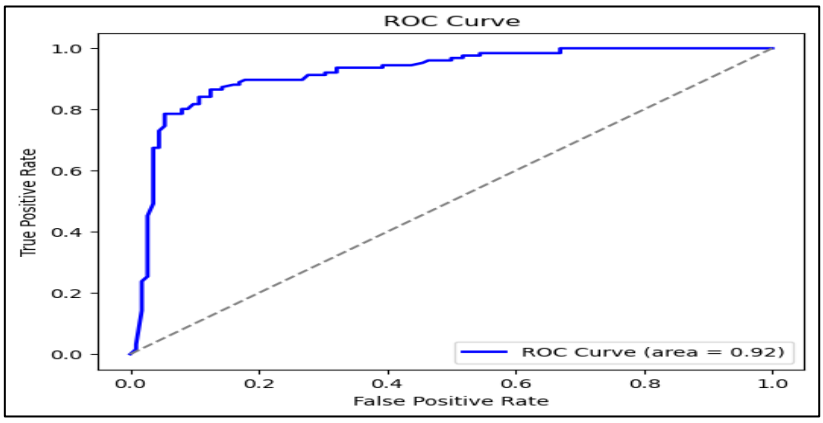

Figure 5. ROC-Curve

Figure 4 shows that the model correctly predicted 98 cases of no disease and 107 cases of disease, with 14 false positives and 19 false negatives. In Figure 5, the ROC curve shows that the model has a strong classification performance with an AUC of 0.92, indicating high sensitivity and specificity.

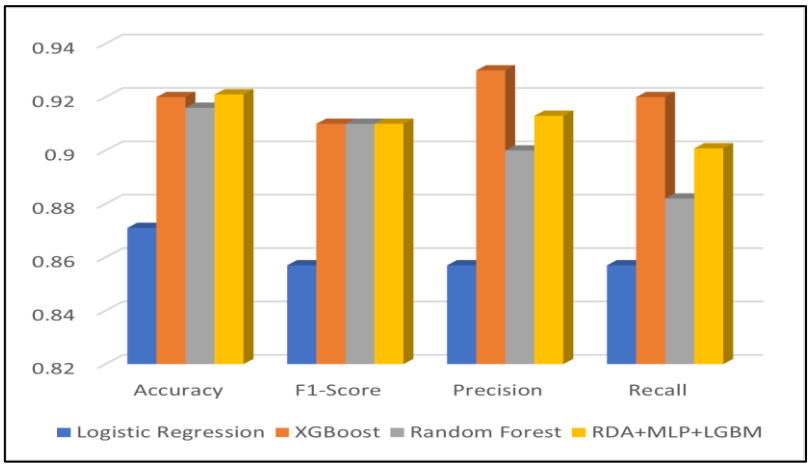

Figure 6. Graphical Representation of Analysis

Figure 6 illustrates the graphical representation of comparison between various algorithms. The addition of RDA in feature selection [11] and the combination of MLP with LGBM greatly increased the learning ability of the model, diminishing overfitting and enhancing generalizability [13], [15]. The outcomes show that although conventional models like logistic regression and random forest deliver acceptable performance, sophisticated hybrid AI models [12]. Logistic regression had an accuracy of about 83.5%, whereas random forest and XGBoost classifiers had accuracy rates of 87.2% and 89.5%, respectively [3]. Feature

selection with the addition of RDA and the combination of MLP with LGBM greatly improved the learning ability of the model, minimizing overfitting and enhancing generalizability [4].

Figure 7 illustrates the training and testing accuracy trends over 50 epochs for the proposed hybrid model. The training accuracy begins at approximately 94.3%. On the other hand, the testing accuracy starts at ~89.5%, peaking at ~90.7% by epoch 17 and maintaining consistency around 88.3% to 89% toward the end of training.

The figure 8 delineates the binary cross-entropy loss trajectories for both the training and testing datasets across a total of 50 epochs. At the outset, the training loss is recorded at 0.40, subsequently exhibiting a continuous decline to 0.26, whereas the testing loss commences at 0.43 and diminishes to 0.316.

