Journal of Current Science and Technology, January-March 2026 Vol. 16 No. 1, Article 166
Copyright ©2018-2026, Rangsit University ISSN 2630-0656 (Online)

Cite this article: Tuion, T., Chumpong, K., & Samart, K. (2026). An ensemble machine learning strategy for accurate and timely prediction of heart disease. *Journal of Current Science and Technology*, *16*(1), Article 166. https://doi.org/10.59796/jcst.V16N1.2026.166

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

An Ensemble Machine Learning Strategy for Accurate and Timely Prediction of Heart Disease

Tassaneeporn Tuion[1, 3], Kittisak Chumpong[1, 2, 3], and Klairung Samart[1, 2, 3, *]

[1]College of Digital Science, Prince of Songkla University, Songkhla 90110, Thailand
[2]Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand
[3]Research Center in Mathematics and Statistics with Applications, Prince of Songkla University, Songkhla 90110, Thailand

*Corresponding author; E-mail: klairung.s@psu.ac.th

Received 3 August 2025; Revised 20 September 2025; Accepted 6 October 2025; Published online 20 December 2025

Abstract

Heart disease is a complicated disorder that is becoming increasingly common. Finding the best treatment plan for any heart disease patient requires early detection. The main goal of this study is to improve the prediction of heart disease by implementing ensemble learning techniques with machine learning (ML) models. The dataset used in this study, comprising 319,755 records from the Centers for Disease Control and Prevention, was downloaded from the Kaggle website. The SMOTE-ENN hybrid approach was used to address class imbalance. To increase the consistency and distribution of numerical variables, data preprocessing entailed standardization and the creation of dummy variables. The machine learning (ML) models Random Forest, Logistic Regression, Support Vector Machine, and Extra Trees were applied to the processed dataset without and with bagging, boosting, and stacking ensemble methods. Stacking, which combined SVM and Logistic Regression, outperformed the baseline models and other ensemble techniques, returning the highest recall score of 0.837731. This study underlines the significance of data balancing and ensemble learning for accurate forecasts based on medical datasets, which are typically large. The findings highlight how ML can enhance early diagnosis and intervention in the treatment of cardiac disease.

Keywords: *heart disease; classification; machine learning; ensemble; bagging; boosting, stacking*

1. Introduction

Heart disease affects a large number of people worldwide. The early and effective detection of heart disease is essential for successful medical care. The complex nature of heart disease demands careful treatment. Failure to identify the correct treatment plan may result in more severe cardiac problems or early death (Li et al., 2020; Mohan et al., 2019). According to the World Health Organization, cardiovascular diseases (CVDs) killed an estimated 19.8 million people in 2022, representing approximately 32% of all global deaths (WHO, 2025). The United States in particular is facing significant increases in the prevalence and cost of CVDs, necessitating effective strategies for management and prevention (Heidenreich et al., 2011).

Assessing the risk of heart disease in its early stages requires analysis of either lifestyle or behavioral patterns, such as stress management, exercise routines, and smoking habits (Franklin et al., 2021; Van Trier et al., 2022). Additionally, common symptoms of CVD, such as chest pain, shortness of breath, fatigue, palpitations, and peripheral symptoms, are crucial indicators in the early detection and management of the disease. These symptoms directly impact quality of life and often prompt patients to seek timely medical intervention. Raising awareness of lifestyle behaviors and disease symptoms, and addressing risk factor management, are essential strategies in reducing the global burden of CVD (Jurgens et al., 2022).

Nevertheless, the present diagnostic techniques for CVD frequently fail to detect the disease early enough. Problems arise during the drawn-out procedures and the intricate, costly study of patient information. Conventional methods, which rely on medical history, physical examination, and the evaluation of symptoms by medical professionals, frequently fail to correctly diagnose CVD. There is an urgent need for more sophisticated diagnostic tools that can get around these restrictions and improve the early diagnosis of CVD in order to improve patient outcomes (Restrepo Tique et al., 2024; Baghdadi et al., 2023).

In the medical field, machine learning (ML) plays a crucial role in analyzing medical data and reducing the workload of healthcare professionals. ML enables the rapid and accurate analysis of large datasets to identify patients at risk of heart disease using predictive models developed from medical information. ML is highly effective at predicting risks in heart disease patients (Westcott & Tcheng, 2019; Badawy et al., 2023; Alowais et al., 2023; Das et al., 2024).

The remainder of this paper is organized as follows. Section 1.1 explains the literature review of the most recent studies that have been suggested in this area. Section 2 outlines the objective of this study to improve model performance. Section 3 describes the research methodology, including the datasets used, the measures used to preprocess the data, the methods used to balance the data, and the classification models used. Section 4 presents the research findings, during which we analyze the performances of the baseline models and their efficiency improvements due to the ensemble techniques applied. Performance was evaluated by focusing on the key metrics of accuracy, precision, recall, and F1-score. Finally, in section 5, we present our conclusions and make suggestions for additional research.

1.1 Literature Review

To highlight the significance of the proposed research, this section reviews some existing ML approaches to medical diagnostics.

Das and Sengur (2010) employed ensemble methods, including bagging, boosting, and random subspace techniques, to improve valvular heart disease detection. Using a dataset of 215 samples, they demonstrated the efficacy of the methods, achieving high levels of sensitivity and specificity. For heart disease classification, Fida et al. (2011) proposed homogeneous ensemble methods optimized by a genetic algorithm. The method outperformed standalone classifiers, achieving notable improvements in accuracy, sensitivity, and specificity. Latha and Jeeva (2019) used ensemble classification techniques to increase the accuracy of their heart disease prediction. In line with the results of previous research, they found that bagging and boosting helped improve the accuracy of heart disease prediction. Using bagging, boosting, and stacking on a CVD dataset, Shorewala (2021) also achieved improvements over basic prediction techniques. Specifically, the results showed that the boosting model achieved the highest AUC (0.73), while the stacking model had the best accuracy at 75.1%. The study confirmed that ensemble techniques can effectively improve the accuracy of heart disease prediction. Rajendra and Latifi (2021) applied logistic regression along with feature selection and ensemble techniques to predict the onset of diabetes. Their study demonstrated that Max Voting and stacking increased accuracy to as high as 93%, and that enhanced data preprocessing improved the performance of the predictive models. In Gao et al. (2021), ensemble techniques were used to improve the accuracy of heart disease prediction by selecting relevant features from the dataset. Two extraction approaches were used: linear discriminant analysis (LDA) and principal component analysis (PCA). The results indicated that bagging with the decision tree model performed the best.

