OWE-CVD: AN OPTIMIZED WEIGHTED ENSEMBLE FOR HEART DISEASE PREDICTION

Fahad Almalki [1] and Abdullah A Sheikh [2]

[1] Computer Science Department College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia
[2] Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, Kingdom of Saudi Arabia

ABSTRACT

Cardiovascular disease (CVD) remains the leading cause of death globally, with roughly 17.9 million fatalities each year. Early and accurate diagnosis of heart disease is critical to improving patient outcomes. We propose OWE-CVD (Optimized Weighted Ensemble for Cardiovascular Disease), a new predictive framework that combines a weighted voting ensemble of three gradient boosting classifiers (XGBoost, LightGBM, and CatBoost) with explainable AI (XAI) techniques. We first addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and optimized each model's hyperparameters using the Optuna framework with stratified 10-fold cross-validation. Ensemble weights were derived from cross-validated accuracy scores. On an independent test set, OWE-CVD achieved 94.44% accuracy with balanced precision, recall, and F1-scores across both classes. We applied SHAP and LIME to interpret the ensemble's predictions at global and local levels. Overall, OWE-CVD demonstrates strong predictive performance while providing transparent decision support for heart disease diagnosis in clinical settings.

KEYWORDS

Heart Disease Prediction, OWE-CVD, Optimized Weighted Ensemble, Explainable AI (XAI), Machine Learning, Healthcare AI

1. INTRODUCTION

The Saudi Ministry of Health reports that a large proportion of deaths in the Kingdom is attributed to CVD (cardiovascular diseases) as the condition often has many different risk factors including obesity, hypertension, diabetes, smoking, physical inactivity [1]. Globally, around 17.9 million people died from cardiovascular diseases in 2019, making up 32% of all deaths that year. Of these, heart attacks and strokes were responsible for roughly 85% [2]. The global concern of CVD indicates the need for novel and accurate methods to improve early detection of CVD in order to limit the number of deaths and drastically improve patient outcomes.

Advances in artificial intelligence (AI), specifically machine learning (ML), are changing the field of healthcare by enabling systems to learn from clinical data, identify patterns and make predictions without being coded [3], [4]. The availability of clinical data and powerful algorithms allow ML to now take and manage large amounts of complex data leading to increased performance. This enables ML to identify patterns and risk factors that may be complicated and hard for traditional diagnostic methods to find [5], [6]. Several supervised learning models have been tested for predicting heart disease, including decision trees, support vector machines, neural networks, and different ensemble methods. While these methods are showing promising results,

there is no one ensemble model that consistently provides optimal predictive performance in these varied populations.

Ensemble learning in machine learning combines predictions from multiple base models to boost accuracy, reduce overfitting, and enhance generalizability in medical diagnosis tasks [7]. For example, HeartEnsembleNet leverage ensemble techniques to improve the accuracy of CVD detection and outperform individual models [8]. While performance is essential, there are critical considerations for AI models to be adopted in practice, such as transparency and interpretability.

Explainable Artificial Intelligence (XAI) refers to methodologies that indicate how machine learning participates in its processes, thereby increasing transparency, accountability, and trust for users, including in health care [9], [10]. Commonly used XAI methods include SHAP (SHapley Additive Explanations), which uses cooperative game theory to provide importance scores of features [11], and LIME (Local Interpretable Model-agnostic Explanations), which allows for local interpretability based on surrogate models [12].

In this study, we propose a model called OWE-CVD (Optimized Weighted Ensemble for Cardiovascular Disease), a predictive model using a XGBoost, LightGBM and CatBoost using a weighted soft voting combination method. The ensemble was optimized using cross-validation and hyperparameter tuning (Optuna). OWE-CVD uses SHAP and LIME to ensure transparency by enabling users to understand feature importance, making OWE-CVD a reliable and interpretable tool for heart disease detection.

