

Comparative Analysis of Ensemble Learning Models for Cardiovascular Disease Using Explainable Artificial Intelligence

Syifa Alfariani ^{1*}, Fathoni ^{2*}

* Sistem Informasi, Universitas Sriwijaya
09031282328096@student.unsri.ac.id¹, fathoni@unsri.ac.id²

Article Info

Article history:

Received 2026-09-10

Revised 2026-09-25

Accepted 2026-10-06

Keyword:

Cardiovascular Disease,
Ensemble Learning,
Stacking,
Explainable Artificial
Intelligence,
SHAP,
LIME.

ABSTRACT

Cardiovascular disease is one of the leading causes of death worldwide, making early risk detection essential to support clinical decision-making. Although machine learning models have been widely applied to cardiovascular disease prediction, challenges remain in achieving consistent performance improvements through ensemble learning and providing interpretable prediction results. This study aims to develop a cardiovascular disease risk prediction model by integrating ensemble learning and Explainable Artificial Intelligence (XAI) using the Cardiovascular Disease Dataset from Kaggle, consisting of 70,000 patient records. The research process includes Exploratory Data Analysis, data preprocessing, and model development using Random Forest, CatBoost, and LightGBM combined using a stacking approach with Logistic Regression as the meta-learner. Model interpretability was analyzed using SHAP and LIME, while performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The results show that LightGBM achieved the highest accuracy of 74.13%, recall of 70.48%, and F1-score of 72.95%, while Random Forest achieved the highest precision of 76.21%. The stacking model achieved the highest AUC-ROC of 80.92%, although it did not consistently outperform individual models across all evaluation metrics. The interpretability analysis identified systolic blood pressure (ap_hi), age, cholesterol, diastolic blood pressure (ap_lo), and body mass index (BMI) as the most influential features in the predictions. This study demonstrates the integration of stacking ensemble learning and XAI using SHAP and LIME, resulting in a model with competitive predictive performance and improved interpretability for cardiovascular disease risk assessment.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Cardiovascular disease (CVD) refers to a group of health conditions that affect the heart and blood vessel system [1]. According to the World Health Organization (2025), cardiovascular disease is the leading cause of death worldwide, with approximately 19.8 million people losing their lives to the disease in 2022. Of these deaths, 85% were attributed to heart attacks and strokes. CVD is not a single disease but encompasses a wide range of conditions, including coronary heart disease, cerebrovascular disease, peripheral artery disease, and congenital heart defects [2]. According to the Survei Kesehatan Indonesia (SKI) in 2023, hypertension and diabetes mellitus are two conditions that

require serious attention, given their high prevalence in the population and their significant contribution as major risk factors for CVD [3].

Despite ongoing efforts to develop early detection methods, CVD risk prediction still faces several fundamental challenges. Medical data are generally heterogeneous and high-dimensional, with complex and not necessarily linear relationships among variables. Consequently, conventional statistical approaches that rely on assumptions of linear relationships between variables may fail to capture the underlying patterns of cardiovascular risk [4]. In addition, cardiovascular disease datasets commonly suffer from class imbalance, in which the number of healthy individuals substantially exceeds the number of individuals at risk. This

condition can cause predictive models to favor the majority class and consequently reduce their ability to identify high-risk cases [5]. Furthermore, machine learning models that demonstrate high predictive performance, particularly ensemble and deep learning models, often function as black-box systems. This makes their predictions difficult for medical professionals to interpret and poses a barrier to their implementation in clinical settings, where transparency in decision-making is essential [4].

To address the challenges posed by nonlinear relationships and data complexity, ensemble learning approaches have been widely used because of their ability to capture complex patterns more effectively than individual models and their consistent performance across various CVD studies. Alqahtani et al. (2022) combined machine learning and deep learning models using a majority voting scheme on the Kaggle Cardiovascular Disease dataset and achieved a best accuracy of 88.70% [6]. Another popular ensemble learning approach is stacking, which combines multiple base models and subsequently integrates their predictions through a meta-learner. Tiwari et al. (2022) developed a stacking model consisting of Random Forest, Gradient Boosting, XGBoost, and Extra Trees, achieving an accuracy of 92.34% [7]. Navita et al. (2025) also developed a stacking approach using Random Forest, K-Nearest Neighbors (KNN), and AdaBoost, with Logistic Regression serving as the meta-learner. By incorporating the SMOTEENN data-balancing technique, their model achieved an accuracy of 97.8% [8]. Meanwhile, Yazdi and Asadi (2025) demonstrated that an ensemble model specifically designed to address class imbalance could achieve superior performance compared with other models when evaluated on seven heart disease datasets [5]. However, these three studies did not incorporate an Explainable Artificial Intelligence (XAI) approach, leaving their predictions difficult to interpret and explain.

Several other studies have incorporated XAI into CVD prediction models. Shah et al. (2025) developed a stacking model using Gradient Boosting, CatBoost, LightGBM, SVM, Random Forest, and Neural Network, with XGBoost serving as the meta-learner, and explained the model predictions using SHAP [4]. Kiran et al. (2025) employed two XAI techniques, namely SHAP and LIME, on the Kaggle Cardiovascular Disease dataset and achieved an accuracy of 98%, however, the model did not employ a stacking approach [9]. Dammak et al. (2025) also combined SHAP and LIME, but their study used a considerably smaller dataset consisting of only 1,025 patients and did not employ a stacking model [10]. Uddin et al. (2026) applied a voting ensemble approach incorporating SHAP and LIME, achieving an accuracy of 98.63% and evaluating the model on 303 data instances [11]. Skouteli et al. (2024) investigated an XAI-based approach using the Kaggle Cardiovascular Disease dataset and achieved an accuracy of approximately 72% [12].

Among the previous studies, three studies (Alqahtani et al., 2022; Kiran et al., 2025; Skouteli et al., 2024) used the same Kaggle Cardiovascular Disease dataset but reported varying

accuracy results, ranging from 72% to 98%. These differences may have been influenced by variations in the models employed, preprocessing schemes, feature selection methods, or even the potential risks of overfitting and data leakage. This condition indicates the need for a comparative evaluation of models on the Kaggle Cardiovascular Disease dataset to determine which models provide more consistent performance across multiple evaluation metrics. One approach that can be investigated is a stacking ensemble combining Random Forest, CatBoost, and LightGBM as base learners, with Logistic Regression serving as the meta-learner.

In addition to differences in model performance, previous studies also indicate limitations in the simultaneous integration of ensemble learning and XAI. Studies reporting high ensemble performance have generally not been accompanied by comprehensive XAI analysis, whereas studies incorporating XAI have often used limited model combinations, applied only a single XAI technique, or relied on relatively small datasets. Based on the reviewed literature, studies that compare individual models with a stacking ensemble consisting of Random Forest, CatBoost, and LightGBM, with Logistic Regression as the meta-learner, while simultaneously using SHAP and LIME to interpret prediction results on a large dataset such as the Kaggle Cardiovascular Disease dataset, remain limited.

