Developing a Novel Framework using Optimized Active Stacking and Explainable AI for Heart Disease Prediction

Aymin Javed[a], Nadeem Javaid[a,*,1], Abdul Khader Jilani Saudagar[b,*] and Dragan Pamucar[c,d]

[a]ComSens Lab, International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Douliu, Yunlin 64002, Taiwan

[b]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[c] Department of Applied Mathematical Science, College of Science and Technology, Korea University, Sejong 30019, Republic of Korea

[d] Department of Industrial Engineering & Management, Yuan Ze University, Taoyuan 320315, Taiwan

ARTICLE INFO

Keywords:
Heart Disease
Machine Learning
Stacking Model
Entropy-based Active Learning
Bayesian Optimization
Paired t-test
SHapley Additive exPlanations
Local Interpretable Model-agnostic Explanations
10-Fold Cross Validation

ABSTRACT

Background and Objective: Heart disease is still the top driver of death worldwide, and developing accurate, interpretable, and efficient predictive systems is essential to enable early diagnosis in time for effective intervention. Although significant efforts have been made, the existing machine learning approaches are subject to problems including class imbalance, high-dimensional input features, complex hyperparameter tuning problems, low classification accuracy, and a limited amount of labeled data available. This article intends to overcome these challenges with a componental framework that is capable of robust heart disease prediction.

Methods: The proposed framework is composed of three components. First, the proximity-weighted random affine shadow sampling technique is applied to mitigate class imbalance by generating synthetic samples for the minority class. Second, principal component analysis reduces feature dimensionality while preserving essential information. Third, three novel models are developed: (i) a stacking model that combines k-nearest neighbors and naïve Bayes as base learners with logistic regression as the meta-learner; (ii) an Optimized Stacking Model with Bayesian Optimization (OSM-BO) for systematic hyperparameter tuning; and (iii) an Entropy-based Active Learning Optimized Stacking Model (EAL-OSM), which uses entropy-based sampling to select the most informative samples for annotation. A paired t-test and 10-fold cross validation are applied for statistical evaluation. Local interpretable model-agnostic explanations and Shapley additive explanations are employed to ensure interpretability and support decision transparency.

Findings: The stacking model improves accuracy by 3.66%, precision by 6.33%, recall by 3.57%, and Precision–Recall Area Under the Curve (PR-AUC) by 6.90%, while reducing Hamming loss by 12.50%. OSM-BO yields further improvements: 7.32% in accuracy, 7.59% in precision, 11.90% in recall, and 9.20% in PR-AUC, with a 31.25% reduction in Hamming loss. EAL-OSM achieves the best results with gains of 9.76% in accuracy, 11.39% in precision, 13.10% in recall, 11.49% in PR-AUC, and Hamming loss decreases by 37.50%.

Conclusions: The proposed componental framework exhibits promising gain in classification performance, statistical robustness, and explainability, thus providing a clinically practical solution to predict heart disease.

1. Introduction

Heart disease, another name for Cardiovascular Disease (CVD), is a collection of disorders that affect the way the heart functions and is structured. It is one of the leading causes of illness and mortality globally, responsible for almost 26 million deaths per year [1], according to the World Health Organization (WHO). The most prevalent form of the illness is known as coronary artery disease, which happens when plaque accumulates in the coronary arteries, narrowing or obstructing them from supplying blood to the heart. This process is referred to as atherosclerosis [2].

This condition can cause angina (chest pain), myocardial infarction (heart attack), and, in extreme situations, sudden cardiac death [3]. Heart failure, which results from the heart's incapacity to pump blood efficiently and builds up fluid in the lungs and other body tissues, is another important component of heart disease [4]. Numerous modifiable and non-modifiable risk factors contribute to the development of heart disease. Modifiable factors include hypertension, diabetes, hyperlipidemia, smoking, poor diet, and excessive alcohol consumption. Family history, age, and gender are considered non-modifiable factors [5]. WHO reported 31% of deaths are caused globally due to heart disease. Heart disease continues to pose a serious threat to public health despite advancements in medical science and therapy. The threat of heart disease is particularly less in undeveloped countries as compared to developed countries. Moreover, if substantial action is not taken, it is predicted that the number of deaths from heart disease will rise to 22 million by 2030 [6]. People who are suffering from CVD usually undergo depression, high cholesterol, and high blood pressure. To control heart disease, we need a high-accuracy system that can predict it in its early stages [7].

The prior studies used different techniques and heart disease datasets for heart disease prediction [8]. But those

*Corresponding Author: (javaidn@yuntech.edu.tw, www.njavaid.com), (aksaudagar@imamu.edu.sa)

ORCID(s):

[1] https://orcid.org/0000-0003-3777-8249

Table 1
Abbreviations and Symbols

Notations	Description
AI	Artificial Intelligence
CVD	Cardiovascular Disease
DL	Deep Learning
EAL-OSM	Entropy-based Active Learning Optimized Stacking Model
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
NB	Naïve Bayes
OSM-BO	Optimized Stacking Model with Bayesian Optimization
PCA	Principal Component Analysis
ProWRAS	Proximity Weighted Random Affine Shadow Sampling
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling TEchnique
b	Translation vector
$d(x_i, x_j)$	Euclidean distance between x_i and x_j
$\hat{f}(\theta)$	Surrogate model approximation of $f(\theta)$
$f(\theta)$	Unknown objective function in Bayesian Optimization
k	Number of nearest neighbors
n	Total number of features or instances
x_i	Minority class instance
x_j	Majority class instance
$\hat{y}$	Final binary class prediction

techniques are not efficient enough for early disease predictions. Deep Learning (DL) and Machine Learning (ML) play an important part in healthcare, smart grids, and nonlinear dynamics. Artificial Intelligence (AI) is improving healthcare services' accuracy, efficiency, and accessibility by combining ML algorithms and DL models. AI helps the health sector in providing immediate patient care, security, reduced healthcare costs, and better treatment. Active learning is a unique technique also used in the detection of heart disease. Labeling data can be costly and time-consuming in many domains, particularly those that require expert knowledge [9]. By lowering the quantity of labels required to train a high-performance model, active learning is beneficial. Choosing the most informative data points through active learning allows for quicker model improvement than randomly labeling data.

In this study, we have used active learning with ML techniques to improve the models' performance with limited labeled data for training. Active learning has many advantages, as it reduces model overfitting and improves model generalizability. Active learning trains the model on minimum labeled data, and from unlabeled data it checks the uncertainty score. If the uncertainty score matches the desired condition, it is given a label, removed from the unlabeled set, and added to a training set. The model is trained until all the unlabeled data is removed. The list of abbreviations and symbols used in this study is mentioned in Table 1.

Developing ML and DL models for predicting heart disease has gained significant attention recently. These models employ diverse datasets and methodologies to address feature complexity and data imbalance challenges. Their goal is to improve the accuracy of diagnosis and early intervention efforts. To address the challenges and improve the precision of early heart disease prediction, Alam, Afroj, and Mohd Muqeem propose a novel Recurrent Neural Network based on Chaos Game Optimization (CGO-RNN) [2]. The Kernel Principal Component Analysis (KPCA) methodology reduces the computational load and complexity of the recommended methods. The features are then extracted and the cardiac samples are categorized for precise and timely prediction.

Manikandan, G., *et al.* in [5] examine the performance of Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) techniques with and without Boruta feature selection. The authors makes use of the Cleveland Clinic Heart Disease dataset from Kaggle, which contains 303 cases and 14 features. The Boruta feature selection algorithm, which chooses six important characteristics, enhanced the algorithm's performance. LR outperformed other classification methods, with an accuracy rate of 88.52%.

Ogunpola, Adedayo, *et al.* in [10] used ML techniques to detect cardiac illnesses early, including myocardial infarction. It addresses the issue of imbalanced datasets. Seven ML and DL classifiers, including eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), SVM, LR, Convolutional Neural Network (CNN), Random Forest (RF), and Gradient Boosting (GB), are used to improve the accuracy of heart disease predictions. The study investigates several classifiers and their performance, providing useful information for constructing strong prediction models for myocardial infarction. The results indicate that XGBoost beats the other models with 99.14% accuracy.

Hasib, Khan Md *et al.* in [11] design a Hybrid Sampling-based DL Model (HSDLM) that can handle effectively the issue of imbalanced datasets. First, to encode the categorical features label encoder is utilized and then to remove noise from them the auhors used Edited Nearest-Neighbors (ENN) under sampling method for better data quality. The Synthetic Minority Over-sampling Technique (SMOTE) technique is used to handle the issue of unbalanced class, synthetic samples for minority classes is generated. Then, three Long Short-Term Memory (LSTM) networks with parallel connections are adopted to learn temporal dependencies and high order feature representations from the balanced data. Results of empirical studies showed that the

Table 2
Summary of Related Work and Proposed Solutions

Method	Limitations	Proposed Solution
CGO-RNN [2]	High computational complexity and may require large datasets for optimal performance.	Use PCA for feature extraction.
SVM, LR, and DT with Boruta feature selection [5]	Performance may degrade with high-dimensional data or overfitting in small datasets.	Employs ProWRAS for class imbalance handling.
XGBoost [10]	Can be prone to overfitting and may require fine-tuning of hyperparameters for optimal results.	Introduced stacking model with KNN, NB, and LR as the meta-learner.
RF, KNN, NB for heart attack prediction [12]	May not generalize well for unseen data, leading to reduced performance for new cases.	Use of PCA for dimensionality reduction and ProWRAS for balancing the dataset.
CNN with BiLSTM [13]	High computational resource requirements and training time. Can be complex to implement.	OSM-BO to tune hyperparameters for better performance and generalization.
Bagged DT with augmented dataset [14]	May require substantial data preprocessing and may not perform well with imbalanced datasets.	ProWRAS is applied.
Multi-label Active Learning Selection (Random adaptive, QUIRE, MMC, AUDI) [15]	Requires iterative querying and may be sensitive to label noise in the dataset.	EAL-OSM to select the most informative data points efficiently.
Fuzzy Membership Functions for Active Learning [16]	May struggle with real-time applications or large-scale datasets. Limited by fuzzy function choice.	EAL-OSM to improve the selection of informative data instances using entropy-based sampling.
ALRFC for gene expression classification [17]	Struggles with very high-dimensional data and may not be efficient for large datasets.	Enhanced stacking model with KNN and NB classifiers to improve generalization on high-dimensional data.
Bayesian Deep-Active-Learning for RUL prediction [18]	High computational cost for Bayesian inference and may not work well with sparse or limited data.	Application of BO for hyperparameter tuning to reduce computational cost and improve accuracy.

proposed HSDLM framework outperforms other compared methods, and hence proved to be an effective robust solution for highly imbalanced datasets.

Stonier, Alexander., *et al.* in [12] propose a method for predicting if a patient is likely to have a heart attack by examining several data sources, including electronic health records and clinical diagnostic reports from hospital clinics. This paper provides a summary of various ML algorithms and tactics used to predict the likelihood of a heart attack. This research examines numerous methods for improving medical diagnosis, including RF, KNN, and Naive Bayes (NB). The RF algorithm is discovered to have a higher accuracy of 88.52% in forecasting heart attack risk.

Ahmad, Shakeel, *et al.* in [13] propose a DL-based system, CNN with Bidirectional Long Short Term Memory (LSTM), for accurately predicting cardiovascular illness from patient data. Only the most relevant characteristics are chosen via feature selection, which involves prioritizing and selecting aspects that are highly ranked in the provided illness dataset. Next, the CNN+BiLSTM hybrid DL approach is used to predict CVD. In contrast to previous research in predicting heart disease, the experimental results of the proposed hybrid DL approach show promising results with 94.507% accuracy.

Al-Ssulami, Abdulrakeeb M., *et al.* in [14] propose a novel approach for creating an augmented dataset by selecting replicating misclassified examples during the leave-one-out Cross Validation (CV) procedure in order to overfit a model. To assess different classifiers, authors employed a paired ML model in conjunction with an enhanced dataset. The starting point is the comprehensive cardiac disease dataset. The technique outperformed the basis dataset in terms of accuracy, with the bagged DT algorithm surpassing cutting-edge models and scoring 97.1% in the 10-Fold Cross Validation (10-FCV) test. The Cleveland dataset is also used for experiments with the same 10-FCV and achieves an accuracy score of 99.2%.

El-Hasnony, Ibrahim M., *et al.* in [15] propose five multi-label active learning selection algorithms—Random Adaptive, Querying Informative and Representative Examples (QUIRE), Maximum Margin Criterion (MMC), and Active Uncertainty and Diversity Integration (AUDI)—to iteratively select the most relevant data samples for querying, thereby reducing labeling costs. Predictive modeling is performed on the heart disease dataset using a label ranking classifier, with its parameters optimized through grid search. The experimental evaluation reports both accuracy and F1-scores, comparing results with and without hyperparameter

optimization.

de Oliveira, Heveraldo R., *et al.* in [16] propose a unique approach to data transformation using fuzzy membership functions as a preprocessing phase for active learning. The goal is to approximate comparable cases for the purpose of prediction, reducing the effort of human specialists in labeling data for training models. The study's findings show that this strategy is helpful in predicting heart disease and provide an argument for employing membership functions to improve ML models in medical data analysis.

Halder, Anindya, and Ansuman Kumar in [17] propose a unique active learning approach for cancer sample classification based on gene expression data called the Active Learning Rough-Fuzzy Classifier (ALRFC). The suggested approach can deal with the ambiguity, overlaps, and indiscernibility that are common in gene expression data subtypes. The proposed algorithm is tested using publicly available cancer datasets, and its kappa, precision, recall metrics, and prediction accuracy are compared to three other active learning methods: two traditional supervised learning methods and one semi-supervised classification method. The experimental results indicate that the suggested model outperformed the baseline models.

Zhu, Rong, *et al.* in [18] propose a novel active learning strategy along with a bayesian deep-active-learning framework for Remaining Useful Life (RUL) prediction that goes beyond standard passive learning. Using Monte Carlo dropout inference and Bayesian Neural Networks (BNN) for samples that don't have run-to-failure labels, the authors predict Remaining Useful Life (RUL) while also measuring the uncertainty of their predictions. The researchers create an acquisition function that actively selects target samples to generate run-to-failure labels based on the prediction uncertainty. Then to maintain accuracy while reducing the amount of training data, a recursive model training and active data selection procedure is developed.

