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

Enhancing Cardiovascular Disease Detection with SMOTE-Boosted Stacking Ensembles and Hybrid Feature Selection

Thoutireddy Shilpa[1], Priyanka T.[2]

[1]Research Scholar, [2]Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India.

E-mail: [1]shilpathoutireddy39@gmail.com, [2]priyadatta969@gmail.com

Abstract

Cardiovascular disease (CVD) is the number one cause of death worldwide and highlights the need for reliable early detection models. In this study, we introduce an integrated machine learning framework that implements efficient data preprocessing, hybrid feature selection (through Chi-square, ANOVA F-test, RFE, and LassoCV), and class balancing using SMOTE, within a stacking ensemble classifier consisting of a Random Forest, XGBoost, LightGBM, MLP, and Logistic Regression classifiers. Our proposed model was evaluated on three unique datasets: an artificially generated large synthetic dataset; a merged public dataset; and actual hospital data from Indian hospitals. Each evaluation demonstrated high levels of performance, with accuracy measures approaching 98.86% and ROC AUC reaching as high as 99.9%. We efficiently addressed class imbalance, non-linear feature interaction and data heterogeneity, achieving excellent and generalizable predictive performance. Based on the findings from this work, ensemble-based hybrid methods demonstrated reliability and may be an efficient clinical decision support system for early detection of cardiovascular risk.

Keywords: Cardiovascular Disease (CVD) Detection, SMOTE, Hybrid Feature Selection, Ensemble Learning, Stacking Ensemble.

1. Introduction

The World Health Organization (WHO, 2023) [27] reports that cardiovascular disease (CVD) is the number one global cause of death, with an estimated 20.5 million deaths in 2021 which accounts for nearly 32% of all deaths worldwide. The term CVD is an umbrella term for many diseases related to the heart and blood vessels including coronary artery disease, cerebrovascular disease, rheumatic heart disease and congestive heart failure. In the United States, cardiovascular disease accounts for a national health crisis, with the Centers for Disease Control and Prevention (CDC) [28] indicating that irresponsible lifestyle behaviors lead to one person's death every 33 seconds (CDC, 2024). Hypertension, diabetes, obesity, smoking, and physical inactivity are all modifiable risk factors contributing to a growing global burden of cardiovascular disease, especially in low- and middle-income countries where preventative care may not be accessible. The research developed a deep learning model, based on VideoMAE, to help assess cardiovascular risk as using over 31,000 carotid ultrasound videos from the Gutenberg Health Study. It used hypertension as a marker to assess visual arterial damage. We demonstrated that individuals with high predicted visual damage - with or without clinical hypertension- had a worse cardiovascular profile and a greater rate of future events. The study

DOI: https://doi.org/10.36548/jiip.2025.3.004
Received: 23.06.2025, received in revised form: 22.07.2025, accepted: 06.08.2025, published: 16.08.2025

showed the model could capture subtle features from the sonography that correspond to known cardiovascular risk factors, and should be explored further as a potential digital biomarker of early cardiovascular risk [4]. In this systematic review and meta-analysis, the authors examined the performance of machine learning (ML) models in predicting cardiovascular disease (CVD) risk using electronic health records (EHRs), as compared to standard risk scores, such as QRISK3 and ASCVD. The authors directly evaluated 20 studies, which covered 32 ML models, and concluded that models, such as random forest and deep learning, had significantly higher discrimination performance (AUC > 0.84) in predicting risk than standard methods. However, the authors noted that, although ML models may offer promising accuracy, there were severe inconsistencies in methodological definitions and a risk of bias in the studies examined; therefore more standardized and clear methods need to be established before clinical implementation [5]. In this study, six algorithms of machine learning methods to predict cardiovascular disease (CVD) risk used both clinical features and geographical features including temperature, humidity, and education level. The authors completed two experiments, one with geographical features and one without. They found that the models performed significantly better when geographical features were included. The highest accuracy was achieved by the XGBoost algorithm, at 95.24%, among all algorithms. This study demonstrates that including environmental data is useful for predicting CVD risk [6]. This paper performed a systematic survey of machine learning and deep learning approaches to heart disease prediction and reported the methods' performance across a variety of datasets and evaluation metrics. In general, the results show hybrid and ensemble models performed better than traditional classifiers, especially when deep learning was used with optimized feature selection to make the selection of features more robust and accurate [7]. This study systematically reviewed 37 studies published between 2018 and 2022 that applied deep and machine learning algorithms for cardiovascular disease prediction. The aim of the study was to review the studies regarding classification accuracy, datasets, tools, and algorithms employed. This review determined that the most applied algorithms were CNN, SVM, KNN, and RNN; the overwhelming majority used supervised learning algorithms. The study also noted limited use of ensemble or reinforcement learning algorithms, as well as the need for larger and more diverse datasets, along with more studies in areas of the world with the least published literature [8]. As the incidence of global mortality caused by cardiovascular diseases continues to rise, it is necessary to create a reliable and efficient detection model for the same. Hospitals, clinics, diagnostic centers, and other healthcare institutions are grappling with indecipherable unstructured data from patients. Furthermore, the ever-increasing reliance on digital data collection frameworks has caused new challenges including missing values, irrelevant features, weakly correlated features, class imbalance, noise and deteriorating data quality, all of which are known to affect prediction models within health contexts. As health-related data is extremely messy and unintentional, data preprocessing is paramount before engaging in classification and predictions. Motivated by the above context, the aim of this study is to propose a comprehensive machine learning-based cardiovascular disease detection model that emphasizes accurate data treatment in order to develop a reliable and clinically valid disease detection method. Previous research on predicting cardiovascular disease predominantly using a single classifier (Random Forest, SVM, Logistic Regression) or standard ensembles has faced several fundamental limitations.

In the first case, these approaches suffer because they are not able to appropriately capture the non-linear and complex associations between heterogeneous clinical, demographic, and fine-grained lifestyle features, leading to overfitting and lack of generalizability.

Second, real-world datasets are almost always imbalanced because it is very unlikely that a large number of patients with higher risk will exist in a healthy cohort. Patients display non-random tendencies leading the model to be unable to identify high-risk patients when they are present in small proportions. Academics have deployed SMOTE or other oversampling approaches to address class imbalance in prior studies, but few studies incorporate these methods on class-imbalanced datasets in conjunction with feature optimization. Most other studies identify and rely on one feature selection approach, thus missing out on unique and subtle predictors and being less generalized across datasets. With these identified knowledge gaps in mind, this study will generate and validate all of the parts proposed pipeline.

The proposed pipeline will entail combined hybrid feature selection (Chi-square, ANOVA F-test, RFE, LassoCV), SMOTE oversampling, and a stacking ensemble. Overall, the proposed stacking model will combine base-level classifiers Random Forest, XGBoost, LightGBM, MLP and Logistic Regression as part of a stacking ensemble, in an effort to capture complexity in the modeling. Each classifier will be linked through a meta-learner. This method combines several techniques to avoid overfitting, correct class imbalance, and increase robustness. As we demonstrate effective generalization across synthetic, public merged and real hospital datasets, we have thus addressed some of the major issues with prior models and developed a more trustworthy decision support tool for the early prediction of cardiovascular risk.

The key contributions of this study can be summarized as follows in Figure 1:

1. The machine learning pipeline adopted in this study consists of preprocessing steps including label encoding of binary and categorical features and scaling data with either StandardScaler or MinMaxScaler.

2. The results of the previous step were followed by the application of feature selection methods (Chi-square, ANOVA F-test, RFE, and LassoCV) and a voting mechanism to pick the most important features based on the outcomes of the methods applied previously to improve model performance and the complexity of the model.

3. Following the previous step, SMOTE was used to oversample the minority class to effectively address class imbalance. we passed through dimensionality reduction through Principal Component Analysis (PCA) to 95% variance in order to visualize the results and process the data efficiently.

4. The modeling phase incorporated a stacking classifier predicting the target classes consisting of various base learners (RandomForest, XGBoost, Logistic Regression, MLPClassifier, LightGBM) including Logistic Regression as the meta-learner.