Based on these research gaps, this study focuses on a comparative analysis of the performance of Random Forest, CatBoost, LightGBM, and a stacking ensemble for CVD prediction, while integrating an XAI approach using SHAP and LIME. The stacking ensemble is not only intended to obtain higher performance but also to examine whether combining multiple base learners can improve performance compared with individual models. The performance of each model is compared based on accuracy, precision, recall, F1-score, and AUC-ROC to provide a more comprehensive assessment of their classification capabilities. Furthermore, SHAP and LIME are used to identify the contributions of features to prediction outcomes and to examine the consistency of the interpretation patterns produced by the two XAI methods.

Accordingly, this study aims to develop a CVD prediction model using a stacking ensemble learning approach integrated with XAI and to evaluate its performance based on accuracy, precision, recall, F1-score, and AUC-ROC. In addition to producing a model with competitive classification performance, this study is expected to provide insights into the contributions of features that influence prediction outcomes through SHAP and LIME.

II. METHODS

This study focuses on developing a cardiovascular disease risk prediction model by employing an ensemble learning approach combined with Explainable Artificial Intelligence (XAI). The dataset was sourced from the Cardiovascular

Disease dataset available on Kaggle [13], which consists of 70,000 patient records collected from medical examinations.

The methodology applied in this study comprises several stages, including Exploratory Data Analysis (EDA), data preprocessing, model development using an ensemble learning approach, and model interpretation through Explainable Artificial Intelligence (XAI).

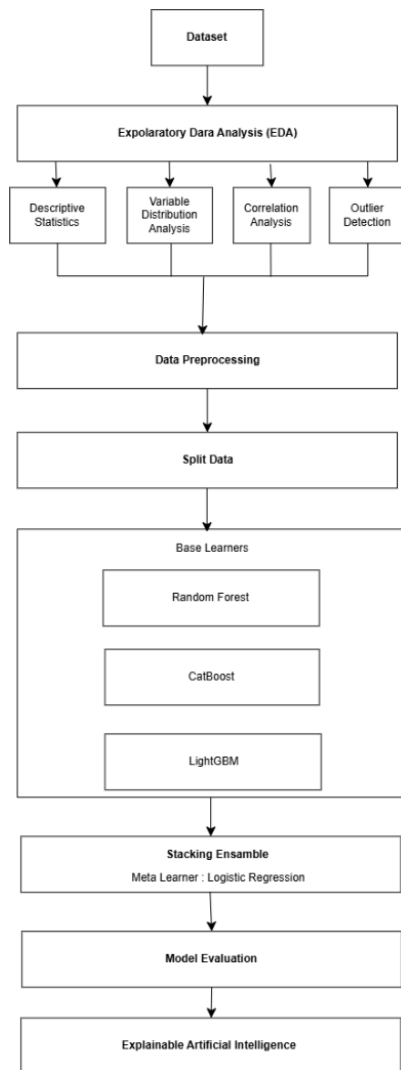


Figure 1. Research Flowchart

A. Data Source

The dataset utilized in this study was sourced from the Kaggle Cardiovascular Disease dataset [13] and comprises 70,000 patient records collected during medical examinations. It contains 13 columns, including 12 input features and one target variable. The input features consist of id, age (patient age in days), gender (patient sex), height (measured in centimeters), weight (measured in kilograms), ap_hi (systolic blood pressure), ap_lo (diastolic blood pressure), cholesterol (cholesterol level, categorized into three levels: 1 for normal, 2 for above normal, and 3 for well

above normal), gluc (glucose level, categorized into three levels: 1 for normal, 2 for above normal, and 3 for well above normal), smoke (smoking status), alco (alcohol consumption), and active (physical activity). The target variable, cardio, is binary and represents the patient's CVD risk status, where 0 indicates that the patient is not at risk and 1 indicates that the patient is at risk.

The dataset was chosen due to its relatively large number of records and the range of factors associated with cardiovascular conditions, including age, blood pressure, cholesterol, glucose levels, smoking habits, alcohol consumption, and physical activity. Its substantial size and diverse features allow for the comparison of multiple machine learning models and the examination of features that contribute to prediction outcomes. Furthermore, the dataset is publicly accessible through Kaggle.

Despite the relatively large number of records, the dataset used in this study consists of secondary data obtained from a publicly available source. Consequently, detailed information on the patients' population characteristics and clinical conditions is limited. Therefore, the findings of this study reflect the models' classification performance on the dataset used and should not yet be regarded as representative of their performance in a broader patient population.

B. EDA

Exploratory Data Analysis (EDA) was conducted to assess the quality of the dataset used in this study and to provide a basis for the subsequent data preprocessing stage according to the research requirements [14]. In this study, EDA was performed through four stages: target variable distribution analysis, descriptive statistics, outlier detection using boxplots, and feature correlation analysis.

- 1) Target Variable Distribution Analysis: This analysis was conducted to determine the distribution of data across each target variable category. It provides both visual and quantitative insights into the data distribution, allowing the characteristics of the dataset to be understood prior to the modeling process [15].
- 2) Descriptive Statistics: This analysis was performed to provide an overview of the characteristics and distribution of each numerical attribute in the dataset. The statistical measures used included the mean, median, standard deviation (std), and minimum and maximum values [16].
- 3) Outlier Detection: Outlier detection was conducted using boxplots to identify extreme values that could potentially affect the modeling process. Boxplots were applied to all attributes in the dataset to illustrate the distribution of values for each attribute [17].
- 4) Correlation Analysis: Correlation analysis was used to measure the relationships among attributes in the dataset and visualize them in the form of a correlation matrix, thereby facilitating the observation of associations between features [18].

C. Preprocessing data

Data preprocessing was conducted to prepare the data for the modeling process. This stage consisted of data inspection and cleaning, with the aim of ensuring the quality of the data used for model training [8].

- 1) Feature Engineering: This stage aimed to create new features that were more suitable for use in the modeling process [19].
- 2) Outlier Handling: Outliers were handled by establishing value thresholds for several key variables based on the characteristics and conditions of the data [20].
- 3) Duplicate Data Removal: Duplicate data were removed by identifying and eliminating rows containing identical information. This step was performed to prevent repeated observations and maintain data quality before the data were used in the subsequent modeling process [20].

D. Split Data

The dataset was then divided into two subsets: training data and testing data [21]. The training data were used to develop and train the model, whereas the testing data were reserved for evaluating its performance in predicting new, previously unseen data that were not involved in the training process [21].