Despite remarkable progress in applying ML and DL techniques for heart disease prediction [2, 5, 10, 12, 13], existing studies face persistent challenges that limit their generalization and clinical applicability. Many models rely on static datasets, such as the Cleveland and Kaggle repositories [5, 14], which restrict their adaptability to diverse population samples. Furthermore, several approaches exhibit overfitting tendencies due to inadequate hyperparameter tuning or the absence of effective optimization mechanisms [2, 13]. Although ensemble and hybrid models such as CNN-BiLSTM and bagged DT demonstrate improved accuracy [13, 14], they still lack active data selection and uncertainty handling, leading to inefficiencies in learning from limited labeled samples. Additionally, most ML models fail to effectively manage class imbalance and complex feature dependencies inherent in medical datasets [10, 12]. Recent works employing active learning frameworks [15, 16, 17, 18] have attempted to address labeling inefficiencies, but they are often limited to specific domains or fail to integrate with optimization-driven ensemble architectures. Therefore, there remains a critical need for a robust, optimization-based active learning framework that can jointly enhance feature learning, handle imbalance, and improve model generalization for reliable and early heart disease prediction. Beyond the enhancement of accuracy, the scope of this study is to provide an interpretable, resource-efficient and clinically applicable framework that responds to the increasing demand for explainable AI in healthcare. By jointly tackling data imbalance, sparse labeling, and hyperparameter optimization, this study closes the gap between algorithmic development and medical usability.

The rest of this paper is organized as follows: Section 2 details the proposed methodology; Section 3 presents the results and discussion; and Sections 4 and 5 concludes the study with future research directions. The summary of related work is given in Table 2.

1.1. Research Gaps

Despite the growing application of ML in heart disease prediction, several research gaps remain in the existing literature.

- A significant number of existing studies overlook the issue of class imbalance, which can cause biased learning toward the majority class and compromise the reliability of predictions, particularly for underrepresented cases.
- Previous work often neglects the problem of high-dimensional data, which can increase computational complexity and reduce generalization by introducing noise and redundancy in the feature space.
- Most existing models are limited in scope and do not leverage ensemble or layered architectures to improve predictive accuracy and robustness.
- Hyperparameter tuning is frequently performed manually or through trial-and-error, resulting in suboptimal model configurations and inconsistent performance across different datasets.
- The challenge of limited labeled data is insufficiently addressed in earlier studies, restricting model learning and reducing applicability in real-world, resource-constrained environments.
- Although some previous works have used CV techniques for performance assessment, they generally rely on only one validation split or a small number of folds, resulting in biased estimates of model generalization.
- Interpretability remains a major concern, as many models operate as black boxes, offering little insight into how predictions are made—an essential requirement in sensitive domains such as healthcare.

1.2. Contributions

The key contributions of this study are as follows:

- The data imbalance issue is effectively handled using the Proximity Weighted Random Affine Shadow Sampling (ProWRAS) technique, which synthetically balances minority classes without adding noise.
- Principal Component Analysis (PCA) is applied to reduce the feature space dimensionality while preserving important discriminatory information, thus improving computational efficiency.
- Three novel classification models are proposed:
  - A stacking model combining KNN and NB as base learners and LR as the meta-learner is introduced for heart disease prediction through ensemble learning.
  - The Optimized Stacking Model with Bayesian Optimization (OSM-BO) is proposed to automatically tune stacking model hyperparameters using Bayesian Optimization (BO), thereby enhancing the model's generalization and predictive capabilities.
  - The proposed Entropy-based Active Learning Optimized Stacking Model (EAL-OSM) is introduced to address the efficient use of a limited amount of labeled data by incorporating Entropy-based Active Learning (EAL) sampling for informative instance selection.
- Extensive end-to-end evaluation with 10-FCV and paired t-test verifies the statistical credibility and superiority of the proposed models.
- Model interpretability is enhanced using Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), enabling transparent and explainable decision-making suitable for clinical environments.

2. Componental Framework

In this study, we propose a complete modeling pipeline, referred to as the "framework" or "componental framework", which are used alternatively throughout the manuscript to describe the proposed system for improving the accuracy and interpretability of heart disease predictions. In the presence of data imbalance, the ProWRAS method is adopted, which could fairly balance the dataset without introducing artificial noise while maintaining original data distribution. In addition, PCA is used to decrease the dimensionality of the feature space and to reduce redundancy among features and computational complexity. Based on such preprocessing methods, three advanced models, named stacking, OSM-BO, and EAL-OSM, are then proposed to obtain robust and reliable predictions. The detailed workflow of the proposed framework is illustrated in Figure 1, while a summary of the adopted methodology is provided in Table 3. .

2.1. Heart Disease Health Indicator Dataset Description

The Behavioral Risk Factor Surveillance System (BRFSS) is a survey that gathers information from U.S. residents regarding their use of preventive services for chronic conditions and health-related behaviors. The BRFSS dataset, which is openly accessible on a public repository, provided the heart disease risk prediction information for the year 2021 [19]. The HDHI contains a total of 253,680 instances and 22 features. These features provide crucial information about how different conditions and behaviors may affect the chance of developing heart disease. The dataset is clean and does not require any preprocessing steps. The dataset is divided using a stratified sampling approach to ensure that the proportion of each class remained consistent with the original data distribution. A total of 10,000 instances are selected from the original dataset while maintaining the same class balance. This stratified sampling technique ensures that the sampled subset accurately reflects the characteristics of the complete dataset, allowing fair model training and evaluation. The list of features in the dataset is mentioned in Table 4.

2.2. Component 1: Data Balancing of HDHI Dataset using ProWRAS

Over-sampling is a technique that artificially enlarges the minority class by generating or duplicating samples, thereby reducing the imbalance ratio and enabling classification algorithms to perform more effectively on the dataset [20]. Many techniques, such as SMOTE [21] and Adaptive Synthetic Sampling (Adasyn), are used to handle class imbalance problems [11], but they have some limitations. SMOTE cannot deal with borderline instances, and does not distinguish between instances that are well inside the minority class region and those that are close to the decision boundary. Because all instances are treated equally, it may produce synthetic samples that are close to border areas. To overcome this limitation, we have used the ProWRAS balancing technique.

The purpose of ProWRAS is to create synthetic samples for the minority class while taking into account the complex data distribution, especially in the areas where the majority and minority classes have decision boundaries [22]. ProWRAS divides the minority class according to its distance from the majority class. The Euclidean distance between a minority class instance x_i and a majority class instance x_j is computed as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{1}$$

ProWRAS calculates the appropriate number of synthetic samples for each partitioned region after partitioning. The number of synthetic samples S_r allocated to region r can be expressed as:

$$S_r = \lfloor W_r \times N_s \rfloor \tag{2}$$

Table 3
Summary of the Research Methodology Followed in this Study

Method	Description	Key Features
ProWRAS	ProWRAS, a technique for handling class imbalance in datasets by generating synthetic samples based on proximity.	Balances dataset by considering decision boundaries, creating synthetic samples for minority class.
PCA	A statistical method for dimensionality reduction that transforms correlated variables into a smaller set of uncorrelated variables called Principal Components (PCs).	Reduces feature space while retaining maximum variance.
Stacking Model	A ML ensemble method that combines multiple base models using a meta-model to improve prediction accuracy.	Combines the strengths of different base classifiers KNN, NB using a meta-learner LR.
OSM-BO	An enhanced version of the stacking model, where hyperparameters are optimized using BO.	Automatically tunes hyperparameters to enhance model performance.
EAL-OSM	A model that combines EAL with the optimized stacking model to minimize labeled data usage by selecting the most informative data points.	Reduces labeling costs by querying the most uncertain data points for labeling.
10-FCV	A model validation technique where the dataset is split into 10 subsets, training and testing are performed 10 times with each subset used as the test set once.	Provides robust performance evaluation by averaging results over multiple splits.
LIME	A technique for explaining ML models' prediction by approximating them locally with interpretable models.	Helps in understanding the model's predictions by approximating it locally with simpler, interpretable models.
SHAP	A method for explaining the output of ML models by assigning Shapley values to each feature, quantifying its contribution to the prediction.	Provides a fair and consistent explanation of the importance of features.

where W_r is the normalized weight of region r, and N_s is the total number of synthetic samples to generate. It take more synthetic samples to properly balance the dataset in regions that are more similar to the majority class or have fewer instances of minority groups. Allocates weights to the minority class instances according to their partition and proximity to the majority class. The weight w_i for an instance x_i is inversely proportional to its average distance from the nearest majority class neighbors:

$$w_i = \frac{1}{\frac{1}{k}\sum_{j=1}^{k} d(x_i, x_j)} \tag{3}$$

Only those instances are selected from the minority class, which has higher weights. After selecting the instances, shadow sampling is applied, in which each instance shadow is created by transforming the original instance. A shadow sample $\tilde{x}_i$ is generated by applying an affine transformation:

$$\tilde{x}_i = Ax_i + b \tag{4}$$

where A is a random affine matrix and b is a translation vector. The new generated instance is not a duplicate of the original instance and is added to the feature space.
In healthcare applications characterized by data imbalance, such as heart disease prediction, the ProWRAS approach becomes particularly advantageous. ProWRAS helps the model learn important patterns from the less common cases by creating realistic and varied synthetic samples that are similar to the decision boundary. The prediction performance and sensitivity are thus improved. Thus, it promotes the early and accurate identification of heart disease situations, finally allowing for timely medical intervention.

2.3. Component 2: Feature Engineering with PCA

PCA is a statistical approach utilized for feature extraction and dimensionality reduction [23]. It converts a dataset of potentially correlated variables into PC, which are a collection of linearly uncorrelated variables. The main purpose of PCA is to keep as much information as possible while reducing the number of features in a dataset [24]. It extracts the most contributing feature and reduces high-dimension data into low dimensions. PCA involves several steps for dimensionality reduction. First, the covariance matrix $\mathbf{C}$ for dataset $\mathbf{X}$ with n observations and p features is computed using:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^{\top}\mathbf{X} \tag{5}$$

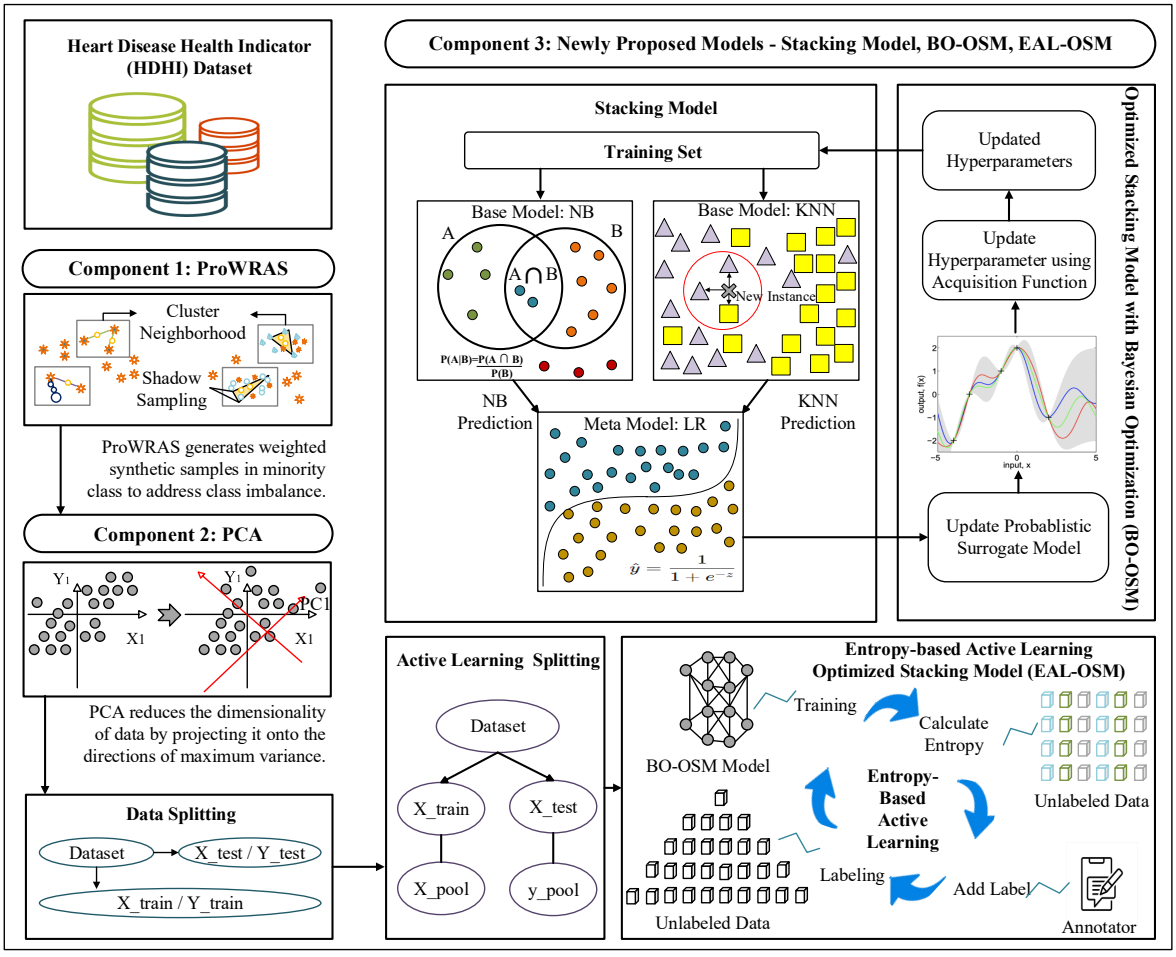


Figure 1: Componental Framework for Heart Disease Prediction

Next, PCA searches for eigenvectors that maximize the variance in the data. Each eigenvector is associated with an eigenvalue that represents the variance along that direction. The eigenvalue equation is:

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v} \tag{6}$$

where $\mathbf{v}$ is the eigenvector and λ is the corresponding eigenvalue. Next, we apply a decreasing order of the eigenvalues and corresponding eigenvectors. To construct a new matrix $\mathbf{W}$, the top k eigenvectors, corresponding to the highest eigenvalues, are selected.:

$$\mathbf{W} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k] \tag{7}$$

The original data is projected onto the new subspace using:

$$\mathbf{X}_{\text{new}} = \mathbf{X}\mathbf{W} \tag{8}$$

The final output $\mathbf{X}_{\text{new}}$ is the transformed dataset with reduced dimensions. PCA assists in removing unnecessary features by lowering the number of dimensions, which can enhance model performance. When the data is less, it takes less computational time [25].

2.4. Component 3: Newly proposed Models - Staking Model, BO-OSM, EAL-OSM

This section introduces the novel models proposed in this study. The models are designed to address the limitations of traditional classifiers through the integration of stacking, BO, and active learning. Each component contributes uniquely to improving the model's generalization ability and robustness. The subsequent subsections elaborate on the architecture and working mechanisms of each proposed model.