5. The model created was assessed against a number of metrics including Accuracy, Precision, Recall, F1 Score, ROC AUC, FPR, FNR and TNR. A test/training split with stratified 5-fold CV as the evaluation strategy was completed to assess the model's performance. The novelty of this work lies in integrating a four-level hybrid feature selection (Chi-Squared, ANOVA, RFE, and LASSO), SMOTE-based class balancing, and PCA-based visualization with a stacking ensemble of RF, XGB, LGBM, MLP, and LR. This unified pipeline for cardiovascular disease prediction has not been previously explored in existing literature. The organization of the paper is as follows: Section 2 describes the literature on cardiovascular

disease (CVD) detection and prediction. Section 3 describes the model that was created. Section 4 provides a description of the dataset, experimental setup, results and discussion. Section 5 concludes the paper and Section 6 covers the future scope of the paper.

2. Related Work

Ahmad et al. [1] The aim of the study was to understand how three feature selection methods, Mutual Information, ANOVA, and Chi-Square, affect the predictive performance of several machine learning and deep learning algorithms for predicting heart disease. Results indicated that Mutual Information improved the accuracy and recall of more advanced models (neural networks) while simple models benefited from the efficient application of ANOVA and Chi-Square. The authors noted throughout the study that the choice of an appropriate feature selection method was an implicit driver of improvement in diagnostic accuracy and computational efficiency.

Lübeck, Frederike, et al.,[2] This study introduced AdaCVD, a cardiovascular disease (CVD) risk prediction framework based on large language models that had been fine-tuned from ove 500,000 participants in the UK Biobank. The model was able to achieve state-of-the-art predictive accuracy and was found to be flexibly adaptable while being able to operate on three different dimensions: the combination of structured and unstructured patient data, the presence of missing information, and the ability to transfer to different clinical populations. AdaCVD outperformed traditional risk scores and machine learning baselines, which included recent work with more than one million patients, suggesting the potential to be an adaptable real-world clinical decision support tool.

Liu, Minyu, et al. [3] This research examined how social determinants may affect the relationship between alcohol use and the risk of cardiovascular disease (CVD) among over 30,000 adults in the United States. The results showed an association between moderate levels of drinking and reduced CVD risk that is at least partly accounted for by favorable social determinants, while heavy drinking was shown to be associated with poorer social determinants that may have obscured the cardiovascular benefits of alcohol. In conclusion, both the consumption of alcohol and social determinants affected CVD health outcomes.

Dritsas et al. [9] Four supervised machine learning models were used in the study (Logistic Regression, SVM, Random forest and Naive Bayes) to predict long term cardiovascular disease (CVD) risk through a balanced dataset of 70,000 participants. Logistic Regression also outperformed the other models in accuracy (72.06%), recall (72.10%), and AUC (78.4%). Logistic Regression seems to be the most effective for predicting CVD at preclinical stages in an elderly cohort.

Trigka et al. [10] This study evaluated five deep learning models for cardiovascular disease prediction using a real world dataset with 20,000 patients. The study utilized a novel enhanced SMOTE technique to handle the class imbalance problem by utilizing correlations in the feature. The convolutional neural network (CNN) model in conjunction with enhanced SMOTE resulted in the best prediction performance, the CNN attained 91% prediction accuracy and 0.90 AUC. The CNN outperformed other models, as well as the traditional SMOTE model, and no balancing approaches. With the use of inter-feature relationships in synthetic data, the results showed that the generalization of prediction increased, making this method a contribution to healthcare analytics.

de Miguel-Díez, Javier, et al [12] This research focused on the often co-occurrence of COPD and cardiovascular disease, which comes with enhanced risk and clinical challenges because of the bidirectional nature of their relationship. This research also provides some recommendations on multidisciplinary management to improve diagnosis and treatment outcomes as well as interdisciplinary care between the pulmonologist and the cardiologist.

Sharma, Narendra Kumar, et al. [13] The study utilized an iterative ensemble machine learning strategy that leveraged classifiers (such as decision trees, logistic regression, K-nearest neighbors, and random forests), because ensemble methods can improve the accuracy of heart disease prediction. The design was able to generate a model that produced higher precision than single models, which indicates that it may be viable to use in support of clinical decision support systems and training programs to help mitigate misdiagnosis.

Mauya, Jannatul, et al [14] The research created a predictive framework using a blended data set from four countries and addressed missing risk factors using multiple linear and Huber regression methods. The proposed stacking classifier, as well as the ensemble-selected features, increased prediction accuracy and improved robustness when predicting cardiovascular disease.

In their research, Jian Yang and Jinhan Guan [15] aimed to develop a heart disease prediction model based on feature optimization and the SMOTE-XGBoost algorithm. They used information gain for feature selection and use SMOTE-ENN to upsample data imbalances in which predictive accuracy was improved significantly. This study shows that this hybrid framework performed better than all of the other baseline machine learning algorithms in this study according to accuracy, precision, recall, and AUC.

de la Brassinne Bonardeaux, Orianne, et al [16] In this retrospective study, the association between high-sensitivity C-reactive protein (hs-CRP) and cancer incidence was examined in 174 cardiovascular disease (CVD) patients. Hs-CRP was not predictive of cancer development but high hs-CRP levels were significantly linked to increased one-year and long-term mortality. Because treatment with antiplatelet and statin therapies was associated with reduced cancer incidence, it suggests that these treatments could have protective effects, which would need to be further investigated.

Talaat, Fatma M [17] This research introduces CardioSentiNet, a novel deep learning framework that integrates convolutional neural networks (CNNs) and self-attention-based transformers to improve cardiovascular disease (CVD) risk prediction from individual health markers. By revealing hidden individual patterns and complex multi-factor interactions within several factors including blood pressure, cholesterol, and lifestyle habits, the model achieved a high prediction accuracy ($R^2 = 0.994$), better than traditional prediction methods. The work also emphasizes the importance of interpretability and ethical implications, thus being beneficial for eventual widespread real-life clinical use, and personalized care.

Cao, Xiyu, et al [18] The analysis purports to establish a position of positive association between airflow obstruction (AO) and cardiovascular disease (CVD) using NHANES III and 2007 - 2012 databases for predictive model building. By implementing AO into their machine learning algorithms especially their XGBoost initiation, the authors advanced predictive risk assessment accuracy, potentially improving health care costs over the long run despite limitations in applicational effectiveness over time. Ultimately, their primary predictors and implicit decision systems partially differed between the general population and those with AO. These findings reinforce the rationale for including respiratory parameters into CVD risk

models to rejuvenate existing interventions and provide reasonable and responsible strategies for early detection.

Tian, Jing, et al. [19] This prospective cohort study shows that an increased estimated glucose disposal rate (eGDR)—a measure of greater insulin sensitivity—is significantly associated with a lower risk for cardiovascular disease (CVD), heart disease, and stroke, among people with CKD stage 0-3. The relationship was partially mediated by BMI, underscoring the important role of controlling for both obesity and insulin resistance in diminishing CVD risk.

Bai, Tiantian, et al [20] This study conducts systematic research on six swarm intelligence feature selection methods (WOA, CSA, FPA, HHO, PSO, and GA) across two cardiovascular disease (CVD) datasets to improve early CVD diagnosis via machine learning. We have demonstrated that optimal feature subsets improved prediction accuracy within classifiers, including Random Forest, XGBoost, AdaBoost, and KNN. The findings show that CSA and WOA performed best on balanced datasets and imbalanced datasets. We conclude that swarm intelligence methods can be used to improve CVD prediction models using various datasets.

Ganie et al [21] Utilizing ensemble learning methods, this work applied stacking and voting approaches combined with explainable AI in ensemble models to improve heart disease prediction from different datasets. The new models, made up of six optimized base classifiers, outperformed the accuracy and robustness of individual models, achieving as much as 98% accuracy. Interpretability from the SHAP modeling led to meaningful understanding of how features contributed, thus allowing for transparent decision-making that is relevant for clinical context.

Mittal, Pooja, et al [22] This study presents a new hybrid machine learning model to accurately detect cardiovascular disease,which has a 4-stage pipeline. It begins with SMOTE-ENN to balance the data, then Chi-square to select a feature set among 30 total features, and finally, it stacks a Random Forest Tree, K-Nearest Neighbor, and AdaBoost with a Meta Learner of Logistic Regression. The proposed model significantly outperforms all other traditional models, with an accuracy of 97.8% and ROC AUC of 98.6%, illustrating that it is a viable method for CVD prediction, effectively and reliably for healthcare practitioners.