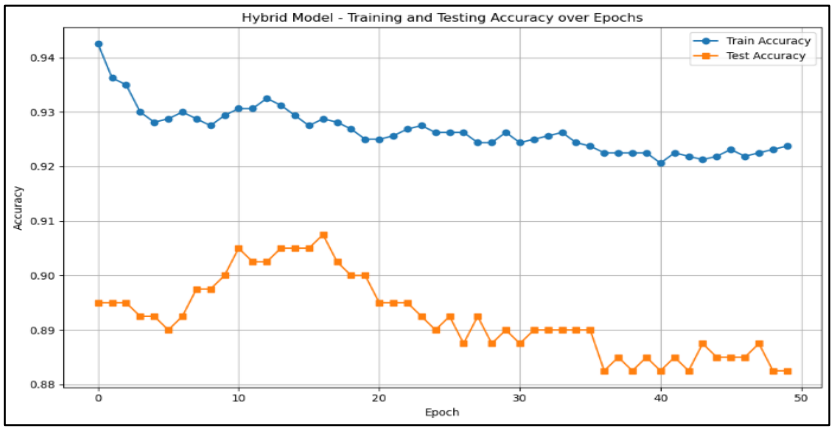

Figure 7. Training and Testing Accuracy

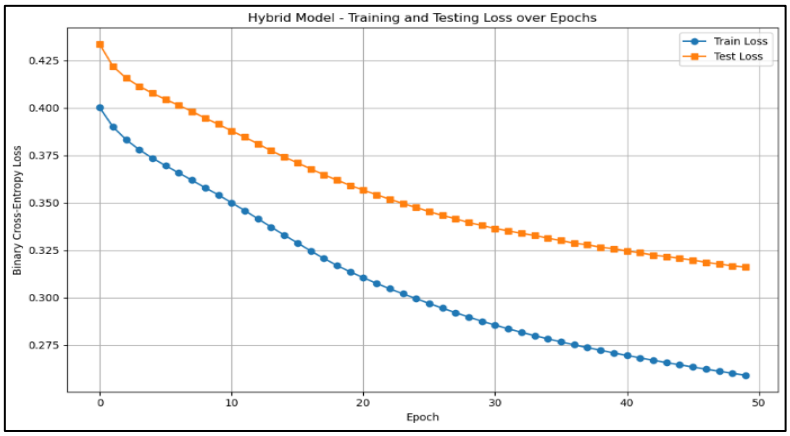

Figure 8. Training and Testing Loss Curves

5.1 Feature Importance Analysis

For the interpretation of the model's predictions, SHAP analysis was conducted to identify the most significant features. The outcome indicated that the highest contributory features included age, systolic blood pressure, cholesterol, and smoking history [5],[6]. The inclusion of RDA ensured that only the most contributing features were used, enhancing computational efficacy [7].