In this study, the dataset was split into 80% training data and 20% testing data using a stratified splitting technique. This approach ensures that the proportion of each class is maintained in both subsets [18], allowing the class distribution to remain consistent between the training and testing data. The training data were then used to train the models and perform 5-fold cross-validation for the construction of the stacking ensemble. In contrast, the testing data were excluded from the training process and used only for the final evaluation of model performance.

E. Ensemble Learning

This study utilized a stacking ensemble learning approach that integrates multiple base models into a single predictive model. In this approach, the predictions produced by each base learner are used as input for a second-level model known as the meta-learner. The meta-learner learns from the prediction patterns of the base learners to produce the final prediction [22].

In this study, Random Forest, CatBoost, and LightGBM were employed as base learners, while Logistic Regression was used as the meta-learner. The stacking process started by dividing the training data into five folds through 5-fold cross-validation. During each iteration, one fold was assigned as the validation set, while the other four folds were used to train the models. This process was repeated five times, allowing each fold to be used once as the validation set.

The predictions generated by the three base learners in each iteration were then collected and combined to construct meta-

features. These meta-features were subsequently provided as input to the Logistic Regression model, which learned the relationships among the predictions of the three base learners to produce the final prediction. Through this process, the final prediction was generated from the combined outputs of the three base learners rather than from a single model [23]. The testing data were excluded from both the cross-validation process and meta-feature construction and were used only at the final stage to evaluate the model's performance.

F. Model Evaluation

Model evaluation was performed to assess the model's ability to accurately classify CVD cases. True Positive (TP) represents positive cases that are correctly classified as positive, while False Positive (FP) refers to negative cases that are incorrectly classified as positive. False Negative (FN) represents positive cases that are incorrectly classified as negative, whereas True Negative (TN) refers to negative cases that are correctly classified as negative [24].

- 1) Accuracy is used to measure the overall correctness of the model in classifying the testing data. It indicates the proportion of instances that are correctly classified by the model [25].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- 2) Precision measures the proportion of instances that are correctly classified as positive among all instances predicted as positive by the model [26].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

- 3) Recall (sensitivity) measures the model's ability to correctly identify instances that belong to the positive class among all actual positive instances [27].

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

- 4) F1-score is used to assess the balance between precision and recall by combining both metrics into a single value. This metric reflects the model's ability to correctly identify positive instances while accounting for errors in classification [26].

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

- 5) AUC-ROC is used to assess the model's ability to distinguish between positive and negative classes. The Receiver Operating Characteristic (ROC) curve illustrates the relationship between Recall (True Positive Rate/TPR) and the False Positive Rate (FPR) at different classification thresholds. The Area Under the Curve (AUC) represents the area beneath the ROC curve. A higher AUC value, particularly one closer to 1, indicates that the model has a stronger ability to distinguish between positive and negative classes [28].

G. XAI

Explainable Artificial Intelligence (XAI) is used to provide a better understanding of predictions produced by machine learning models. Although machine learning models have shown strong potential for disease prediction, their complex nature can make the decision-making process difficult to interpret, leading them to be commonly viewed as black-box models. XAI is therefore applied to explain how predictions are generated and to identify the features that influence the model's decisions [9]. In this study, two XAI methods were employed, namely Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), both of which are used to provide interpretations of the predictions generated by the model.

- 1) SHAP is an XAI method used to assess the contribution of each feature to the predictions generated by a model. It is based on the concept of Shapley values from game theory, which is used to calculate the contribution of individual features to a specific prediction. The resulting SHAP values indicate the degree to which each feature influences the model's prediction [29].

$$\phi_i(f, X) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad (5)$$

where $\phi_i(f, X)$ represents the SHAP value of feature i , N denotes the set of all features, S represents a subset of features that does not include feature i , $|S|$ indicates the number of features in subset S , M represents the total number of features, and $f_x(S)$ represents the model output based on the features included in subset S . The difference $f_x(S \cup \{i\}) - f_x(S)$ represents the marginal contribution of feature $\{i\}$ to the model prediction.

- 2) LIME is used to interpret the predictions of specific instances by analyzing the features that contribute to the model's decision. It uses a simple and interpretable model to approximate the behavior of a complex model around the instance being examined. This approach makes the reasoning behind a particular prediction easier to understand [30].

III. RESULT AND DISCUSSION

This section presents the results of the cardiovascular disease risk prediction model developed using the proposed stacking ensemble learning approach integrated with Explainable Artificial Intelligence (XAI). The results cover exploratory data analysis, data preprocessing, model performance evaluation, and interpretation of the model's predictions using SHAP and LIME. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC, while SHAP and LIME were applied to examine the contribution of individual features to the model's predictions.

A. EDA

Exploratory data analysis was performed on a dataset consisting of 70,000 rows and 13 columns to examine the

characteristics and condition of the data before it was used for the modeling process. The initial inspection revealed that no missing values were found in any of the variables, indicating that no imputation was required in this study. For the cardio variable, 35,021 instances (50.03%) belonged to class 0, representing patients without CVD risk, while 34,979 instances (49.97%) belonged to class 1, representing patients at risk of CVD. The very small difference in class proportions indicates that the two classes were relatively balanced, as shown in Figure 2.

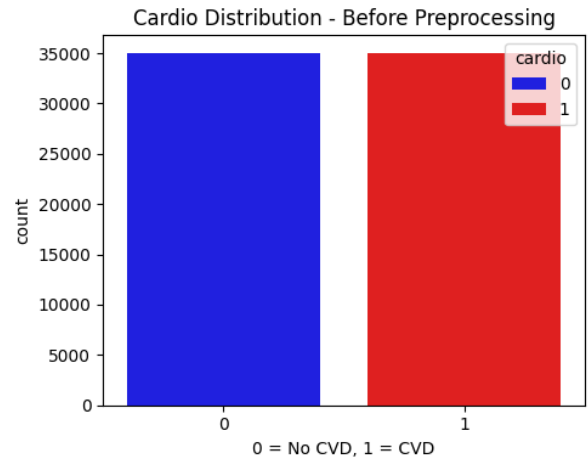


Figure 2. Distribution of the CVD Risk Classes

Although the class distribution was relatively balanced, descriptive statistical analysis revealed extreme values in several features, particularly systolic blood pressure (ap_hi) and diastolic blood pressure (ap_lo). The ap_hi feature had a minimum value of -150 and a maximum value of 16,020, while ap_lo had a minimum value of -70 and a maximum value of 11,000. These values fall outside the plausible range and indicate the presence of outliers or potential data recording errors. Similar conditions were also observed in the height and weight features, which exhibited relatively extreme value ranges. The presence of these extreme values was further confirmed through the boxplot visualization shown in Figure 3, which revealed several data points located far beyond the whisker boundaries.