2.4.1. Stacking Model

Stacking, sometimes referred to as stacked generalization, is a potent ensemble learning method that minimizes generalization error by combining several classification models to enhance prediction performance. Unlike traditional ensemble methods such as bagging and boosting, stacking integrates predictions from multiple base-level models using a meta-model that learns how best to combine these predictions. The meta-model is trained to make the final prediction using the outputs of the base models, which are trained on the same input dataset. This layered architecture allows stacking model to harness the

Table 4
Dataset Features Description

Features	Description
HeartDiseaseorAttack	Target variable indicating the presence of heart disease (0 = No, 1 = Yes)
HighBP	Indicates if the individual has high blood pressure (0 = No, 1 = Yes)
HighChol	Indicates if the individual has high cholesterol (0 = No, 1 = Yes)
CholCheck	Indicates if the individual had their cholesterol checked (0 = No, 1 = Yes)
BMI	Body Mass Index (calculated from height and weight)
Smoker	Indicates if the individual is a smoker (0 = No, 1 = Yes)
Stroke	Indicates if the individual has had a stroke (0 = No, 1 = Yes)
Diabetes	Indicates if the individual has diabetes (0 = No, 1 = Yes)
PhysActivity	Indicates if the individual engages in physical activity (0 = No, 1 = Yes)
Fruits	Indicates if the individual consumes fruits (0 = No, 1 = Yes)
Veggies	Indicates if the individual consumes vegetables (0 = No, 1 = Yes)
HvyAlcoholConsump	Indicates if the individual engages in heavy alcohol consumption (0 = No, 1 = Yes)
AnyHealthcare	Indicates if the individual has access to any healthcare (0 = No, 1 = Yes)
NoDocbcCost	Indicates if the individual couldn't see a doctor due to cost (0 = No, 1 = Yes)
GenHlth	Self-reported general health status (1 = Excellent, 5 = Poor)
MentHlth	Mental health status (number of days the individual felt mentally unhealthy in the past 30 days)
PhysHlth	Physical health status (number of days the individual felt physically unhealthy in the past 30 days)
DiffWalk	Indicates if the individual has difficulty walking (0 = No, 1 = Yes)
Sex	Gender of the individual (0 = Male, 1 = Female)
Age	Age of the individual (in years)
Education	Highest level of education attained by the individual (1 = Less than high school, 5 = Graduate or professional education)
Income	Income level of the individual (1 = Less than $10,000, 8 = $75,000 or more)

strengths of different learning algorithms, potentially leading to improved predictive accuracy.

This study presents a stacking model utilizing KNN and NB as base-level classifiers, with LR functioning as the meta-classifier. The selection of these models is intentional and founded on their complementing attributes. KNN is a non-parametric, distance-based technique that effectively identifies local structures in data without supposing any underlying distribution [26]. It is especially useful for datasets with nonlinear decision boundaries. NB is a probabilistic model that presumes feature independence and can generate calibrated probability estimates rapidly, even with limited datasets [27]. Despite its simplicity, it performs well in many high-dimensional problems. LR is selected as the meta-classifier because of its simplicity, robustness, and ability to linearly separate the decision space based on the outputs of the base learners [28]. This is the optimal choice for integrating predictions from many models, as it quantifies the probability of the positive class and provides interpretable weights for each input features.

The working of stacking is shown in Figure 1. Let the original dataset be $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{0, 1\}$ is the binary class label. We define two base classifiers h_1 and h_2, where h_1 is KNN and h_2 is NB. These classifiers are trained on the dataset D, and for each instance $\mathbf{x}_i$, we compute their predictions as given in Algorithm 1:

$$h_1(\mathbf{x}_i), \quad h_2(\mathbf{x}_i) \tag{9}$$

A new dataset is then constructed for the meta-classifier, where each instance is:

$$\mathbf{z}_i = (h_1(\mathbf{x}_i), h_2(\mathbf{x}_i)) \tag{10}$$

The LR meta-model H is trained on $\{(\mathbf{z}_i, y_i)\}_{i=1}^{n}$, and computes the probability of the positive class using the *sigmoid* activation function:

$$H(\mathbf{z}) = \sigma(\mathbf{w}^\top \mathbf{z} + b) = \frac{1}{1 + e^{-(w_1 h_1(\mathbf{x}) + w_2 h_2(\mathbf{x}) + b)}} \tag{11}$$

where $\mathbf{w} = (w_1, w_2)$ are the weights learned by LR, and b is the bias term. For any new input instance $\mathbf{x}_{\text{new}}$, the final prediction $\hat{y}$ is given by:

$$\hat{y} = \begin{cases} 1, & \text{if } H(h_1(\mathbf{x}_{\text{new}}), h_2(\mathbf{x}_{\text{new}})) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

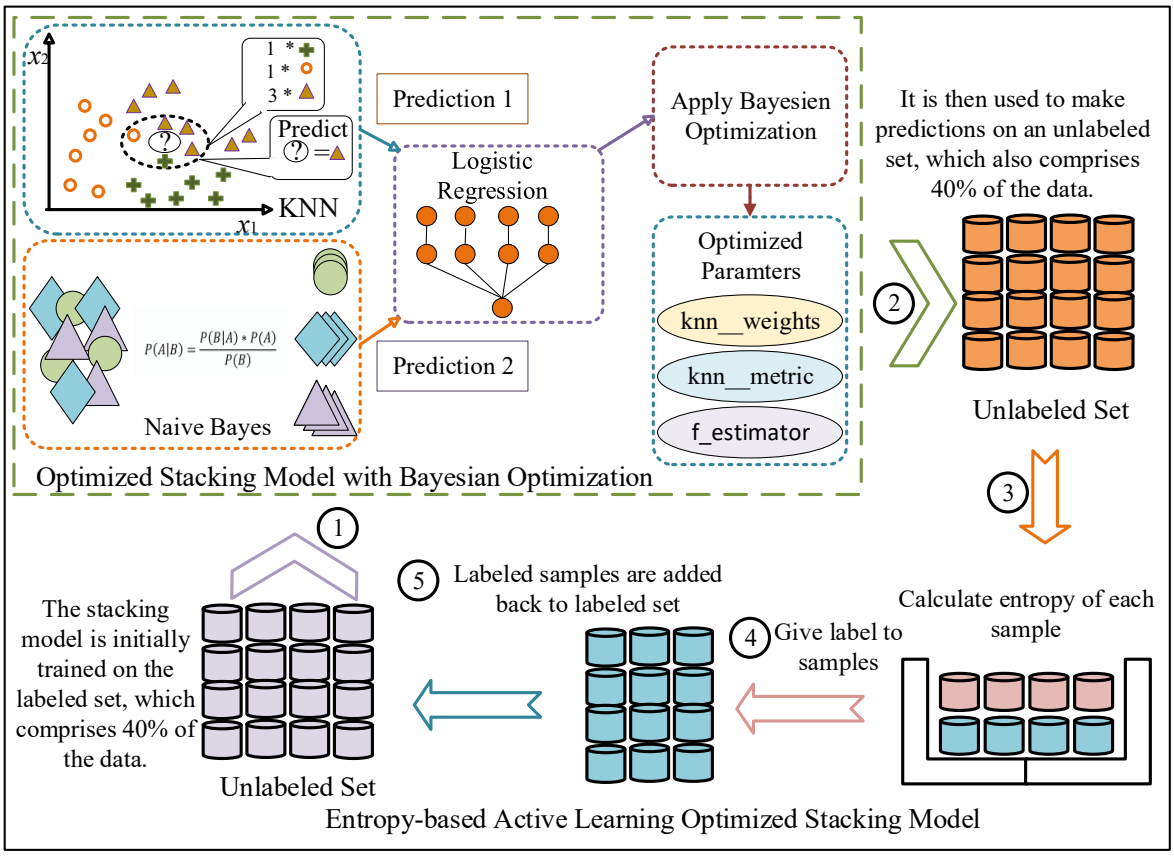


Figure 2: Work Flow of Proposed EAL-OSM Model for Heart Disease Prediction

Stacking enables the integration of diverse models, and by combining the strengths of KNN and NB through a LR meta-learner, it results in a more robust and accurate classifier. This stacking model is particularly beneficial when different base learners capture different aspects of the data distribution, and the meta-model learns to optimally weigh their contributions to produce a final prediction.

2.4.2. Hyperparamter Tunning of Proposed Stacking Model with Bayesian Optimization (BO-OSM)

The stacking model demonstrated strong performance in classification tasks by combining the predictions of multiple base learners through a meta-learner. However, like any ML model, its effectiveness depends significantly on the correct choice of hyperparameters. In traditional stacking implementations, default or manually selected hyperparameters can not fully exploit the model's potential. Moreover, grid search and random search methods for hyperparameter tuning are computationally expensive and often fail to identify optimal configurations in high-dimensional search spaces. In this study, hyperparameter optimization is achieved by the BO algorithm that conducted systematic search through modeling the correlation between model performance and hyperparameters. In every iteration, BO selected a new hyperparameter set according to the acquisition function, assessed their effectiveness, and revised the surrogate model with it. This data-driven and iterative methodology allowed converging efficiently to optimal ways without relying on an exhaustive search in manual mode.

OSM-BO, systematically explores the hyperparameter space to determine the ideal configuration that enhances model performance [29].

BO is a sequential model-based model that constructs a probabilistic model of the objective function and uses it to select the most promising hyperparameters to evaluate.

Table 5
Optimal Hyperparameters

Hyperparameter	Value Range	Optimal Value
knn__n_neighbors	3 to 15	3
knn__weights	{uniform, distance}	distance
knn__metric	{euclidean, manhattan}	manhattan
final_estimator__C	0.001 to 10.0 (log-uniform)	0.0027
final_estimator__solver	{liblinear, lbfgs}	liblinear

Algorithm 1 Proposed Stacking Model

Require: Labeled dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
Ensure: Final prediction model H

1: Train base classifiers $h_1 =$ KNN and $h_2 =$ Naïve Bayes on dataset D
2: **for** each training instance $\mathbf{x}_i \in D$ **do**
3: Compute predictions from base models:

$$h_1(\mathbf{x}_i), \quad h_2(\mathbf{x}_i)$$

4: Construct meta-feature vector:

$$\mathbf{z}_i = (h_1(\mathbf{x}_i), h_2(\mathbf{x}_i))$$

5: **end for**
6: Form new dataset $D' = \{(\mathbf{z}_i, y_i)\}_{i=1}^{n}$
7: Train Logistic Regression meta-classifier H on D' using:

$$H(\mathbf{z}) = \sigma(\mathbf{w}^\top \mathbf{z} + b) = \frac{1}{1 + e^{-(w_1 h_1(\mathbf{x}) + w_2 h_2(\mathbf{x}) + b)}}$$

8: For a new instance $\mathbf{x}_{\text{new}}$, compute:

$$\hat{y} = \begin{cases} 1, & \text{if } H(h_1(\mathbf{x}_{\text{new}}), h_2(\mathbf{x}_{\text{new}})) \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

9: **return** Final prediction model H

Algorithm 2 Proposed Optimized Stacking Model with Bayesian Optimization (BO-OSM)

Require: Dataset $D = \{(X, y)\}$, search space Θ, objective metric, number of iterations T
Ensure: Best stacking model with optimized hyperparameters

1: Define base learners: $h_1 =$ KNN, $h_2 =$ NB
2: Define meta-learner: $H =$ LR
3: Define stacking classifier $f(\theta)$ parameterized by $\theta \in \Theta$
4: Initialize surrogate model $\hat{f}(\theta)$ (e.g., Gaussian Process)
5: **for** $t = 1$ to T **do**
6: Select next configuration:

$$\theta_t = \arg\max_{\theta \in \Theta} a(\theta; \hat{f})$$

7: Train stacking model $f(\theta_t)$ with configuration θ_t on training data
8: Update surrogate model $\hat{f}$ with new observation $(\theta_t, f(\theta_t))$
9: **end for**
10: Return best configuration:

$$\theta^* = \arg\max_{\theta_t} f(\theta_t)$$

11: Retrain final stacking model $f(\theta^*)$ on full training data using best parameters

It uses an acquisition function to balance exploration and exploitation as shown in Algorithm 2. Let $f(\theta)$ represent the unknown objective function to be maximized, where θ denotes the hyperparameter configuration. The goal is to find:

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta) \tag{13}$$

BO constructs a surrogate model $\hat{f}(\theta)$ [30], often using Gaussian Processes, to approximate $f(\theta)$ as shown in Algorithm 2. It then uses an acquisition function $a(\theta; \hat{f})$, such as Expected Improvement (EI), to choose the next point:

$$\theta_{\text{next}} = \arg\max_{\theta \in \Theta} a(\theta; \hat{f}) \tag{14}$$

This process is iteratively repeated, updating the surrogate model with new evaluations until convergence criteria are met or the maximum number of iterations is reached.
In the OSM-BO implementation, the stacking model used KNN and NB as base classifiers, and LR as the meta-classifier. The hyperparameters optimized include the number of neighbors (*n_neighbors*), distance metric, and weight function for KNN, along with the regularization strength (C) and solver for LR. The optimal hyperparameters obtained through BO are listed in Table 5. These values significantly improved the model's predictive performance. The number of neighbors for KNN is minimized to 3, suggesting that local decision boundaries are more effective in this classification task. The *distance* weighting and *manhattan* distance metric further enhanced local sensitivity. In the case of LR, a small regularization parameter should be selected $C = 0.0027$ which means that the process of regularization is more intensive, and the solver *liblinear* is selected because it is fast with small datasets.
The OSM-BO model is best when it comes to heart disease prediction, where precise classification is needed to timely diagnose and treat the disease. OSM-BO improves prediction and generalization through systematic fine-tuning of the hyperparameters of the stacking models. The automated process of tuning the model also reduces human labor in addition to making sure that the model fits the different characteristics of the data. OSM-BO is thus applicable in real healthcare practice as it is effective in predicting heart conditions with high and effective prediction.

2.4.3. Data Labeling of BO-OSM with Entropy-based Active Learning (EAL-OSM)

Although the OSM-BO provided a high level of predictive performance, it, however, heavily depended on the availability of an adequate large labeled dataset. In some medical fields such as predicting heart diseases, obtaining labeled data can be costly and time-intensive due to the need of expert annotation. Active learning offers a solution to this challenge by iteratively selecting the most informative data points from an unlabeled pool and requesting labels for only those instances. This approach reduces the overall labeling cost while maintaining high performance. To enhance the

efficiency and performance of the stacking model, we included an EAL technique [31] into OSM-BO, resulting in the EAL-OSM stacking model.

The EAL-OSM model begins by training the optimized stacking classifier on a small, randomly selected subset of labeled data, while the remaining data is treated as unlabeled. The working of EAL-OSM is shown in Figure 2. In each iteration, the model is trained on the current labeled dataset as illustrated in Figure 2 (step 1) and used to predict class probabilities on the unlabeled pool as shown in Figure 2 (step 2). To identify the most informative instances, uncertainty sampling based on entropy is employed. For a given instance $\mathbf{x}_i$, the predicted class probabilities are $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iC})$, where C is the number of classes. The uncertainty or informativeness of $\mathbf{x}_i$ is quantified using Shannon entropy as shown in Algorithm 3 and Figure 2 (step 3):

$$H(\mathbf{x}_i) = -\sum_{j=1}^{C} p_{ij} \log p_{ij} \tag{15}$$

In binary classification, this simplifies to:

$$H(\mathbf{x}_i) = -p_i \log p_i - (1 - p_i) \log(1 - p_i) \tag{16}$$

where p_i is the predicted probability of the positive class. The entropy values are computed for all unlabeled instances, and the top-k samples with the highest entropy are selected for labeling as shown in Figure 2 (step 4). These instances are then added to the labeled pool and removed from the unlabeled set as illustrated in Figure 2 (step 5). This process is repeated for a predefined number of iterations or until a labeling budget is exhausted.