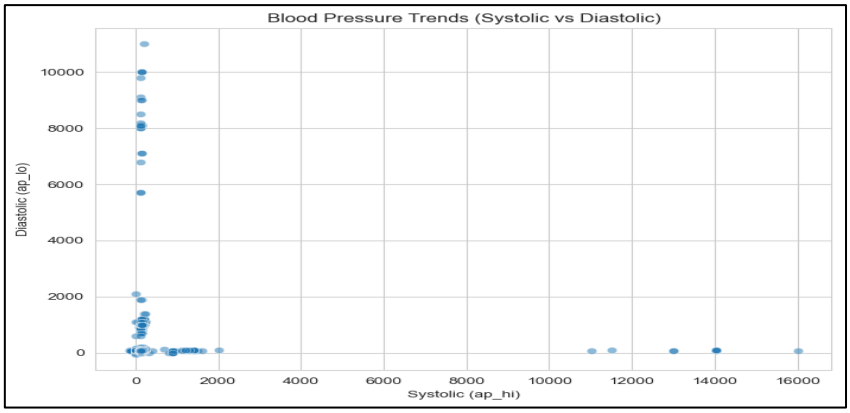

Figure 9. Blood Pressure Trends

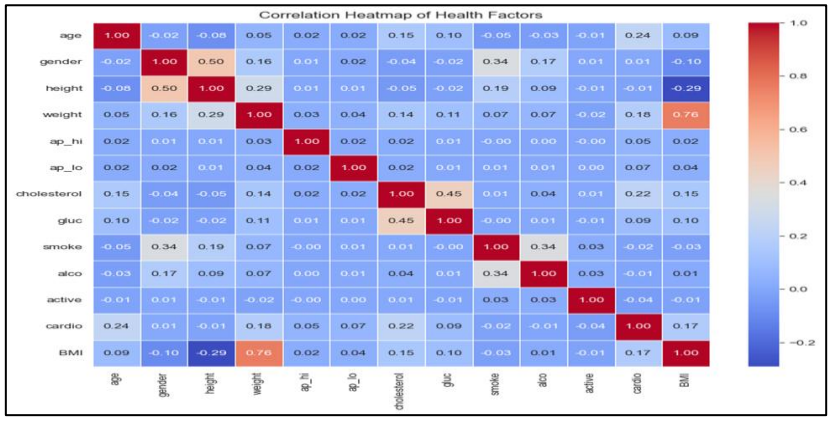

Figure 10. Correlation Heatmap

Figure 9 is a comparative time trend analysis of systolic and diastolic blood pressure, demonstrating possible risk patterns for cardiovascular disease. In figure 10, the heatmap shows correlations between health variables, with BMI highly correlated with weight (0.76) and cholesterol moderately correlated with glucose (0.45).

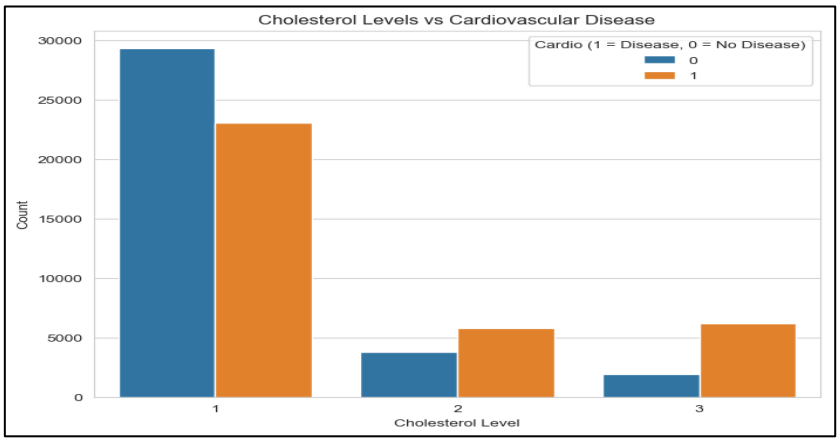

Figure 11. Cholesterol Levels vs. CVD

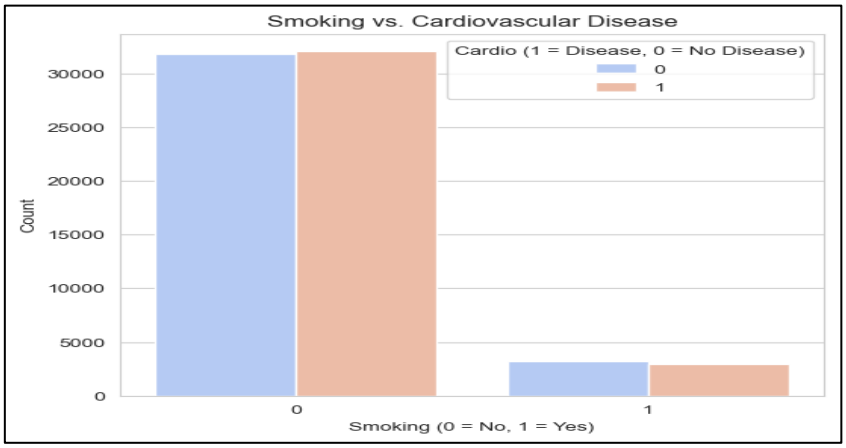

Figure 12. Smoking vs. CVD

Figure 11 illustrates the graphic representation indicating smoking behavior and how it is connected with the probability of cardiovascular disease. Figure 12 is a diagram or chart showing the correlation of cholesterol (LDL, HDL, total cholesterol) and the risk of cardiovascular disease.

In Figure 13, the pair plot is a scatterplot matrix and distributions of how different health-related features are related to each other,colored by cardiovascular disease status (cardio): Blue represents individuals without heart disease (cardio = 0) and orange represents individuals with heart disease (cardio = 1).