Xia, Biao, et al [23] Researchers have introduced a novel model, ICVD-ACOEDL, that utilizes ACO for feature selection and Bayesian optimization for hyperparameter adjustment. The model enhances deep learning-based diagnosis of cardiovascular disease (CVD). It achieved a maximum classification accuracy of 99.71% on benchmark datasets, outperforming previous traditional techniques. This model is promising to pave the way for an effective, scalable, and practical method for early and accurate detection of CVD.

Dorraki, Mohsen, et al [24] In this prospective study, the authors developed an ensemble machine learning model that considerably improved cardiovascular disease (CVD) prediction accuracy by considering mental health data along with traditional risk factors. Using data from 375,145 UK Biobank participants, the model's accuracy increased from 71.3% to 85.1% when psychological variables of depression, anxiety, and stress were included. This work illustrates the importance of incorporating mental health assessments to improve the prediction of CVD risk.

Zheng, Dongze, et al [25] This cross-sectional research evaluated the association of the triglyceride-glucose (TyG) index and the three respective measures of obesity (TyG-BMI,

TyG-WC and TyG-WHtR) with chest pain and cardiovascular disease (CVD) in individuals with diabetes and pre-diabetes. Findings revealed positive associations between the TyG index and obesity measures and chest pain risk and total-CVD risk. The TyG-WC and TyG-WHtR had the greatest predictive ability in identifying and stratifying risk in pathologically hyperglycemic individuals. Asadi, Fariba, et al [26] This research paper examines the effectiveness of advanced tree-based machine learning methods (Random Forests, XGBoost and LightGBM) in detecting cardiovascular disease (CVD) using clinical data. Of the models considered in this paper, XGBoost provided the best prediction accuracy - which can drive possibilities for robust, data-driven CVD diagnosis and practice in the health sector.

Recent research on predicting cardiovascular disease using machine learning approaches has demonstrated good, but inconsistent predictive performance. Trigka et al. [10] reported a predictive performance of 91% and 0.90 AUC using CNN with enhanced SMOTE; Dritsas et al. [9] reported 72.06% and 0.78 AUC using logistic regression; Mittal et al. [22] also reported a predictive performance of 97.8% and 98.6% AUC using a hybrid stacking model; Ganie et al. [21] reported combining stacking with explainable AI and achieving up to 98% accuracy.

3. Proposed Work

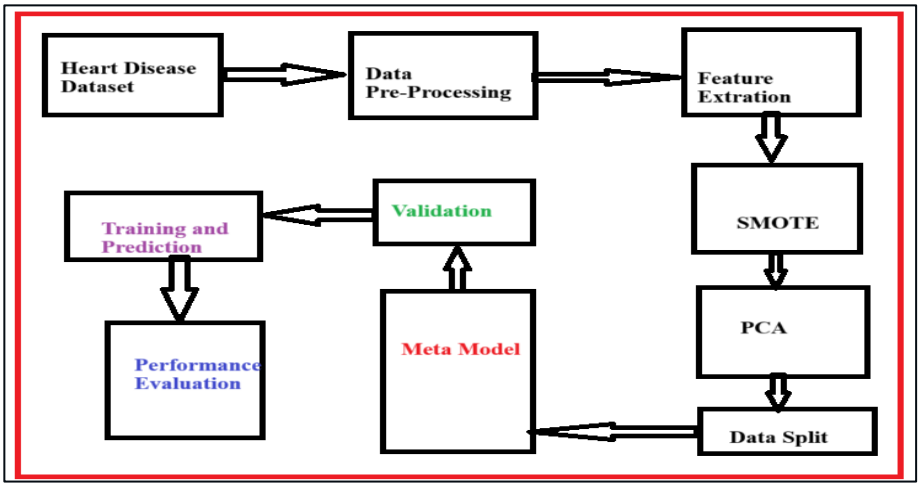


Figure 1. Flow Diagram

3.1 Dataset

In the study three unique datasets were utilized to investigate heart disease risk and prediction. The first dataset was a synthetic heart disease risk dataset provided by EarlyMed–VIT-AP, which has 70,000 records. The dataset was created to mimic the relationships between symptoms database and risk factors. The dataset contains 18 features, which are either binary for symptom or binary/continuous for risk factor (for example symptoms of chest pain, shortness of breath and fatigue, risk factors of age, diabetes, cholesterol, sedentary lifestyle, and others). The target variable identifies the level of heart disease risk (0: Low Risk, 1: High Risk). The dataset is available online at Dataset 1[29].

The second dataset is a real-world heart disease dataset that has been merged from five publicly available datasets, containing 1,888 records. There are 14 features which include

clinical and demographic features such as age, sex, type of chest pain, cholesterol level, fasting blood sugar, max heart rate, and whether they have heart disease. This dataset is meant to be engaging and provide a great real-world perspective, you can find the second dataset through Dataset 2[30].

This study uses a third dataset sourced from an Indian multispecialty hospital consisting of 1,000 records. It includes real clinical records from Indian patients, accounting for variations in clinical incident severity, and contains 14 attributes such as patient ID, gender, resting blood pressure, serum cholesterol, electrocardiogram (ECG) results, and other risk markers. This dataset offers localized perspectives on cardiovascular risk classification and is published at Dataset 3 [31].

Each dataset was pre-processed and the clinical traits relevant to the investigation were extracted for analysis and modeling. The integration of synthetic and real data empowers the authors to form a strong foundation for a predictive model for cardiovascular disease.

3.2 Data Pre-Processing

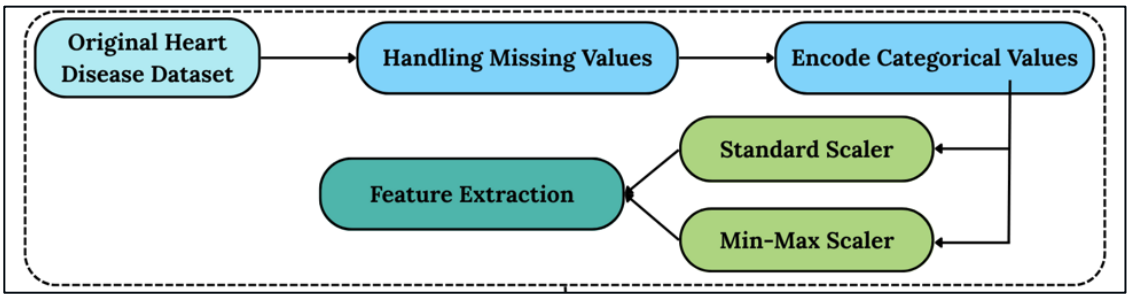


Figure 2. Diagrammatic View of the Pre-Processing

A solid data pre-processing pipeline is vital for improving model performance and reliability in many fields, and in healthcare-based applications it is essential for data quality. In the above Figure 2, the following tasks illustrate the pre-processing process that we undertook on the original heart disease datasets:

3.2.1 Original Heart Disease Dataset

The raw datasets contain both synthetic and real clinical data, with a diverse number of features (symptoms, demographics, biomedical parameters). The raw input often contains missing data elements, formatting issues, and varying and unscaled numerical ranges, representing an extensive pre-processing stage

3.2.2 Missing Values

The first step of pre-processing is to identify and address missing values as they have the potential to introduce practitioner bias or decrease model accuracy rates. For the numerical features, mean/mode imputation will be used and for the categorical features most-frequent imputation is applied. Excessive missing data is removed through the exclusion of entire rows when necessary to protect data integrity.

3.2.3 Encode Categorical Values

Categorical Value Encoding Because machine learning models require numerical inputs, categorical features need to be encoded. Categorical features (gender, chest pain type, exercise-induced angina) will be transformed with appropriate encoding techniques. For example, ordinal variables will be label encoded and nominal variables will be subject to one-hot encoding. Textual features will undergo model transformation, in which transformed variables will conform to standard (custom) representations.