2. LITERATURE REVIEW

2.1. Machine Learning Models for Heart Disease Prediction

Machine learning (ML), a subfield of AI, has seen increasing application in cardiovascular disease prediction due to its ability to process complex clinical datasets and generate predictive models without explicit programming [13]. As clinical datasets are now available in increasing amounts in structured form, researchers began to evaluate a number of ML algorithms to detect heart disease using patient data. Many datasets available for heart disease prediction (including UCI Heart Disease dataset) have a relatively small sample size, which is also limited by missing values making it difficult to develop robust predictive ML models.

Past studies used models including neural networks, K-nearest neighbours (KNN), Naive Bayes, and logistic regression, on the UCI Cleveland dataset, and noted that performance varied greatly depending on the approach of feature selection in conjunction with data cleaning and data preprocessing strategies [14]. While neural networks achieved reasonable accuracy, a drawback with this method was the extensive amount of parameter tuning required with small sample sizes. Some studies note that KNN and Decision Trees achieved better overall performance than more complex models, as KNN used effective feature selection and reduced dimensionality [15].

Other approaches have proposed stacking ensembles of multiple learners, for example, a two-level model including Random Forest, Multilayer Perceptron (MLP), and XGBoost predicted a 92% accuracy once cross-validation and preprocessing was applied [16]. Similarly, they examined combining datasets from different sources (Cleveland, Statlog, Long Beach) which achieved higher ensemble accuracy: one model with Random Forest, XGBoost, and Extra Trees had a score of 92.34%, and used ROC and MCCs scores [17].

A recent study proposed a stacking ensemble for heart disease prediction using Logistic Regression as the meta-classifier and base learners including SVM, Random Forest, and CatBoost. SHAP values were used for interpretability, highlighting key contributing features. The model also had better performance than each individual classifier when combining multiple datasets and when there was noisy and imbalanced data [18].

2.2. Ensemble Techniques for Cardiovascular Diagnosis

Ensemble learning techniques, particularly stacking and weighted voting models, have consistently outperformed single classifiers in cardiac diagnostics. In one study, a hybrid model ensemble that combined Random Forest, SVM, and LightGBM predicted ICU mortality with an accuracy of 95.25%. Among many features, blood urea nitrogen and glucose were important to the prediction [19]. Another study used an ensemble model to predict carotid atherosclerosis. This ensemble model combined Random Forest, XGBoost, and a gradient boosting classifier, and also used SHAP to demonstrate interpretability, and feature ranking [20].

The SDKABL hybrid architecture used KNN, decision tree, and SVM as base learners, as well as a BiLSTM-based attention mechanism at the meta-layer. The architecture also performed PCA to dimensionally reduce the input to improve predictive performance, ultimately outperforming baseline ensemble models [21]. Similarly, the NCDG stacking framework used Naive Bayes, CatBoost, and decision trees as the base-level classifiers, and a gradient boosting classifier as a meta-classifier. They combined SMOTE and Borderline-SMOTE to address class imbalance when predictive uncertainty in the individual base-level classifiers. The NCDG average accuracy was close to 91%, and they used SHAP tables to add model interpretability [22].

These studies and results emphasize an increasing importance of ensemble models in clinical prediction tasks. When combining various classifications models and utilizing preprocessing strategies (for example, SMOTE), ensemble learners outperform individual learners with more stable and accurate performance.

2.3. Explainable AI in Healthcare

Predictive accuracy is important, but trust and transparency are also critical for medical applications. Explainable AI (XAI) frameworks have been an attempt to mitigate the "black-box" nature of many ML models. LIME explains individual predictions using interpretable local models, whereas SHAP uses game-theoretic principles to get consistent importance scores among features [11], [12].

While SHAP is theoretically grounded, some scholars have noted pitfalls using it when used for clinical studies—specifically, the lack of alignment with mechanistic goals of human interpretation [23]. Still, using combined explanations using both feature-based explanations, and example-based explanations to improve the clinician's ability to understand complex models [24]. In a high-risk environment such as healthcare, present models that mitigate uncertainty by applying transparency and accountably by using XAI [25].