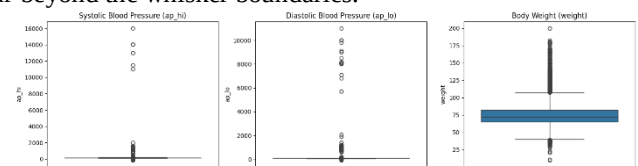


Figure 3. Boxplot Visualization of Feature Distributions

Furthermore, correlation analysis of the data prior to preprocessing showed that age had the strongest correlation with the cardio variable, with a correlation coefficient of 0.24, followed by cholesterol at 0.22 and weight at 0.18. Meanwhile, systolic and diastolic blood pressure showed relatively weak correlations with the target variable, with correlation coefficients of 0.05 and 0.07, respectively. The low correlation values may have been influenced by the

presence of extreme values in both blood pressure features. In addition, a moderate correlation was observed between cholesterol and gluc, with a correlation coefficient of 0.45, whereas the other features showed relatively weak relationships with the target variable, as shown in Figure 4. These findings indicate that preprocessing is necessary to address outliers before the data are used in the modeling stage.

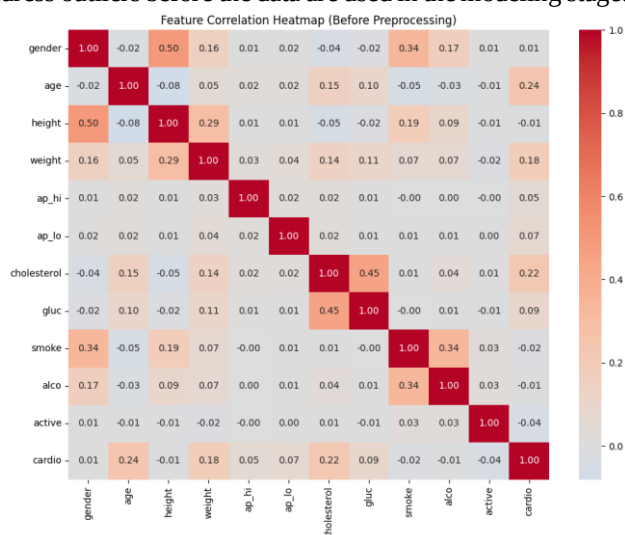


Figure 4. Heatmap Correlation before preprocessing

B. Preprocessing data

Based on the results of the EDA, the next stage involved data preprocessing to prepare the data for the modeling process. This stage began with feature engineering, which involved converting age from days to years (*age_years*) and creating a new feature, BMI, based on body weight and height.

Next, data filtering was performed to remove implausible values. The *ap_hi* feature was restricted to a range of 80–250 mmHg, while *ap_lo* was restricted to a range of 40–200 mmHg. This filtering was based on previous research that performed blood pressure data quality checks to identify invalid measurements [31]. In this study, stricter lower bounds were applied to exclude extremely low blood pressure values.

Filtering was also applied to the height and weight variables, with height restricted to a range of 120–220 cm and weight restricted to a range of 30–200 kg. In addition, the condition $ap_hi > ap_lo$ was applied to ensure that systolic blood pressure was higher than diastolic blood pressure [32]. Duplicate records were subsequently removed to prevent repeated observations. The *id* column was excluded because it was used only as a data identifier, while the *age* column was removed because age information was already represented by the *age_years* feature.

After the cleaning process, the number of records decreased from 70,000 to 68,611. A total of 1,389 records, or 1.98% of the entire dataset, were removed because they did not meet the filtering criteria or were identified as duplicate records. Despite the reduction in the number of records, the

target class distribution remained relatively balanced. Class 0, representing patients without cardiovascular disease risk, consisted of 34,666 records (50.53%), whereas class 1, representing patients at risk of cardiovascular disease, consisted of 33,945 records (49.47%). The small difference in class proportions indicates that the two classes remained balanced. therefore, no additional class-balancing techniques, such as oversampling or undersampling, were required.

After cleaning, correlation analysis was performed to examine the relationships between the numerical variables and the *cardio* target variable. The results are presented in Figure 5. The correlation between *ap_hi* and *cardio* increased from 0.05 to 0.43, while the correlation between *ap_lo* and *cardio* increased from 0.07 to 0.34. These increases suggest that the presence of extreme values in the initial data may have influenced the relationship between blood pressure and cardiovascular disease risk. Meanwhile, *age_years*, *cholesterol*, and BMI showed correlations of 0.24, 0.22, and 0.19, respectively, with *cardio*. Strong correlations were also observed between *ap_hi* and *ap_lo* (0.73) and between weight and BMI (0.86). The strong relationship between weight and BMI is expected, given that body weight is one of the components used to calculate BMI.

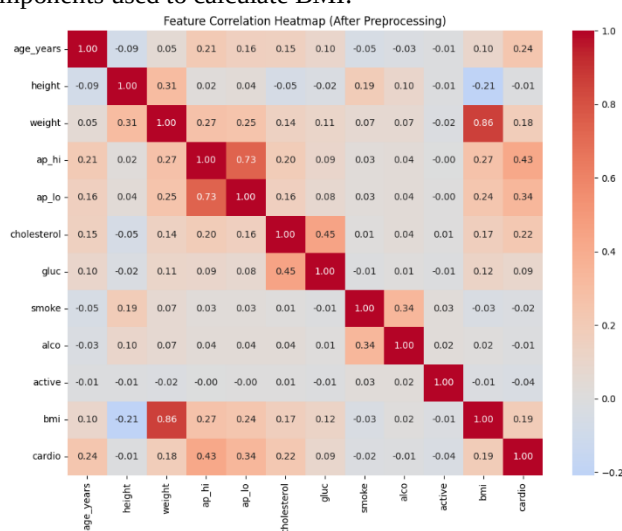


Figure 5. Heatmap Correlation after preprocessing

After the preprocessing stage, the data were divided into training and testing sets using an 80:20 ratio with a stratified split. This approach was selected to maintain the proportion of the target classes across both subsets. The resulting split produced 54,888 training records and 13,723 testing records, with a total of 12 features used as model inputs: *gender*, *height*, *weight*, *ap_hi*, *ap_lo*, *cholesterol*, *gluc*, *smoke*, *alco*, *active*, *age_years*, and BMI.

C. Model Evaluation

The model evaluation process was conducted to compare the performance of Random Forest, CatBoost, LightGBM, and the *stacking ensemble* approach in predicting CVD risk.

