International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 3, 2025, pp. 512-526.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

SRD-GDT²: A SMOTE-Regularized Dual-Guided Deep Tabular Framework for Accurate Cardiovascular Disease Prediction

A. Devi [a], T. NirmalRaj [b]

[a] *Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Kancheepuram, Tamil Nadu, India. E-mail: adevics21@gmail.com, Orcid: https://orcid.org/0009-0000-5285-1278*

[b] *Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Kancheepuram, Tamil Nadu, India. E-mail: tnirmalraj@kanchiuniv.ac.in*

Abstract

Cardiovascular disease is one of the major causes of death in the world, and therefore, effective and dependable prediction models are necessary. The models can assist doctors in making informed choices and allow the identification of heart-related conditions in a timely manner. Machine learning systems have shown promise in this field; however, their effectiveness is frequently constrained by class imbalance, noisy samples, and unreliable probability estimations in real-world medical datasets. The suggested SRD-GDT² is a system for predicting cardiovascular disease utilizing large-scale tabular data. It is an SMOTE-Regularized Dual-Guided Deep Tabular Distillation framework. The suggested method combines thorough data preprocessing, SMOTE-ENN-based imbalance regularization, and a dual-guided learning strategy in which two tree-based models help a deep tabular learner. An agreement-based weighting system is used to stress dependable recommendations, and Youden's index is used to find the best choice threshold that balances sensitivity and specificity. After preprocessing and resampling a balanced dataset of both majority and minority class samples, experiments were done on a cardiovascular dataset. The suggested model did quite well on the independent test set, with an accuracy of 95.63%, an AUROC of 0.9905, an AUPRC of 0.9870, a sensitivity of 93.79%, and a specificity of 97.51%. The low Brier score of 0.0324 shows that the probability estimations are accurate. Decision curve analysis further illustrates that the suggested framework yields a superior overall treatment advantage when compared with the treat-all and treat-none approaches. Overall, the findings indicate that SRD-GDT² is an effective and clinically valuable method for predicting the risk of cardiovascular disease. It addresses issues related to data imbalance and improves learning stability.

Keywords: Cardiovascular Disease Prediction, Deep Tabular Learning, SMOTE-ENN, Dual-Guided Learning, Knowledge Distillation, Medical Machine Learning, Clinical Decision Support.

DOI: 10.71086/IAJSE/V12I3/IAJSE1284
Received: 26 July 2025/Revised: 30 August 2025/Accepted: 16 September 2025/Published: 30 September 2025

Introduction

Approximately 17.9 million people die annually from cardiovascular diseases, with heart attacks and strokes accounting for nearly 80% of these deaths. Obesity, hypertension, dyslipidaemia, and hyperglycemia are measurable risk factors that can be addressed in a timely manner to lower death rates. When people with heart disease are diagnosed early, they have more time to make lifestyle changes or get medical help. However, electrocardiograms and other traditional diagnostic tools used in hospitals and clinics to identify irregular heart rhythms do not reliably identify patients experiencing a heart attack (Karna et al., 2025). One of the most pressing problems in clinical data analysis is the prediction of CVD. By carefully selecting relevant features, a reliable model for predicting cardiac events can be constructed. Data normalization, missing value imputation, data transformation, and data cleaning are the initial steps in the data pre- processing process (Wankhede et al., 2020). Cardiovascular disease (CVD) ranks high among the world's leading killers. Medical practitioners can better assess their patients' health with the help of CVD prediction. Data

mining is a method for extracting useful information from massive datasets, in this case medical records, in order to improve prediction and decision-making (Wankhede et al., 2020). Predicting diseases earlier helps improve healthcare quality and allows for improved decision-making in emergency situations, thanks to the tremendous advances in the medical field. A cardiovascular disease risk variable automated prediction model has been the primary focus of most of the previous research. However, current classifiers have a harder difficulty running due to the high-dimensional data management involved in diagnosing cardiac disease, making accurate classification a critical task (Kannan, 2022).

Heart disease occurs when the blood cannot be pumped to the body by the heart at the necessary volume. Worldwide, heart disease (HD) accounts for the majority of deaths. Predicting cardiac illnesses at an early stage can save lives. By being able to foretell the onset of cardiovascular or heart illness, one can take preventative measures. Classifying the population as healthy or sick with heart disease has been greatly aided by machine learning (ML) (Goyal, 2022). A wide range of medical issues can be classified as "heart diseases." Problems can arise from different parts of the cardiovascular system, including the heart, blood vessels, valves, and the internal electrical pathways that control muscle contractions. This study introduces a new method for predicting cardiovascular disease. The proposed approach consists of three main stages: feature extraction, dimensionality reduction, and an optimal ensemble-based model for heart disease prediction (Sheeba et al., 2022). The consequences of heart disease are far-reaching and significant, impacting people all around the globe. One of the main killers on a global scale is cardiovascular disease, making early diagnosis all the more important. Emerging technologies such as machine learning (ML) are now widely applied in biomedical research, healthcare, and health prediction systems (Noor et al., 2023). Several surveys indicate that heart disease remains one of the major causes of death worldwide. The main causes of this are the high likelihood of incorrect diagnoses and the complexity of these disorders. This significant problem is faced by those who earn a living providing medical care. The creation of a technology that can identify cardiac disease is imperative, given the remarkable performance of ML in estimation and decision-making. Not only may early detection of heart disease help patients avoid heart attacks, but it also allows doctors to learn more about the main causes of these events and provide preventative measures (Prabhavathi & Mareeswari, 2023).

Heart disease has risen to prominence as a major cause of death among individuals around the world due to the increasing prevalence of pandemics and generalized stress. The best way to improve people's possibility of survival is to catch this illness in its early stages. There are a lot of ways that use ML algorithms to predict cardiac disease, but they all share the same problem of overfitting (Gupta & Singh, 2023). The incidence of HD is unfortunately on the rise, and it is already the leading cause of death globally. It is difficult to detect early signs of heart illness before a cardiac event occurs. The health care industry, including hospitals, clinics, and private practices, has access to massive amounts of data pertaining to heart disease. Nevertheless, the data is not being processed correctly in order to uncover the underlying trends. To make sense of this mountain of medical data, machine learning techniques are a lifesaver (Almazroi et al., 2023). One of the most serious health issues affecting individuals today is coronary HD and the health of the public at large. Reducing its detrimental impact requires early prediction and diagnosis. An effective method for detecting cardiac diseases using data processing and cutting-edge optimization techniques is the primary goal of this effort (Aswathi et al., 2024). Globally, heart disease continues to be the leading cause of mortality in most countries. The only way to reduce mortality rates caused by heart disease is to discover it early. The use of big data in medical diagnostics has simplified the development of complex models based on ML and DL for the early automated detection of cardiac illness (Majhi & Kashyap, 2024).

Early detection and timely treatment of heart disease can help reduce mortality rates associated with this condition, which is the leading cause of death worldwide. Clinicians would benefit from the ability to predict which patients are at elevated risk of getting HD so they can provide them with specific preventative and therapeutic interventions. While there are a number of approaches to cardiac disease, ensuring

accurate assessments is still a major obstacle (Pati et al., 2024). There is an immediate need for reliable prediction tools and effective therapies for cardiovascular disease because it is the main killer on a global scale. Predicting heart illness using a hybrid architecture that combines autoencoder and DenseNet (Alghamdi et al., 2024). A prominent cause of death for all humans in this century is heart-related disease. Heart disease accounts for almost 17.5 million fatalities annually on a global scale. Over three quarters of those with middle-class incomes experience cardiovascular disease and its consequences. Predicting cardiac infections and their consequences is so necessary. "Data mining" means to glean useful information from massive databases (Natarajan et al., 2024).