In a regulatory context, counterfactual explanations can provide more benefits from XAI models because this type of explanation can align with legal privacy regulations, like the GDPR. Counterfactual explanations can inform patients of an actionable steps and insights without exposing those sensitive details of the algorithm [19]. One of the main obstacles to the widespread integration of AI into clinical workflows is the lack of an integrated healthcare structure [21].

Table 1. Literature Review Comparison Table.

Reference	Methodology	Key Findings
Sadr et al. [5]	CNN + LSTM + KNN + XGBoost	Combined deep learning and machine learning models improved heart disease prediction.
Rezk et al. [26]	Voting Ensemble (LightBoost + XGBoost)	Voting ensemble model gave better results than using LightBoost or XGBoost alone.
Tiwari et al. [17]	Stacking Ensemble (ExtraTrees, Random Forest, XGBoost)	Proposed a stacked ensemble framework that predicted heart disease better than older models.
Sultan et al. [22]	NCDG Model (Naive Bayes, CatBoost, Decision Tree + Gradient Boosting arner)	NCDG stacking model achieved better prediction accuracy than single models.
Mohapatra et al. [16]	Stacking Classifiers Model	Two-level stacking model improved accuracy by combining different classifiers.
Zhang et al. [20]	Norks Arbitration Model (Feedforward MLP + SVM)	Compared two models; SVM had better performance than MLP.

The recent methods for predicting heart disease are summarized in Table 1, and ensemble learning is used most frequently and seems to be the most widely used. Studies by Rezk et al. [26], Tiwari et al. [17], Sultan et al. [22], and Mohapatra et al. [16] demonstrated that combining models often produced better results than using a single model. For instance, Rezk's simple voting ensemble produced superior results than when he used single learners, and Tiwari's and Mohapatra's methods used stacking to improve generalization. Sadr et al. [5] used the hybrid CNN-LSTM ensemble learning, without interpretability. Sultan's new model NCDG achieved a more balanced output than other methods, and Zhang [20], concluded that SVMs outperformed many MLPs. While ensemble learning and hybrid methods are used increasingly but lack of transparency, an issue OWE-CVD's addresses through using explainable artificial intelligence.

3. METHODOLOGY

This study offers a high performing, interpretable, heart disease prediction model using a weighted ensemble framework that is optimized and enhanced with explainable AI methods. The workflow includes five core stages: gathering data, preprocessing it, building the model, evaluating its performance, and interpreting the results using explainable AI (XAI).

3.1. The Dataset

The dataset used was acquired by merging the Cleveland and Hungarian subsets from the UCI Heart Disease datasets [27]. The whole dataset consists of 1,190 patient records each with 11 input features and one binary outcome variable which describes the presence (1) or absence (0) of heart disease.

Table 2. Features Description.

Feature	**Description**
age	Age of the patient
sex	Biological sex
chest pain type	Type of chest pain experienced
resting bp s	Resting blood pressure
cholesterol	Serum cholesterol level
fasting blood sugar	Fasting blood sugar level (>120 mg/dl)
resting ecg	Resting ECG results
max heart rate	Maximum heart rate achieved
exercise angina	Exercise-induced angina
oldpeak	ST depression induced by exercise
ST slope	Slope of ST segment
target	Presence of heart disease

These features were chosen for their clinical significance and their role in diagnosing cardiovascular conditions [27].

3.1.1. Data Exploration

Most features provided for statistical analysis and machine learning are categorical, but these features can also be used with several algorithms as they have integer encoding. Exercise and ECG parameters including chest pain type, ST slope, and oldpeak can be utilised to enhance predictive models (diagnostic quality) reflecting their clinically relevant values.