3.2.4 Scalers

Standard Scaler performs standardization by removing the mean and scaling to unit variance. This is appropriate to use for a model that makes use of a normal hypothesis about the data, like logistic regression or a neural network. Min-Max Scaler scales the features to a specific range, generally [0, 1]. This is good for algorithms that are sensitive to absolute magnitude, or the independently scaled value of features (e.g., k-NN, neural networks). Depending on the distribution and nature of each feature, both scalers were applied to specific instances to ensure that no individual feature dominates the learning simply because of its scale.

3.3 Hybrid Feature Selection (FS)

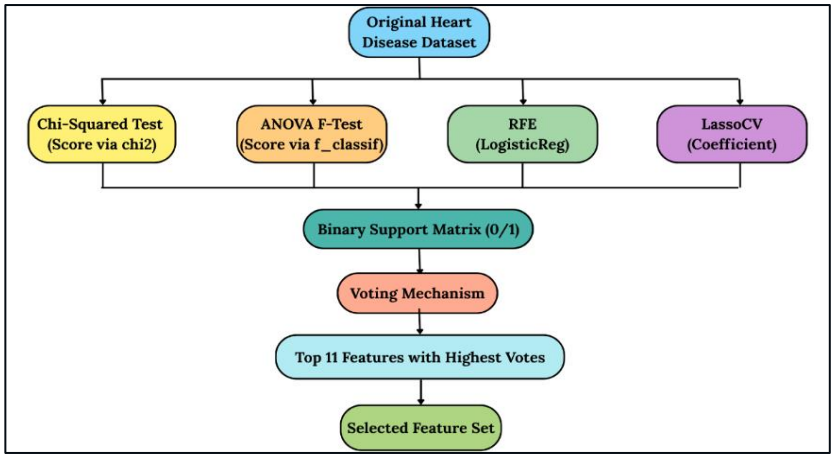


Figure 3. Diagrammatic View of the Feature Selection

In the above Figure 3, the original heart disease dataset will go through four different hybrid feature selection procedures: Chi-Squared Test (chi2), ANOVA F-Test (f_classif), Recursive Feature Elimination (RFE utilized with Logistic Regression), and LassoCV (based on coefficients reading). Each of these approaches provides a binary support vector (0/1), which indicates whether the feature was selected. The support vectors will be combined into a Binary Support Matrix format and a voting mechanism will be implemented to simply count the number of methods that selected each feature. The top 11 features with the most votes will be included in the final selected feature sets for further analysis or modeling.

3.4 SMOTE and Dimensionality Reduction (DR)

In this analysis, different supervised machine learning classifiers were employed to classify various indicators of student academic performance on an imbalanced dataset. The

addition of SMOTE ultimately improved the overall classification performance of the classifiers, especially for the minority classes. Naïve Bayes and Random Forest improved dramatically. The experimentation confirmed that oversampling techniques are effective in alleviating class imbalance in educational data mining [11].

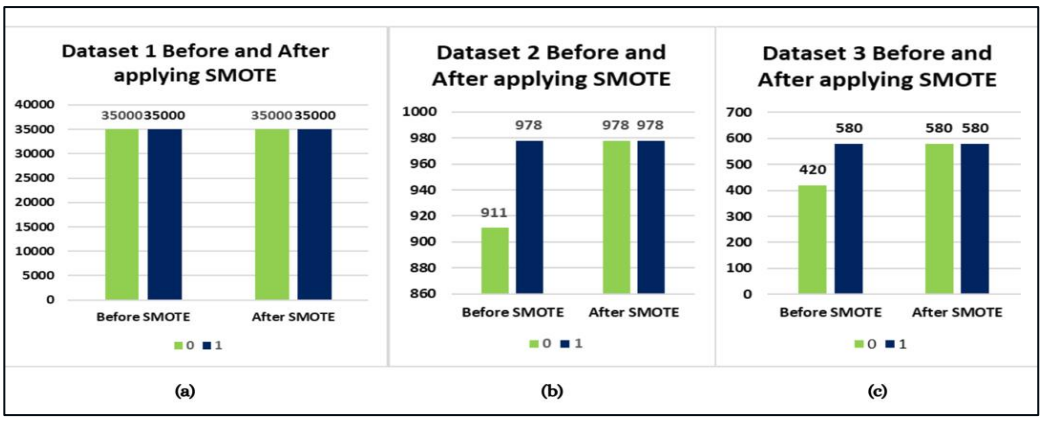


Figure 4(a). SMOTE

The bar charts above show Figure. 4(a) the effect of SMOTE on class balance in three datasets.

(a) Dataset 1: The dataset has a perfect 1:1 ratio, that is, both class 0 and class 1 have exactly 35000 samples. Since there are no class per-trained examples needed to sample, we see that SMOTE preserves the class balance. (b) Dataset 2: The two pre-SMOTE classes were imbalanced, i.e., class 0 had 911 samples and class 1 had 978 samples. In order to create a more balanced class distribution, SMOTE oversampled class 0 until it reached 978 (the same number of examples as with class 1). (c) Dataset 3: The two pre-SMOTE classes had 420 samples for class 0 and 580 samples for class 1, which was not completely balanced. After SMOTE was applied, class 0 underwent enough sampling to match class 1 (580 samples across both classes). These charts demonstrate how effective SMOTE can be in producing balanced datasets, and removing the effects of class imbalance. This is important in stopping any bias in the model and more accurately measuring the performance of the classifier on the minority class (class 0). It will allow for fairer representation of both classes during training and will lead to models that predict with more reliability and generalizability.

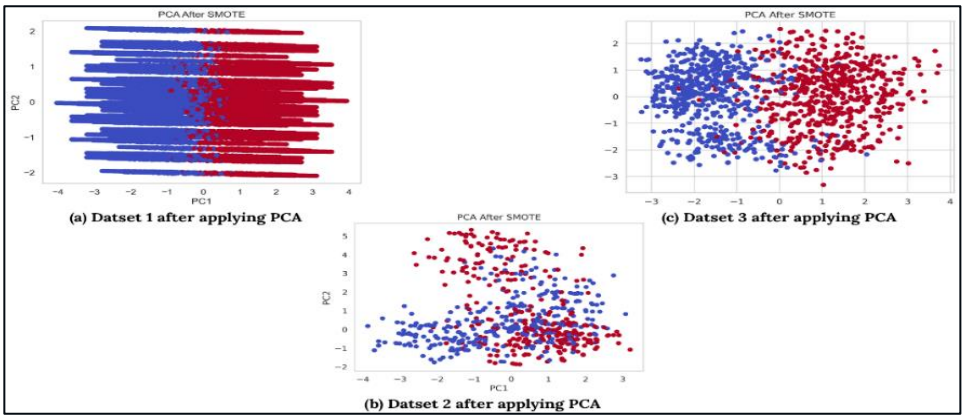


Figure 4(b). Dimensionality Reduction

The plots above show distributions of classes in three different datasets (Dataset 1, Dataset 2, Dataset 3) Figure 4(b) after performing SMOTE and dimensionality reduction with PCA. In each subfigure, there are red and blue dots representing samples from two different classes.

(a) shows that Dataset 1 is still very overlapped along the principal components, indicating that the classes remain hard to separate. (b) displays better separation between classes in Dataset 2 which shows improved class separability after SMOTE was performed. (c) confirms better class separation in Datasets 3 than in Dataset 1 and 2 as the distribution shows vertical (less overlapped area) separation on the PCA plot and more distinct cubic clustering. Essentially, thesefigure have confirmed the usefulness of SMOTE and PCA to examine the distributions and separability of classes in imbalanced datasets.