Literature Survey

Heart disease is more difficult to diagnose when additional risk factors, such as diabetes, hypertension, irregular heart rate, and high cholesterol, are present. Predicting cardiac sickness accurately before to a heart attack is crucial for the effective treatment of cardiac patients. An optimized ensemble classifier along with an integrated filter–evolutionary search-based feature selection (iFES-FS) method is applied for the diagnosis of heart disease. The suggested feature selection method uses a hybrid of evolutionary gravity-search-based feature selection (EGS-FS) and adaptive threshold information gain-based feature selection (aTIG-FS) to extract the most useful information from the original collection of features. By optimizing the classifier's hyperparameters using a firefly-driven FF-MOMVO (Multi-Objective Multi-Verse Optimizer) algorithm, this technique improves ML performance (Venkata MahaLakshmi & Rout, 2024). Since machine learning algorithms aren't very good at understanding complex relationships and don't scale well, they often require greater precision when applied to larger datasets. The ensemble framework SENNERF uses DenseNet topologies in conjunction with a revised random forest model to enhance cardiac disease prediction. A model that works well with big datasets is used in many different situations. Using the DenseNet128, DenseNet256, and CompactNet64 architectures (Thumu et al., 2025) makes feature extraction and random forest prediction better.

Heart prediction system that makes a feature fusion model by combining a bidirectional gated recurrent unit (BiGRU) with a bidirectional temporal convolutional network (BiTCN). This architecture captures the spatiotemporal dependencies of the heart data. The model is much more generalizable due to the integration of the Bayesian optimization and the Arctic Lemming Algorithm (ALA). This hybrid approach addresses the issue of maximizing the model hyperparameters (Jiang et al., 2025). Two parts of the deep learning-based heart disease prediction system, Harris Hawking Optimization (HHO) and K-mode Clustering, are to find important characteristics and improve data preprocessing. For training and validation, a cardiovascular disease dataset consisting of 70,000 patient records and multiple clinical attributes was utilized (Isabella et al., 2025). Heart disease prediction was carried out using an ensemble model based on Deep Learning (DL), SMOTE, and FST (correlation features selection technique). SMOTE was used to balance the Cleveland dataset and get it ready for analysis. SMOTE also cuts down on noise. The Pearsons Correlation Coefficient (PCC) was used to find the most relevant and helpful characteristics since the level of correlation of 0.2 between the input qualities and the target classes was strong (Sofi et al., 2025). HD prediction a hybrid XGBoost Classifier that enhances hyper parameter optimization eliminates outliers. This strategy used the "Optuna" framework to find the best hyperparameters and the z-score and interquartile range (IQR) methods to deal with outliers. The Cleveland HD dataset was also used to different train-test ratios (70:30, 80:20, and 90:10) affected the models' performance, both with and without outliers (Dhanka & Maini, 2025).

The main reason for using HEART and Stacked Meta Neural Networks (SMNN) to find HD risk factors is to raise awareness of these problems. Researchers employed various statistical methodologies, including Shapley values, McFadden's pseudo-R-squared, mutual information, Friedman's and McNemar's tests, and odds ratio analysis, to identify and rank pertinent risk factors.

These significant risk factors (Bandyopadhyay et al., 2025) are what SMNN uses with 5-fold cross-validation to predict HD. A Grow Network (GrowNet) is a type of architecture that can make things more stable by changing its settings in real time based on the data. Adaptive Refinement sampling and TriDyn Dependence feature selection to improve GrowNet's stability. These techniques have been effective in managing the class imbalance problem in training as well as in enhancing the pertinent characteristics that are chosen from unbalanced datasets (Akter et al., 2025). Advanced ensemble learning techniques, including Stacking, Bagging, Voting, and Boosting, can be applied to achieve accurate and early prediction of heart disease. To validate the proposed approaches, performance

evaluation was conducted on three datasets: the Framingham Heart Disease dataset, the Cleveland Heart Disease dataset, and the Indicators of Heart Disease Dataset 2020 (Rabbi et al., 2025). The study also examines the most effective ensemble learning methods for predicting cardiac conditions. For feature extraction, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are utilized, which help improve prediction accuracy. By incorporating these advanced feature extraction techniques within ensemble learning models, the approach enhances the effectiveness of heart disease prediction (Yadav et al., 2025).

A classification model for predicting heart disease uses a Genetic Algorithm (GA) to choose the best features and an ensemble deep learning technique that is improved by the Tunicate Swarm Algorithm (TSA). GA selects the most relevant features to enhance model performance while reducing the number of dimensions. Then, a group of deep learning models is trained using the chosen traits. The TSA changes the weights of all the models in the ensemble to make the predictions more accurate. This hybrid approach addresses several common challenges in the field, including high dimensionality, redundant features, and low classification accuracy. It does this by enhancing the ensemble weighting of deep learning models and adding a good feature selection mechanism (Al-Mahdi et al., 2025). In order to detect early-stage cardiac disease, an automated model is suggested that makes use of clinical characteristics and risk variables. Optimal feature extraction using a bio-inspired meta-heuristic algorithm called BSSA, which incorporates the Wolf Optimizer and opposition-based learning. The three salps that help an improved BSSA maximize feature selection are omega, beta, and alpha. The BSSA-SC framework enhances accuracy and surpasses state-of-the-art approaches with low misclassification rates through the use of an Ensemble Stacking Classifier (SC) (Sowmiya et al., 2025).

To optimize the autoencoder's weights and select the best features, an improved equilibrium optimizer based on uniform random variables is employed. In order to create the best weighted features, the features that were chosen optimally are combined. Multi Serial Cascaded Adaptive Dilated Network receives these characteristics as input. A Deep Temporal Convolution Network, a One-Dimensional Convolution Neural Network, and a Long Short-Term Memory all come together to make this network (Sowmiya et al., 2025). Data preprocessing tool DGPN, feature selection tool STHIO, and prediction tool SADNet are all introduced using this method. In order to better depict features, DGPN normalizes data based on distribution characteristics. A method called Tuna Swarm Optimization is used for feature exploitation by STHIO, whereas the Sheep Flock Optimization method is used for feature exploration (Sowmiya et al., 2025).

For this reason, we devised a hybrid model that combines the strengths of SCNN and LFGOA, the Levy Flight-based Grasshopper Optimization Algorithm. While LFGOA enhances the model by reducing overfitting, improving classification accuracy, and optimizing hyperparameters, SCNN ensures effective feature extraction. The proposed method is evaluated using four publicly available cardiovascular disease datasets. Various demographic and clinical factors are reflected in these databases (Bukhari et al., 2025). A state-of-the-art convolutional neural network (CNN) is used for effective extraction of global and local features, while a bidirectional long short-term memory network (BiLSTM) is incorporated to enhance the design with upcoming and past contextual knowledge. The model is able to isolate crucial audio segments by combining the spatial attention of the CNN with the temporal attention of the RNN. It uses SMOTE to enhance data representation.