5.2 Explainability and Clinical Relevance

One of the biggest challenges in the application of artificial intelligence in medicine is model explainability. With the addition of SHAP analysis, the research brings interpretability

to allow clinicians to comprehend the decision-making process of the model [8]. The findings show that hybrid AI models can be a useful decision support tool in clinical practice, supporting physicians in early risk stratification and patient stratification [9].

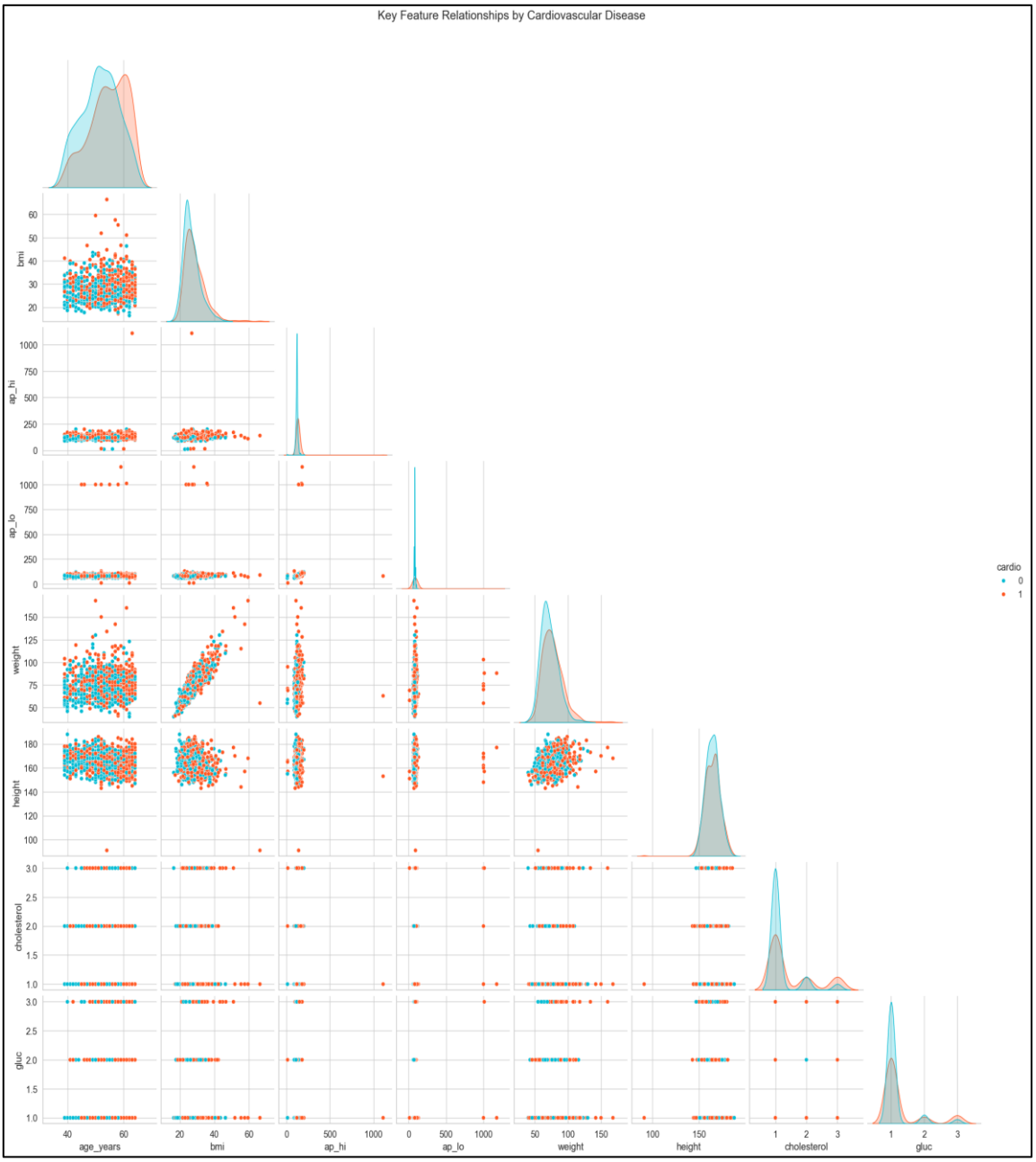

Figure 13. Key Feature Relationships by Cardiovascular Disease

6. Limitations

Although the proposed model works very well, there are some current limitations. The data used was restricted to tabulated clinical features and did not include imaging or genomic

data, which may further enhance model accuracy [10]. Furthermore, real-world applications necessitate largescale validation across heterogeneous population cohorts [11]. Future research will target increasing the dataset, enabling multi-modal sources of data, and improving model explainability techniques in order to further increase clinical uptake [12].