3.5 Machine Learning Techniques

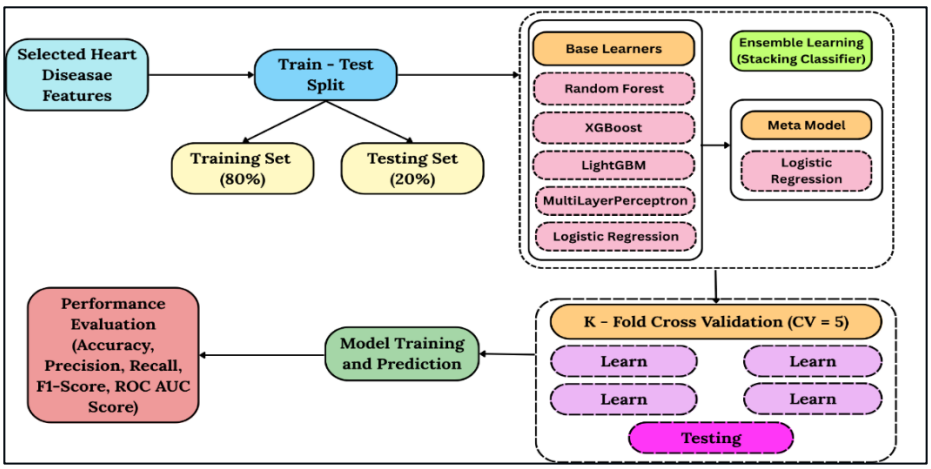


Figure 5. Workflow Depicting the Steps Followed by Stacking Ensemble Techniques

After the pre-processing stage, various commonly used machine learning (ML) approaches were researched to find the appropriate technique suitable for the data analysis process. This chapter contains a brief description of these common ML algorithms, in addition to the comparative performance measurements, selecting the best approach for the data set. We proceed to perform feature selection, by applying a hybrid feature selection approach. These features will then be used for predicting the models. In dataset splitting, the dataset is split into training (80%) and testing (20%) sets using a Train-Test Split method. This ensures the model is trained and evaluated on different data, which can help avoid overfitting. Ensemble Model Architecture The base learners' outputs are combined through a stacking ensemble model that includes Random Forest, XGBoost, LightGBM, Multi-Layer Perceptron (MLP), and Logistic Regression. The base models pass their outputs into a Logistic Regression classifier, which is the meta-model, to make the final prediction.

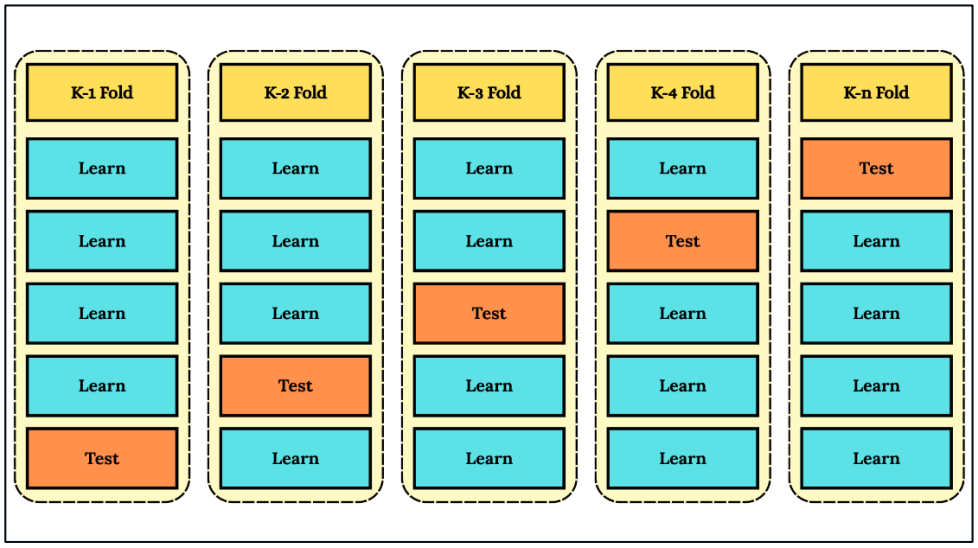


Figure 6. K-fold Cross Validation

The training set is subjected to a cross-validation process (CV = 5), where the training set is segmented into 5 equal components to use 4 for training and 1 for testing the model for each fold. This involves executing the training process on 4 of the segments 5 times, using each of the folds as testing data once. This approach provides a solid inference of the model's generalization performance.

In Model Training and Prediction, the base learners and meta-learner are trained on the training data (inside each fold). Predictions are made on the test set (unseen data) after training.

4. Results and Discussion

To analyze the robustness, scalability, and reliability of the presented machine learning models, experiments were conducted on three unique cardiovascular datasets. To assess the generalization ability and robustness of the proposed model, experiments were performed on three separate datasets.

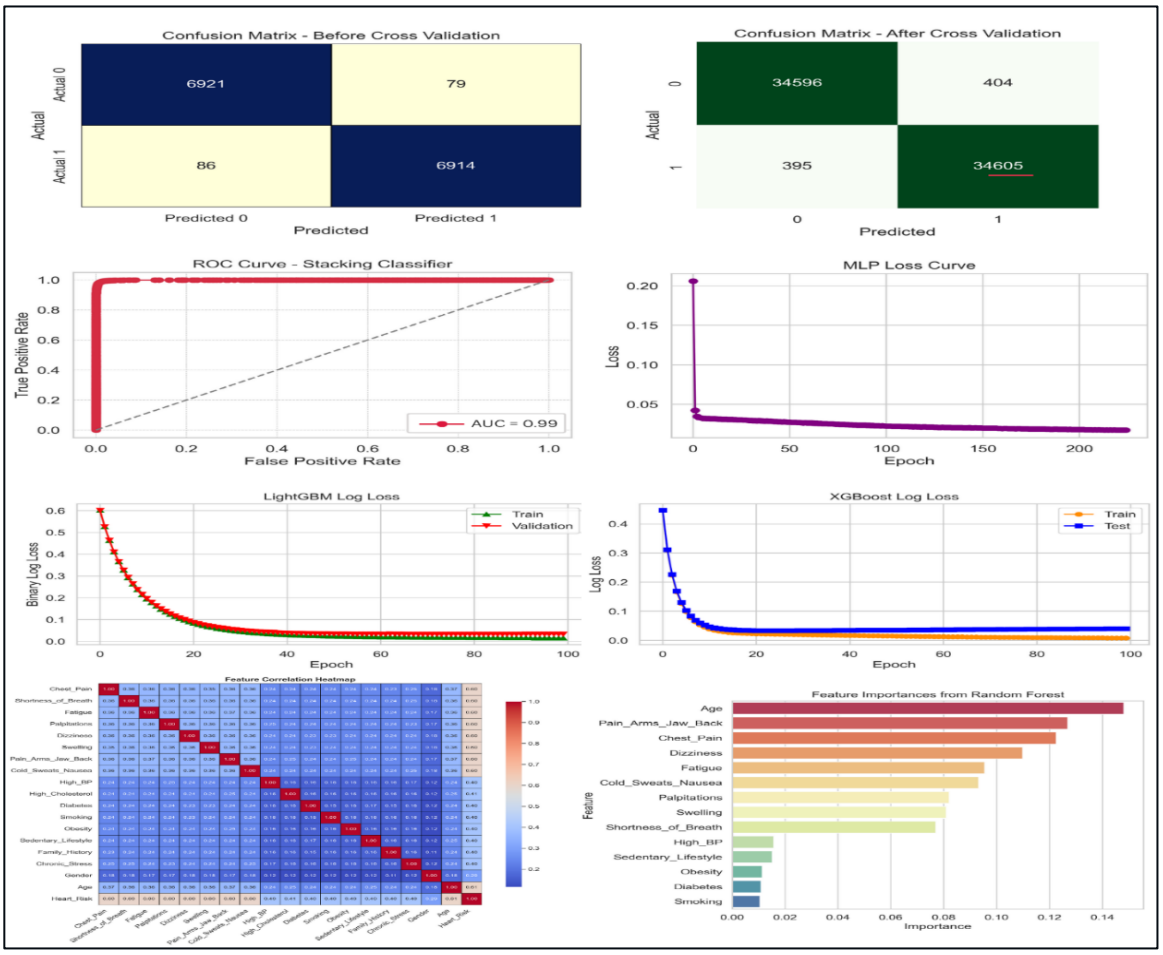


Figure 7. Dataset1[29] Results

There appears to be a dataset that is relatively large, well balanced (tens of thousands of instances), and feature-rich in relation to clinical symptoms and indicators based on lifestyle (Chest Pain, Shortness of Breath, Age, Fatigue).