This is useful for avoiding issues that might diminish the effectiveness of deep learning systems, such as imbalanced and noisy datasets (Lilhore et al., 2025). SMOTEENN is a hybrid resampling method that combines SMOTE and ENN. Its goal is to stabilize the dataset while also decreasing noise. Both the initial unbalanced dataset and the balanced dataset generated by SMOTEENN were used for training (Shahid & Jayanth, 2025). By performing preparatory phases like data segmentation using DBSCAN, mean imputation, and min-max normalization, we can ensure data integrity and successful clustering. To improve the model's performance, AHP-VIKOR chooses and ranks important features optimum. The MSFMR-IncepDenseNet classifier improves DenseNet's classification accuracy by representing features more accurately by using mechanisms for attention on several scales, ASPP modules, and blocks from Inception (Pazhanikumar, 2025).

A smartphone app that uses deep convolutional neural networks (DCNNs) to automatically categorize heart rates and alert users to potential cardiac problems in real time. Two datasets were utilized to extract features of heart sounds using the Log-Mel spectrogram and Mel-frequency cepstral coefficients (MFCCs). An ensemble technique to merge the estimated probability from five separate DL models (Huang et al., 2025). Finch hunt optimization modified BiLSTM (FHO modified BiLSTM) is an IoT-enabled model for cardiac illness prediction that gets beyond these limitations. Smart IoT-based framework integration also helps with heart disease patient monitoring and providing efficient, fast, high-quality healthcare. In addition, biomedical data is stored on a cloud server supported by decentralized blockchain technology, which improves mobility, privacy, security, low latency, and bandwidth

efficiency (Chichani & Kasar, 2025). Table 1 summarizes studies related to cardiovascular disease prediction using machine learning and deep learning techniques. Take notice of the algorithms, datasets, performance indicators (such F1-score, recall, and accuracy), and primary contributions of each approach while comparing them. Many different methods are shown in the articles that were evaluated. Some of these methods include deep neural networks, feature selection using meta-heuristic optimization, hybrid ensemble models, and imbalance managing techniques based on resampling. Optimal feature selection and deep model construction have been the main foci of prior research in an effort to improve prediction accuracy on small or benchmark datasets, such as the UCI Cleveland heart disease dataset. Although some studies report high accuracy values, scalability and generalizability may be compromised due to the use of small datasets, heavy feature engineering, or complicated hybrid optimization procedures. The suggested SRD-GDT² system, on the other hand, stands out by combining a large-scale cardiovascular dataset with probability calibration, dual-guided learning, and imbalance regularization to produce balanced and clinically relevant predictive results.

Table 1: Comparison of Algorithm and its Metrics for the Literatures

Ref	Algorithm	Dataset	Accuracy	Recall	F1-Score	Contribution
(Wankhede et al., 2020)	DFCSS– IESFO Algorithm	UCI Cleveland dataset.	98.7	98.5	98.7	A decision function–based chaotic salp swarm (DFCSS) technique is applied for feature selection. An Improved Elman Neural Network (IENN) is then trained using the selected features to classify the data. In this approach, the Sailfish Optimization (SFO) algorithm is used to compute the ideal weights of the IENN. Accurate prediction of heart disease is possible with the help of the IESFO method, which integrates DFCSS and IENN.
(Noor et al., 2023)	Proximity Weighted Random Affine Shadowsampl ing (ProWRAS)	UCI Cleveland dataset.	97	67	80	A stacking model called PaRSEL is introduced. In the base layer, it combines four classifiers, including PAC, RC, SGDC, and XGBoost, and the final predictions are formed by using the LogitBoost in the meta layer. The presence of irrelevant and imbalanced features in the dataset increases the complexity of classification models.
(Aswathi et al., 2024)	Clustering using a Hybrid K-means-Fuzzy-C-means (HKMFCM) algorithm for segmentation	Heart Disease dataset	96.09%	94.3%	95%	The classifier employs Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) networks to improve its ability to identify complex temporal patterns in patient data. These networks are bidirectional, meaning they capture both past and future contexts.
(Pati et al., 2024)	Recursive feature elimination technique and ensemble learning approaches	Five heart disease datasets from the UCI-ML warehouse	98.32%	98.33%	98.33%	The UCI-ML warehouse's five heart disease datasets combined into one, using a two-stage experiment to inform an ensemble hybrid machine learning and deep learning approach.
(Alghamdi et al., 2024)	Dense Autoencoder and DenseNet architecture	Heart Disease UCI Cleveland dataset.	99.67%	99.97%	99.96%	Predicting cardiac disease and illness using an autoencoder and DenseNet architectures on the thirteen-variable Heart Disease UCI Cleveland dataset. When applied to cardiovascular care, this paradigm can enhance both clinical decision-making and patients.
(Isabella et al., 2025)	Harris Hawks Optimization (HHO)	Cleveland HD dataset	87.64	88	88	A HD prediction system developed using DL incorporates Harris Hawks Optimisation (HHO)
			86.14	86	86	
			87.94	88	88	
	MLP CNN LSTM Bidirectional LSTM Gated Recurrent Unit		87.78	88	88	for critical feature selection and K-mode clustering for optimal data preprocessing.
(Dhanka & Maini, 2025)	Hybrid XGBoost Classifier using "Optuna" algorithm	Cleveland HD dataset	95.45	100	96.3	HD prediction using a hybrid XGBoost Classifier framework that optimizes hyperparameter tuning and removes outliers. This method optimized hyperparameters using the "Optuna" algorithm and used z-score and interquartile range to deal with outliers.

(Akter et al., 2025)	Grow Network architecture	UCI heart disease dataset	92	90	90	Grow Network, which can adapt its configuration on the fly according to the properties of the data. Adaptive Refinement sampling and TriDyn Dependence feature selection efficiently handle class imbalance during training by selecting relevant features from uneven data.
(Yadav et al., 2025)	A hybrid classifier called HybBPF-ELM uses an optimization technique called firefly	Heart disease prediction dataset	96.56	96.53	99.55	Machine learning classifiers are used to enhance heart disease prediction through the integration of feature engineering and feature selection methods. A combined dataset was analyzed using a Random Forest (RF) model to first determine four key variables. Afterwards, a feature method was employed to enhance the dataset and boost predictive performance. This was achieved by generating 36 extra features using basic mathematical procedures.
(Sowmiya et al., 2025)	Binary Salp Swarm Algorithm	UCI heart disease dataset	95%	96%	95%	A model for feature selection based on meta-heuristics tuned for the prediction of heart illness. It employs a refined binary salp swarm technique to further improve feature selection.
(Bukhari et al., 2025)	Neural Network with Stacked Convolutional Layers and Levy Grasshopper Optimization Algorithm Based on Flight Data	Four databases—Heart-Data, Framingham, Heart-UCI, and cardiovascular disease	98.6%	98%	98%	The SCNN-LFGOA model integrates advanced feature extraction with an optimization technique that effectively tunes hyperparameters. This ensures that the model works well on different datasets while reducing the chance of overfitting.
(Shahid & Jayanth, 2025)	SMOTEENN Technique	Original imbalanced dataset and the SMOTEENN-balanced dataset	93.3%	96.4%	96.3%	SMOTEENN is a hybrid resampling method that uses Edited Nearest Neighbours (ENN) and Synthetic Minority Oversampling Technique (SMOTE) to minimize noise and achieve dataset balance.
(Chichani & Kasar, 2025)	Optimization for the Finch Hunt using Bi-LSTM classifier	1602 samples from the CACHET-CADB dataset	97.24%	93.86%	95.17%	The approach integrates multiple predictive models, including Bi-LSTM, which improves predictive performance and identifies complex temporal patterns in patient data.