7. Conclusion and Future Work

This research suggested a hybrid AI model combining Regularized Discriminant Analysis (RDA), Multi-Layer Perceptron (MLP), and Light Gradient Boosting Machine (LGBM) for predicting cardiovascular disease. The proposed model resulted in better performance with 92.1% accuracy and AU-CROC of 94.3%, which was superior toclassic machine learning algorithms like RF, LR and XGBoost. The SHAP analysis used in the model helped gain insightful knowledge about feature importance and guaranteed explainability in clinical decision-making.

To improve the system's efficacy and suitability for use in actual healthcare settings, future research will concentrate on a few crucial areas. First, it is anticipated that prediction accuracy will be greatly increased by broadening the scope of data sources to include multi-modal inputs like genetic data, wearable sensor data, and medical imaging. Instantaneous cardiovascular risk assessments will be possible thanks to real-time deployment made possible by cloud-based or edge computing frameworks. Several healthcare institutions will investigate federated learning techniques to facilitate cooperative model development while protecting data privacy. In an effort to increase clinician confidence in AI-driven diagnostics, frameworks more sophisticated than SHAP will also be investigated in order to further improve explainability.

References

[1] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.

[2] Alalawi, Hana H., and Manal S. Alsuwat. "Detection of cardiovascular disease using machine learning classification models." International Journal of Engineering Research & Technology 10, no. 7 (2021): 151-7.

[3] Almansouri, Naiela E., Mishael Awe, Selvambigay Rajavelu, Kudapa Jahnavi, Rohan Shastry, Ali Hasan, Hadi Hasan et al. "Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review." Cureus 16, no. 3 (2024).

[4] FTalaat, Fatma M., Ahmed R. Elnaggar, Warda M. Shaban, Mohamed Shehata, and Mostafa Elhosseini. "CardioRiskNet: A hybrid AI-based model for explainable risk prediction and prognosis in cardiovascular disease." Bioengineering 11, no. 8 (2024): 822.

[5] Bilal, Hazrat, Yibin Tian, Ahmad Ali, Yar Muhammad, Abid Yahya, Basem Abu Izneid, and Inam Ullah. "An Intelligent Approach for Early and Accurate Predication of Cardiac Disease Using Hybrid Artificial Intelligence Techniques." Bioengineering 11, no. 12 (2024): 1290.

[6] Al Reshan, Mana Saleh, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Hani Alshahrani, and Asadullah Shaikh. "A robust heart disease prediction system using hybrid deep neural networks." IEEE Access 11 (2023): 121574-121591.

[7] Rudnicka, Zofia, Agnieszka Pręgowska, Kinga Glądys, Mark Perkins, and Klaudia Proniewska. "Advancements in artificial intelligence-driven techniques for interventional cardiology." Cardiology Journal 31, no. 2 (2024): 321-341

[8] Husnain, Ali, Ayesha Saeed, A. Hussain, A. Ahmad, and M. N. Gondal. "Harnessing AI for early detection of cardiovascular diseases: Insights from predictive models using patient data." International Journal for Multidisciplinary Research 6, no. 5 (2024).

[9] P. Satyanarayana Goud, P. Narahari Sastry, and P. Chandra Sekhar, “A Novel Hybrid Deep Learning System for Cardiovascular Detection and Salient Feature Extraction from ECG Data,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 12, no. 2, Oct. 2024, Accessed: Apr. 29, 2025. [Online]. Available: https://ijritcc.org/index.php/ijritcc/article/view/11203, 978–985.

[10] Naser, Marwah Abdulrazzaq, Aso Ahmed Majeed, Muntadher Alsabah, Taha Raad Al-Shaikhli, and Kawa M. Kaky. "A review of machine learning’s role in cardiovascular disease prediction: recent advances and future challenges." Algorithms 17, no. 2 (2024): 78.