Chaurasia & Pal (2021) developed a stacking-based ensemble technique and feature selection method to enhance the accuracy of breast cancer diagnosis. They utilized four base models: SVM, K-Nearest Neighbors, Naive Bayes, and perceptrons, and combined their results using the stacking technique with logistic regression as the meta-model. Their findings demonstrated the significant potential of stacking to improve breast cancer diagnostic systems. Pal and Gangwar (2023) tested the performances of models created using algorithms. Specifically, ET feature selection processes were used to pick classification algorithms and appropriate attributes. While bagging, SVM, Multilayer Perceptron, and Gradient Boosting increased the accuracy of heart disease prediction, the bagging method performed better in terms of accuracy. Using a well-defined search strategy, Mahajan et al. (2023) reviewed and analyzed the use of bagging, boosting, stacking and voting, for five highly researched diseases: diabetes, skin disease, kidney disease, liver disease and heart disease. They found that stacking performed particularly well for predicting skin disease and diabetes, while bagging demonstrated strong performances for kidney and liver disease predictions.

Furthermore, Gabriel (2024) focused on developing a web application for heart disease prediction using a dataset from the CDC. Five machine learning models-XGBoost, RF, SVM, LR, and LightGBM-were trained. The findings demonstrated that the XGBoost model outperformed the others, particularly in predicting heart disease patients, with an F1-score of 34%, recall of 81%, and precision of 21%. The model's precision, recall, and F1-score for patients without cardiac disease were 97%, 72%, and 83%, respectively.

Nissa et al. (2024) investigated boosting ensemble techniques, focusing on CatBoost, RF, Gradient Boosting, AdaBoost, and LightGBM. Their results underscored the effectiveness of AdaBoost, which showcased its superiority over the other algorithms by achieving a predictive accuracy of 95% for CVD. Bhagat et al. (2025) explored stacking-based ensemble methods, employing several classifiers. Their results demonstrated that stacking was particularly effective, achieving a maximum accuracy of 98.53%, making it a robust approach for early heart attack prediction. Sultan et al. (2025) proposed a stacking learning model named NCDG for heart disease prediction using a dataset from the CDC. The model employed Naive Bayes, Categorical Boosting, and Decision Tree as base learners and Gradient Boosting as the meta-learner. To address the issue of data imbalance, the study applied SMOTE and Borderline-SMOTE techniques. In addition, the SHAP method was utilized to explain the model's decision-making process. The experimental results showed that the NCDG model achieved the best performance, and its reliability was validated using K-Fold Cross-Validation.

Despite the depth of the existing research, applications of ensemble learning techniques for early heart disease detection are lacking. Primarily, these techniques have been studied to improve the accuracy of predicting heart disease and related conditions. Given the complexity of patient data frequently encountered in the medical profession today, data preparation is typically overlooked in current research, which frequently concentrates on general datasets or traditional models. The potential of ensemble methods to diagnose and predict early-stage cardiac diseases, especially when addressing unbalanced datasets, remains underexplored. This study proposes a novel approach that integrates ensemble techniques with data balancing methods to enhance model development for early cardiac disease prediction. By improving diagnostic precision, expanding treatment options and reducing associated risks, this approach aims to facilitate early detection and timely intervention, leading to significantly better patient outcomes.

2. Objectives

This research aims to:

1. To compare the predictive performance of heart disease classification using different machine learning models, including RF, LR, SVM, and ET.

2. To improve the predictive performance of heart disease classification using ensemble techniques.

3. Materials and Methods

The proposed methodology for heart disease prediction comprises five major stages, as illustrated in Figure 1. The dataset obtained from the Centers for Disease Control and Prevention (CDC) was preprocessed through data cleaning, duplicate removal, cluster sampling, transformation, and standardization to ensure data quality and consistency. Second, SMOTE-ENN was utilized to resolve the problem of class imbalance by generating synthetic minority instances and discarding misclassified or noisy samples. Third, several machine learning (ML) models, including RF, LR, SVM, and ET were trained as base classifiers. Fourth, ensemble techniques namely bagging, boosting and stacking were implemented to enhance predictive performance. Finally, the models were evaluated using accuracy, precision, recall, and F1-score to determine their effectiveness in the early detection of heart disease. The entire process was implemented using Python version 3.11.5 on a Jupyter Notebook utilizing basic libraries from Scikit-Learn.

3.1 Dataset

The Centers for Disease Control and Prevention (CDC) provided the heart disease datasets used in this study.

They are accessible on the Kaggle website at https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. Each of the 319,755 records in the original dataset includes one dependent characteristic and 17 independent variables that indicate whether cardiac disease is present. The attribute of each variable is described in Table 1.