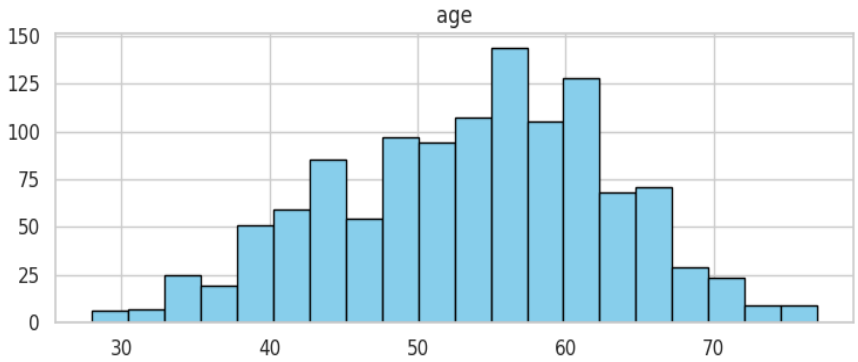

Figure 1. Age Distribution.

In figure 1 we can see the age feature has a normal distribution with its center point between 55 and 60 years of age. The majority of participants are age 45 to 65 years, which is at the normal age range of likely patients for heart problems.

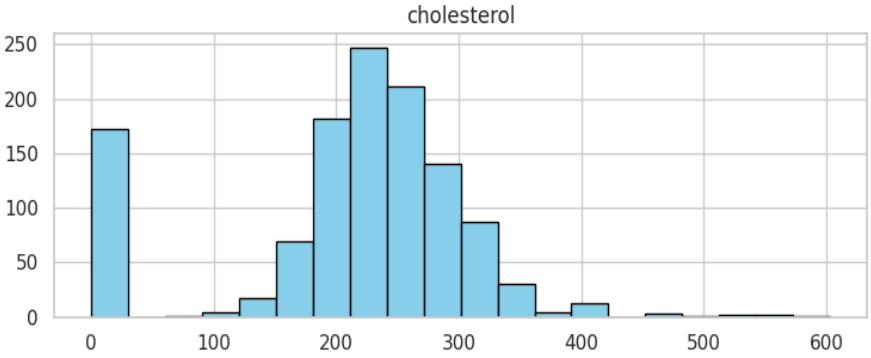

Figure 2. Cholesterol Level.

The cholesterol level shown in figure 2 displays distribution with a peak at 0, which indicates the possibility of missing or invalid values encoded as zeros. Many patients are in the 200 to 300 mg/dL range which is consistent with borderline or high cholesterol levels that are clinically relevant to heart disease risk.

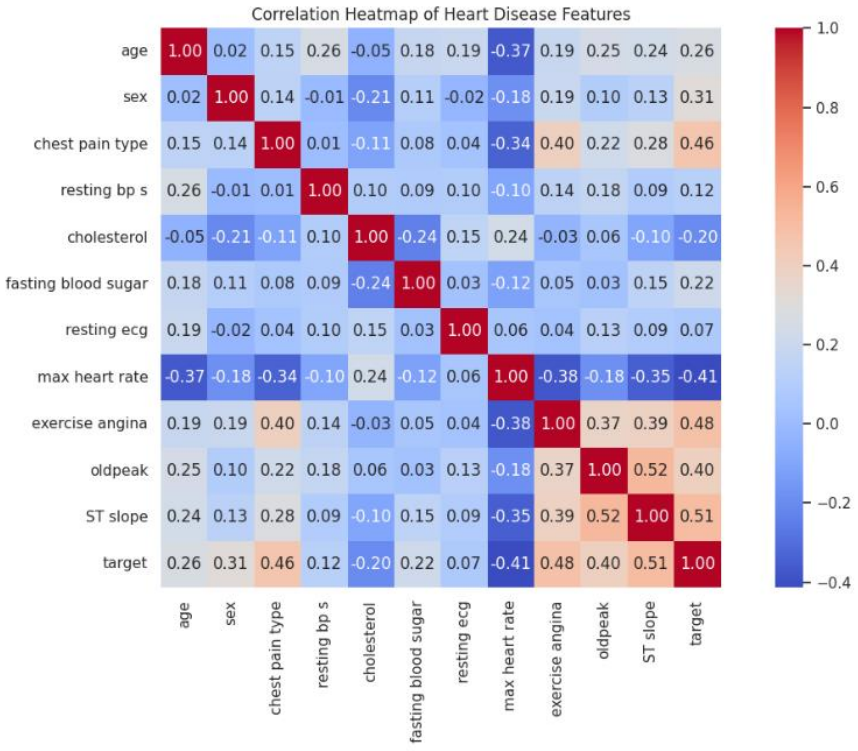

Figure 3. Correlation heatmap.