Methodology

One of the most important therapeutic issues is the early and accurate prediction of cardiovascular disease risk, as it is one of the most common causes of mortality in the world. Machine learning techniques offer a viable method for assessing patient data and enhancing clinical decision-making, particularly with the increasing prevalence of electronic health records. Conventional machine learning methods do not perform well with real-world medical data due to issues such as class imbalance, noisy measurements, and intricate connections between health factors. To overcome these problems, this study suggests a structured learning approach to predicting cardiovascular disease by integrating guided deep learning, class balance, and painstaking data preparation. The suggested method is powerful and easy to understand, and it can easily handle large amounts of tabular medical data. The methodology ensures reliable model training and evaluation through a step-by-step approach when utilized on a cardiovascular dataset of 70,000 patient records. In the suggested procedure, the first step is data preparation, while model evaluation is the final step. The cardiovascular dataset is loaded and checked to ensure that it has all the necessary properties. Then, feature engineering is done to make the data more clinically useful by adding more health indicators like blood pressure and body mass index. Numerical features are cleaned and changed as required, e.g., changing age into years and applying logarithmic transformations to some variables. In order to keep the general data distribution, missing values are filled in using the median. Once the preprocessing is done, all the characteristics are scaled so that discrepancies in measurement units do not interfere with the learning process. The dataset has an unequal number of individuals with and without cardiovascular disease; therefore, the SMOTE-ENN technique is used for resampling. This procedure makes the dataset better for training by adding more examples of scenarios that aren't well represented and getting rid of noisy samples.

The processed data is then utilized to teach the proposed SMOTE-Regularized Dual-Guided Deep Tabular Distillation (SRD-GDT²) model. Within this framework, there are two tree-based models that serve as guiding references, which supply supplementary learning signals that assist the deep model to derive meaningful insights from the data. Deep learning model and guidance information are trained using real patient labels and other sources of data. This makes them more stable and better at predicting. During training, a validation set is used to find the best decision threshold that strikes a balance between finding diseases and false alarms. Then, the trained model is tested using different test data. Common performance metrics and visualization tools are used to ensure that the model is not only accurate but also useful in clinical settings. This structured approach helps provide reliable cardiovascular disease predictions while maintaining clarity and interpretability.

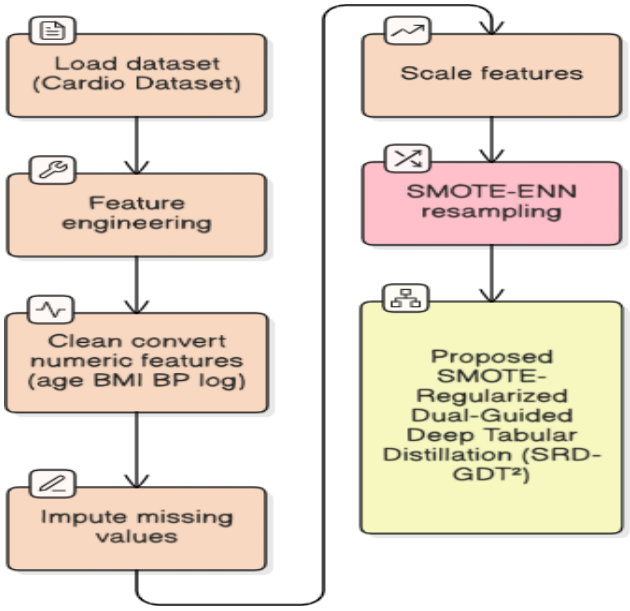


Fig. 1: Overview of the Data Preparation Steps and the Proposed SRD-GDT² Framework

Fig. 1 shows the order of the actions taken to prepare the cardiovascular disease dataset for model training. The first step is to import the cardiac information, and then feature engineering is used to find therapeutically useful variables. Numeric features such as age, body mass index, and blood pressure values are cleaned and transformed to ensure consistent data representation. Missing values are handled through imputation to avoid information loss. The features are then scaled so that all variables contribute equally during learning. To address class imbalance between diseased and non- diseased patients, SMOTE-ENN resampling is applied. The resulting balanced and processed data is finally provided as input to the proposed SMOTE-Regularized Dual-Guided Deep Tabular Distillation (SRD-GDT²) model.

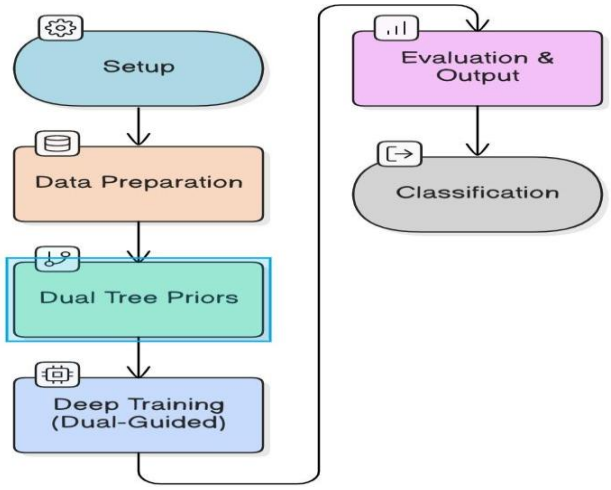


Fig. 2: Overall System Architecture of the Proposed Dual-Guided Deep Learning Framework for Cardiovascular Disease Classification

The overall framework of the proposed cardiovascular disease prediction system is shown in fig 2. The process starts with an initial setup stage, followed by data preparation where the input dataset is cleaned and transformed. Two tree-based models are then trained to act as guiding components, referred to as dual tree priors. These guiding models provide supportive learning signals to the deep learning model during training. The deep model is trained under this dual guidance to

improve prediction stability and accuracy. At last, the trained model is put through its paces by performing classification. Performance metrics and output visualizations are then used to assess the findings. A coherent and organized process that combines deep learning with traditional machine learning methods is illustrated by this layout.

From data preparation to agreement-weighted distillation, threshold selection, and final classification, the suggested SRD-GDT² framework is detailed step-by-step in Algorithm 1. In order to enhance the data quality and clinical relevance, the raw dataset is first processed with a feature engineering function. To make it easier to understand, we eliminate identifiers that do not carry any useful information and change age numbers from days to years. Additional health indicators, such as body mass index and blood pressure–related metric, are calculated in order to obtain a more precise picture of cardiovascular risk. In order to maintain the original data distribution, median-based imputation is used to manage missing or inconsistent values. Before the learning process begins, all numerical features are scaled to ensure that variables that are recorded on various ranges do not have an outsized impact. An imbalance regularization strategy is used to resolve the imbalance between patients with and without cardiovascular disease. This technique removes samples that are noisy or unclear and increases minority class representation. The data is then split into training, validation, and testing sets after preprocessing.