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.740 363	0.762106	0.690 823	0.724 716	0.808 833
CatBoost	0.739 343	0.754275	0.701 723	0.727 051	0.808 618
LightGBM	0.741 383	0.755924	0.704 817	0.729 476	0.807 234
Stacking	0.740 436	0.754455	0.704 669	0.728 713	0.809 188

Based on Table I, the four models achieved relatively similar performance. LightGBM obtained the highest accuracy of 74.13%, while Random Forest achieved the highest precision of 76.21%. In terms of recall, LightGBM obtained the highest value of 70.48%, indicating that this model was slightly better at predicting patients at risk of CVD. LightGBM also achieved the highest F1-score of 72.94%, indicating that the model provided the best balance between precision and recall compared with the other models.

Meanwhile, the stacking ensemble achieved the highest AUC-ROC of 80.91%, which was slightly higher than Random Forest at 80.88%, CatBoost at 80.86%, and LightGBM at 80.72%. These results indicate that combining the three base learners through Logistic Regression provided slightly better discriminative ability than each individual model. However, the difference was relatively small and did not indicate an improvement across all evaluation metrics. The stacking ensemble also did not achieve the highest values for accuracy, precision, recall, and F1-score.

Therefore, the use of stacking in this study did not provide consistent improvements across all metrics, although it achieved the highest AUC-ROC. The advantage of stacking was more evident in the model's ability to distinguish between the two classes based on predicted probabilities. These prediction results are consistent with the findings of Shah et al. (2025), who applied a stacking ensemble for CVD prediction and achieved an AUC-ROC of 82% [4].

Confusion Matrix - Stacking (Logistic Regression Meta-Learner)

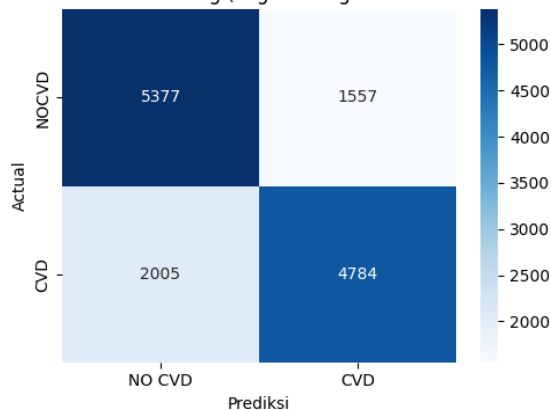


Figure 6. Confusion Matrix Stacking Ensemble

The confusion matrix in Figure 6 shows that the stacking ensemble correctly classified 5,377 instances as not at risk (true negative) and 4,784 instances as at risk (true positive). On the other hand, 1,557 instances that were actually not at risk were incorrectly classified as at risk by the model (false positive), while 2,005 instances that were actually at risk were incorrectly predicted as not at risk (false negative).

In CVD risk prediction, false negatives are particularly critical because they represent patients at risk who are not identified by the model. This is reflected in the recall value of 0.7047, indicating that the model correctly identified approximately 70% of the positive cases. Therefore, although the model achieved an accuracy of 0.7404, challenges remain, particularly regarding its sensitivity in detecting individuals who are truly at risk of CVD.

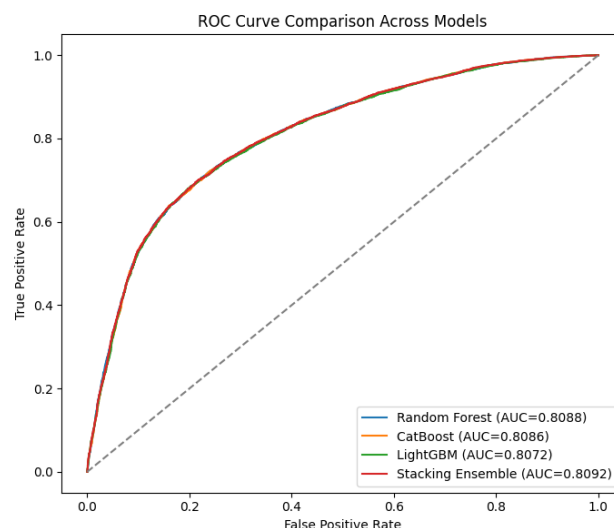


Figure 7. ROC Curves

The comparison of the ROC curves in Figure 7 shows that the curves of the four models are positioned very closely to one another. This is consistent with the AUC-ROC values, which range from 80.72% to 80.92%. The very small differences indicate that the discriminative ability of the four models was relatively similar on the testing data.

Overall, the evaluation results show that there were no substantial differences in performance between the three base learners and the stacking ensemble. LightGBM achieved the best performance on most classification-based metrics, whereas the stacking ensemble achieved the highest AUC-ROC. Therefore, the results indicate that the main advantage of the stacking ensemble was its ability to discriminate between the not-at-risk and at-risk classes based on predicted probabilities, while LightGBM provided better results on most classification metrics.

D. XAI

To understand how the model generates predictions, an analysis was conducted using SHAP and LIME. These two methods were employed to identify the features that had the greatest influence on the model's predictions and to examine the factors contributing to the model's decisions for patients at risk and not at risk of CVD. The results of the analysis are presented in the following figures.

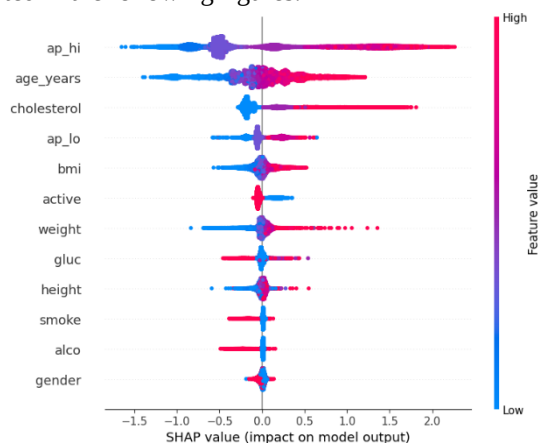


Figure 8. SHAP Summary Plot

Figure 8 presents the SHAP summary plot, which ranks the features based on the magnitude of their average contribution, measured by the mean absolute SHAP value, to the model's predictions [33]. The figure shows that *ap_hi* is the most influential feature in the model's predictions. The features with relatively substantial influence are *age_years*, *cholesterol*, *ap_lo*, and *BMI*. Meanwhile, *active*, *weight*, *gluc*, *height*, *smoke*, *alco*, and *gender* show relatively smaller contributions to the model's predictions.

In addition to showing the contribution of each feature, the color and position of the points in the plot also provide information about the direction of their influence on the predictions. For the *ap_hi* feature, for example, higher values are represented by red points and tend to be located on the positive side of the SHAP values. This indicates that higher systolic blood pressure tends to push the model toward a risk prediction. Conversely, lower *ap_hi* values, represented by blue points, are more frequently located on the negative side, indicating a tendency to support predictions of no risk.