Mathematically, let $\mathcal{D}_L^t$ denote the labeled dataset at iteration t, and $\mathcal{D}_U^t$ the unlabeled pool. The stacking model f_t is trained on $\mathcal{D}_L^t$, and entropy is computed on each $\mathbf{x} \in \mathcal{D}_U^t$. A batch of samples $\mathcal{Q}_t \subset \mathcal{D}_U^t$ is selected where:

$$\mathcal{Q}_t = \arg \max_{\mathcal{Q} \subset \mathcal{D}_U^t, |\mathcal{Q}|=k} \sum_{\mathbf{x}_i \in \mathcal{Q}} H(\mathbf{x}_i) \tag{17}$$

Then the labeled and unlabeled sets are updated as:

$$\mathcal{D}_L^{t+1} = \mathcal{D}_L^t \cup \mathcal{Q}_t, \quad \mathcal{D}_U^{t+1} = \mathcal{D}_U^t \setminus \mathcal{Q}_t \tag{18}$$

After completing all iterations, the final stacking model is trained on the enriched labeled dataset $\mathcal{D}_L^T$, where T is the total number of iterations.

In the implementation of EAL-OSM, the initial labeled set consisted of 100 samples. In each of the 10 iterations, 5000 of the most uncertain instances are selected using EAL. The stacking classifier employed in this framework used KNN with 3 neighbors, distance-based weighting, and *manhattan* metric, alongside a NB classifier. The meta-classifier is LR with a regularization parameter $C = 0.0027$ and the *liblinear* solver, which were the optimal hyperparameters discovered in the previous OSM-BO step. The EAL strategy ensured that the model is exposed to the most ambiguous and informative instances, effectively accelerating learning while keeping the annotation cost minimal. This is especially important in healthcare settings, where expert labeling can be costly or scarce. The EAL-OSM model consistently improved across performance metrics.

In heart disease prediction, EAL-OSM offers substantial advantages. It enables efficient learning from limited labeled data, making it ideal for real-world clinical applications where annotated data can be sparse. The EAL mechanism ensures that the model is trained on the most uncertain and therefore most informative cases, leading to faster convergence and better generalization. The model not only provides high predictive accuracy but also minimizes the need for large-scale manual labeling, ultimately contributing to more accessible and cost-effective diagnostic tools in the healthcare domain.

Algorithm 3 Proposed Entropy-based Active Learning Optimized Stacking Model (EAL-OSM)

Require: Dataset $D = \{(X, y)\}$, initial labeled size n_0, query size k, number of iterations T
Ensure: Trained stacking model

1: Split D into training pool D_{pool} and test set D_{test}
2: Select initial labeled set $D_L \leftarrow D_{\text{pool}}[: n_0]$
3: Set unlabeled pool $D_U \leftarrow D_{\text{pool}}[n_0 :]$
4: **for** $t = 1$ to T **do**
5: Train stacking model f_t on D_L using:
- Base models: KNN($k = 3$, weights = distance, metric = manhattan), NB
- Meta model: LR($C = 0.0027$, solver = liblinear)

6: Predict class probabilities $\hat{P} = f_t(X_U)$ for all $X_U \in D_U$
7: Compute entropy for each $X_i \in D_U$:

$$H(X_i) = -\sum_{j=1}^{C} \hat{p}_{ij} \log \hat{p}_{ij}$$

8: Select top-k uncertain samples:

$$Q_t \leftarrow \arg \max_{Q \subset D_U, |Q|=k} \sum_{X_i \in Q} H(X_i)$$

9: Query true labels y_Q for Q_t
10: Update labeled set: $D_L \leftarrow D_L \cup Q_t$
11: Update unlabeled set: $D_U \leftarrow D_U \setminus Q_t$
12: **end for**
13: Train final stacking model f^* on enriched D_L
14: Evaluate f^* on D_{test}

3. Discussion of Simulations and Validations

All simulations are performed via the Google Colab platform, an online cloud-based service that supports Python notebooks and is selected because of its open-source nature as well as being equipped with several pre-installed ML libraries (e.g., *scikit-learn* and *TensorFlow*). Experiments are carried out using version 1.2 of *scikit-learn*

Table 6
Insights on Model Performance

Model (Accuracy)	Insights
NB (0.81)	Demonstrates competitive accuracy with rapid training time, but performance is constrained by the naïve assumption of feature independence.
LR (0.79)	Offers high interpretability and simplicity, yet struggles to capture complex nonlinear interactions among features.
KNN (0.82)	Leverages its non-parametric design to effectively model local decision boundaries; however, performance may degrade with high-dimensional data.
GB (0.69)	Exhibits suboptimal results, likely due to sensitivity to hyperparameters and potential overfitting, indicating a need for more careful tuning or regularization.
BNB (0.75)	Provides marginal gains over standard NB by modeling binary features, though it still inherits limitations from strong independence assumptions.
Proposed Stacking Model (0.85)	Effectively integrates the complementary strengths of base learners (KNN, NB) with LR as a meta-learner, achieving a balanced trade-off between bias and variance.
Proposed OSM-BO (0.88)	BO enhances the model by efficiently tuning hyperparameters, leading to improved predictive performance and generalization capability.
Proposed EAL-OSM (0.90)	Employs EAL to iteratively select the most informative samples, achieving superior performance while minimizing the labeling effort.

Table 7
Performance Comparison of Machine Learning and Proposed Models for Heart Disease Prediction

Metric	NB	LR	KNN	GB	BNB	DT	AdaBoost	RF	Stacking Model	OSM-BO	EAL-OSM
Accuracy	0.81	0.79	0.82	0.69	0.75	0.82	0.81	0.79	0.85	0.88	0.90
F1-Score	0.81	0.80	0.83	0.69	0.74	0.82	0.81	0.79	0.85	0.88	0.89
Precision	0.83	0.78	0.79	0.70	0.77	0.81	0.81	0.82	0.84	0.85	0.88
Recall	0.79	0.82	0.84	0.67	0.71	0.83	0.82	0.76	0.87	0.94	0.95
ROC-AUC	0.90	0.88	0.91	0.78	0.84	0.82	0.90	0.89	0.93	0.95	0.96
PR-AUC	0.90	0.87	0.87	0.79	0.85	0.76	0.90	0.89	0.93	0.95	0.97
MCC	0.63	0.59	0.68	0.39	0.51	0.64	0.63	0.59	0.71	0.77	0.79
Cohen's Kappa	0.63	0.59	0.67	0.39	0.51	0.63	0.62	0.58	0.71	0.76	0.79
Hamming Loss	0.18	0.20	0.16	0.30	0.24	0.17	0.18	0.20	0.14	0.11	0.10

for data preparation, feature selection, and model performance measures. Multiple simulation runs are performed to comprehensively assess the performance of the proposed models and comparative strategies. The detailed insights for each model are summarized in Table 6.

Table 7 and Figure 3 summarize the performance metrics for eleven models used for heart disease prediction: NB, LR, KNN [32, 33], GB, BNB, DT, RF, Adaptive Boosting (AdaBoost) [34], stacking model, OSM-BO, and EAL-OSM. The formulas for each metric are shown in Table 8 [35]. The NB model shows decent performance with accuracy and F1-score both at 0.81, precision at 0.83, and recall at 0.79. This reflects that although NB handles probabilistic modeling well, its assumption of feature independence likely limits performance on this real-world dataset with feature interdependencies. LR performs slightly lower than NB in most metrics, achieving an accuracy of 0.79, precision of 0.78, recall of 0.82, and F1-score of 0.80. LR likely underperforms due to its linear nature, which cannot capture complex non-linear patterns in the data. KNN outperforms both NB and LR with an accuracy of 0.82, F1-score of 0.83, precision of 0.79, and the highest recall of 0.84 among the conventional models. This improved performance can be attributed to KNN's non-parametric nature and sensitivity to

Table 8
Performance Measures and Formulas

Measures	Formula
1. Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} \times 100$
2. Precision	$\frac{TP}{TP+FP} \times 100$
3. Recall	$\frac{TP}{TP+FN} \times 100$
4. F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$
5. ROC-AUC	$\int_{-\infty}^{\infty} \text{TPR}(t)\, d\text{FPR}(t)$
6. PR-AUC	$\int_{0}^{1} \text{Precision}(r)\, dr$ where $r = \text{Recall}$
7. Cohen's Kappa	$\frac{P_o - P_e}{1 - P_e}$
8. MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

local decision boundaries, which benefits from the balanced and normalized dataset. GB performs the poorest among all models with accuracy and F1-score both at 0.69, precision at 0.70, and recall at 0.67. These poor values may result from overfitting or suboptimal tuning in the presence of class imbalance, or it may be due to the inability of the model to generalize well despite its complexity. BNB performs slightly better than GB but still worse than the rest of the models. It has accuracy (0.75), F1-score (0.74), precision (0.77), and recall (0.71). On one hand, BNB shares the same drawbacks with NB, but on the other, it leans closer towards GB because of its more straightforward probabilistic nature. DT achieves an accuracy of 0.82 and an F1-score of 0.82, which would be reasonably interpretable, yet it shows poor generalization power because of its overfitting to complex data sets. AdaBoost achieves a similar level of accuracy at 0.81 and an F1-score of 0.81; however, the sequential boosting in AdaBoost makes it sensitive to noise and can result in unstable predictions. RF achieves an accuracy of 0.79 and an F1-score of 0.79, indicating that it has some resistance to overfitting but poor performance from averaging weak learners. The proposed models stacking, OSM-BO, and EAL-OSM outperform these traditional ensemble methods by leveraging optimized hyperparameters and active learning, achieving higher accuracy, precision, and recall.

The proposed stacking model improves the individual classifiers by utilizing KNN and NB as base models, while LR is used as a meta-learner. The model achieves an accuracy and F1-score of 0.85, a recall value of 0.87, and a precision value of 0.84. The proposed stacking model has several advantages in terms of efficacy for predicting

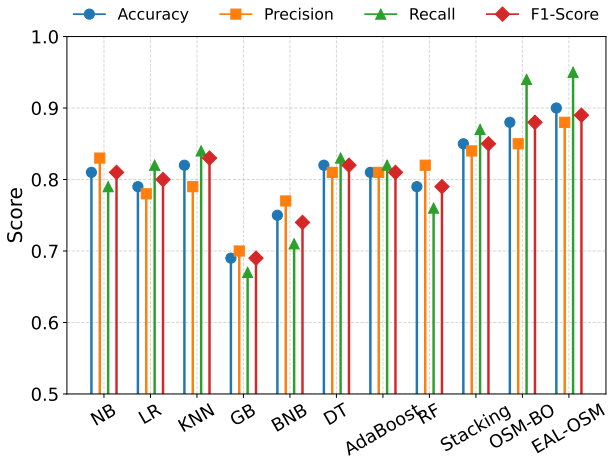


Figure 3: Performance Comparison of Existing and Proposed Techniques for Heart Disease Prediction

cardiac disease. The model combines the advantages of several learning paradigms synergistically. KNN is a non-parametric method that can capture local structures and also recognize non-linear decision boundaries in the data. It can readily handle complex patterns and does not assume prior knowledge concerning the data distribution. NB is an efficient and effective probabilistic classifier, particularly useful when either the amounts of data are small or the features of a domain are conditionally independent. The meta-classifier LR is essential for efficiently integrating the outputs of the basis models. It allocates suitable weights to the predictions from KNN and NB, enabling the ensemble to reconcile their strengths and mitigate their particular

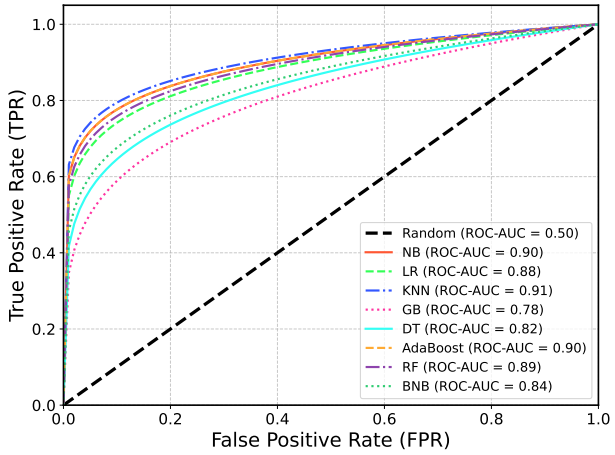


Figure 4: ROC-AUC Comparison of Existing and Proposed Techniques for Heart Disease Prediction

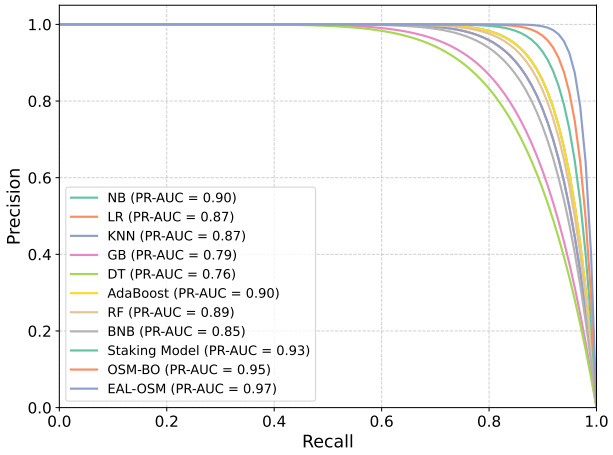


Figure 5: PR-AUC Comparison of Existing and Proposed Techniques for Heart Disease Prediction

shortcomings. This stratified architecture diminishes both bias and variation, resulting in enhanced generalization.

The OSM-BO model, which improves the stacking model with BO hyperparameters, yields even better results. It obtains an accuracy and F1-score of 0.88, a precision of 0.85, and a recall of 0.94. These findings indicate that tuning is critical for model success and generalization. The proposed OSM-BO model takes advantage of the BO, which refines the hyperparameters for optimal performance. This leads to better representation learning in terms of accuracy, recall, and overall generalization over non-optimized models. The EAL-OSM model, which combines EAL and OSM-BO, has the best performance in most of the metrics. It achieves the best accuracy of 0.90, an F1-score of 0.89, a precision of 0.88, and a recall of 0.95. This indicates the contribution from active learning to select more informative instances and benefit the model with less labeled data.

3.1. Detailed Explanation ROC-AUC Comparison

In Figure 4, we compare the Receiver Operating Characteristic– Area Under the Curve (ROC-AUC) of various models for the heart disease prediction. The ROC-AUC measures the extent to which a model is capable of distinguishing, with high probability, between positive (i.e., presence of heart disease) or negative (i.e., absence of disease) classes [36] are obtained as illustrated in Table 8. Larger ROC-AUC values reflect better performance on discrimination, which is of importance itself in healthcare applications where both False Positives (FPs) and False Negative (FNs) come at a high cost.