Algorithm 1 **SRD-GDT² — SMOTE-Regularized Dual-Guided Deep Tabular Distillation**

Input: Dataset $D = \{(x_i, y_i)\}$ where $y_i \in \{0,1\}$

Feature engineering function $\varphi(\cdot)$

Feature scaler $S(\cdot)$

Imbalance regularizer $R(\cdot)$

Two Guided models: T_1 (Random Forest) and T_2 (XGBoost / CatBoost)

Learner deep model S_θ

Distillation weight $\alpha \in [0,1]$

Fusion weights $w_f, w^1, w^2 with\ w_f + w^1 + w^2 = 1$

Agreement exponent $\gamma \geq 0$

Number of epochs E and batch size B

Output: Trained Learner model $_{\{\theta*\}}$ Optimal decision threshold τ^*

1 Data preparation

$D \leftarrow \varphi(D)$

$X \leftarrow S(X)$

$(X, y) \leftarrow R(X, y)$

2 Train dual guided models on training data

$(1)(x) = T^1(x)$

$(2)(x) = T^2(x)$

3 Fused guided probability

$(f)(x) = 0.5\ [\ (1)(x) + p^{(2)}(x)]$

4 Distillation target

$\tilde{p}(x) = w_f p^{(f)}(x) + w^1 p^{(1)}(x) + w^2 p^{(2)}(x)$

5 Agreement weight:

$(x) = (1 - |^{(1)(x)} - p^{(2)(x)}|)^{\gamma}$

6 Learner training

$z = S_{\theta(x)}$

$\hat{p} = \sigma(z)$

$L_{hard} = BCE(z, y)$

$L_{distill} = \|\hat{p} - \tilde{p}(x)\|^2$

$L = [(x)((1 - \alpha)L_{hard} + \alpha L_{distill})]$

7 Validation

$$Sensitivity(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}$$

$$\text{Specificity}(\tau) = \frac{\text{TN}(\tau)}{\text{TN}(\tau) + \text{FP}(\tau)}$$

$(\tau) = \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1$

Threshold Selection Rule

$\tau^* = \arg \max (\text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1)$

$\tau \in [0,1]$

8 Predict

$$\hat{y} = I[\hat{p} \geq \tau *]$$

Step 2 of Algorithm 1 involves training two guided models, which are tree-based, on the training data. An XGBoost or CatBoost-style gradient-boosted decision tree model and a Random Forest serve as the first and second guiding models, respectively. A patient's risk of cardiovascular disease is estimated using a probability estimate in each guided model. The deep learner relies on these probability outputs, which show various viewpoints learnt from the same data. In Step 3, the fused guided probability is calculated by combining the outputs of the two guided models. This indicates the average belief of both models. There is less bias in the separate models and a more consistent indication of direction because to this fused probability. In Step 4, the fused probability and the individual directed probabilities are weighted to produce a final distillation target. The weights make sure that the guided models keep both shared and unique information by controlling the influence of each component. Step 5 introduces a system for weighing agreements; this mechanism is used to find out the degree to which both guided models agree on their predictions. A larger agreement weight indicates more faith in the guidance when the two guided models give comparable probability estimates. The decreasing effect of the guidance can be achieved when the models differ. This safeguard ensures that the learner model increases its reliance on credible guidance only. Sixth, the two guided models and the actual patient labels are used to train the learner deep model. The student generates probability outputs by use of a sigmoid function. The difference between the learner's prediction and the guiding target is measured by a distillation loss, and a hard loss is used to measure the loss based on the true disease label. These losses are aggregated with a weighted formula, and the agreement weight is applied to ensure that samples with stronger guide agreement are used to contribute more to learning. We reduce the total loss through gradient-based optimization over a large number of training epochs, averaging over mini-batches. Step 7 is about choosing a threshold and validating it. Patient disease classification requires a decision threshold since the learner model produces probability labels instead of straight class labels. The ideal threshold is selected using Youden's index, which considers both sensitivity and specificity. In order to prevent the potential bias of the test data, this threshold is calculated solely on the validation set. By default, we keep the learner model with the highest validation performance. Step 8 concludes the process, which involves applying the trained learner model and threshold to the independent test set to produce final predictions. The patients are categorized based on whether their projected likelihood surpasses the chosen cutoff. Some of the clinically important criteria that are used to evaluate a model's performance include accuracy, sensitivity, specificity, and area-based measurements. This final evaluation reflects the real-world predictive potential of the proposed SRD-GDT² framework.

Results and Discussion

The cardiovascular disease dataset, which contains data from 70,000 patients and is open to the public, has been used to test the suggested methodology. Personal information such as demographics, medical history, and lifestyle choices are all part of each record. Cardio, a binary target label in the dataset, takes on the values 0 for no cardiovascular disease and 1 for the presence of the condition. Table 2 provides an overview of the study's important qualities and the dataset's minimum, maximum, and usual ranges for these attributes. Cardiovascular risk assessment can benefit from the therapeutically useful information provided by these features. Strong preprocessing and learning techniques are required since these ranges show substantial variation among patients.

Table 2: Description of Cardiovascular Dataset Attributes

Attribute	Description	Minimum	Maximum
Age (years)	Patient age	~29	~65
Gender	1- Female 2- Male	1	2
Height (cm)	Patient height	~55	~250
Weight (kg)	Patient weight	~10	~200
BMI	Body mass index (derived)	~15	~60
Systolic BP	Upper blood pressure	~80	~240
Diastolic BP	Lower blood pressure	~40	~160
Pulse Pressure	Systolic – Diastolic BP	~10	~120
Mean Arterial Pressure	Average blood pressure	~60	~160
Cholesterol	Cholesterol category	1	3
Glucose	Glucose category	1	3
Smoking	Smoking status 0-No 1-Yes	0	1
Alcohol Intake	Alcohol consumption 0-No 1-Yes	0	1
Physical Activity	Physical activity status 0-No 1-Yes	0	1
Physical Activity	Physical activity status 0-No 1-Yes	0	1

The original cardiovascular dataset has 70,000 samples with 20 different attributes. Using the SMOTE-ENN approach, the dataset was balanced. The number of characteristics was kept constant and the sample size was decreased to 24,257. By included 12,031 samples of individuals without cardiovascular illness and 12,226 samples of patients with cardiovascular disease, the resultant dataset attained near-perfect class balance. Because there was less bias against the dominant class, model learning was improved by this balanced distribution. Following preprocessing, the distribution of the dataset is shown in table 3.

Table 3: Dataset Distribution After Preprocessing

Dataset Split	Samples	Percentage
Training	19,405	80%
Validation	2,426	10%
Test	2,426	10%

The proposed framework was used to train the deep tabular learner for 25 epochs. To find the best decision threshold and stop overfitting, we watched how well the model did on the validation set. Table 4 shows the validation performance for certain epochs.

Table 4: Validation Performance During Training

Epoch	Validation Accuracy	Threshold	AUROC	AUPRC
1	0.9468	0.65	0.9842	0.9800
5	0.9563	0.48	0.9900	0.9838
10	0.9596	0.60	0.9909	0.9838
15	0.9608	0.59	0.9911	0.9814
20	0.9617	0.64	0.9922	0.9922
25	0.9625	0.53	0.9935	0.9942

As training continues, the data reveal a consistent rise in accuracy and discrimination capability. The high AUROC and AUPRC values show that the disease detection capability is robust and that there is significant class reparability. Models were evaluated on separate datasets using optimum thresholds derived from validation data after being chosen

based on their accuracy during validation. The results of the final tests conducted on the proposed model are displayed in table 5.