Table 1 Attributes of the dataset

Attribute	Score Information
Heart Disease	0= No, 1 = Yes
BMI (kg/m²)	Min = 12.02, Max = 94.85
Smoking	0= No, 1 = Yes
Alcohol Drinking	0= No, 1 = Yes
Stroke	0= No, 1 = Yes
Physical Health (day)	Min = 0, Max = 30
Mental Health (day)	Min = 0, Max = 30
Difficulty Walking	0= No, 1 = Yes
Sex	0= Female, 1 = Male
Age Category	1 = 18–29, 2 = 30–34, 3 = 40–49, 4 = 50–59, 5 = 60–69, 6 = 70–79, 7 = 80 or older
Race	0 = American Indian, 1 = Asian, 2 = Black, 3 = Hispanic, 4 = Other, 5 = White
Diabetic	0 = No, 1 = No (borderline diabetes), 2 = Yes, 3 = Yes (during pregnancy)
Physical Activity	0= No, 1 = Yes
General Health	0 = Excellent, 1 = Fair, 2 = Good, 3 = Poor, 4 = Very good
Sleep Time (hr)	Min = 1, Max = 24
Asthma	0= No, 1 = Yes
Kidney Disease	0= No, 1 = Yes
Skin Cancer	0= No, 1 = Yes

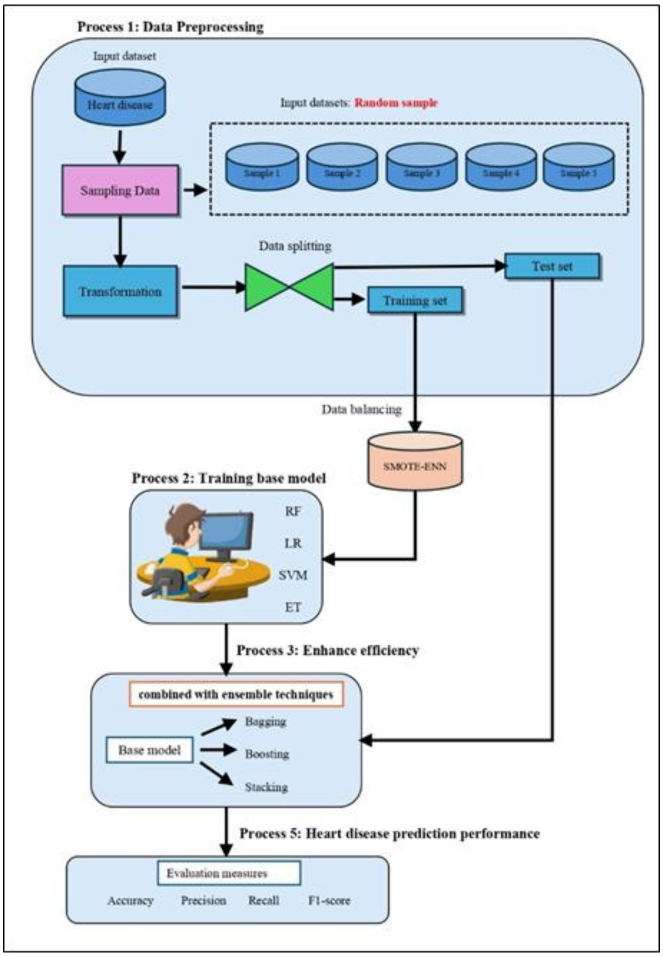

Figure 1 Heart disease prediction workflow

3.2 Data Preprocessing

The data preprocessing process comprised four steps: data cleaning, cluster sampling, data transformation and splitting, and data balancing. The overall methodology is illustrated in Figure 1.

3.2.1 Data Cleaning

Data cleaning involved checking for missing values and duplicating data. No missing values were found. However, we identified 18,078 duplicate records, which were removed, resulting in a cleaned dataset of 301,717 records.

3.2.2 Cluster Sampling

The cleaned dataset of 301,717 records included 27,261 cases of heart disease (class 1) and 274,456 cases of no heart disease (class 0). This distribution was considered a notable class imbalance, which was rectified by using a straightforward cluster sampling to downsample the majority class while keeping all occurrences of the minority class in order. To simplify data management and balance the classes for the model training, and to mitigate bias (Matloff, 2017), we partitioned class 0 into five groups of 54,891 records, ensuring that the groups of class 0 instances were evenly distributed and did not dominate the learning process. Since there were no significant differences between the five new groups, we randomly selected one of them to make up the training dataset, which consisted of 27,261 records of heart disease and 54,891 records of no heart disease. This procedure produced a balanced dataset of 82,152 records.

3.2.3 Data Transformation and Splitting

Data transformation comprised two steps. In the first step, categorical variables of age, race, diabetic status and general health were converted into dummy variables using one-hot encoding, producing 22 new variables. The transformation converted categorical data into numeric form (0 or 1). As a result, the number of independent variables increased to 30.

Standardizing the data into a suitable format is crucial when dealing with numeric variables such as BMI, physical health, mental health and sleep time. Standardization plays a vital role in data preparation, ensuring that all variables contribute equally and minimizing skewness in the data distribution. It was accomplished here by applying a standard normal distribution using Eq. (1),

$$Z=\frac{x-\mu}{\sigma}, \tag{1}$$

where

x is the value of each data point in a variable,
μ is the mean of the variable x,
σ is the standard deviation of the variable x.

This operation gave the dataset a mean of 0 and a variance of 1.

After standardizing the numeric variables, we performed an outlier detection. We found that the attributes BMI (2,326 records) and sleep time (1,511 records) contained a total of 3,768 overlapping outliers. These outlier records were removed, resulting in a dataset of 78,384 records for analysis. The primary objective of these preparatory procedures was to minimize bias in both numerical and categorical features. Specifically, categorical variables were transformed into dummy variables using one-hot encoding, thereby eliminating artificial ordering and ensuring independent contributions from each category. For numerical variables, standardization was applied to place features on a comparable scale, reducing skewness and preventing dominance by variables with larger ranges. In addition, outlier detection and removal were performed for variables such as sleep duration and BMI to mitigate the influence of extreme values on the learning process. Collectively, these preprocessing techniques reduced potential sources of bias and produced a more balanced and representative dataset, providing a robust foundation for developing machine learning models that are both reliable and generalizable. This was the dataset used in the subsequent stages of our study. The 78,384 records were randomly split into a training set and a testing at a 70:30 ratio. The training set, used to build and train the ML model, comprised 54,868 records. The testing set, used to evaluate the performance of the trained model, comprised 23,516 records that had not been used during the learning process.