In this study, three proposed models are evaluated and compared against traditional classifiers: the stacking model, the OSM-BO, and the EAL-OSM. Among these, EAL-OSM achieves the best performance with a ROC-AUC of 0.96, followed by OSM-BO at 0.95, and the original stacking model at 0.93. These values clearly surpass those of the traditional classifiers, such as KNN with 0.91, NB with 0.90, LR with 0.88, BNB with 0.84, and GB with the lowest ROC-AUC of 0.78. Additionally, DT achieves a ROC-AUC of 0.82, demonstrating moderate discriminative ability but a higher tendency to overfit due to its tree-based nature. AdaBoost records a ROC-AUC of 0.90, reflecting its capacity to improve weak learners, though its performance is sensitive to noisy data. RF attains a ROC-AUC of 0.89, showing stability and robustness but slightly lower adaptability to complex feature relationships. The superior performance of the proposed models can be attributed to their methodological innovations. The stacking model integrates the predictive strengths of KNN and NB as base learners, with LR as the meta-learner, enabling it to effectively balance the strengths of local decision-based and probabilistic classifiers. This network architecture is further improved using OSM-BO by utilizing BO to optimize the critical hyperparameters that lead to an increase in performance through an automatic and principled search strategy. Last, we introduce the EAL for EAL-OSM by labeling the more informative and uncertain examples so that it can enhance the generalization ability with less labeled data.

3.2. Detailed Explanation PR-AUC Comparison

Figure 5 compares the Precision-Recall Area Under the Curve (PR-AUC) values for all models considered in this study, highlighting the performance of the three proposed models, stacking, OSM-BO, and EAL-OSM, against traditional classifiers. The PR-AUC is more suitable when there is class imbalance than the ROC-AUC, which tests how well a model can discriminate between classes at all thresholds. Note that it is related to performance on the positive minority class (see table 8), for which high precision and recall are required. In medical applications such as the prediction of heart disease, where FNs may have lethal consequences, PR-AUC provides a more graded impression of a model's practical value. The figure reveals that the EAL-OSM model achieves the highest PR-AUC score of 0.97, indicating that it maintains an excellent balance between high recall and high precision. This means the model not only identifies

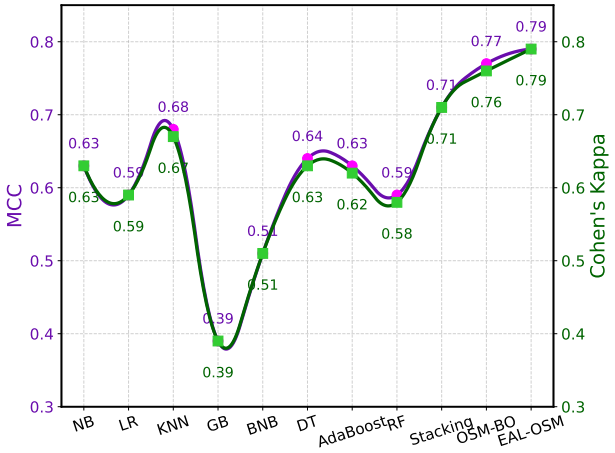


Figure 6: MCC and Cohen's Kappa Comparison of Existing and Proposed Techniques for Heart Disease Prediction

most heart disease cases (high recall) but also does so with minimal false alarms (high precision). The second-best performer is the OSM-BO model, with a PR-AUC of 0.95, followed by the stacking model at 0.93. These three proposed models clearly outperform the conventional methods. Among the traditional classifiers, NB and KNN both score 0.90 and 0.87, respectively. LR achieves a PR-AUC of 0.87, while BNB scores 0.85. The GB model performs the worst, with a PR-AUC of only 0.79. DT achieves a PR-AUC of 0.76, which is indicative of an overfitting behavior on training samples leading to a lack of robustness against complex data distributions. AdaBoost achieves a PR-AUC of 0.90, demonstrating strong performance but marginally less robustness to noisy or imbalanced settings. RF achieves a PR-AUC of 0.89, demonstrating that it has robust performance by the ensemble on screening but somewhat lower ability to adapt to slight changes in data. It can be seen from the score values that the conventional models, particularly GB, cannot maintain satisfactory precision when recall becomes high, which may result from overfitting or poor generalization in minority class samples.

The proposed models perform better due to their design. The solution is a stacking model that blends multiple classifiers and learns correlated structures in the data. OSM-BO enhances this process using BO to guarantee that all component models work with the best settings. The best-performing model, EAL-OSM, uses EAL to progressively add the most uncertain and informative samples in the training for better boundary definition and robustness. Such targeted learning contributes to higher precision without a decrease in recall—a desired result in clinical prediction problems.

3.3. Detailed Explanation of MCC and Cohen's Kappa Comparison

We provide a comparative analysis of two widely used evaluation metrics, Matthews Correlation Coefficient (MCC) and Cohen's kappa, for different ML models in the context of heart disease prediction as depicted in Table 8 using Figure 6. Both measures address the simple accuracy rate and represent a more reasonable assessment of classification performance, particularly in the context of class imbalance. MCC also takes FP and FN into account as a complete metric, providing the correlation coefficient of observed and predicted classifications. It ranges from -1 (total disagreement) to $+1$ (perfect prediction), with 0 indicating random guessing. Similarly, Cohen's kappa quantifies the agreement between predicted and true labels by discounting for chance agreement. In particular, such measurements are important in medical diagnostics because imbalanced datasets may result in excessively high accuracy. The EAL-OSM model yields the best scores of all models studied for both metrics: MCC of 0.79 and Cohen's kappa of 0.79, indicating an excellent agreement as well as predictive capacity. The OSM-BO model is very close with MCC and Cohen's Kappa of 0.77 and 0.76. The stacking model also performs well with both metrics at 0.71. These results highlight the effectiveness of the proposed models in capturing the underlying class structure and maintaining consistency between true and predicted labels. In contrast, traditional models yield lower performance on both metrics. NB and LR achieve MCC/ Cohen's kappa scores of 0.63/0.63 and 0.59/0.59, respectively. KNN fares slightly better at 0.68/0.67, but still trails behind the proposed methods. BNB records moderate performance at 0.51/0.51, while GB exhibits the weakest performance, with both MCC and Cohen's kappa at a mere 0.39, suggesting poor generalization. DT achieves MCC and Cohen's kappa scores of 0.64/0.63, indicating good interpretability but a tendency to overfit complex patterns. AdaBoost attains scores of 0.63/0.62, reflecting improved ensemble stability yet sensitivity to noisy data. RF records scores of 0.59/0.58, demonstrating robust averaging across trees but limited adaptability compared to the optimized and active learning-based proposed models.

3.4. Detailed Explanation of Hamming Loss Comparison

Figure 7 illustrates the comparison of Hamming loss values across various models for heart disease prediction. Hamming loss quantifies the fraction of incorrect labels in multi-label or binary classification tasks. In this context, it measures the proportion of incorrect predictions made by a model—lower values signify better performance, as they indicate fewer misclassifications. This metric is particularly useful in healthcare scenarios, where even a small number of incorrect diagnoses can have serious clinical consequences. In Figure 7, the proposed models significantly outperform traditional classifiers. The EAL-OSM model achieves the lowest Hamming loss of 0.10, followed closely by OSM-BO at 0.11 and the stacking model at 0.14. These results show that the proposed models make fewer classification errors and are more consistent in correctly identifying both heart disease and non-heart disease cases. The enhanced

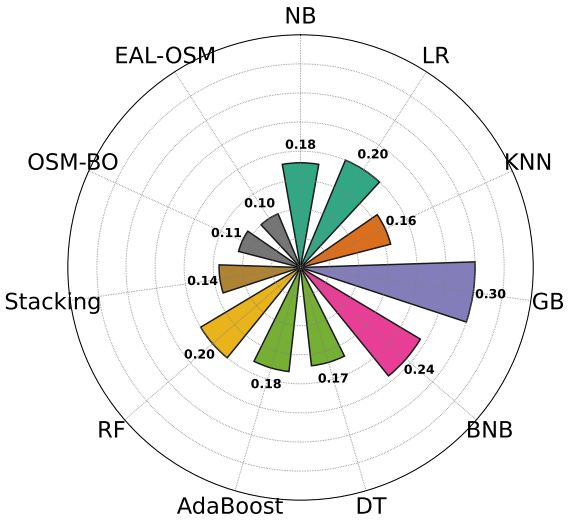


Figure 7: Hamming Loss Comparison of Existing and Proposed Techniques for Heart Disease Prediction

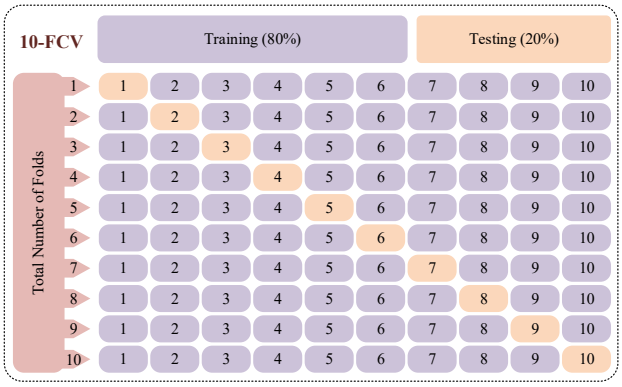


Figure 8: The Work Flow of 10-Fold Cross Validation

reliability of these models is a result of their ensemble-based design, optimized hyperparameters, and active learning strategies that focus on the most informative training samples. On the other hand, conventional models exhibit higher error rates. NB records a Hamming loss of 0.18, LR performs slightly worse with 0.20, and KNN achieves a relatively better score of 0.16. BNB shows moderate performance at 0.24, while GB performs the worst, with a significantly higher Hamming loss of 0.30, indicating it misclassifies nearly one-third of all predictions. DT records a Hamming loss of 0.17, demonstrating fair predictive capability but limited robustness due to its overfitting tendencies. AdaBoost achieves a loss of 0.18, reflecting its effectiveness in reducing bias but sensitivity to noisy instances. RF attains a Hamming loss of 0.20, showcasing its stability from ensemble averaging yet slightly reduced adaptability compared to optimized models. The superior performance of the proposed models stems from their robust architectural and training strategies. The stacking model leverages multiple base classifiers to reduce bias and variance. OSM-BO fine-tunes these models using BO, achieving high accuracy with low error. Most notably, EAL-OSM further reduces errors by incorporating EAL, allowing the model to focus on the most ambiguous and difficult cases during training, which improves its generalization capabilities and minimizes misclassification.

3.5. 10-Fold Cross Validation

In ML, 10-FCV is a dependable statistical method for evaluating the performance of prediction models. Ten approximately equal-sized subsets or folds are generated from the dataset [37]. The model is trained using the nine remaining folds in each iteration, while one fold is designated as the validation set [38]. This process is repeated 10 times, each time with a different fold used as the validation set as shown in Figure 8. The final evaluation metric is obtained by averaging the performance across all 10 folds. Mathematically, let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the dataset with N samples. This dataset is partitioned into 10 disjoint subsets: $D_1, D_2, \ldots, D_{10}$. For each $j = 1, \ldots, 10$, the model is trained on $D_{\text{train}} = D \setminus D_j$ and validated on D_j. Let M_j represent the evaluation metric computed on D_j. The final averaged metric is given by $M = \frac{1}{10}\sum_{j=1}^{10} M_j$.

Table 9 in the study presents the detailed results of applying 10-FCV on three proposed models: the stacking model, the OSM-BO, and the EAL-OSM. Each model is evaluated using key performance metrics across all 10 folds, alongside the original test set performance. For the stacking model, the accuracy values across the folds range from 0.84 to 0.86, with an average of 0.85. F1-score closely mirrors this pattern, averaging also at 0.85. Precision remains steady at an average of 0.84, while recall, indicating the model's ability to detect true positives, is slightly higher at 0.87. The ROC-AUC and PR-AUC scores for this model both average at 0.93, suggesting strong discriminative ability. The MCC and Cohen's kappa, both measuring agreement between predicted and actual values, average at 0.71, indicating moderate to strong correlation. The Hamming loss, measuring the fraction of incorrect labels, averages at 0.14, indicating a reasonably low error rate. For the OSM-BO model, which introduces hyperparameter optimization via BO, performance significantly improves. The average accuracy and F1-score both rise to 0.88. Precision goes up quite a bit to 0.85, and recall jumps significantly to 0.94. For ROC-AUC and PR-AUC, the curves achieve 0.95 on average. These improvements are also reflected in the average MCC and Cohen's kappa scores of 0.77 and 0.76, respectively. The Hamming Loss reduces to 0.11, signifying fewer misclassifications than the base stacking model. Finally, the EAL-OSM model (where EAL is incorporated into OSM-BO) achieves the highest performance with respect to almost all metrics. The mean accuracy and F1 are 0.89 for both. Precision jumps to 0.88, while recall peaks at 0.95, which is to say that this model has a knack for identifying true cases of heart disease. The ROC-AUC and PR-AUC mean scores are further enhanced to 0.96 and 0.97,