Performance

- **Confusion Matrix:** Excellent accuracy from the MLP approaches both prior to (TP = 6914, TN= 6921, FP = 79, FN = 86) and after cross-validation (TP = 34,605, TN =34,596, FP = 404, FN = 395) with low average false positives (FP = 404) and low average false negatives (FN = 395) provided by our approach (nearly 70000 predictions) corresponding to low misclassification rates.
- **AUC Score:** 0.99
- **MLP Loss Curve:** Sufficiently smooth convergence was obtained with stable training within 200 epochs.
- **Log Loss:** LightGBM and XGBoost provided slightly decreasing binary log loss, indicating little to no overfitting.
- **Feature Importances:** The Random Forest approach identified Age, Chest Pain, and Pain in Arms/Jaw/Back as the top 3 features.

- **Correlation Heatmap:** A strong positive correlation was established between chest-related indicators and the target label, indicating strong clinical relevance.

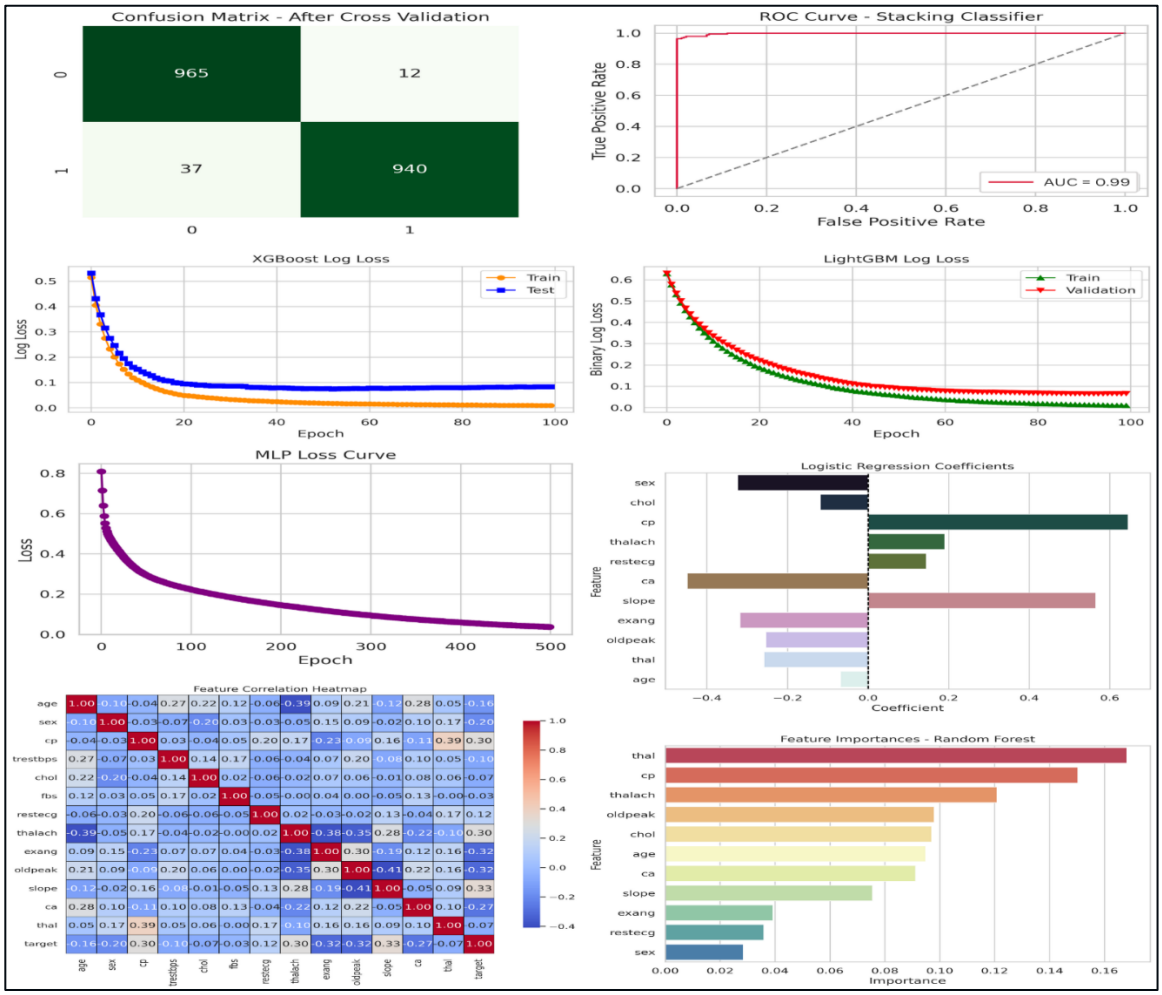


Figure 8. Dataset2[30] Results

The dataset consists of standard heart disease characteristics (e.g., sex, age, cholesterol, tha lach, ca, cp, and others). The dataset is of moderate size and frequently appears in research.

Performance

- **Confusion Matrix:** After performing cross validation, achieved very high classification accuracy and with 12 false positives and 37 false negatives.
- **AUC Score:** 0.99
- **MLP Loss Curve:** Although convergence is longer (~500 epochs), the MLP converges in a smooth and stable way.
- **Log Loss:** Both XGBoost and LightGBM log-loss curves low error and consistent convergence.

- **Feature Importance:** According to Random Forest, the most important predictors were thal, chest pain (cp), and thalach. Logistic Regression showed ca, cp, and old peak as the most important predictors (positive/negative weights).
- **Correlation Heatmap:** The correlation heatmap demonstrates expected clinical relationships (e.g., cp and thalach were inversely associated with heart disease).