3.2.4 Data Balancing Techniques

Class imbalance significantly impacts the performance of ML models by introducing bias toward the majority class, which contains more samples. Since most of the training data comes from the majority class, the model tends to return more accurate predictions from the majority class while underperforming on the minority class (Ketu & Mishra, 2022; Chapakiya et al., 2025). To mitigate the effects of class imbalance, which often causes ML models to favor the majority class, we addressed this issue by evaluating six data balancing techniques: SMOTE-Tomek, SMOTE-ENN, ADASYN, SMOTE, TomekLinks, and Random Under Sampling (RUS). Based on previous experimental results (Tui-on et al., 2024), the SMOTE-ENN technique demonstrated the highest overall performance across various evaluation metrics. Therefore, SMOTE-ENN was selected as the preferred method for handling class imbalance in the training dataset in this study. SMOTE generates synthetic samples for the minority class, while ENN removes noisy or misclassified instances from both classes. This combination enhances class balance while reducing noise in the dataset (Muntasir Nishat et al., 2022). After applying the SMOTE-ENN method, the final training dataset comprised 42,192 records, as shown in Table 2.

Table 2 Class distribution before and after data balancing with SMOTE-ENN

	Class 0	Class 1	Total
Imbalanced Data	36,837	18,031	54,868
Balanced Data	20,221	21,971	42,192

3.3 Classification Models

Predictive models for heart disease datasets are built from classification models. The classification models analyzed in this research included the following:

3.3.1 Random Forest

RF is an ensemble algorithm that combines multiple decision trees to improve classification performance. It creates a random forest of decision trees by randomly selecting and repeating subsets of the training data multiple times. The outcomes of each decision tree are then aggregated using majority voting for classification (Latha & Jeeva, 2019).

3.3.2 Logistic Regression

LR is an ML method within the category of supervised learning. It is used to predict outcomes represented as categorical variables based on independent variables. The output of LR ranges between 0 and 1, making the method suitable for predicting probabilities between two classes. Since the method is computationally efficient and highly interpretable, it is particularly well-suited to forecasting medical events and for real-world healthcare applications (Ahmad, 2021; Dissanayake & Md Johar, 2021; Hasanin & Khoshgoftaar, 2018).

3.3.3 Support Vector Machine

Support vectors are the closed data points to the hyperplane that help maximize the margin between classes and can affect the orientation and position of the hyperplane. The dimensions of the hyperplane are determined by the amount of input features. SVM works well for classifying high-dimensional data. The aim of SVM is to find a hyperplane with the largest margin that separates data points into two groups (Schölkopf et al., 1996). Because SVM can handle large, high-dimensional datasets, it performs well in disease prediction, especially diabetes, cancer, and heart disease. Although its restricted interpretability can be problematic, its computational economy and resilience to overfitting make it a dependable option (Soman et al., 2009).

3.3.4 Extra Trees

ET is an ensemble technique that emphasizes randomness in the tree building process (Ahmad, 2021). The primary distinction between RF and ET lies in how they choose the root node for decision making. Whereas ET randomly selects k features, RF chooses the optimal feature. ET shows less correlation and greater variability between features as a result of the randomization in feature selection (Sharaff & Gupta, 2019). Another difference is that RF uses bootstrap replicas to generate subsets of size N for training ensemble members (decision trees), whereas ET uses the whole original sample.

3.3.5 Ensemble Techniques

Instead of using a single model to generate predictions, ensemble learning generates predictions by combining multiple classification models (Latha & Jeeva, 2019). This approach typically improves performance by utilizing the following three techniques.

Bagging

Bagging classification enhances and improves the performance of models by combining multiple instances of the same model, such as decision trees. The process begins with bootstrap sampling, where subsets of the training set are randomly selected. Each subset is used to create individual models, and the final prediction is determined using a majority vote method (Breiman, 1996).

Boosting

Boosting classification is an ensemble technique designed to improve the performance of weak learners. New models are created by taking into account the errors of the previous models. It is an iterative process that depends on prior predictions, adjusting the weights of misclassified samples in subsequent rounds (Ghosh et al., 2021). The AdaBoost algorithm was applied in this research.

Stacking

Stacking is a technique that enhances predictive performance by combining two or more models through a meta-classifier that is suitable for classification or regression tasks (Natarajan et al., 2024). The outputs from base classifiers are aggregated to form the input for the meta-classifier. In this work, RF, LR, SVM, and ET were the base-level classifiers and LR was the designated meta-classifier.

The parameters of the four base models, RF, LR, SVM, and ET, were set as shown in Table 3. These parameters were mostly drawn from the Scikit-learn library's default settings, with small tweaks made to ensure consistency and reproducibility. Such configurations were purposefully chosen to give a transparent and reliable baseline for assessing the proposed methodology.

Similarly, the ensemble methods bagging, boosting, and stacking were also configured based on default settings with minor adjustments, as shown in Table 4.

3.4 Performance Metrics

On the testing set, model performances were evaluated based on the classifications obtained from the confusion matrix in Table 5. The results were used to calculate the performance metrics accuracy, precision, recall and F1-score for the final evaluation of model performances. Patients defined as having heart disease were indicated by a positive prediction, whilst those classed as not having heart disease were indicated by a negative prediction.

Table 3 Parameter settings for bagging, boosting and stacking

Model	function	Parameters
RF	**from** sklearn.ensemble **import** RandomForestClassifier	Random Forest Classifier (n_estimators = 73, max_depth = 12, min_samples_split = 2, random_state = 0, bootstrap = False)
LR	**from** sklearn. linear_model **import** LogisticRegression	LogisticRegression (random_state=0)
SVM	**from** sklearn.svm **import** SVC	SVC (probability = True, kernel = 'linear')
ET	**from** sklearn.ensemble **import** ExtraTreesClassifier	Extra Trees Classifier (n_estimators = 100, random_state = 0)

Table 4 Parameter settings for bagging, boosting and stacking

Technique	Ensemble function	Parameters
bagging	**from** sklearn.ensemble **import** BaggingClassifier	BaggingClassifier (estimator = *model*, n_estimators = 10, max_samples = 0.8, max_features = 1.0, random_state = 42, bootstrap = True)
boosting	**from** sklearn.ensemble **import** AdaBoostClassifier	AdaBoostClassifier(estimator = *model*, n_estimators = 50, learning_rate = 1.0, random_state = 42)

stacking	**from** sklearn.ensemble **import** StackingClassifier	*.1set estimators*
		Model = [1st base classifier]
		.2create meta-model
		StackingClassifier(estimators = Model, final_estimator = LogisticRegression())

Table 5 Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive)TP(	False Negative)FN(
Actual Negative	False Positive)FP(	True Negative)TN(

The terms used in the confusion matrix define the following classifications

TP: The model correctly predicted that the patient had heart disease.