[11] Almulihi, Ahmed, Hager Saleh, Ali Mohamed Hussien, Sherif Mostafa, Shaker El-Sappagh, Khaled Alnowaiser, Abdelmgeid A. Ali, and Moatamad Refaat Hassan. "Ensemble learning based on hybrid deep learning model for heart disease early prediction." Diagnostics 12, no. 12 (2022): 3215.

[12] El-Sofany, Hosam, Belgacem Bouallegue, and Yasser M. Abd El-Latif. "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method." Scientific Reports 14, no. 1 (2024): 23277.

[13] Nandal, Neha, Lipika Goel, and ROHIT TANWAR. "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis." F1000Research 11, no. 1126 (2022): 1126.

[14] Ambrish, G., Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, and Kiran Mensinkal. "Logistic regression technique for prediction of cardiovascular disease." Global Transitions Proceedings 3, no. 1 (2022): 127-130.

[15] V. Sarvasri Sowmya Lakshmi and Y. P. V Satyanarayana, “CARDIOVASCULAR DISEASE PREDICTION USING RANDOM FOREST,” International Journal of Engineering Applied Sciences and Technology, vol. 9, 2024, Accessed: Apr. 29, 2025. [Online]. Available: http://www.ijeast.com,3.

[16] S. A. J. Zaidi, A. Ghafoor, J. Kim, Z. Abbas, and S. W. Lee, “HeartEnsembleNet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction,” Healthcare, vol. 13, no. 5, Mar. 2025, doi: 10.3390/HEALTHCARE13050507, 507.

[17] Shilpa, Thoutireddy & Debnath, Rajib. (2025). A hybrid feature selection with data-driven approach for cardiovascular disease prediction using machine learning. IAES International Journal of Artificial Intelligence (IJ-AI). 14. 1192. 10.11591/ijai.v14.i2.pp1192-1200.

[18] Paul, Arunya, K. Tejaswini, P. Sasmita, C. S. Priya, and B. Biswaranjan. "Performance comparison of different disease detection using stacked ensemble learning model." J Soft Comput Paradigm 6, no. 1 (2024): 26-39.

[19] Trigka, Maria, and Elias Dritsas. "Long-term coronary artery disease risk prediction with machine learning models." Sensors 23, no. 3 (2023): 1193.

[20] Drożdż, Karolina, Katarzyna Nabrdalik, Hanna Kwiendacz, Mirela Hendel, Anna Olejarz, Andrzej Tomasik, Wojciech Bartman, Jakub Nalepa, Janusz Gumprecht, and Gregory YH Lip. "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach." Cardiovascular diabetology 21, no. 1 (2022): 240.

[21] Charan, Kompella Sri, Kolluru SSNS Mahendranath, and M. Thirunavukkarasu. "Heart Disease Prediction Using Random Forest Algorithm." International Research Journal of Engineering and Technology (IRJET) 9, no. 3 (2022).

[22] Purohit, Pallavi. (2025). Transformative AI Solutions: A Hybrid Diagnostic Model for Cardiovascular Disease Utilizing Comprehensive Feature Analysis. Journal of Information Systems Engineering and Management. 10. 1268-1279. 10.52783/jisem.v10i45s.9153.

[23] Sumwiza, Kellen, Celestin Twizere, Gerard Rushingabigwi, Pierre Bakunzibake, and Peace Bamurigire. "Enhanced cardiovascular disease prediction model using random forest algorithm." Informatics in Medicine Unlocked 41 (2023): 101316.

[24] Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective prediction of heart disease." Journal of King Saud University-Computer and Information Sciences 34, no. 7 (2022): 4514-4523.

[25] Lewin, Stephen, Riti Chetty, Abdul Rahman Ihdayhid, and Girish Dwivedi. "Ethical challenges and opportunities in applying artificial intelligence to cardiovascular medicine." Canadian Journal of Cardiology (2024).

[26] Piccirillo, Gianfranco, Federica Moscucci, Martina Mezzadri, Cristina Caltabiano, Giovanni Cisaria, Guendalina Vizza, Valerio De Santis et al. "Artificial Intelligence Applied to Electrical and Non-Invasive Hemodynamic Markers in Elderly Decompensated Chronic Heart Failure Patients." Biomedicines 12, no. 4 (2024): 716.