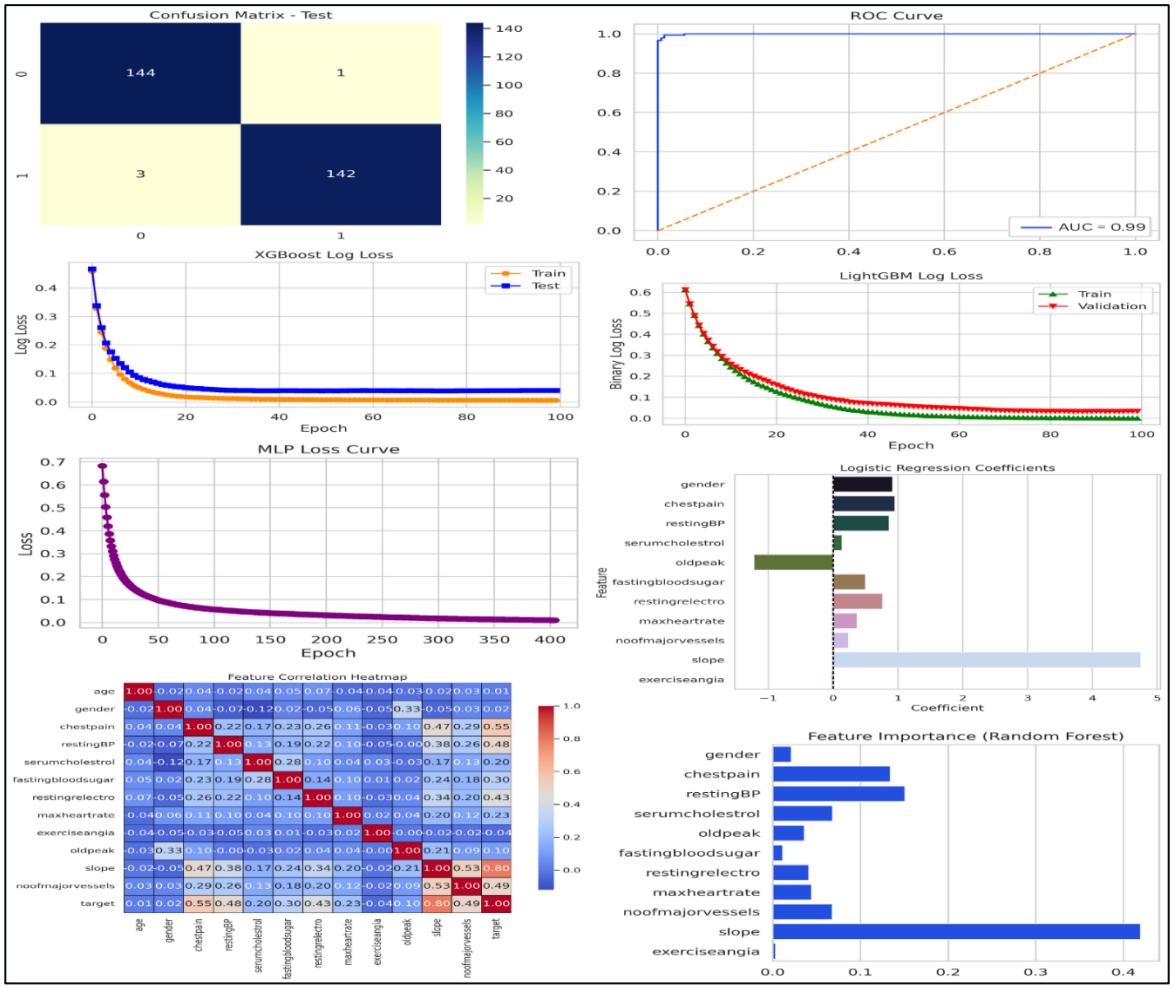


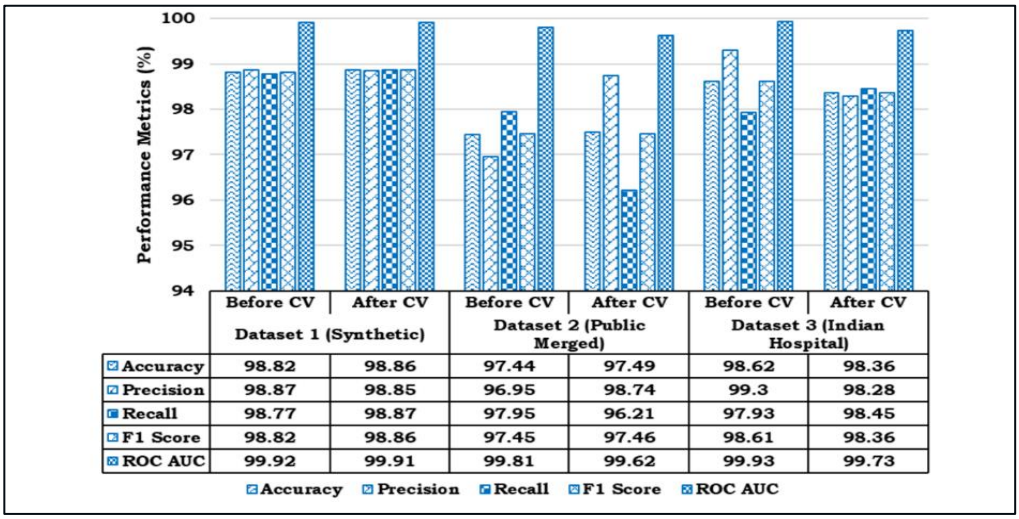
Figure 9. Dataset3[31] Results

This dataset appears to be extracted from more specific lab-based indicators including resting BP, cholesterol, fasting blood sugar, max heart rate, exercise angina and the number of major vessels.

Performance

- **Confusion Matrix:** Excellent test data accuracy with 4 misclassifications out of 290 (144+142 correct).
- **AUC Score:** 0.99
- **MLP Loss Curve:** Consistent decline over ~400 epochs suggesting a learning process that appears steady with just enough complexity for basic learning.

- **Log Loss:** Once again both LightGBM and XGBoost models performed well with controlled learning behavior.
- **Feature Importance:** In Random Forest the top predictors were Exercise Angina, Major Vessels and Slope. In Logistic Regression the highest predictor Exercise Angina was given high positive weight.
- **Correlation Heatmap:** Shows very strong associations between the predictor variables of age, resting BP, cholesterol and the heart disease target.



	Dataset 1 (Synthetic)		Dataset 2 (Public Merged)		Dataset 3 (Indian Hospital)	
	Before CV	After CV	Before CV	After CV	Before CV	After CV
Accuracy	98.82	98.86	97.44	97.49	98.62	98.36
Precision	98.87	98.85	96.95	98.74	99.3	98.28
Recall	98.77	98.87	97.95	96.21	97.93	98.45
F1 Score	98.82	98.86	97.45	97.46	98.61	98.36
ROC AUC	99.92	99.91	99.81	99.62	99.93	99.73

Figure 10. Comparison Model Performance

Figure 10: Comparison of model performance across datasets. The bar chart illustrates the comparison of the proposed ensemble model across three datasets—Dataset 1 (Synthetic), Dataset 2 (Public Merged), and Dataset 3 (Indian Hospital)—on a before and after cross-validation (CV) basis. The performance comparisons were carried out using five standard measures: Accuracy, Precision, Recall, F1 Score, and ROC AUC.

- Dataset 1 (Synthetic): Overall, the model achieved the best performance across all measures before and after CV. Assessments of Accuracy, Precision, Recall, and F1 Score remained stable around the value of 98.8%. The ROC AUC was ranked at 99.92% before CV and 99.91% after cross-validation, which indicates an outstanding level of class separation. Questions can be raised about the minor changes in predictions pre and post CV; the relative certainty in the generalizations may be attributed to the large sample size creating noise and very capable features.
- Dataset 2 (Public Merged): The overall performance was quite strong but slightly lower than Dataset 1. The notable change in Precision (a strong modifier for false positives) increased from 96.95% to 98.74% after cross-validation. Recall did decrease somewhat from 97.95% to 96.21% after CV, but did not change the F1 Score value to a notable amount, which reflects the balanced classification ability. The ROC AUC for this dataset was very strong, at 99.81% before CV and 99.62% after CV, again demonstrating distinctly separate values.
- Dataset 3 (Indian Hospital): This real dataset performed comparably to the synthetic dataset and was very good. The accuracy, F1 Score, and Recall were all

approximately 98.36% after CV, and Recall improved to 98.45% after CV. ROC AUC remained very high at 99.73% after CV, indicating excellent sensitivity and specificity. There were slight drops in Precision, indicating a small increase in false positives. However, these drops did not detract from the overall effectiveness of the model.

In comparison, the proposed SMOTE-boosted stacking ensemble for this research consistently achieved superior performance across the three datasets, with approximately 98.86% accuracy and 99.92% AUC on the large synthetic dataset; approximately 98.74% accuracy and 99.62% AUC on the merged public dataset; and approximately 98.45% accuracy with 99.73% AUC on the real Indian hospital dataset. This superior performance outperforms most of the single and ensemble models and demonstrates the advantages of hybrid feature selections, systematic SMOTE balancing, and a stacking ensemble of diverse classifiers for more robust and generalizable cardiovascular risk prediction.

Model Predictions

- Dataset 1 (Synthetic) - Evaluation: All models performed extremely well, with Accuracy, Precision, Recall, and F1 Score all near or in excess of 98.6%. The suggested model gained the highest performance in ROC AUC with a value of 99.91% and was able to produce balanced output across the metrics. LGB and LR also performed very well, with very similar outputs to the proposed model, but still performed slightly better in ROC AUC, while RF, XGB, and MLP performed the same or slightly worse than these two models.

- Dataset 2 (Public Merged) - Exploratory Analysis: Performance was quite mixed among the models. The LR model was underwhelming, with an Accuracy of 70.59%, low Precision at 67.86%, and an F1 Score of 72.55%. The other models, RF, XGB, LGBM, and MLP, all performed well, generally achieving accurate values around 96-97% for Accuracy, high Recall, and ROC AUC at or close to 99%. The proposed model had the best balance for Accuracy, Precision, Recall, F1 Score, and ROC AUC: Accuracy (97.49%), Precision (98.74%), Recall (96.21%), F1 Score (97.46%), and ROC AUC (99.62%).

- All models performed well, with very high Precision, Recall, and ROC AUC. XGB and LGBM scored perfect or nearly perfect Precision (100%, 99.3%) and ROC AUC (99.9%). The proposed model performed slightly worse on Precision (99.3%) and Recall (97.93%) but still performed similarly on Accuracy (98.62%) and ROC AUC (99.93%).
  In all datasets, the proposed model consistently exhibits high Accuracy, Precision, Recall, and exceeds or closely achieves the highest ROC AUC. In synthetic and hospital datasets, the majority of models perform well; however, on the real, merged public dataset, the proposed model is clearly superior to classic models like LR. The proposed model excels in its combination of feature selection and addressing class imbalance, delivering better and more consistent performance across many datasets. XGB and LGBM algorithms exhibit tremendously high scores across the hospital dataset with the use of tree-based ensemble learning. They seem to perform equally well across synthetic and public datasets; however, some overall balanced scores were marginally lower than the proposed model. The LR algorithm worked well across synthetic and hospital datasets but performed

poorly against public merged data. The RF and MLP algorithms showed decent performance, but none generally performed better than the proposed model, XGB, or LGBM.