Table 5: Final Test Performance of the Proposed Model

Metric	Value
Accuracy	0.9563
F1-score	0.9558
AUROC	0.9905
AUPRC	0.9870
Sensitivity	0.9379
Specificity	0.9751
Brier Score	0.0324
Selected Threshold	0.62

The model's overall prediction performance is strong, as it reaches an accuracy of 95.63%. The high sensitivity means that people with cardiovascular disease may be effectively detected, and the high specificity means that healthy persons can be reliably identified. To aid in clinical decision-making, a low Brier score indicates that the probability projections have been well-calibrated. Using large-scale cardiovascular data, the results show that the suggested SRD-GDT² framework successfully learns complicated patterns. The combination of guided deep learning, imbalance regularization, and feature engineering, is suitable for achieving more reliable and accurate predictions. The model has outstanding discrimination capacity after balancing and noise removal, as confirmed by the strong values of the AUROC and AUPRC. The proposed method is superior to conventional single-model approaches of evaluation because it uses guided learning and agreement-based weighting to improve generalization and reduce overfitting. The model presents a fair trade-off between specificity and sensitivity, using Youden's index for threshold selection, thus it is most suitable to be used in clinical screening and decision-support scenarios. Taken together, the results of the experiments demonstrate the practical healthcare uses and excellent predictive power of the proposed framework for cardiovascular illness.

The confusion matrix results of the independent test set are depicted in fig. 3. The matrix shows the number of samples that were correctly and falsely identified by the respective classes. The cases that were properly labeled as non-disease were 1,173, and the cases that were wrongly labeled as disease were only 30 cases. We also correctly identified 1,147 disease cases and falsely classified 76 of them as non-disease. False positives and false negative rates are low, which suggests that the diagnostic reliability is strong, and there is a reasonable balance between sensitivity and specificity.

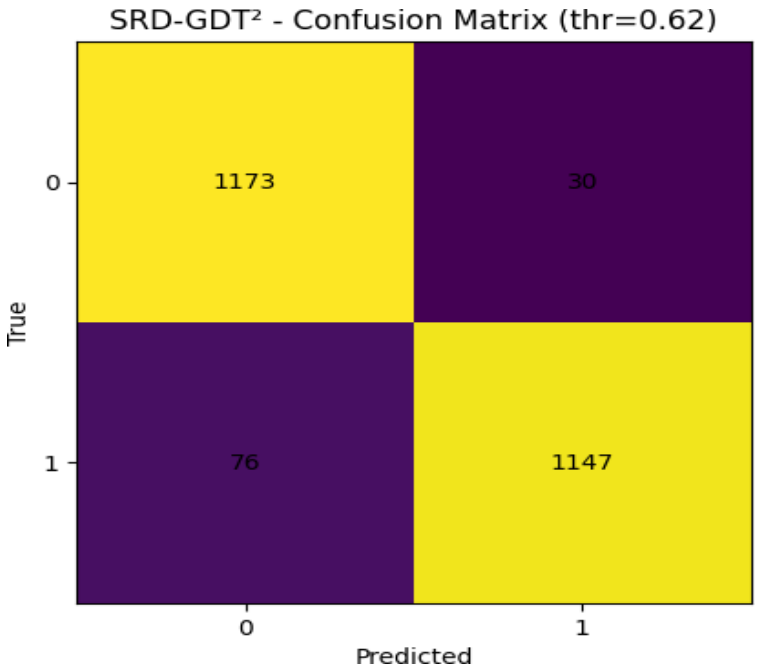


Fig. 3: Confusion Matrix of the Proposed SRD-GDT² Model on the Test Set Using the Validation-Optimized Threshold ($\tau = 0.62$)

The ROC curve of the proposed model in fig. 4 reveals a comparison of the true positive rate and the false positive rate at different decision thresholds. This is shown by the curve's tenacity in the upper left corner, which is indicative of exceptional discriminatory ability. The area under the ROC curve (AUROC) of 0.990 obtained indicates that this model is useful in effectively distinguishing between persons with and without cardiovascular illness across a wide range of thresholds. Fig. 5 demonstrates the model's accuracy in identifying illness cases using the Precision-Recall curve. High precision is maintained across a wide range of recall values, and the area under the precision-recall curve (AUPRC) has reached a value of 0.987. Even when focusing on disease detection, an important consideration in medical decision-making scenarios, the proposed model's outstanding prediction accuracy is apparent. Fig. 6 is used to compare the expected risk probabilities with observed event frequency to examine the calibration of the proposed model. The curve's proximity to the diagonal reference line is an indication of accurate calibration. The model's low Brier score of 0.032 shows that the projected probabilities are in good agreement with actual outcomes, making it suitable for risk evaluation and clinical decision-making. Fig. 7 shows the Decision Curve Analysis (DCA) to determine the clinical utility of SRD-GDT 2 model by graphing net benefit against different levels of threshold probabilities. These findings illustrate that the proposed model retains a greater net benefit than the two approaches, namely, the treat-all and treat-none approach at all risk levels. This high ground implies that cardiovascular risk assessment using the SRD-GDT 2 framework will be better in assisting in achieving better clinical outcomes in that it will be able to recognize high-risk cases, and at the same time reduce the harm caused by unnecessary medical treatment.

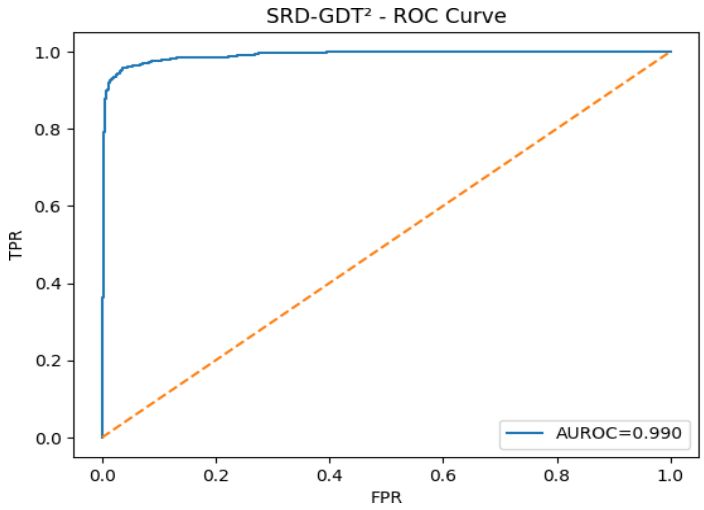


Fig. 4: Receiver Operating Characteristic (ROC) Curve of the Proposed SRD-GDT² Model on the Test Set