A similar pattern can also be observed for *age_years* and *cholesterol*, where higher values tend to make positive contributions to the model's predictions. Thus, higher age and higher cholesterol levels tend to push the model toward predicting cardiovascular risk. This finding is consistent with a previous study that used the same dataset and showed that *ap_hi*, *age*, *ap_lo*, and *cholesterol* were the features with the greatest influence based on SHAP analysis. In that study, higher values of *ap_hi*, *age*, *ap_lo*, and *cholesterol* also tended to be located on the positive side of the SHAP values, indicating a tendency of the model to predict cardiovascular risk [9].

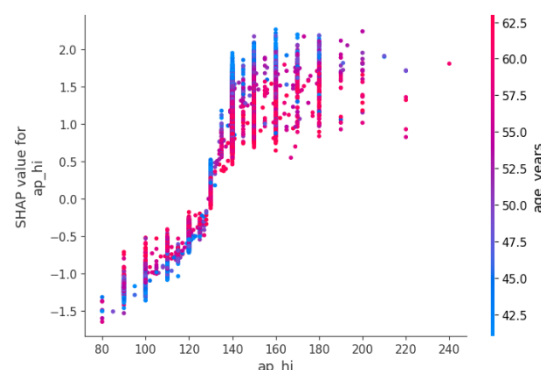


Figure 9. SHAP Dependence Plot

Figure 9 shows the relationship between the *ap_hi* values and the SHAP values generated by the model. The plot shows that the contribution of *ap_hi* to the model's predictions follows a positive and nonlinear pattern. At *ap_hi* values of approximately 80–120 mmHg, the SHAP values are generally located on the negative side. This indicates that systolic blood pressure within this range tends to push the model toward a not-at-risk prediction. As the *ap_hi* value increases, the SHAP value also increases and begins to approach zero at approximately 120–135 mmHg.

A more pronounced change can be observed in the range of 135–145 mmHg, where the SHAP values increase and mostly begin to appear on the positive side. At *ap_hi* values above approximately 140 mmHg, the SHAP values are generally positive, indicating that higher systolic blood pressure tends to push the model toward a cardiovascular risk prediction. However, considerable dispersion of SHAP values can still be observed at higher *ap_hi* values. This indicates that the contribution of *ap_hi* to the predictions is not determined solely by blood pressure values but may also be influenced by other features used in the model.

The color of the points represents *age_years*. Points representing both lower and higher ages are distributed across various *ap_hi* and SHAP values and do not form a clear interaction pattern between *ap_hi* and *age_years*. Therefore, the interaction relationship between systolic blood pressure and age cannot be strongly concluded based solely on this plot, although age itself remains one of the important features based on the SHAP summary plot results.

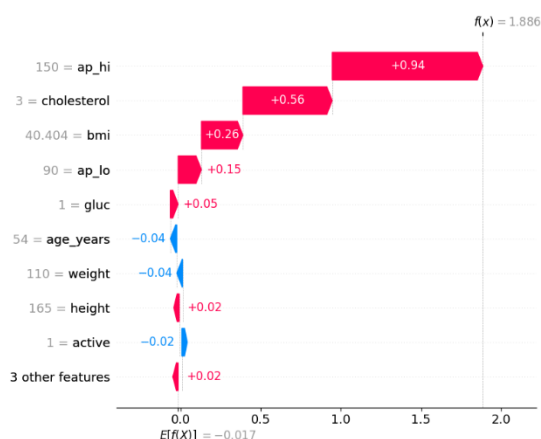


Figure 10. SHAP Waterfall Plot for a High-Risk Patient

Figure 10 shows the SHAP explanation for one patient who was predicted to be at risk of CVD. The model output for this patient was $f(x) = 1.886$, while the base value was $E(f(x)) = -0.017$. The difference between these values was influenced by several features that made both positive and negative contributions to the prediction. ap_hi was the feature with the greatest influence, followed by cholesterol and BMI. ap_lo and $gluc$ also made substantial contributions to the prediction.

On the other hand, age_years and $weight$ showed small contributions that slightly decreased the model output for this patient. This condition indicates that the contribution of a feature is not necessarily the same for all patients. Although age_years was one of the features with a relatively substantial contribution in the overall SHAP analysis, it provided a negative contribution for this particular patient. This shows that the model's prediction is not determined by a single feature but results from the contributions of multiple features considered simultaneously.

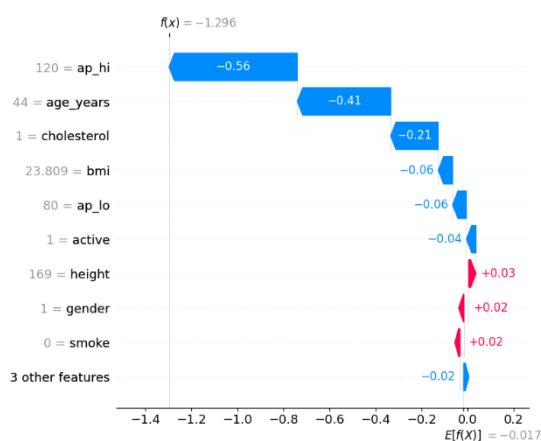


Figure 11. SHAP Waterfall Plot for a Patient Not at Risk of CVD

Figure 11 shows the SHAP explanation for one patient who was predicted to be not at risk of CVD. The model output for this patient was $f(x) = -1.296$, while the base value was $E(f(x)) = -0.017$. The difference between these two values

was influenced by several features that made both positive and negative contributions to the prediction. The feature with the greatest contribution was ap_hi , followed by age_years , cholesterol, BMI, and ap_lo . The active feature also made a substantial contribution toward directing the model's prediction to the not-at-risk class. Several other features, such as height, gender, and smoke, made relatively small positive contributions.

These results indicate that, for this patient, several key features contributed toward the not-at-risk prediction. However, the contribution of each feature is specific to the patient data being analyzed and does not necessarily follow the same contribution pattern observed in the overall SHAP analysis.

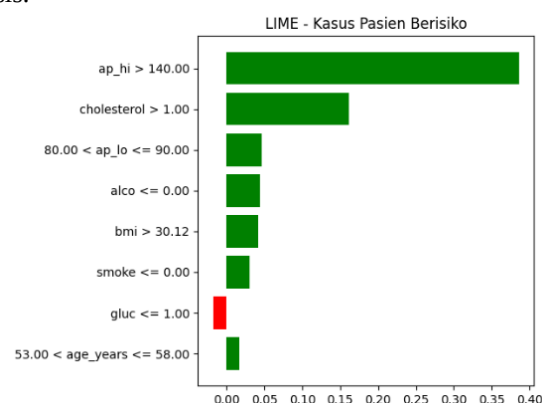


Figure 12 LIME Explanation for a Patient at Risk of CVD

Figure 12 shows the LIME interpretation for one patient who was predicted to be at high risk. ap_hi was the feature with the greatest contribution to the prediction, followed by cholesterol. The features ap_lo , $alco$, $smoke$, and age_years also made positive contributions to the prediction. Meanwhile, $gluc$ made a relatively small contribution, indicating that its influence on the model's prediction was not substantial.