TN: The model correctly predicted that the patient did not have heart disease.

FP: The model incorrectly predicted that the patient had heart disease, when the patient did not have heart disease.

FN: The model incorrectly predicted that the patient did not have heart disease, when the patient did have heart disease.

Accuracy, precision, recall and F1-score were calculated as follows (Sokolova & Lapalme, 2009).

Accuracy: This is a measure of the overall performance of a model in predicting outcomes. It represents the ratio of the number of correct predictions to the total number of samples and was calculated as:

$$\text{Accuracy}= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Precision: This is a measure of model accuracy, specifically its ability to correctly identify samples predicted as positive. It focuses on minimizing the number of False Positives, and was calculated as:

$$\text{Precision}= \frac{\text{TP}}{\text{TP+FP}}$$

Recall: This is a measure of the model's ability to correctly identify all Positive samples, and was calculated as:

$$\text{Recall}= \frac{\text{TP}}{\text{TP+FN}}$$

F1-score: This is a measure of balance between precision and recall, calculated using their harmonic mean. The F1-score is an important metric for evaluating precision and recall simultaneously, and was calculated as:

$$\text{F1-score}= 2\,\frac{(\text{Precision x Recall})}{(\text{Precision+Recall})}$$

4. Results and Discussion

The performances of the RF, LR, SVM, and ET classification models on the heart disease dataset were compared based on accuracy, precision, recall and F1-score. The results before and after balancing the data with SMOTE-ENN are shown in Table 6.

In medical diagnostics, a high recall is crucial as it minimizes FNs, when patients with heart disease are misclassified as disease-free. Failing to diagnose a case of heart disease can lead to delayed treatment, which increases the risk of complications and death (Pope et al., 2000; Behnke, 2022). All recall values were considerably higher after the data were balanced. After balancing the data, the LR model achieved the highest recall value of 0.833377, along with accuracy, precision and F1-score values of 0.726909, 0.550453 and 0.662993, respectively.

The enhancement of model performance due to bagging, boosting and stacking ensemble techniques were compared. The results in Table 7 show that bagging SVM achieved the highest recall value of 0.835092, along with accuracy, precision, and F1-score values of 0.728738, 0.552404 and 0.664951, respectively. The next best performance was achieved by bagging LR, which had a recall value of 0.833641, along with accuracy, precision, and F1-score values of 0.727547, 0.55115 and 0.663586, respectively.

Table 6 Performances of base model classification

Model	**Before balancing with SMOTE-ENN**			
	accuracy	**precision**	**recall**	**F1-score**
RF	0.772708	0.689825	0.535752	0.603103
LR	0.773643	0.683049	0.555541	0.612731
SVM	0.773303	0.689853	0.539050	0.605199
ET	0.733968	0.591614	**0.563984**	0.577469
Model	**After balancing with SMOTE-ENN**			
	accuracy	**precision**	**recall**	**F1-score**
RF	0.737916	0.604843	0.810775	0.692830
LR	0.726909	0.550453	**0.833377**	0.662993
SVM	0.728695	0.552548	0.832322	0.664175
ET	0.727377	0.554489	0.784697	0.649806

Table 7 Performances of base model classification with bagging

Bagging model	Performance			
	accuracy	precision	recall	F1-score
Bagging RF	0.733926	0.560405	0.809631	0.662349
Bagging LR	0.727547	0.551155	0.833641	0.663586
Bagging SVM	0.728738	0.552404	**0.835092**	0.664951
Bagging ET	0.728015	0.553955	0.801847	0.655239

Table 8 Performances of base model classification with boosting

Boosting model	Performance			
	accuracy	precision	recall	F1-score
Boosting RF	0.740219	0.571015	0.780211	0.659419
Boosting LR	0.726314	0.551887	**0.802639**	0.654053
Boosting SVM	0.724528	0.552129	0.769921	0.643085
Boosting ET	0.725251	0.552158	0.781398	0.647075

Table 9 Performance of base model classification with stacking

Stacking model	Performance			
	accuracy	precision	recall	F1-score
RF+LR	0.733075	0.559471	0.808575	0.661343
RF+SVM	0.733245	0.559697	0.808311	0.661413
RF+ET	0.728865	0.556558	0.781530	0.650132
LR+SVM	0.726697	0.549926	**0.837731**	0.663983
LR+ ET	0.727717	0.555284	0.779815	0.648669
SVM+ ET	0.727845	0.555419	0.780079	0.648853
RF+LR+ SVM	0.733075	0.559460	0.808707	0.661379
RF+LR+ ET	0.728653	0.556424	0.779947	0.649492
LR+ SVM+ ET	0.727802	0.555399	0.779683	0.648702
SVM+ ET+RF	0.728016	0.556559	0.780211	0.649676
RF+LR+ SVM+ ET	0.728780	0.556623	0.779419	0.649445

Models were processed through the boosting framework using the AdaBoost algorithm. The results shown in Table 8 indicate that boosting LR returned the highest recall value of 0.802639, with accuracy, precision, and F1-score values of 0.726314, 0.551887 and 0.654053, respectively. The second best performance was achieved by boosting ET, which gave a recall value of 0.781398, along with accuracy, precision, and F1-score values of 0.725251, 0.552158 and 0.647075, respectively.

Two or more models were stacked, with all four base classifiers being utilized. The meta-model used for final prediction was LR. The results are presented in Table 9. Stacking LR+SVM returned the highest recall value of 0.837731, along with accuracy, precision, and F1-score values of 0.726697, 0.549926, and 0.663983, respectively. Stacking RF+LR+SVM produced the next best scores, with a recall value of 0.808707, and accuracy, precision and F1-score values of 0.733075, 0.559460 and 0.661379 respectively.