Figure 3 demonstrates the strength of the associations between each feature and the target. The target and a number of features also show moderate association: Heart disease has a positive relationship with ST Slope, Exercise Angina, Chest Pain Type, and Old peak; participants with heart conditions are likely to have higher values in these characteristics. However, the negative

association between maximum heart rate and heart disease indicates that a low heart rate during exercise may be associated with having heart disease. The chart shows that no strong redundancies are present within the data because most features have low correlation with each other. This will lower the chance for overfitting or redundant data when the data is used to create machine learning models.

3.1.2. Features Relationships

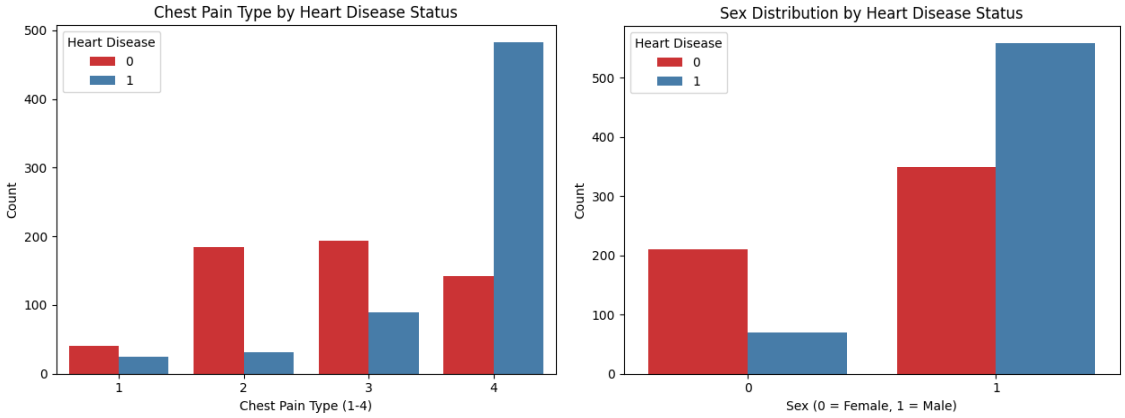

Figure 4. Distribution of Heart Disease Status by Chest Pain Type and Sex

The dataset outlines patterns that make medical sense: Asymptomatic chest pain (Type 4) is most associated with heart disease, while Types 2 and 3 appear in non-disease cases more often. Moreover, males display a greater percentage of heart disease cases, indicating a possible risk trend associated with sex. These medical findings allow a level of clinical validity for the dataset, which allows for relevance of ranked feature importance in further model explanations.

3.1.3. Class Balancing

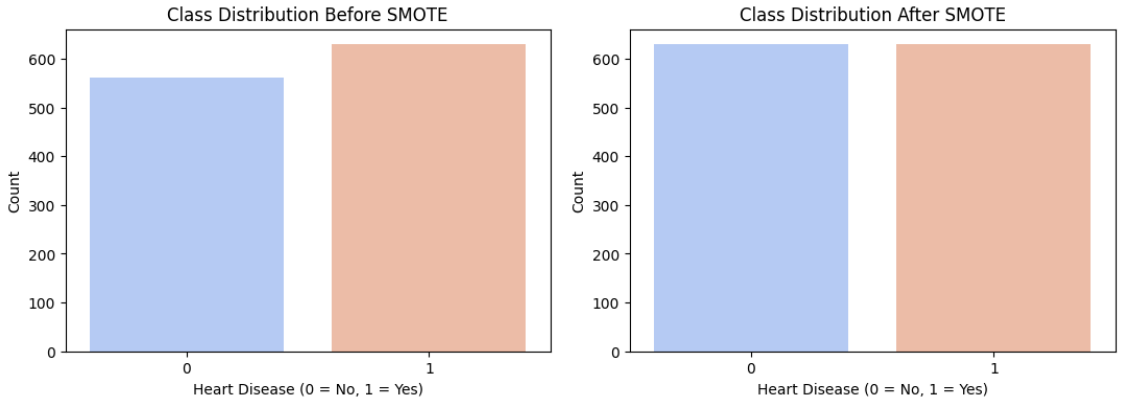

Figure 5. Class Distribution Before and After SMOTE.