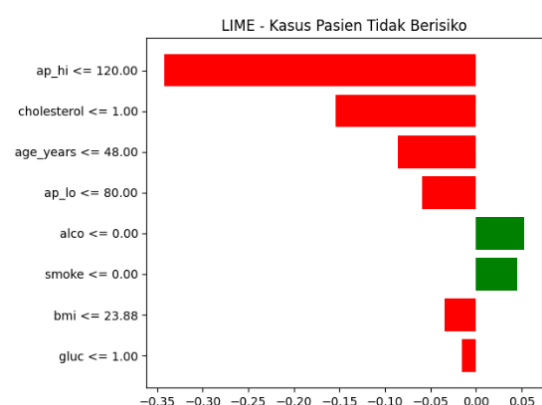


Figure 13 LIME Explanation for a Patient Not at Risk of CVD

Figure 13 shows the LIME interpretation for one patient who was not at risk of CVD. ap_hi was the feature with the greatest contribution to the model's prediction, followed by

cholesterol and age_years. The features ap_lo, BMI, and gluc also contributed toward the prediction of the not-at-risk class. In contrast, alco and smoke had positive weights toward the at-risk class, although their contributions were relatively small. These results indicate that, for this patient, the contributions of several features were directed toward the not-at-risk class, while some other features contributed toward the at-risk class.

Based on the SHAP and LIME analyses, several features made substantial contributions to the model's predictions. Overall, ap_hi was the feature with the greatest contribution, followed by age_years, cholesterol, ap_lo, and BMI. Higher ap_hi values tended to direct the model's predictions toward the at-risk class, whereas lower values more often directed the predictions toward the not-at-risk class. A similar pattern was also observed for age_years and cholesterol.

When analyzing individual patients, the influence of the features was not always the same as that observed in the overall analysis. For patients predicted to be at risk, ap_hi, cholesterol, and BMI made relatively substantial contributions to the predictions. Meanwhile, for patients predicted to be not at risk, ap_hi, age_years, and cholesterol had greater influence toward the not-at-risk class. This indicates that the model's predictions were not determined by a single feature but resulted from several features being considered simultaneously.

The LIME results also showed a similar pattern. For patients predicted to be at risk, ap_hi and cholesterol were the features that most strongly contributed toward the at-risk prediction. Conversely, for patients predicted to be not at risk, ap_hi, cholesterol, age_years, and ap_lo more strongly supported the not-at-risk prediction. The occurrence of several key features in both the SHAP and LIME results indicates a relatively consistent interpretation pattern, although the ranking of feature importance and the magnitude of feature contributions generated by the two methods were not completely identical. These differences may occur because SHAP and LIME use different approaches to explain the model's predictions.

Therefore, the use of SHAP and LIME in this study helped explain the features that contributed to the model's predictions, both overall and at the individual-patient level. These interpretations were used to examine how the model utilized each feature to generate predictions, rather than to determine causal relationships between the features and CVD [34]. Therefore, features with high contributions based on SHAP or LIME cannot be directly regarded as factors that cause CVD.

IV. CONCLUSION

This study compared the performance of Random Forest, CatBoost, LightGBM, and a stacking ensemble for cardiovascular disease risk prediction using accuracy, precision, recall, F1-score, and AUC-ROC. The results showed that LightGBM achieved the best performance on

most classification metrics, with an accuracy of 74.14%, recall of 70.48%, and F1-score of 72.95%, while Random Forest achieved the highest precision of 76.21%. Meanwhile, the stacking ensemble achieved the highest AUC-ROC of 80.91%. However, the difference in AUC-ROC compared with the individual models was relatively small, and the stacking ensemble did not achieve the best results on the other metrics. Therefore, combining Random Forest, CatBoost, and LightGBM through a stacking ensemble did not demonstrate consistent performance improvements over the individual models on the dataset used in this study.

The interpretation results using SHAP and LIME showed that ap_hi, age_years, cholesterol, ap_lo, and BMI were among the features that contributed substantially to the model's predictions. Both XAI methods showed relatively similar patterns for several key features, although differences were observed in the ranking and magnitude of their contributions. These results indicate that SHAP and LIME can help provide an understanding of how the model utilizes features to generate predictions, both overall and at the individual-data level. However, the feature contributions indicated by the two methods cannot be used to infer causal relationships with cardiovascular disease.

Overall, this study shows that individual models, particularly LightGBM, provided better results on most classification metrics than the stacking ensemble, whereas the stacking ensemble had an advantage in terms of AUC-ROC. The use of SHAP and LIME complemented the model evaluation by providing interpretations of feature contributions in the prediction process. Future studies may consider hyperparameter optimization, the use of datasets from different sources or real-world clinical data, and external validation to evaluate the model's generalizability.

REFERENCES

- [1] T. Apriliyani, N. A. O. Pusparini, K. Karmilah, Z. Rohmah, W. A. S. Tunjung, and A. Nuriliani, "Mekanisme Penyakit Kardiovaskular terkait Penuaan," *Bioma : Berkala Ilmiah Biologi*, vol. 26, no. 2, pp. 80–93, 2024, doi: 10.14710/bioma.2024.59115.
- [2] World Health Organization, "Cardiovascular diseases (CVDs)," World Health Organization. Accessed: Jul. 23, 2026. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- [3] Kemenkes BPKP, "Potret Sehat Indonesia dari Kacamata SKI 2023," Kemenkes.
- [4] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-01650-7.
- [5] F. Yazdi and S. Asadi, "Enhancing Cardiovascular Disease Diagnosis: The Power of Optimized Ensemble Learning," *IEEE Access*, vol. 13, no. February, pp. 46747–46762, 2025, doi: 10.1109/ACCESS.2025.3550015.
- [6] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular Disease Detection using Ensemble Learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Aug. 2022, doi: 10.1155/2022/5267498.
- [7] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput. Biol. Med.*, vol. 146, p. 105624, Jul. 2022, doi: 10.1016/j.combiomed.2022.105624.