Table 10 presents a comparison of recall scores returned by the four base models on imbalanced data and balanced data, as well as when applied with bagging, boosting and stacking. Recall values were significantly better on the balanced data than the imbalanced data. The recall score of LR increased from 0.555541 to 0.833377 the highest recall among the four base models.

The recall results returned after bagging indicated that bagging enhanced recall for certain models, notably SVM and ET. However, the recall values returned by RF and LR exhibited only minor changes after bagging. When boosting was applied, all the models returned lower recall scores compared to the base model. The recall score of SVM decreased from 0.832322 to 0.769921 while the recall score of LR decreased from 0.833377 to 0.802639. The results suggested that boosting does not improve the performance of the four models used, possibly due to the asymmetrical influence of misclassified samples, which affects the boosting mechanism’s learning process. When stacking was applied, the LR+SVM model returned the highest recall value of 0.837731, surpassing the recall values of all the other techniques. This result demonstrated that stacking was the most effective method of enhancing the vital recall metric, especially when integrating the LR and SVM models. Consequently, stacking emerged as the most efficient approach to the problem of resolving class imbalance issues.

To further validate the proposed LR+SVM model's performance, a classification report was created to evaluate precision, recall, and F1-score across both classes. As shown in Table 11, the model produced balanced results with notably high recall for the positive (heart disease) class, suggesting its great capacity to reduce false negatives, which is especially important for early disease identification.

The classification report in Table 11 shows the proposed LR+SVM stacking model's performance in both classes. In particular, the model had a recall of 0.84 for the positive class (heart disease), which is crucial for reducing false negatives and guaranteeing quick diagnosis. Furthermore, the macro and weighted averages validate the model's robustness, producing an overall accuracy of 0.73, indicating that it performs consistently even under imbalanced data conditions.

Table 10 Comparison of recall between imbalanced data, balanced data and ensemble techniques

Imbalanced data	**Recall**
RF	0.535752
LR	0.555541
SVM	0.539050
ET	**0.563984**
Balanced data	**Recall**
RF	0.810775
LR	**0.833377**
SVM	0.832322
ET	0.784697
Bagging	**Recall**
Bagging RF	0.809631
Bagging LR	0.833641
Bagging SVM	**0.835092**
Bagging ET	0.801847
Boosting	**Recall**
Boosting RF	0.780211
Boosting LR	**0.802639**
Boosting SVM	0.769921
Boosting ET	0.781398
Stacking	**Recall**
RF+LR	0.808575
RF+SVM	0.808311
RF+LR+ SVM	0.808707
LR+SVM	**0.837731**

Table 11 Classification report for the LR+SVM model

Model Accuracy	**Classes**	**precision**	**recall**	**F1-score**	**support**
	No Disease (0)	0.90	0.67	0.77	15,936
	Heart Disease (1)	0.55	0.84	0.66	7580
accuracy				0.73	23,516
Macro average		0.72	0.76	0.72	23,516
Weighted average		0.79	0.73	0.74	23,516
Micro average		0.73	0.73	0.73	23,516

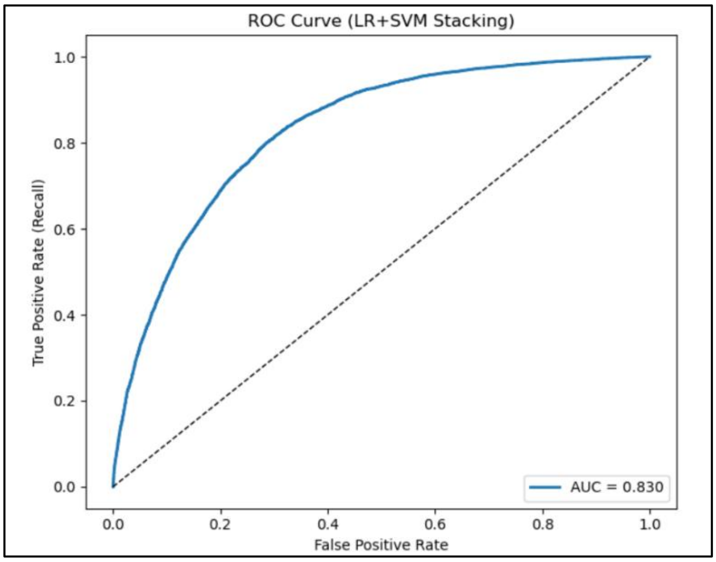

Figure 2 AUC–ROC of the LR+SVM stacking model

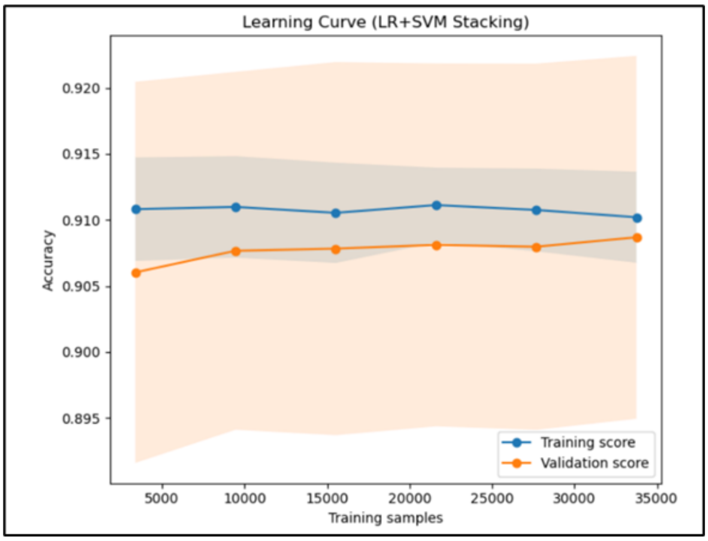

Figure 3 Learning curve of the LR+SVM stacking model

Furthermore, the ROC curve shown in Figure 2 demonstrates the model's balance between sensitivity and specificity, with an AUC of 0.830. This demonstrates a high level of discrimination between individuals with and without cardiac disease. These findings corroborate the LR+SVM stacking framework's clinically relevant prediction accuracy and highlight its critical role in early identification and medical decision-making.