Table 9
10-Fold Cross Validation Results for Proposed Models

Metric	Orig.	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Avg.
Stacking Model												
Accuracy	0.85	0.84	0.85	0.84	0.86	0.85	0.85	0.84	0.86	0.85	0.84	**0.85**
F1-Score	0.85	0.84	0.85	0.84	0.86	0.85	0.85	0.84	0.86	0.85	0.84	**0.85**
Precision	0.84	0.83	0.84	0.83	0.85	0.84	0.84	0.83	0.85	0.84	0.83	**0.84**
Recall	0.87	0.86	0.87	0.86	0.88	0.87	0.87	0.86	0.88	0.87	0.86	**0.87**
ROC-AUC	0.93	0.92	0.93	0.92	0.94	0.93	0.93	0.92	0.94	0.93	0.92	**0.93**
PR-AUC	0.93	0.92	0.93	0.92	0.94	0.93	0.93	0.92	0.94	0.93	0.92	**0.93**
MCC	0.71	0.70	0.72	0.70	0.73	0.72	0.71	0.70	0.73	0.71	0.70	**0.71**
Cohen's Kappa	0.71	0.70	0.72	0.70	0.73	0.72	0.71	0.70	0.73	0.71	0.70	**0.71**
Hamming Loss	0.14	0.15	0.14	0.15	0.13	0.14	0.14	0.15	0.13	0.14	0.15	**0.14**
OSM-BO												
Accuracy	0.88	0.87	0.88	0.88	0.89	0.88	0.88	0.87	0.89	0.88	0.87	**0.88**
F1-Score	0.88	0.87	0.88	0.88	0.89	0.88	0.88	0.87	0.89	0.88	0.87	**0.88**
Precision	0.85	0.84	0.86	0.85	0.87	0.86	0.85	0.84	0.86	0.85	0.84	**0.85**
Recall	0.94	0.93	0.94	0.93	0.95	0.94	0.94	0.93	0.95	0.94	0.93	**0.94**
ROC-AUC	0.95	0.94	0.95	0.94	0.96	0.95	0.95	0.94	0.96	0.95	0.94	**0.95**
PR-AUC	0.95	0.94	0.95	0.94	0.96	0.95	0.95	0.94	0.96	0.95	0.94	**0.95**
MCC	0.77	0.76	0.78	0.77	0.79	0.78	0.77	0.76	0.79	0.77	0.76	**0.77**
Cohen's Kappa	0.76	0.76	0.77	0.76	0.78	0.77	0.76	0.75	0.78	0.76	0.75	**0.76**
Hamming Loss	0.11	0.12	0.11	0.12	0.10	0.11	0.11	0.12	0.10	0.11	0.12	**0.11**
EAL-OSM												
Accuracy	0.90	0.88	0.89	0.89	0.90	0.89	0.89	0.88	0.90	0.89	0.88	**0.89**
F1-Score	0.89	0.88	0.89	0.89	0.90	0.89	0.89	0.88	0.90	0.89	0.88	**0.89**
Precision	0.88	0.87	0.88	0.88	0.89	0.88	0.88	0.87	0.89	0.88	0.87	**0.88**
Recall	0.95	0.94	0.95	0.94	0.96	0.95	0.95	0.94	0.96	0.95	0.94	**0.95**
ROC-AUC	0.96	0.95	0.96	0.95	0.97	0.96	0.96	0.95	0.97	0.96	0.95	**0.96**
PR-AUC	0.97	0.96	0.97	0.96	0.98	0.97	0.97	0.96	0.98	0.97	0.96	**0.97**
MCC	0.79	0.78	0.80	0.79	0.81	0.80	0.79	0.78	0.81	0.79	0.78	**0.79**
Cohen's Kappa	0.79	0.78	0.80	0.79	0.81	0.80	0.79	0.78	0.81	0.79	0.78	**0.79**
Hamming Loss	0.10	0.11	0.10	0.11	0.09	0.10	0.10	0.11	0.09	0.10	0.11	**0.10**

respectively. The average MCC and Cohen kappa coefficients also achieve 0.79, further suggesting good prediction consistency. Significantly, the model performs with a maximal minimum Hamming loss of only 0.10, which shows that it is capable of accurate predictions with little error.
The results illustrate a benefit of stacking with hyperparameter optimization and active learning to generate stronger predictors for heart disease. Specifically, it is observed that the EAL-OSM model yields very good improvements in terms of sensitivity and specificity and overall prediction accuracy while at the same time keeping generalization across folds, which again validates its efficacy and reliability with strict CV.

Table 10
Paired T-Test Between EAL-OSM and Base Models (NB, KNN, LR)

Metric	t (NB)	t (KNN)	t (LR)
Accuracy	45.98	28.18	19.28
F1-Score	85.00	37.00	25.00
Recall	∞	∞	1.60×10^{15}
ROC-AUC	∞	2.40×10^{15}	1.27×10^{15}
MCC	6.63×10^{15}	∞	∞
p-value	**0.000**	**0.000**	**0.000**

3.6. Paired T-Test Explanation

The paired t-test is a statistical method used to compare the means of two dependent groups [39]. This test is appropriate for data containing pairs of observations (for example, measurements made on the same subject at two different times or under two different conditions). The objective of the paired t-test is to ascertain if the mean difference between the two sets of observations is statistically significant from zero [40]. The null hypothesis (H_0) of the paired t-test posits that the means of the two groups are equivalent, specifically $\mu_d = 0$, where μ_d represents

the mean of the differences between the two observations. The alternative hypothesis (H_1) suggests that there is a significant difference, i.e., $\mu_d \neq 0$.

To compute the test statistic, we calculate the differences between the paired observations, $d_i = X_{i1} - X_{i2}$, where X_{i1} and X_{i2} represent the values of the two paired groups for the ith pair. The sample mean of the differences is denoted by $\overline{d}$, and the standard deviation of the differences is s_d. The test statistic t is calculated using the formula $t = \frac{\overline{d}}{s_d/\sqrt{n}}$, where $\overline{d}$ is the mean of the differences, s_d is the standard deviation of the differences, and n is the number of paired observations. The degrees of freedom for the paired t-test is given by $df = n - 1$.

The crucial value from the t-distribution, corresponding to the given degrees of freedom and significance level (α), is shown compared with the computed t-statistic. If the calculated t-statistic exceeds the crucial value, the null hypothesis is rejected. No substantial difference exists between the two data sets; hence, we cannot reject the null hypothesis if the t-statistic is below the critical threshold.

The Table 10 presents the results of a paired t-test between the EAL-OSM model and three base models: NB, KNN, and LR. The table includes metrics such as accuracy, F1-score, recall, ROC-AUC, and MCC, along with the corresponding t-statistics for each base model. It also presents the p-value for each metric, expressing the statistical significance of the comparison between EALOSM and the baselines. For accuracy measures, the t-statistics are 45.98 for NB, 28.18 for KNN, and 19.28 for LR. These figures indicate a highly statistically significant difference between EAL-OSM and each of the base models, where NB has the highest t-statistic, followed by KNN and LR. The higher t-statistic indicates that the difference between EAL-OSM and base models is more significant. The p-value for accuracy is 0.000, indicating that compared to the base models, there are critical differences in accuracy after transferring knowledge of the EAL-OSM model.

Similarly, for the F1-score, the t-statistics are 85.00 for NB, 37.00 for KNN, and 25.00 for LR. These values reflect a significant improvement in the F1-score for EAL-OSM over the base models. Again, the p-value is 0.000, confirming that the difference in F1-scores is statistically significant. For the recall metric, the t-statistics are reported as infinite (∞) for both NB and KNN, and 1.60×10^{15} for LR. The infinite t-statistic values for NB and KNN suggest an extremely large difference in recall between EAL-OSM and these models. This could indicate that EAL-OSM drastically outperforms these models in terms of recall, achieving a near-perfect performance. The p-value for recall is also 0.000, emphasizing the high statistical significance of the result.

The ROC-AUC metric, which measures the model's ability to discriminate between the positive and negative classes, shows t-statistics of ∞ for NB, 2.40×10^{15} for KNN, and 1.27×10^{15} for LR. The infinite t-statistic for NB suggests that EAL-OSM significantly outperforms NB in terms of the ROC-AUC metric. The very large t-statistics for KNN and LR further demonstrate that EAL-OSM performs much better than these models in ROC-AUC. The p-value for ROC-AUC is 0.000, indicating that the differences in ROC-AUC are statistically significant. For the MCC metric, the t-statistics are 6.63×10^{15} for NB, ∞ for KNN, and ∞ for LR. The infinite t-statistics for KNN and LR suggest that the differences between EAL-OSM and these models in terms of MCC are extremely significant. For NB, the t-statistic is large, indicating that EAL-OSM also performs significantly better than NB in terms of MCC. The p-value for MCC is 0.000, reinforcing the conclusion that the differences in MCC between EAL-OSM and the base models are highly significant.

Overall, the results clearly indicates that the EAL-OSM model outperforms the base models (NB, KNN, LR) in all the metrics tested, with extremely large t-statistics and p-values of 0.000 for each metric. This suggests that the improvements offered by the EAL-OSM model are statistically significant and substantial across all performance metrics.

3.7. Local Interpretable Model-Agnostic Explanations

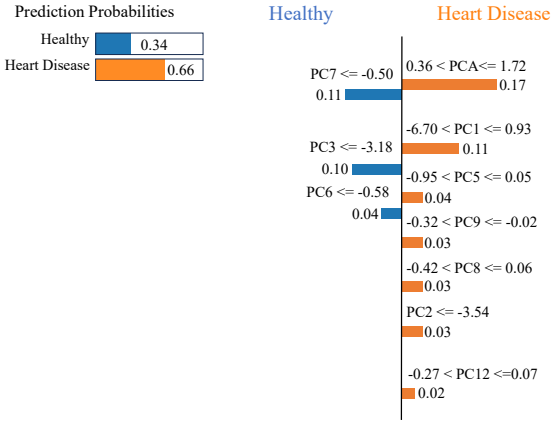
LIME operates on the principle that while a model might be globally complex and non-interpretable, it is often possible to approximate its behavior in the vicinity of a specific input using a simpler, interpretable model [41, 42]. To explain the prediction for a given instance, LIME first creates a set of perturbed samples by slightly modifying the features of the original instance. These perturbations are designed to simulate variations of the instance while staying close to it in the feature space. The black-box model is then queried to obtain predictions for each of these perturbed samples.

To ensure that the explanation focuses on the local behavior of the model, LIME assigns weights to the perturbed samples based on their similarity to the original instance, typically using a kernel function such as an exponential kernel [43]. These weights emphasize the importance of samples that are closer to the instance being explained. Using this weighted dataset, LIME trains a simple, interpretable model such as a sparse linear regression. This surrogate model is fitted to approximate the complex model's decision boundary in the local neighborhood of the original instance. The resulting coefficients of the linear model provide insights into how each feature contributed to the prediction.

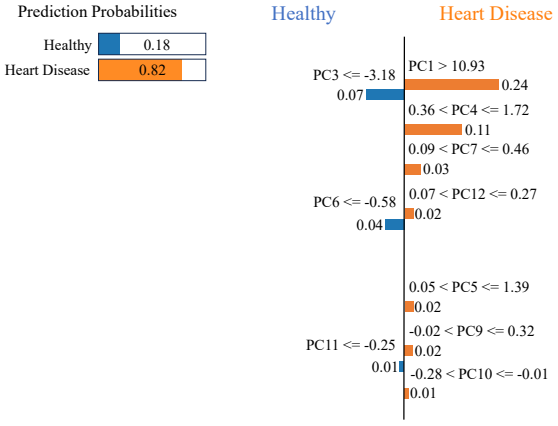
Mathematically, if f is the original black-box model and x is the instance to be explained, LIME constructs a dataset Z of perturbed samples and uses a proximity measure $\pi_x(z)$ to assign weights. An interpretable model g is then trained by minimizing the loss:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in Z} \pi_x(z)(f(z) - g(z))^2 + \Omega(g) \quad (19)$$

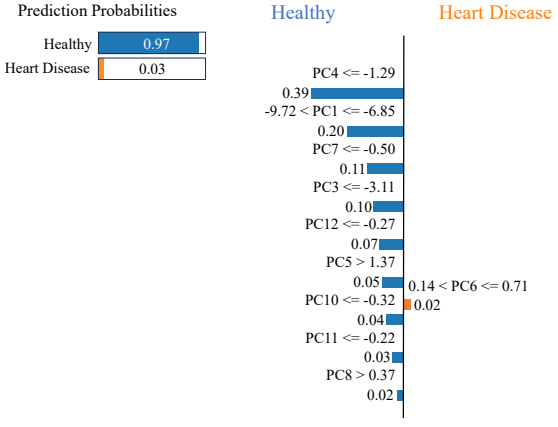
where $\Omega(g)$ is a regularization term that ensures the simplicity of the model g, making it interpretable. LIME is model-agnostic, meaning it does not rely on the internal structure or parameters of the predictive model. It only



(a) LIME Prediction of Stacking Model for Heart Disease Prediction

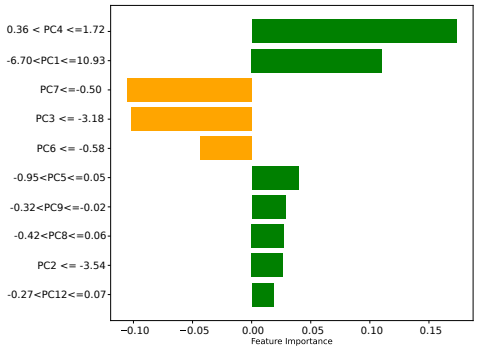


(b) LIME Prediction of OSM-BO for Heart Disease Prediction

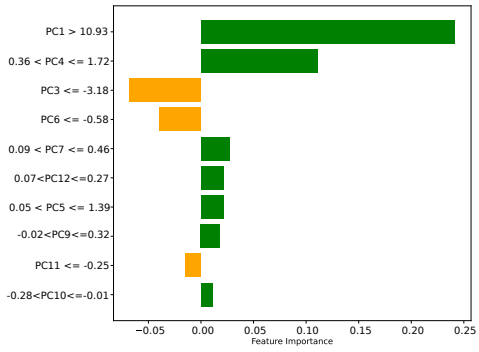


(c) LIME Prediction of EAL-OSM for Heart Disease Prediction

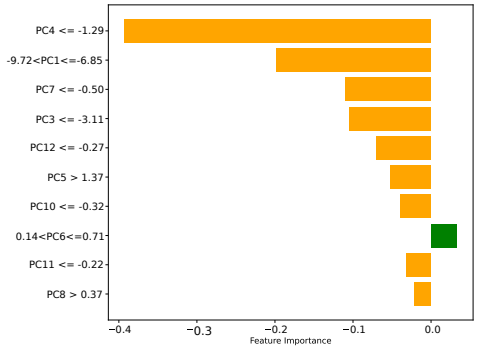
Figure 9: LIME Predictions for Proposed Models



(a) LIME Summary Plot of Stacking Model for Heart Disease Prediction



(b) LIME Summary Plot of OSM-BO for Heart Disease Prediction

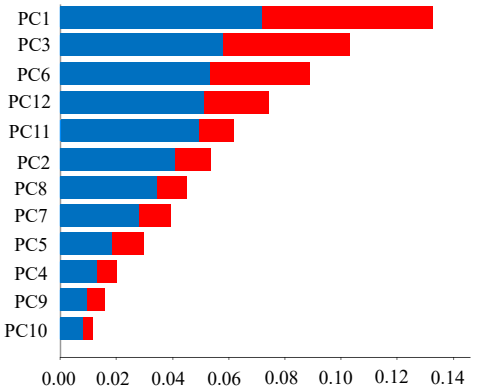


(c) LIME Summary Plot of EAL-OSM for Heart Disease Prediction

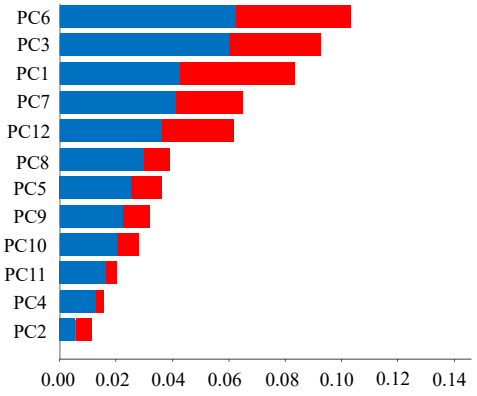
Figure 10: LIME Summary for Proposed Models

requires access to the input-output behavior of the model, making it broadly applicable across different domains and model architectures. While LIME is a powerful tool for interpretability, its explanations can sometimes vary due to the random nature of sample perturbations and may depend on the number of samples generated and the choice of kernel. Despite these limitations, LIME remains a widely used and effective technique for generating understandable, faithful, and actionable explanations of individual predictions.