- [8] Navita *et al.*, "Advanced Hybrid Machine Learning Model for Accurate Detection of Cardiovascular Disease," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 51, Mar. 2025, doi: 10.1007/s44196-025-00771-1.
- [9] I. Kiran, S. Ali, S. ur Rehman Khan, M. Alhussein, S. Aslam, and K. Aurangzeb, "An AI-Enabled Framework for Transparency and Interpretability in Cardiovascular Disease Risk Prediction," *Computers, Materials and Continua*, vol. 82, no. 3, pp. 5057–5078, 2025, doi: 10.32604/cmc.2025.058724.
- [10] N. Z. Dammak, H. Bezine, A. S. Nouri, and N. Derbel, "Innovative XAI Approaches for Predicting Cardiovascular Diseases," *OALib*, vol. 12, no. 10, pp. 1–25, 2025, doi: 10.4236/oalib.1114071.
- [11] N. Uddin, P. Mahmud, M. Ahammed, Md. A. Uddin, M. Kazi, and Md. A. Talukder, "Explainable ensemble learning model for cardiovascular disease prediction with feature optimization and data balancing," *Discover Computing*, vol. 29, no. 1, p. 44, Jan. 2026, doi: 10.1007/s10791-026-09930-0.
- [12] C. Skouteli, N. Prenzias, A. Kakas, and C. S. Pattichis, "Explainable AI Modeling in the Prediction of Cardiovascular Disease Risk," 2024, doi: 10.3233/SHTI240574.
- [13] Svetlana Ulianova, "Cardiovascular Disease dataset," Kaggle.
- [14] S. M. Ganie and P. K. D. Pramanik, "A comparative analysis of boosting algorithms for chronic liver disease prediction," *Healthcare Analytics*, vol. 5, Jun. 2024, doi: 10.1016/j.health.2024.100313.
- [15] W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," *PLoS One*, vol. 20, no. 8 August, Aug. 2025, doi: 10.1371/journal.pone.0330899.
- [16] J. Liang *et al.*, "Predictive Framework for Membrane Fouling in Full-Scale Membrane Bioreactors (MBRs): Integrating AI-Driven Feature Engineering and Explainable AI (XAI)," *Processes*, vol. 13, no. 8, p. 2352, Jul. 2025, doi: 10.3390/pr13082352.
- [17] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Front. Genet.*, vol. 14, Oct. 2023, doi: 10.3389/fgene.2023.1252159.
- [18] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, p. 144, Jan. 2024, doi: 10.3390/diagnostics14020144.
- [19] D. Yadav, D. Rani, and O. P. Verma, "Impact of feature selection and feature engineering in prediction of cardiovascular diseases," *Comput. Biol. Med.*, vol. 197, p. 111027, Oct. 2025, doi: 10.1016/j.combiomed.2025.111027.
- [20] P. Iacobescu, V. Marina, C. Anghel, and A.-D. Anghel, "Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 12, p. 396, Dec. 2024, doi: 10.3390/jcdd11120396.
- [21] O. D. Yanti, S. Alfariani, S. N. Milla Celesta, K. D. Tania, and A. Rifai, "Comparative Analysis of the Performance of Machine Learning Methods and Text Embedding Techniques in Classifying Toxic Conversations in the Roblox Game," *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 2775–2789, Jun. 2026, doi: 10.30871/jaic.v10i3.12646.
- [22] L. Chmielowski, M. Kucharak, and R. Burduk, "Novel method of building train and test sets for evaluation of machine learning models related to software bugs assignment," *Sci. Rep.*, vol. 13, no. 1, p. 21512, Dec. 2023, doi: 10.1038/s41598-023-48617-0.
- [23] A. Ashraf *et al.*, "A preprocessing-enhanced stacking classifier for generalized cardiovascular disease detection across diverse datasets," *Sci. Rep.*, vol. 16, no. 1, p. 16206, Mar. 2026, doi: 10.1038/s41598-026-41042-z.
- [24] J. Liu, X. Dong, H. Zhao, and Y. Tian, "Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion," *Processes*, vol. 10, no. 4, p. 749, Apr. 2022, doi: 10.3390/pr10040749.
- [25] A. Hussain and A. Aslam, "Hate speech against women and immigrants: A comparative analysis of machine learning and text embedding techniques," 2024. [Online]. Available: www.jart.icat.unam.mx
- [26] S. A. Atimbire, J. K. Appati, and E. Owusu, "Empirical exploration of whale optimisation algorithm for heart disease prediction," *Sci. Rep.*, vol. 14, no. 1, p. 4530, Feb. 2024, doi: 10.1038/s41598-024-54990-1.
- [27] F. Ocampo Osorio, S. Alzate-Ricaurte, T. E. Mejia Vallecilla, and G. A. Cruz-Suarez, "The anesthesiologist's guide to critically assessing machine learning research: a narrative review," *BMC Anesthesiol.*, vol. 24, no. 1, p. 452, Dec. 2024, doi: 10.1186/s12871-024-02840-y.
- [28] A. E. Korial, I. I. Gorial, and A. J. Humaidi, "An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection," *Computers*, vol. 13, no. 6, p. 126, May 2024, doi: 10.3390/computers13060126.
- [29] O. Rainio, J. Teuhio, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.
- [30] A. Bozorgpanah and V. Torra, "Explainable machine learning models with privacy," *Progress in Artificial Intelligence*, vol. 13, no. 1, pp. 31–50, Mar. 2024, doi: 10.1007/s13748-024-00315-2.
- [31] M. U. Sreeja, A. O. Philip, and M. H. Supriya, "Towards explainability in artificial intelligence frameworks for heartcare: A comprehensive survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 6, p. 102096, Jul. 2024, doi: 10.1016/j.jksuci.2024.102096.
- [32] J. E. Ebinger *et al.*, "Blood pressure variability supersedes heart rate variability as a real-world measure of dementia risk," *Sci. Rep.*, vol. 14, no. 1, p. 1838, Jan. 2024, doi: 10.1038/s41598-024-52406-8.
- [33] I. Naskinova, M. Kolev, D. Karova, and M. Milev, "Machine Learning-Based Blood Pressure Prediction Using Cardiovascular Disease Data: A Comprehensive Comparative Study," *Electronics (Basel)*, vol. 15, no. 2, p. 312, Jan. 2026, doi: 10.3390/electronics15020312.
- [34] J. A. Otieno, J. Häggström, D. Darehed, and M. Eriksson, "Developing machine learning models to predict multi-class functional outcomes and death three months after stroke in Sweden," *PLoS One*, vol. 19, no. 5 May, May 2024, doi: 10.1371/journal.pone.0303287.
- [35] P. Neirz, H. Allende, and C. Saavedra, "Epistemic frontiers and the distinction between causality, information, and predictability in pattern recognition," *Sci. Rep.*, vol. 16, no. 1, p. 21803, May 2026, doi: 10.1038/s41598-026-52883-z.