The learning curve in Figure 3 shows that the LR+SVM stacking model performs consistently as the number of training samples increase. The close alignment of training and validation accuracy, with validation converging to training accuracy indicates low overfitting or underfitting. The suggested methodology for large-scale clinical datasets is resilient, as seen by its consistent validation accuracy of around 0.91 across varied sample sizes. Importantly, this robustness can be attributed to the stringent preprocessing pipeline and the careful application of SMOTE-ENN exclusively within the training folds, which together mitigate class imbalance and prevent data leakage that might otherwise bias the evaluation.

In contrast, the study by Sultan et al. (2025) may have been prone to data leakage, as indicated by their uniformly high-performance measures (accuracy, precision, recall, and F1-score all reported at 0.91), which may not reflect true predictive performance. By comparison, our technique demonstrates more realistic and dependable improvements, thereby strengthening the credibility of early cardiac disease diagnosis in real-world settings. Furthermore, these findings highlight our methodology's practical potential for clinical translation, particularly in supporting medical decision-making via clinical decision support tools.

5. Conclusion

This study applied the machine learning models Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Extra Trees (ET) to the CDC heart disease dataset, available on the Kaggle platform. Before training the models, we applied the SMOTE-ENN technique to address class imbalance in the dataset. The implementation of bagging, boosting, and stacking ensemble learning methods significantly improved the prediction of heart disease, particularly in terms of recall, which is crucial for accurately identifying individuals with heart disease. Stacking LR and SVM with LR as the meta-learner returned the highest recall value of 0.837731, outperforming the baseline models and the bagging and boosting methods. Stacking the LR and SVM models produced a significant improvement in recall on balanced data, returning the highest recall score. This study highlighted the effectiveness of ensemble techniques, especially stacking, in enhancing recall performance, and the significance of balancing data in large medical datasets.

The study offers a promising approach to improved heart disease prediction, which can result in better patient management and treatment outcomes by enabling early detection. The approach can identify high-risk patients earlier, even before they show obvious symptoms, by picking up on fewer signs and behavior. By correcting class imbalance and emphasizing recall, the system helps reduce missed diagnoses among high-risk patients. Future research can investigate other enhancements, such as feature selection integration and model optimization, to strengthen and increase the predictive power of machine learning models used in diagnosis of heart disease.

6. Abbreviations

Abbreviation	Full Term
CVD	Cardiovascular diseases
CDC	The Centers for Disease Control and Prevention
ML	Machine learning
RF	Random Forest
LR	Logistic Regression
SVM	Support Vector Machine
ET	Extra Trees
SMOTE-ENN	Synthetic Minority Over-sampling Technique and Edited Nearest Neighbours
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

7. CRediT Statement

Tassaneeporn Tuion: methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, visualization
Kittisak Chumpong: Conceptualization, methodology, validation, investigation, writing review and editing, supervision
Klairung Samart: Conceptualization, methodology, validation, investigation, writing review and editing, supervision, project administration, funding acquisition

8. Acknowledgements

We extend our sincere thanks to the Faculty of Science, Prince of Songkla University for providing the facilities. We also wish to acknowledge Thomas Duncan Coyne for his invaluable assistance in refining the English language of this manuscript. Any remaining errors are solely the responsibility of the authors. This research was supported by the National Science, Research and Innovation Fund (NSRF) and Prince of Songkla University (Grant No. SCI6801200S).

9. References

Ahmad, M. N. (2021). Comprehensive analysis of heart disease prediction using Scikit-Learn. *International Research Journal of Modernization in Engineering Technology and Science*, *3*(3), 67–80.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), Article 689. https://doi.org/10.1186/s12909-023-04698-z

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. *Journal of Electrical Systems and*

Information Technology, *10*(1), Article 40. https://doi.org/10.1186/s43067-023-00108-y

Baghdadi, N. A., Farghaly Abdelaliem, S. M., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, *10*(1), Article 144. https://doi.org/10.1186/s40537-023-00817-1

Behnke, L. M. (2022). The danger of underdiagnosing coronary microvascular disease in women. *Journal of the American Association of Nurse Practitioners*, *34*(5), 780-783. https://doi.org/10.1097/JXX.0000000000000703

Bhagat, M., Sharma, A., & Agarwal, P. (2025). An efficient stacking-based ensemble technique for early heart attack prediction. *Multimedia Tools and Applications*, *84*(30), 36351-36375. https://doi.org/10.1007/s11042-024-19293-7

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/10.1023/A:1018054314350

Chapakiya, I., Traisuwan, A., Chumpong, S., & Chumpong, K. (2025). Follow-up period classification of type 2 diabetes patients using data mining techniques. *Journal of Health Science and Medical Research*, *43*(2), Article 20241083. https://doi.org/10.31584/jhsmr.20241083

Chaurasia, V., & Pal, S. (2021). Stacking-based ensemble framework and feature selection technique for the detection of breast cancer. *SN Computer Science*, *2*(2), Article 67. https://doi.org/10.1007/s42979-021-00465-3

Das, R., & Sengur, A. (2010). Evaluation of ensemble methods for diagnosing of valvular heart disease. *Expert Systems with Applications*, *37*(7), 5110-5115. https://doi.org/10.1016/j.eswa.2009.12.085

Das, S., Nayak, S. P., Sahoo, B., & Nayak, S. C. (2024). Machine learning in healthcare analytics: A state-of-the-art review. *Archives of Computational Methods in Engineering*, *31*(7), 3923-3962. https://doi.org/10.1007/s11831-024-10098-3

Dissanayake, K., & Md Johar, M. G. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Applied Computational Intelligence and Soft Computing*, *2021*(1), Article 5581806. https://doi.org/10.1155/2021/5581806