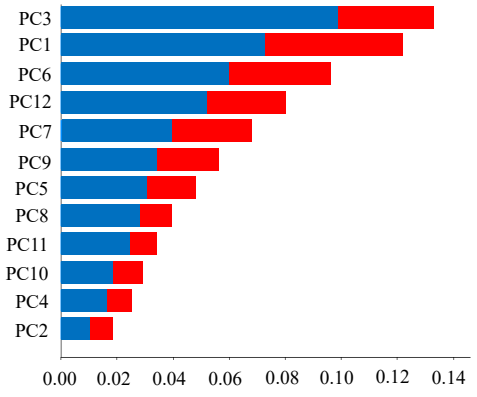
The LIME-based interpretability plot illustrates in Figure 9a the probabilistic reasoning behind the model's prediction for heart disease classification. The model predicts a 0.66 probability of *Heart Disease* and a 0.34 probability of a *Healthy* condition. The major contributing features influencing the prediction include *PC5* in the range $-0.95 < PC5 \leq 0.05$, *PC7* with $PC7 \leq -0.50$, and *PC3* with $PC3 \leq -3.18$, which collectively push the prediction toward the *Heart Disease* class. Conversely, features such as *PC1* ($-6.70 < PC1 \leq 0.93$), *PC12* ($-0.27 < PC12 \leq 0.07$), and *PC6* ($PC6 \leq -0.58$) provide minor opposing



(a) Summary Plot of Stacking Model for Heart Disease Prediction



(b) Summary Plot of OSM-BO for Heart Disease Prediction



(c) Summary Plot of EAL-OSM for Heart Disease Prediction

Figure 11: SHAP Summary for Proposed Models

influence toward the *Healthy* outcome. The feature intervals reveal how subtle variations in principal components shift the model's decision boundary, emphasizing that *PC5*, *PC7*, and *PC3* play dominant roles in increasing disease likelihood, while the remaining features contribute less significantly to the prediction confidence.

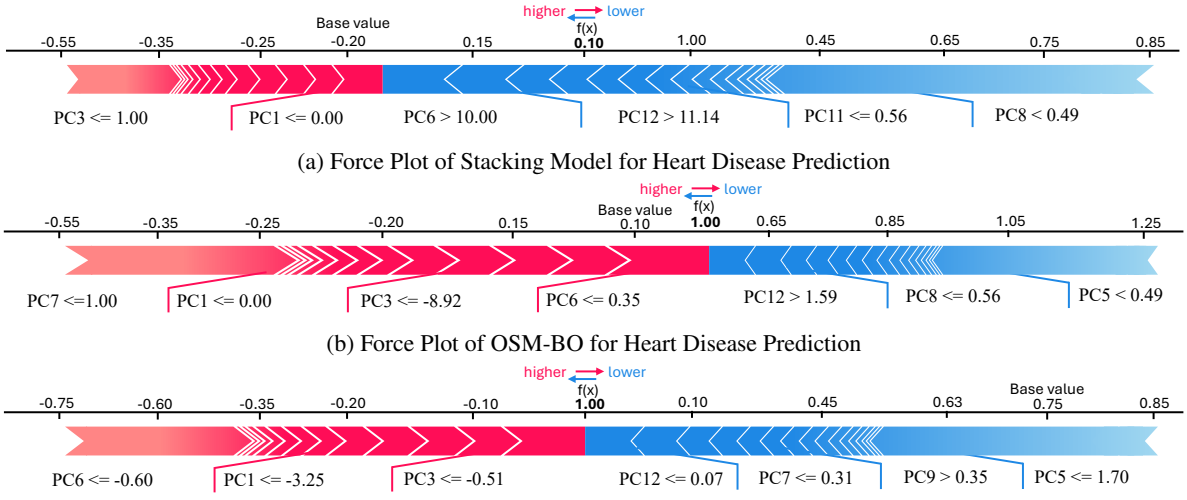
The optimized LIME plot illustrates in Figure 9b presents the model's refined reasoning for heart disease classification. The prediction shows a 0.82 probability for Heart Disease and 0.18 for Healthy. The most influential features contributing positively to heart disease prediction are PC7 in the range $0.09 < PC7 \leq 0.46$, PC4 within $0.36 < PC4 \leq 1.72$, and PC12 in the interval $0.07 < PC12 \leq 0.27$. Additional contributing features include PC5 ($0.05 < PC5 \leq 1.39$) and PC9 ($-0.02 < PC9 \leq 0.32$), which slightly reinforce the disease likelihood. In contrast, PC3 ($PC3 \leq -3.18$), PC6 ($PC6 \leq -0.58$), and PC11 ($PC11 \leq -0.25$) exert a small influence toward the healthy outcome. The results indicate that the optimized model strengthens the importance of PC7 and PC4 while reducing the uncertainty in feature attribution, leading to a more confident disease prediction.

The LIME explanation plot for the EAL-OSM model illustrates in Figure 9c presents the contribution of each principal component feature towards the predicted class. The model predicts a 97% probability for the Healthy class and a 3% probability for Heart Disease. Among the influential features, $PC4 \leq -1.29$ contributes the most towards the Healthy prediction with a weight of 0.39, followed by *PC1* ($-9.72 < PC1 \leq -6.85$) with a weight of 0.20. Other features such as $PC7 \leq -0.50$, $PC3 \leq -3.11$, and $PC12 \leq -0.27$ exhibit smaller positive influences with values of 0.11, 0.10, and 0.07, respectively.

The LIME summary plot in Figure 10a presents an aggregated interpretation of feature contributions across multiple predictions for heart disease classification. It highlights that *PC5*, *PC7*, and *PC3* are the most influential features, consistently driving the model's predictions toward the *Heart Disease* class. In contrast, features such as *PC1*, *PC12*, and *PC6* exhibit relatively lower influence, occasionally supporting the *Healthy* outcome. The distribution of feature effects in the summary plot indicates that higher absolute values of *PC5* and *PC7* correspond to an increased probability of heart disease, reaffirming their importance in shaping the model's decision pattern.

The optimized LIME summary plot as shown in Figure 10b of OSM-BO model provides an overall view of feature importance across multiple predictions. It demonstrates that *PC1, PC4*, and *PC7* have the highest contribution toward the heart disease classification, consistently increasing the predicted probability of the disease. Features such as *PC5* and *PC9* also show moderate influence, supporting the same outcome in several instances. On the other hand, *PC3*, *PC6*, and *PC11* appear less impactful and occasionally push the prediction toward the healthy class. The feature distribution in the summary plot confirms that the optimized model emphasizes more discriminative components, particularly PC7 and PC4, resulting in improved interpretability and stable decision behavior across different samples.

The summary plot for the EAL-OSM model presents the overall influence of each principal component on the model's predictions. The most significant features include *PC4*, *PC1*, and *PC6*. Among them, *PC4* and *PC1* strongly contribute towards the Healthy class, indicating that their lower values are associated with a reduced risk of heart disease. In contrast, *PC6* is the only feature that contributes towards the Heart Disease class, where higher values



(a) Force Plot of Stacking Model for Heart Disease Prediction

(b) Force Plot of OSM-BO for Heart Disease Prediction

(c) Force Plot of EAL-OSM for Heart Disease Prediction

Figure 12: SHAP Force Plot for Proposed Models

increase the likelihood of disease prediction. Features such as *PC7*, *PC3*, and *PC12* have moderate effects, while *PC5*, *PC10*, *PC11*, and *PC8* show relatively minor influence on the model's output. Overall, the summary plot emphasizes that the model's decisions are primarily driven by variations in *PC4*, *PC1*, and *PC6*, with *PC6* being a key indicator for heart disease risk in the EAL-OSM framework.

3.8. SHapley Additive exPlanations

A cohesive framework grounded in cooperative game theory SHAP allocates a prediction to each feature [44] to elucidate the output of any ML model. It is founded on game theory's Shapley values, which allocate the reward among the players (the input features) equally. The primary objective is to assess the impact of each feature by considering all potential feature combinations [45]. The SHAP value for a feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{20}$$

where, N is the set of all features, S is a subset of features not containing feature i, $f(S)$ is the model's prediction using only the features in subset S, and ϕ_i is the SHAP value for feature i. The equation computes the average marginal contribution of feature i over all possible feature subsets S. SHAP works by evaluating the model multiple times, once for each subset, to estimate each feature's impact on the prediction [46]. This makes SHAP model-agnostic and interpretable, as it provides a global understanding of feature importance and local explanations for individual predictions.

The SHAP summary plot illustrated in 11a is the stacking-based heart disease prediction reveals that the features *PC1*, *PC3*, *PC6*, and *PC12* exhibit the highest SHAP values, indicating that these components contribute most significantly to the model's predictions. Their strong influence suggests that these PCs capture critical latent patterns related to heart disease risk. Other features such as *PC8*, *PC7*, *PC4*, *PC5*, *PC2*, *PC9*, and *PC10* demonstrate relatively lower SHAP magnitudes, implying a smaller yet complementary role in refining decision boundaries. Overall, the stacking model effectively integrates these features to enhance interpretability and achieve robust discrimination between healthy and heart disease cases.

The SHAP summary plot shown in Figure 11b for the OSM-BO model demonstrates that the features *PC6*, *PC3*, and *PC1* have the highest SHAP values, indicating their dominant contribution to heart disease prediction. These features encapsulate critical latent representations that significantly influence the model's decision-making process. In contrast, features such as *PC2*, *PC4*, and *PC11* exhibit the lowest SHAP values, reflecting their minimal impact on the overall prediction outcome. This distribution highlights that the OSM-BO model primarily relies on a subset of principal components that effectively capture the underlying patterns associated with heart disease risk, while less informative components contribute marginally to the prediction process.

The SHAP summary plot shown in Figure 11c for the EAL-OSM model indicates that the features *PC3*, *PC1*, and *PC6* are the top contributing components, showing the highest SHAP values and thus exerting the strongest influence on the heart disease prediction. These features represent the most informative latent dimensions derived from the input space, effectively capturing the nonlinear dependencies linked with disease risk. Conversely, the features *PC2*, *PC4*, and *PC10* possess the lowest SHAP values, suggesting their limited contribution to the model's output. This pattern reveals that the EAL-OSM model primarily depends on a few dominant principal components to achieve accurate and interpretable predictions of heart disease.

The SHAP force plot as shown in Figure 12a for the stacking

model illustrates how individual feature contributions collectively influence the prediction outcome for heart disease. Starting from the base value, the features $PC8 \leq 0.49$, and $PC6 > 10.00$ exert a strong positive impact, driving the model output toward a higher probability of heart disease. Additionally, the conditions $PC12 > 11.14$ and $PC11 \leq 0.56$ further enhance this effect, highlighting their relevance in identifying high-risk instances. On the other hand, features such as $PC1 \leq 0.00$ and $PC3 \leq 1.00$ contribute negatively, pulling the prediction toward the lower end, thus representing protective or less risky patterns. The combined influence of these features reflects the model's interpretability.

The SHAP force plot AS shown in Figure 12b for the OSM-BO model demonstrates how individual feature values influence the prediction outcome for heart disease. The features *PC12* and *PC8* play dominant roles, where higher values of *PC12* (> 1.59) and lower values of *PC8* (≤ 0.56) notably increase the model's prediction confidence. Similarly, *PC5* (> 0.49) also contribute positively to the predicted outcome. Conversely, conditions such as *PC1* (≤ 0.00) and *PC3* (≤ -8.29) reduce the output probability, indicating their negative influence. Overall, the force plot highlights how feature interactions in the OSM-BO model balance positive and negative contributions to achieve the final prediction outcome.

The Figure 12c shows how each feature shifts the prediction higher or lower from this base value. In this force plot, features such as *PC9* (≤ 0.35), *PC7* (> 0.31), and *PC5* (≤ 1.70) push the prediction higher, contributing positively to the predicted value. On the other hand, features like *PC1* ($\leq -3,25$) and *PC6* reduce the predicted probability, pulling the output lower. The plot effectively demonstrates how each feature's value drives the model's output, either by increasing or decreasing the likelihood of the predicted outcome, providing a clear view of the feature's impact on the final prediction.

3.9. Comparative Analysis of LIME and SHAP

SHAP and LIME are two popular eXplainable AI systems that are used to interpret ML models. Both methods seek to give an insight into the decision making process of the black box models, however, in radically different ways. This section compares SHAP and LIME on the basis of their methodology, their strengths, weaknesses and their application. SHAP relies on Shapley values, which is an idea borrowed off of cooperative game theory. Shapley values give an opportunity to allocate the contribution of every feature to the prediction of the model fairly and taking into account all the possible combinations of features. The most important benefit of SHAP is the fact that this algorithm is theoretically justified, which means that the contribution of features is distributed equally. The additive nature of SHAP values implies that the contribution of features will all add up to the prediction of the model. This offers a universally compact explanations of the influences of each feature to the prediction, both in single cases and globally in the entire dataset. SHAP is model-agnostic, which is why it can be implemented to any ML model, but especially complex models like DL and ensemble models are well-adapted to SHAP. The consistency of SHAP and its capability to give specific explanations, which add up to model output, can be considered one of the greatest benefits of SHAP over LIME. On the other hand, LIME is an approximation-based method that explains the predictions of black-box models by fitting a locally interpretable model around the prediction of interest. LIME perturbs the input data and observes how the model's output changes. It then fits a simple, interpretable model (such as linear regression) to approximate the decision boundary of the complex model within the local vicinity of the observation. This approach allows LIME to provide local explanations, meaning it explains individual predictions without considering global feature importance. The advantage of LIME is that it is computationally less expensive for individual predictions compared to SHAP, making it more suitable for situations where real-time or interactive interpretation is required. However, LIME's explanations may vary depending on the choice of local surrogate model and the randomness in the perturbation process. SHAP and LIME both provide valuable insights into ML models but are suited to different types of problems. SHAP is best used when a theoretical and consistent global and local meaning is required, and when there is sufficient computing power to compute the actual Shapley values. LIME, in its turn, has much quicker and more adaptable explanations but loses some consistency and reliability. SHAP versus LIME is ultimately a matter of trade-off between the needs of the analysis, i.e. between the computational cost and the accuracy of the explanation, or between the global and local interpretability.

Practical implementation of SHAP and LIME in application to a real clinical prediction system can help tremendously to improve the transparency and reliability of AI-helpful decision-making. These interpretable methods of AI are applicable in the healthcare environment where medical practitioners can decode the complicated model predictions and interpret the relative impact of every clinical aspect. SHAP gives global and local interpretability by giving each feature values of importance giving a measure of the contribution made by the variables like; *age, blood pressure, BMI,* and *diabetes* to the risk of a heart disease. On the other hand, LIME creates local surrogate models on individual predictions, so that clinicians can look at them and confirm why one particular patient is considered as high- or low-risk. These techniques can be combined together to provide clinical decision-support dashboards, which will provide insight at both patient and population levels. Such interpretability facilitates trust, supports ethical accountability, and empowers healthcare professionals to make evidence-based decisions by correlating model explanations with medical reasoning, thus ensuring that AI-driven systems augment rather than replace expert clinical judgment.