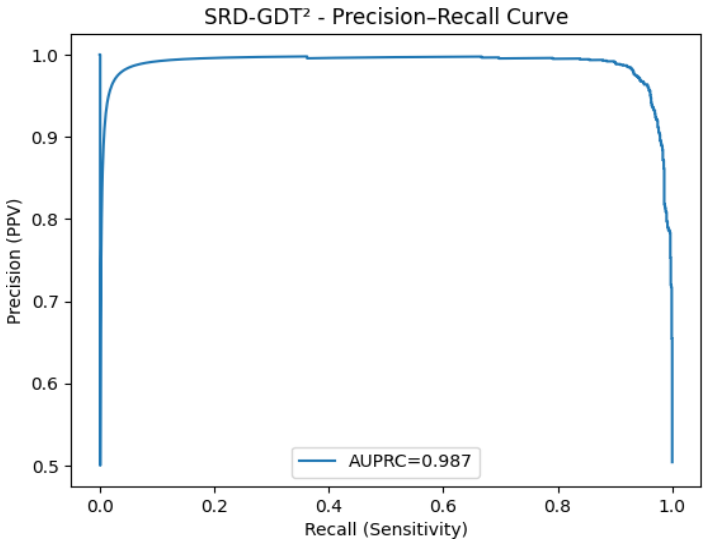


Fig. 5: Precision–Recall Curve of the Proposed SRD-GDT² Model on the Test Set

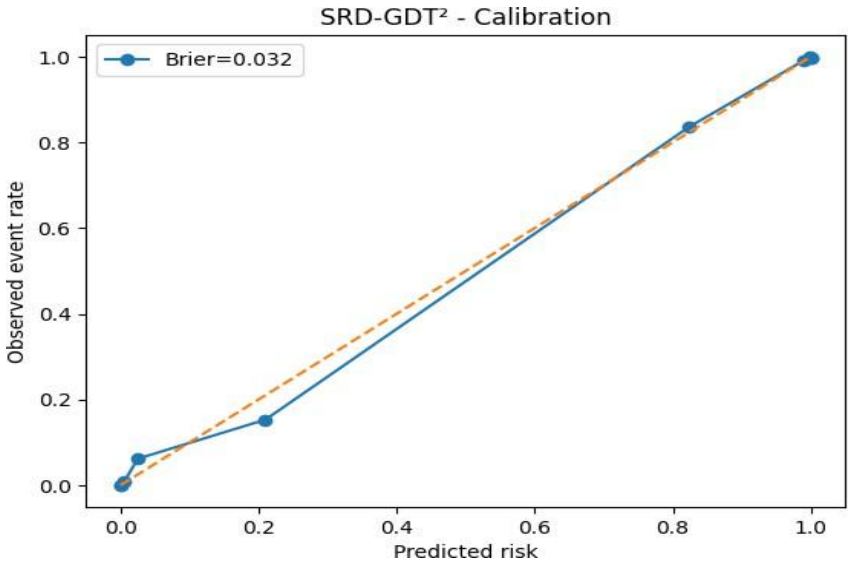


Fig. 6: Calibration Curve of the Proposed SRD-GDT² Model Showing Agreement Between Predicted Risk and Observed Outcomes

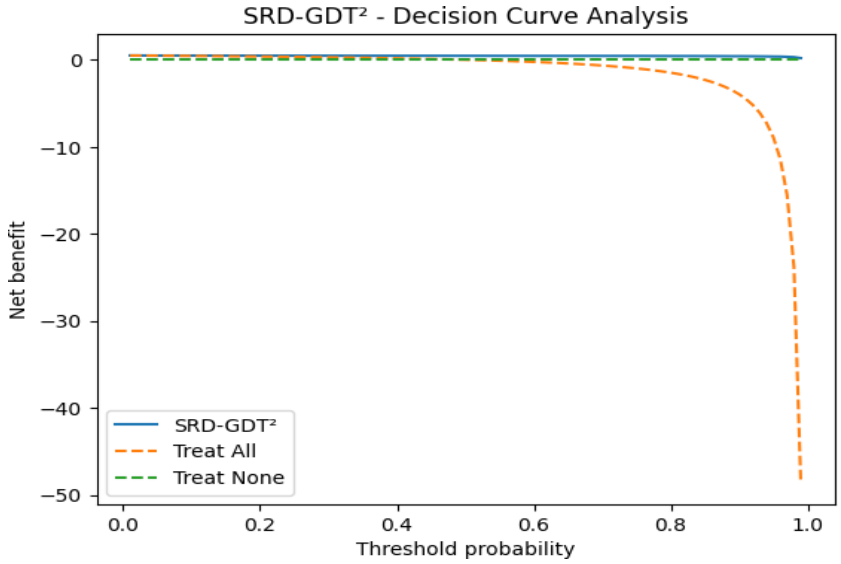


Fig. 7: Decision Curve Analysis of the Proposed Srd-Gdt² Model Compared with Treat- All and Treat-None Strategies

Conclusion

Cardiovascular disease prediction using the SMOTE-ENN regularized dual-guided deep tabular learning framework, SRD-GDT³. Common issues with medical tabular data that the suggested approach aims to resolve include class imbalance, noisy samples, and inaccurate probability estimations. The system achieves steady and reliable learning from large-scale patient data by integrating feature engineering, imbalance regularization, dual tree-based guidance, and agreement-weighted distillation. The suggested method delivers great predictive performance, with balanced sensitivity and specificity, well-calibrated risk estimations, and strong discrimination, according to experimental results on a cardiovascular dataset with 70,000 entries. Both the statistical robustness and the clinical utility are confirmed by the confusion matrix, calibration assessment, ROC and precision-recall assessments, decision curve analysis, and the whole suite of analyses. The SRD-GDT² framework demonstrates high diagnostic precision with a 95.63% accuracy and a 0.9905 AUROC, while the use of Youden's index and Brier score calibration ensures reliable, balanced clinical decision-making. Furthermore, decision curve analysis confirms that the model provides a superior net benefit over traditional treatment strategies, validating its effectiveness for real-world medical applications. The

significant overall improvement observed across different threshold probabilities implies that SRD-GDT² can assist with realistic decision-making in screening and diagnostic situations in the real world. It has never been easier to use tabular medical data to predict the likelihood of cardiovascular disease than with the proposed method. Future studies on the possible applications of the framework could focus on areas such as multi-class disease prediction, real-time decision support through interaction with electronic health record systems, and external clinical validation.

References

Karna, V. V. R., Karna, V. R., Janamala, V., Devana, V. K. R., Ch, V. R. S., & Tummala, A. B. (2025). A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. *Archives of Computational Methods in Engineering*, *32*(3), 1763-1795. https://doi.org/10.1007/s11831-024-10194-4

Wankhede, J., Kumar, M., & Sambandam, P. (2020). Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO. *IET systems biology*, *14*(6), 380-390. https://doi.org/10.1049/iet-syb.2020.0041

Kannan, S. (2022). An automated clinical decision support system for predicting cardiovascular disease using ensemble learning approach. *Concurrency and Computation: Practice and Experience*, *34*(18), e7007. https://doi.org/10.1002/cpe.7007

Goyal, S. (2022). Predicting the heart disease using machine learning techniques. In *ICT Analysis and Applications: Proceedings of ICT4SD 2022* (pp. 191-199). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-5224-1_21

Sheeba, P. T., Roy, D., & Syed, M. H. (2022). A metaheuristic-enabled training system for ensemble classification technique for heart disease prediction. *Advances in Engineering Software*, *174*, 103297. https://doi.org/10.1016/j.advengsoft.2022.103297

Noor, A., Javaid, N., Alrajeh, N., Mansoor, B., Khaqan, A., & Bouk, S. H. (2023). Heart disease prediction using stacking model with balancing techniques and dimensionality reduction. *IEEE Access*, *11*, 116026-116045. https://doi.org/10.1109/ACCESS.2023.3325681

Prabhavathi, K., & Mareeswari, V. (2023). Diagnosis of cardiac disease utilizing machine learning techniques and dense neural networks. *SN Computer Science*, *4*(5), 673. https://doi.org/10.1007/s42979-023-02081-9