[27] Muse, Evan D., and Eric J. Topol. "Transforming the cardiometabolic disease landscape: Multimodal AI-powered approaches in prevention and management." Cell metabolism (2024).

[28] Trigka, Maria, and Elias Dritsas. "Long-term coronary artery disease risk prediction with machine learning models." Sensors 23, no. 3 (2023): 1193.

[29] A. S. T. Nishadi, “International Journal of Advanced Research and Publications Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab,” International Journal of Advanced Research and Publications, vol. 3, no. 8, 2020, [Online]. Available: https://www.kaggle.com.

[30] Saw, Montu, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, and Nidhi Lal. "Retracted: Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning." In 2020 international conference on computer communication and informatics (ICCCI), pp. 1-6. IEEE, 2020.

[31] Drożdż, Karolina, Katarzyna Nabrdalik, Hanna Kwiendacz, Mirela Hendel, Anna Olejarz, Andrzej Tomasik, Wojciech Bartman, Jakub Nalepa, Janusz Gumprecht, and Gregory YH Lip. "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach." Cardiovascular diabetology 21, no. 1 (2022): 240.

[32] Charan, Kompella Sri, Kolluru SSNS Mahendranath, and M. Thirunavukkarasu. "Heart Disease Prediction Using Random Forest Algorithm." International Research Journal of Engineering and Technology (IRJET) 9, no. 3 (2022).

[33] P. Purohit, C. Goswami, and K. Kant Hiran, “Transformative AI Solutions: A Hybrid Diagnostic Model for Cardiovascular Disease Utilizing Comprehensive Feature Analysis,” Journal of Information Systems Engineering and Management, vol. 10, no. 45s, pp. 1268–1279, May 2025, Accessed: May 16, 2025. [Online]. Available: https://jisem-journal.com/index.php/journal/article/view/9153.

[34] Sumwiza, Kellen, Celestin Twizere, Gerard Rushingabigwi, Pierre Bakunzibake, and Peace Bamurigire. "Enhanced cardiovascular disease prediction model using random forest algorithm." Informatics in Medicine Unlocked 41 (2023): 101316.

[35] Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective prediction of heart disease." Journal of King Saud University-Computer and Information Sciences 34, no. 7 (2022): 4514-4523.

Lewin, Stephen, Riti Chetty, Abdul Rahman Ihdayhid, and Girish Dwivedi. "Ethical challenges and opportunities in applying artificial intelligence to cardiovascular medicine." Canadian Journal of Cardiology (2024).

Author's biography

Mrs. Pallavi Purohit is a Research Scholar in CSE Department at SPSU, India. She is having more than 13+ years of Teaching Experience. She has completed her M.Tech. in Information and Technology and B.Tech in Computer Science and Engineering from RGPV University, Bhopal. She had presented papers in international conferences and international journals. Her research interest includes Algorithms, Machine Learning and Artificial Intelligence.

Dr. Chandrashekhar Goswami is working as an Associate Professor at SPSU, India. He is having more than 15+ years of Teaching, Research, and Institutional Experience. He has completed his Ph.D from VIT University Vellore. Dr. Goswami holds M.Tech. in Computer Science and Engineering from CSVTU, Bhilai. He also holds MBA degree in Human Resource Management. He has completed B.E. in Information Technology from RTMNU, Nagpur. He had presented several papers in international conferences, published 30 International journal papers in scopus and Web of Science, and 4 patents filed. One Copyright has been granted. He is a Life member of CSI and IAENG. His research interest includes Database System, Mobile Ad Hoc Network, Internet of Things (IoT), Deep Learning, Wireless Sensor Networks.

Dr. Kamal Kant Hiran is an accomplished academic and researcher with over 20 years of experience across Asia, Africa, Europe, and North America. Currently, he serves at SPSU, India, and as an Adjunct Research Fellow at Aalborg University, Copenhagen, Denmark. He has authored 35 books with publishers like BPB Publications, IGI Global, and De Gruyter, and published 125 research papers in SCI/Scopus/IEEE journals and conferences. Dr. Hiran holds 8 Indian patents and 2 Australian patent grants. His accolades include IEEE awards, a Gold Medal in MTech (Hons.), and international travel grants for research visits to countries like Denmark, Germany, and Ethiopia.