5. Conclusion

This study developed and validated a SMOTE-boosted stacking ensemble approach to predicting cardiovascular disease, showing strong and consistent results on synthetic data, public datasets, and hospital datasets. The model was able to address the issues of data imbalance and complex feature interactions through comprehensive data preprocessing, a hybridized feature selection technique, and balanced sampling. The achieved ROC AUC values (~99.9%) and accuracy levels confirmed that the model is robust and generalizable to a wide spectrum of data situations. Critically, the use of interpretable feature importance maintains a clinical context, indicating that this pipeline offers a strong opportunity for clinical prediction to help support clinical decision making. Overall, this is a clear example of how the methods in this work combining ensemble learning and systematic data treatment are advantageous to the early prediction of cardiovascular disease.

References

[1] Ahmad, Bilal, Jinfu Chen, and Haibao Chen. "Feature selection strategies for optimized heart disease diagnosis using ML and DL models." arXiv preprint arXiv:2503.16577 (2025).

[2] Lübeck, Frederike, Jonas Wildberger, Frederik Träuble, Maximilian Mordig, Sergios Gatidis, Andreas Krause, and Bernhard Schölkopf. "Adaptable Cardiovascular Disease Risk Prediction from Heterogeneous Data using Large Language Models." arXiv preprint arXiv:2505.24655 (2025).

[3] Liu, Minyu, Yuxiong Pan, Ziyong Wang, Jvhong Wang, Yibao Shi, and Jun Chu. "The role of social determinants in alcohol consumption and cardiovascular health: the pathways study." Nutrition, Metabolism and Cardiovascular Diseases 35, no. 5 (2025): 103783.

[4] Balada, Christoph, Aida Romano-Martinez, Vincent ten Cate, Katharina Geschke, Jonas Tesarz, Paul Claßen, Alexander K. Schuster et al. "Deep Learning for Cardiovascular Risk Assessment: Proxy Features from Carotid Sonography as Predictors of Arterial Damage." In Annual Conference on Medical Image Understanding and Analysis, pp. 251-265. Cham: Springer Nature Switzerland, 2025.

[5] Liu, Tianyi, Andrew Krentz, Lei Lu, and Vasa Curcin. "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis." European Heart Journal-Digital Health 6, no. 1 (2025): 7-22.

[6] Sianga, Bernada E., Maurice C. Mbago, and Amina S. Msengwa. "Predicting the prevalence of cardiovascular diseases using machine learning algorithms." Intelligence-Based Medicine 11 (2025): 100199.

[7] Saikumar, K., P. S. Ravindra, M. D. Sravanthi, Abolfazl Mehbodniya, J. L. Webber, and Ali Bostani. "Heart disease prediction using machine learning and deep learning approaches: a systematic survey." Heart Dis 35, no. 2s (2025): 2398.

[8] Alkayyali, Z. K., S. Anuar Bin Idris, and Samy S. Abu-Naser. "A systematic literature review of deep and machine learning algorithms in cardiovascular diseases diagnosis." Journal of Theoretical and Applied Information Technology 101, no. 4 (2023): 1353-1365.

[9] Dritsas, Elias, Sotiris Alexiou, and Konstantinos Moustakas. "Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques." ICT4AWE 1 (2022): 315-321.

[10] Trigka, Maria, and Elias Dritsas. "Improving Cardiovascular Disease Prediction With Deep Learning and Correlation-Aware SMOTE." IEEE Access (2025).

[11] Rattan, Vikas, Ruchi Mittal, Jaiteg Singh, and Varun Malik. "Analyzing the application of SMOTE on machine learning classifiers." In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, 2021, 692-695.

[12] de Miguel-Diez, Javier, Julio Nunez Villota, Salud Santos Perez, Nicolas Manito Lorite, Bernardino Alcazar Navarrete, Juan Francisco Delgado Jimenez, Juan Jose Soler-Cataluna, Domingo Pascual Figal, Patricia Sobradillo Ecenarro, and Juan Jose Gomez Doblas. "Multidisciplinary management of patients with chronic obstructive pulmonary disease and cardiovascular disease." Archivos de Bronconeumología 60, no. 4 (2024): 226-237.

[13] Sharma, Narendra Kumar, Alok Singh Chauhan, Shahnaz Fatima, and Swati Saxena. "Enhancing heart disease diagnosis: Leveraging classification and ensemble machine learning techniques in healthcare decision-making." Journal of Integrated Science and Technology 13, no. 1 (2025): 1016-1016.

[14] auya, Jannatul, Saad Sahriar, Sanjida Akther, Ruhul Amin, Sabba Ruhi, and Md Shamim Reza. "Missing risk factor prediction in cardiovascular disease using a blended dataset and optimizing classification with a stacking algorithm." Engineering Reports 7, no. 1 (2025): e13034.

[15] Yang, Jian, and Jinhan Guan. "A heart disease prediction model based on feature optimization and smote-Xgboost algorithm." Information 13, no. 10 (2022): 475.

[16] de la Brassinne Bonardeaux, Orianne, Manon Deneye, Cecile Oury, Marie Moonen, and Patrizio Lancellotti. "High-Sensitivity CRP and Occurrence of Cancer in Cardiovascular Disease Patients with Cardiovascular." Journal of Clinical Medicine 14, no. 4 (2025): 1193.

[17] Talaat, Fatma M. "Revolutionizing cardiovascular health: integrating deep learning techniques for predictive analysis of personal key indicators in heart disease." Neural Computing and Applications 37.1 (2025): 1-24.

[18] Cao, Xiyu, Jianli Ma, Xiaoyi He, Yufei Liu, Yang Yang, Yaqi Wang, and Chuantao Zhang. "Unlocking the link: predicting cardiovascular disease risk with a focus on airflow

obstruction using machine learning." BMC Medical Informatics and Decision Making 25, no. 1 (2025): 50.

[19] Tian, Jing, et al. "Association between estimated glucose disposal rate and prediction of cardiovascular disease risk among individuals with cardiovascular-kidney-metabolic syndrome stage 0–3: a nationwide prospective cohort study." Diabetology & Metabolic Syndrome 17.1 (2025): 58.

[20] Bai, Tiantian, et al. "Exploration and comparison of the effectiveness of swarm intelligence algorithm in early identification of cardiovascular disease." Scientific Reports 15.1 (2025): 4647.

[21] Ganie, Shahid Mohammad, Pijush Kanti Dutta Pramanik, and Zhongming Zhao. "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets." Scientific reports 15.1 (2025): 13912.

[22] Mittal, Pooja, et al. "Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease." International Journal of Computational Intelligence Systems 18.1 (2025): 1-20.

[23] Xia, Biao, et al. "Intelligent cardiovascular disease diagnosis using deep learning enhanced neural network with ant colony optimization." Scientific Reports 14.1 (2024): 21777.

[24] Dorraki, Mohsen, et al. "Improving cardiovascular disease prediction with machine learning using mental health data: a prospective UK Biobank study." JACC: Advances 3.9_Part_2 (2024): 101180.

[25] Zheng, Dongze, et al. "The association of triglyceride-glucose index and combined obesity indicators with chest pain and risk of cardiovascular disease in American population with pre-diabetes or diabetes." Frontiers in Endocrinology 15 (2024): 1471535.

[26] Asadi, Fariba, et al. "Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms." Scientific Reports 14.1 (2024): 22230.

[27] World Health Organization (WHO). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)].

[28] Centers for Disease Control and Prevention (CDC). [https://www.cdc.gov/heartdisease/facts.htm].

[29] Dataset1:https://www.kaggle.com/datasets/mahatiratusher/heart-disease-risk-prediction-dataset

[30] Dataset2: https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset/data

[31] Dataset3: https://data.mendeley.com/datasets/dzz48mvjht/1.