Gupta, A., & Singh, A. (2023). EDL-NSGA-II: Ensemble deep learning framework with NSGA-II feature selection for heart disease prediction. *Expert Systems*, *40*(7), e13254. https://doi.org/10.1111/exsy.13254

Almazroi, A. A., Aldhahri, E. A., Bashir, S., & Ashfaq, S. (2023). A clinical decision support system for heart disease prediction using deep learning. *IEEE Access*, *11*, 61646-61659. https://doi.org/10.1109/ACCESS.2023.3285247

Aswathi, R. R., Kumar, K. P., & Ramakrishnan, B. (2024). An innovative attention infused-BiRecurrenTwin network assisted hybrid segmentation technique for accurate heart disease prediction. *Engineering Applications of Artificial Intelligence*, *138*, 109389. https://doi.org/10.1016/j.engappai.2024.109389

Majhi, B., & Kashyap, A. (2024). Explainable AI-driven machine learning for heart disease detection using ECG signal. *Applied Soft Computing*, *167*, 112225. https://doi.org/10.1016/j.asoc.2024.112225

Pati, A., Pattanayak, S. R., Agrawal, U. K., Panigrahi, A., Parhi, M., & Pati, A. K. (2024, November). An ensembled hybrid machine and deep learning approach for heart diseases prediction. In *2024 3rd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)* (pp. 1-5). IEEE. https://doi.org/10.1109/ODICON62106.2024.10797624

Alghamdi, N. S., Zakariah, M., Shankar, A., & Viriyasitavat, W. (2024). Heart disease prediction using autoencoder and DenseNet architecture. *Egyptian Informatics Journal*, *28*, 100559. https://doi.org/10.1016/j.eij.2024.100559

Natarajan, K., Vinoth Kumar, V., Mahesh, T. R., Abbas, M., Kathamuthu, N., Mohan, E., & Annand, J. R. (2024). Efficient heart disease classification through stacked ensemble with optimized firefly feature selection. *International Journal of Computational Intelligence Systems*, *17*(1), 174. https://doi.org/10.1007/s44196-024-00538-0

Venkata MahaLakshmi, N., & Rout, R. K. (2024). An intelligence method for heart disease prediction using integrated filter-evolutionary search based feature selection and optimized ensemble classifier. *Multimedia Tools and Applications*, *83*(13), 39841-39865. https://doi.org/10.1007/s11042-023-16924-3

Thumu, M. B., Balajiraja, N., & Yousoof, M. (2025). SENNERF: Stacking Ensemble with Neural Networks and Enhanced Random Forest for improved heart disease prediction. *Neural Computing and Applications*, *37*(31), 26057-26078. https://doi.org/10.1007/s00521-025-11615-6

Jiang, W., Zheng, L., Tian, Y., & Kho, Y. H. (2025). A novel research on a heart disease prediction model based on hybrid intelligent optimization of deep networks. *Discover Networks*, *1*(1), 6. https://doi.org/10.1007/s44354-025-00008-9

Isabella, S., Lokeshraja, B., Nithish Narayan, M., Dharshan, S., & Manikandan, G. (2025). Harris Hawks optimization based deep learning models for heart disease diagnosis. *Sci. Rep*, *15*, 38395. https://doi.org/10.1038/s41598-025-22326-2

Sofi, A. Q., Sharma, M., Teli, T. A., & Kumar, R. (2025). An effective deep learning-based ensemble model for heart disease prediction. *Soft Computing*, *29*(21), 5893-5923. https://doi.org/10.1007/s00500-025-10907-2

Dhanka, S., & Maini, S. (2025). A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. *International journal of cardiology*, *420*, 132757. https://doi.org/10.1016/j.ijcard.2024.132757

Bandyopadhyay, S., Samanta, A., Sarma, M., & Samanta, D. (2025). Novel framework of significant risk factor identification and cardiovascular disease prediction. *Expert Systems with Applications*, *263*, 125678. https://doi.org/10.1016/j.eswa.2024.125678

Akter, S. B., Akter, S., Hasan, R., Hasan, M. M., Eisenberg, D., Azim, R., ... & Pias, T. S. (2025). Optimizing stability of heart disease prediction across imbalanced learning with interpretable Grow Network. *Computer Methods and Programs in Biomedicine*, *265*, 108702. https://doi.org/10.1016/j.cmpb.2025.108702

Rabbi, M. S. H., Bari, M. M., Debnath, T., Rahman, A., Das, A. K., Hossain, M. P., & Muhammad, G. (2025). Performance evaluation of optimal ensemble learning approaches with PCA and LDA-based feature extraction for heart disease prediction. *Biomedical Signal Processing and Control*, *101*, 107138. https://doi.org/10.1016/j.bspc.2024.107138

Yadav, D., Rani, D., & Verma, O. P. (2025). Impact of feature selection and feature engineering in prediction of cardiovascular diseases. *Computers in Biology and Medicine*, *197*, 111027. https://doi.org/10.1016/j.compbiomed.2025.111027

Al-Mahdi, I., Darwish, S., & Madbouly, M. (2025). Heart disease prediction model using feature selection and ensemble deep learning with optimized weight. *Computer Modeling in Engineering & Sciences*, *143*(1), 875. http://dx.doi.org/10.32604/cmes.2025.061623

Sowmiya, M., Rekha, B. B., & Malar, E. (2025). Optimized heart disease prediction model using a meta-heuristic feature selection with improved binary salp swarm algorithm and stacking classifier. *Computers in Biology and Medicine*, *191*, 110171. https://doi.org/10.1016/j.compbiomed.2025.110171

Bukhari, S. M. S., Zafar, M. H., Moosavi, S. K. R., Mansoor, M., & Sanfilippo, F. (2025). An integrated stacked convolutional neural network and the levy flight-based grasshopper optimization algorithm for predicting heart disease. *Healthcare Analytics*, *7*, 100374. https://doi.org/10.1016/j.health.2024.100374

Lilhore, U. K., Simaiya, S., Alhussein, M., Dalal, S., Aurangzeb, K., & Hussain, A. (2025). An attention-driven hybrid deep neural network for enhanced heart disease classification. *Expert Systems*, *42*(2), e13791. https://doi.org/10.1111/exsy.13791

Shahid, S., & Jayanth, T. (2025, August). Optimized Heart Disease Classification using Neural Networks and SMOTEENN Balanced Training. In *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 724-728). IEEE. https://doi.org/10.1109/ICSCDS65426.2025.11167580

Pazhanikumar, K. (2025). A hybrid framework for heart disease prediction using multi-scale feature map reconstruction with inception in DenseNet and AHP-VIKOR feature selection model. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *14*(1), 1-20. https://doi.org/10.1007/s13721-025-00574-6

Huang, Y. P., Mushi, R., Sandnes, F. E., Hu, H. F., & Lee, J. H. (2025). Deep convolutional neural networks to automatically classify human heart diseases. *IEEE Access*, 13, 120672-120683. https://doi.org/10.1109/ACCESS.2025.3586214

Chichani, Y. S., & Kasar, S. L. (2025). An efficient IoT enabled heart disease prediction model using Finch hunt optimization modified BiLSTM classifier. *Biomedical Signal Processing and Control*, *100*, 107170. https://doi.org/10.1016/j.bspc.2024.107170