Fida, B., Nazir, M., Naveed, N., & Akram, S. (2011). *Heart disease classification ensemble optimization using genetic algorithm* [Conference presentation]. 2011 IEEE 14th International Multitopic Conferenc, IEEE, Karachi, Pakistan. https://doi.org/10.1109/INMIC.2011.6151471

Franklin, B. A., Rusia, A., Haskin-Popp, C., & Tawney, A. (2021). Chronic stress, exercise and cardiovascular disease: Placing the benefits and risks of physical activity into perspective. *International Journal of Environmental Research and Public Health*, *18*(18), Article 9922. https://doi.org/10.3390/ijerph18189922

Gabriel, J. (2024). A machine learning-based web application for heart disease prediction. *Intelligent Control and Automation*, *15*(1), 9-27. https://doi.org/10.4236/ica.2024.151002

Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*, *2021*(1), Article 6663455. https://doi.org/10.1155/2021/6663455

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, *9*, 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

Hasanin, T., & Khoshgoftaar, T. (2018). *The effects of random undersampling with simulated class imbalance for big data* [Conference presentation]. 2018 IEEE international conference on information reuse and integration (IRI), IEEE, Salt Lake City, UT, USA. https://doi.org/10.1109/IRI.2018.00018

Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., ... & Woo, Y. J. (2011). Forecasting the future of cardiovascular disease in the United States: A policy statement from the American Heart Association. *Circulation*, *123*(8), 933-944. https://doi.org/10.1161/CIR.0b013e31820a55f5

Jurgens, C. Y., Lee, C. S., Aycock, D. M., Masterson Creber, R., Denfeld, Q. E., DeVon, H. A., ... & American heart association council on cardiovascular and stroke nursing; council on hypertension; and stroke council. (2022). State of the science: the relevance of symptoms in cardiovascular disease and research: a scientific statement from the American Heart Association. *Circulation*, *146*(12), e173-e184. https://doi.org/10.1161/CIR.0000000000001089

Ketu, S., & Mishra, P. K. (2022). Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection. *Arabian Journal for Science and Engineering*, *47*(2), 1447-1469. https://doi.org/10.1007/s13369-021-05972-2

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, Article 100203. https://doi.org/10.1016/j.imu.2019.100203

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using ML classification in E-Healthcare. *IEEE Access*, *8*, 107562–107582. https://doi.org/10.1109/ACCESS.2020.3001149

Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023, June). Ensemble learning for disease prediction: A review. *Healthcare*, 11(12), Article 1808. https://doi.org/10.3390/healthcare11121808

Matloff, N. (2017). *Statistical regression and classification: From linear models to ML.* CRC Press.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., ... & Khan, M. R. H. (2022). A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming*, *2022*(1), Article 3649406. https://doi.org/10.1155/2022/3649406

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), Article 174. https://doi.org/10.1007/s44196-024-00538-0

Nissa, N., Jamwal, S., & Neshat, M. (2024). A technical comparative heart disease prediction framework using boosting ensemble techniques. *Computation*, *12*(1), Article 15. https://doi.org/10.3390/computation12010015

Pal, G. K., & Gangwar, S. (2023). Discovery of approaches by various machine learning ensemble model and features selection method in critical heart disease diagnosis. *International Research Journal on Advanced Science Hub*, *5*(1), 15-21. https://doi.org/10.47392/irjash.2023.003

Pope, J. H., Aufderheide, T. P., Ruthazer, R., Woolard, R. H., Feldman, J. A., Beshansky, J. R., ... & Selker, H. P. (2000). Missed diagnoses of acute cardiac ischemia in the emergency department. *New England Journal of Medicine*, *342*(16), 1163-1170. https://doi.org/10.1056/NEJM200004203421603

Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, Article 100032. https://doi.org/10.1016/j.cmpbup.2021.100032

Restrepo Tique, M., Araque, O., & Sanchez-Echeverri, L. A. (2024). Technological advances in the diagnosis of cardiovascular disease: A public health strategy. *International Journal of Environmental Research and Public Health*, *21*(8), Article 1083. https://doi.org/10.3390/ijerph21081083

Schölkopf, B., Burges, C., & Vapnik, V. (1996). *Incorporating invariances in support vector learning machines* [Conference presentation]. International conference on artificial neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-61510-5_12

Sharaff, A., & Gupta, H. (2019). Extra-tree classifier with metaheuristics approach for email classification. *Advances in computer communication and computational sciences: proceedings of IC4S 2018*. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-6861-5_17

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, *26*, Article

100655. https://doi.org/10.1016/j.imu.2021.100655

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Soman, K. P., Loganathan, R., & Ajay, V. (2009). *ML with SVM and other kernel methods*. Delhi, India: PHI Learning.

Sultan, S. Q., Javaid, N., Alrajeh, N., & Aslam, M. (2025). Machine learning-based stacking ensemble model for prediction of heart disease with explainable ai and k-fold cross-validation: A symmetric approach. *Symmetry*, *17*(2), Article 185. https://doi.org/10.3390/sym17020185

Tui-On, T., Chumpong, K., & Samart, K. (2024). *Comparison of imbalanced data techniques in predicting heart disease* [Conference presentation]. Proceedings of the 7th International Conference on Applied Statistics 2024 (ICAS2024) & The 14th National Conference on Applied Statistics and Information Technology, Chiang Mai, Thailand.

Van Trier, T. J., Mohammadnia, N., Snaterse, M., Peters, R. J. G., Jørstad, H. T., & Bax, W. A. (2022). Lifestyle management to prevent atherosclerotic cardiovascular disease: Evidence and challenges. *Netherlands Heart Journal*, *30*(1), 3-14. https://doi.org/10.1007/s12471-021-01642-y

Westcott, R. J., & Tcheng, J. E. (2019). Artificial intelligence and machine learning in cardiology. *Cardiovascular Interventions*, *12*(14), 1312-1314. https://doi.org/10.1016/j.jcin.2019.03.026

World Health Organization. (2025). *Cardiovascular diseases (CVDs)*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)