4. Discussion

The predictive systems of heart disease development should find a balance between accuracy, efficiency and clarity in an actual clinical setting where it is essential to interpret the predicted results. This study proposed a componental architecture consisting of sophisticated ensemble learning, intelligent sampling, and explainable AI to overcome the long-standing problems in the data imbalance, high-dimension features, few annotations, and automated tuning. The systematic ensemble of synthetic sample generation, dimensionality reduction, hyperparameter optimization and EAL, the proposed framework provides a viable solution that minimizes the annotation expenses and maintains the reliability of the decision-making. Furthermore, explainable AI tools allow clinicians to interpret and have confidence in the model decisions, which can bolster the expansion of ML in delicate medical areas. The framework is also based on modularity and adaptability, which means it can be used in various healthcare data and applicable cases, thereby leading to responsible AI creation in medical diagnostics.

5. Conclusion, Future Work, and Limitations

This study introduces a new framework for predicting heart disease by solving major problems like imbalance of data, excessive dimensionality, scarcity of labeled data, and optimization of hyperparameters. The three most important aspects that are incorporated in the proposed pipeline include ProWRAS based balancing, PCA based dimensionality reduction, and an ensemble of strong classification models. The design of each stage is intended to improve the performance of a model and its interpretability without affecting computational efficiency. The comparative analysis of the proposed models, stacking, OSM-BO, and EAL-OSM, demonstrates notable improvements over existing classifiers across multiple evaluation metrics. The stacking model achieved an accuracy of 0.85, F1-score of 0.85, precision of 0.84, recall of 0.87, ROC-AUC of 0.93, and PR-AUC of 0.93. In addition, it obtained a MCC of 0.71, Cohen's kappa of 0.71, and a Hamming loss of 0.14. These results reflect the strength of ensemble learning in enhancing predictive power. The OSM-BO model further improved these values by incorporating hyperparameter optimization, yielding an accuracy of 0.88, F1-score of 0.88, precision of 0.85, recall of 0.94, ROC-AUC of 0.95, and PR-AUC of 0.95. It also achieved an MCC of 0.77, Cohen's kappa of 0.76, and a Hamming Loss of 0.11, indicating better generalization and reduced classification errors compared to the stacking model. The EAL-OSM model demonstrated the best overall performance by integrating active learning, which efficiently selected informative samples to reduce the need for large labeled datasets. It achieved an accuracy of 0.90, F1-score of 0.89, precision of 0.88, recall of 0.95, ROC-AUC of 0.96, and PR-AUC of 0.97. Moreover, it reached the highest MCC and Cohen's kappa values of 0.79 each, along with the lowest Hamming loss of 0.10 among all the models, validating its superior effectiveness in heart disease prediction. 10-FCV and paired t-tests confirmed the statistical significance and robustness of these results. Furthermore, the incorporation of LIME and SHAP added interpretability and transparency to the proposed framework, allowing for both local and global explanations of the models' behavior, which is essential for clinical acceptance. The proposed framework not only enhances classification performance but also offers practical utility for real-world deployment in digital healthcare systems. Its active learning mechanism minimizes annotation effort, while explainability tools such as SHAP and LIME provide transparent decision support making it suitable for integration into telemedicine platforms and clinical diagnostic tools.

Despite the promising performance of the proposed models, certain limitations remain. First, the framework has been validated on a single publicly available dataset, which may limit its generalizability across diverse demographic or clinical settings. Second, while the active learning mechanism effectively reduces labeling effort, it still depends on the quality and representativeness of the initial labeled samples. Third, the current study focuses on binary classification and does not explicitly address multi-class or multi-label disease prediction scenarios. Finally, the models were evaluated in an offline environment, and real-time deployment and clinical validation were beyond the current study's scope.

The proposed framework is not only effective but also generalizable, offering a scalable solution for real-world clinical applications. Future work may explore the integration of domain-specific medical knowledge, implementation of transfer learning for cross-domain generalization, and extension to multi-class or multi-label disease prediction. This research paves the way toward more reliable, interpretable, and trustworthy AI-driven systems in medical decision-making.

Declarations:

Ethical approval: Not applicable.
Competing interests: Authors declare no competing interests.

Funding:

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2501).

Acknowledgment:

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding this work through (grant number IMSIU-DDRSP2501).

Availability of Data

The dataset used in this study is publicly accessible and can be found at: https://www.kaggle.com/

datasets/alexteboul/heart-disease-health-\ indicators-dataset (Accessed on: 07-11-2025).

References

[1] Ponikowski, P., Anker, S.D., AlHabib, K.F., Cowie, M.R., Force, T.L., Hu, S., Jaarsma, T., Krum, H., Rastogi, V., Rohde, L.E. and Samal, U.C., 2014. Heart failure: preventing disease and death worldwide. ESC heart failure, 1(1), pp.4-25.

[2] Alam, A. and Muqeem, M., 2024. An optimal heart disease prediction using chaos game optimization-based recurrent neural model. International Journal of Information Technology, 16(5), pp.3359-3366.

[3] Lu, L., Liu, M., Sun, R., Zheng, Y. and Zhang, P., 2015. Myocardial infarction: symptoms and treatments. Cell biochemistry and biophysics, 72, pp.865-867.

[4] Rabadia, J.P., Thite, V.S., Desai, B.K., Bera, R.G. and Patel, S., 2024. Cardiovascular System, Its Functions and Disorders. In Cardioprotective Plants (pp. 1-34). Singapore: Springer Nature Singapore.

[5] Manikandan, G., Pragadeesh, B., Manojkumar, V., Karthikeyan, A.L., Manikandan, R. and Gandomi, A.H., 2024. Classification models combined with Boruta feature selection for heart disease prediction. Informatics in Medicine Unlocked, 44, p.101442.

[6] Bizimana, P.C., Zhang, Z., Asim, M., El-Latif, A.A.A. and Hammad, M., 2024. Learning-based techniques for heart disease prediction: a survey of models and performance metrics. Multimedia Tools and Applications, 83(13), pp.39867-39921.

[7] Pitchal, P., Ponnusamy, S. and Soundararajan, V., 2024. Heart disease prediction: Improved quantum convolutional neural network and enhanced features. Expert Systems with Applications, 249, p.123534.

[8] Olsen, C.R., Mentz, R.J., Anstrom, K.J., Page, D. and Patel, P.A., 2020. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. American Heart Journal, 229, pp.1-17.

[9] Khozeimeh, F., Alizadehsani, R., Shirani, M., Tartibi, M., Shoeibi, A., Alinejad-Rokny, H., Harlapur, C., Sultanzadeh, S.J., Khosravi, A., Nahavandi, S. and Tan, R.S., 2023. ALEC: active learning with ensemble of classifiers for clinical diagnosis of coronary artery disease. Computers in Biology and Medicine, 158, p.106841.

[10] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A.M. and Qasem, S.N., 2024. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics, 14(2), p.144.

[11] Hasib, K.M., Towhid, N.A. and Islam, M.R., 2021. Hsdlm: a hybrid sampling with deep learning method for imbalanced data classification. International Journal of Cloud Applications and Computing (IJCAC), 11(4), pp.1-13.

[12] Stonier, A.A., Gorantla, R.K. and Manoj, K., 2024. Cardiac disease risk prediction using machine learning algorithms. Healthcare Technology Letters, 11(4), pp.213-217.

[13] Ahmad, S., Asghar, M.Z., Alotaibi, F.M. and Alotaibi, Y.D., 2023. Diagnosis of cardiovascular disease using deep learning technique. Soft Computing, 27(13), pp.8971-8990.

[14] Al-Ssulami, A.M., Alsorori, R.S., Azmi, A.M. and Aboalsamh, H., 2023. Improving coronary heart disease prediction through machine learning and an innovative data augmentation technique. Cognitive Computation, 15(5), pp.1687-1702.

[15] El-Hasnony, I.M., Elzeki, O.M., Alshehri, A. and Salem, H., 2022. Multi-label active learning-based machine learning model for heart disease prediction. Sensors, 22(3), p.1184.

[16] de Oliveira, H.R., Vieira, A.W., Santos, L.I., Ekel, P.Y. and D'Angelo, M.F.S., 2024. A fuzzy transformation approach to enhance active learning for heart disease prediction. Journal of Intelligent & Fuzzy Systems, (Preprint), pp.1-17.

[17] Halder, A. and Kumar, A., 2019. Active learning using rough fuzzy classifier for cancer prediction from microarray gene expression data. Journal of biomedical informatics, 92, p.103136.

[18] Zhu, R., Chen, Y., Peng, W. and Ye, Z.S., 2022. Bayesian deep-learning for RUL prediction: An active learning perspective. Reliability Engineering & System Safety, 228, p.108758.

[19] Bajpai, A.; Sinha, S.; Yadav, A.; Srivastava, V. Early prediction of cardiac arrest using hybrid machine learning models. In Proceedings of the 2023 17th International Conference on Electronics Computer and Computation (ICECCO), Kaskelen, Kazakhstan, 1–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.

[20] Hasib, K.M., Iqbal, M.S., Shah, F.M., Mahmud, J.A., Popel, M.H., Showrov, M.I.H., Ahmed, S. and Rahman, O., 2020. A survey of methods for managing the classification and solution of data imbalance problem. arXiv preprint arXiv:2012.11870.

[21] Huang, J., Cai, Y., Wu, X., Huang, X., Liu, J. and Hu, D., 2024. Prediction of mortality events of patients with acute heart failure in intensive care unit based on deep neural network. Computer Methods and Programs in Biomedicine, 256, p.108403.

[22] Bej, S., Schulz, K., Srivastava, P., Wolfien, M. and Wolkenhauer, O., 2021. A multi-schematic classifier-independent oversampling approach for imbalanced datasets. IEEE Access, 9, pp.123358-123374.

[23] Ray, P., Reddy, S.S. and Banerjee, T., 2021. Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, 54(5), pp.3473-3515.

[24] Bharadiya, J.P., 2023. A tutorial on principal component analysis for dimensionality reduction in machine learning. International Journal of Innovative Science and Research Technology, 8(5), pp.2028-2032.

[25] Rahmat, F., Zulkafli, Z., Ishak, A.J., Abdul Rahman, R.Z., Stercke, S.D., Buytaert, W., Tahir, W., Ab Rahman, J., Ibrahim, S. and Ismail, M., 2024. Supervised feature selection using principal component analysis. Knowledge and Information Systems, 66(3), pp.1955-1995.

[26] Rajabinasab, M., Pakdaman, F., Zimek, A. and Gabbouj, M., 2025. Randomized PCA forest for approximate k-nearest neighbor search. Expert Systems with Applications, 281, p.126254.

[27] Ravinder, B., Seeni, S.K., Prabhu, V.S., Asha, P., Maniraj, S.P. and Srinivasan, C., 2024, February. Web data mining with organized contents using naive bayes algorithm. In 2024 2nd International Conference on Computer, Communication and Control (IC4) (pp. 1-6). IEEE.

[28] Abbas, M.A., Munir, K., Raza, A., Amjad, M., Samee, N.A., Jamjoom, M.M. and Ullah, Z., 2024. A novel meta learning based stacked approach for diagnosis of thyroid syndrome. PLoS One, 19(11), p.e0312313.

[29] Garouani, M. and Bouneffa, M., 2024. Automated machine learning hyperparameters tuning through meta-guided Bayesian optimization. Progress in Artificial Intelligence, pp.1-12.

[30] Wang, H. and Yang, K., 2023. Bayesian optimization. In Many-Criteria Optimization and Decision Analysis: State-of-the-Art, Present Challenges, and Future Perspectives (pp. 271-297). Cham: Springer International Publishing.

[31] Xie, T., Zhu, J., Ma, G., Lin, M., Chen, W., Yang, W. and Liu, S., 2024. Structural-Entropy-Based Sample Selection for Efficient and Effective Learning. arXiv preprint arXiv:2410.02268.

[32] Khan, H., Javaid, N., Bashir, T., Ali, Z., Khan, F.A. and Pamucar, D., 2025. A novel deep gated network model for explainable diabetes mellitus prediction at early stages. Knowledge-Based Systems, Article 114178.

[33] Ibnath, M.H., Hasib, K.M., Parvez, M.R. and Mridha, M.F., 2025, February. Enhancing Multi-Emotion Detection in Text: A Comparative Study of Feature Extraction Techniques and Machine Learning Models. In 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

[34] Patra, S.C., Maheswari, B.U. and Pati, P.B., 2023. Forecasting coronary heart disease risk with a 2-step hybrid ensemble learning method and forward feature selection algorithm. IEEE Access, 11, pp.136758-136769.

[35] Waheed, Z., Gui, J., Heyat, M.B.B., Parveen, S., Hayat, M.A.B., Iqbal, M.S., Aya, Z., Nawabi, A.K. and Sawan, M., 2025. A novel lightweight deep learning based approaches for the automatic diagnosis of gastrointestinal disease using image processing and knowledge distillation techniques. Computer Methods and Programs in Biomedicine, 260, p.108579.

[36] Rajendran, R. and Karthi, A., 2022. Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers. Expert Systems with Applications, 207, p.117882.

[37] Shaheen, I., Javaid, N., Rahim, A., Alrajeh, N. and Kumar, N., 2025. Empowering early predictions: A paradigm shift in diabetes risk assessment with Deep Active Learning. Knowledge-Based Systems, 315, p.113284.

[38] Elkari, B., Chaibi, Y. and Kousksou, T., 2024. Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems. Energy Reports, 12, pp.988-999.

[39] Yu, Z., Guindani, M., Grieco, S.F., Chen, L., Holmes, T.C. and Xu, X., 2022. Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron, 110(1), pp.21-35.

[40] Karan, E. and Brown, L., 2022. Enhancing Student's Problem-Solving Skills through Project-Based Learning. Journal of Problem Based Learning in Higher Education, 10(1), pp.74-87.

[41] Javed, A., Javaid, N., Hasnain, M., Sarfraz, U., Ahmed, I., Shafiq, M. and Choi, J.G., 2024. Applying advanced data analytics on pregnancy complications to predict miscarriage with explainable AI. IEEE Access.

[42] Lisboa, P.J., Saralajew, S., Vellido, A., Fernández-Domenech, R. and Villmann, T., 2023. The coming of age of interpretable and explainable machine learning models. Neurocomputing, 535, pp.25-39.

[43] Wang, B., Pei, W., Xue, B. and Zhang, M., 2025. Explaining deep convolutional neural networks for image classification by evolving local interpretable model-agnostic explanations. IEEE Transactions on Emerging Topics in Computational Intelligence.

[44] Antwarg, L., Miller, R.M., Shapira, B. and Rokach, L., 2021. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert systems with applications, 186, p.115736.

[45] Shaheen, I., Javaid, N., Ali, Z., Ahmed, I., Khan, F.A. and Dragan, D., 2025. A transparent and robust framework for early diabetes prediction using deep residual networks and proximity-based data balancing. Biomedical Signal Processing and Control.

[46] Liu, Y., Liu, Z., Luo, X. and Zhao, H., 2022. Diagnosis of Parkinson's disease based on SHAP value feature selection. Biocybernetics and Biomedical Engineering, 42(3), pp.856-869.