In Figure 4 we represent class distribution before and after SMOTE (Synthetic Minority Over-sampling Technique) is applied. before the sampling, Class 0 (no heart disease) is slightly underrepresented compared to Class 1, which could bias models towards the class with the most samples. After applying SMOTE both classes are very similar with respect to sample sizes.

SMOTE improves generalization by generating synthetic examples of minority class, which decreases bias, and helps prevent overfitting.

3.2. Data Preprocessing

The data exploration showed some class imbalance and potential outliers (e.g. cholesterol with zero values). Based on our understanding of the dataset, we carried out the following steps before starting the analysis:

1. Low Variance Feature Removal: Any features with variance lower than 0.01 were removed with Variance Threshold, to ensure only informative features were kept.
2. Standardization: all numeric data was standardized using z-score standardization with StandardScaler, through this process all models would view features at the same scale.
3. Class Balancing using SMOTE: To correct the class imbalance, we applied SMOTE (Synthetic Minority Over-Sampling Technique), which generates synthetic samples for the minority class (no disease) to balance the dataset. This will increase generalization and reduce bias in the classification
4. Train-Test splitting: Once we completed the previous steps the data will be divided into 80:20 split for training and testing, maintaining the information of each class.

3.3. OWE-CVD: Model Architecture, Optimization, and Weighting

The OWE-CVD model was built as a weighted soft voting ensemble that combines three gradient boosting algorithms: XGBoost, LightGBM, and CatBoost. We separately trained each model on the same pre-processed dataset to ensure models had no difference in feature representation and learning environment. The ensemble was implemented through soft voting, meaning that based on the weighted average of predicted class probabilities across all models, the final class prediction was determined.

In order to optimize the individual learning performance of each base learner we used Optuna, a state-of-the-art hyperparameter optimization framework that utilizes Tree-structured Parzen Estimators (TPE) to navigate complex search spaces with minimal sampling overhead. For every algorithm there were twenty trials with stratified 10-fold cross-validation during the optimization, ensuring accurate, realistic metric estimates of discriminative learning performance. Hyperparameter tuning was performed for each model within their respective algorithmic search space for each model, then the identified hyperparameter configurations were retained for the ensemble.

Ensemble weights were determined using five-fold cross-validation on the training data. The accuracy scores from this validation process were used to assign weights in the VotingClassifier, allowing models with superior predictive performance to exert greater influence on the final output. This weighted strategy enabled the OWE-CVD ensemble to capitalize on the individual strengths of each model, thereby improving overall predictive accuracy.

3.4. Evaluation and Explainability

Post hyperparameter optimization, the final model of the OWE-CVD ensemble was retrained on the complete set of training data. We assessed the model's performance on the independent test set using several metrics: accuracy, precision, recall, F1-score, confusion matrix, and the area under the ROC curve (AUC). The inclusion of these metrics was intended to provide an overall perspective of classification performance while maintaining care about the balance of both classes with regards to sensitivity and specificity.

To enhance model transparency and support clinical validation, two model-agnostic explainability methods were applied. They were SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) algorithms. Both SHAP and LIME allow the data scientists to establish, both global and local insights into model behavior, with visualizations like SHAP summary and waterfall plots to show the features importance and reasoning behind its predictions. In contrast, LIME allows researchers to deduce the most impactful features of a specific instance using a local approximation. These tools provided the means to understanding how the model decided a path, and thus create a dialogue with domain experts.

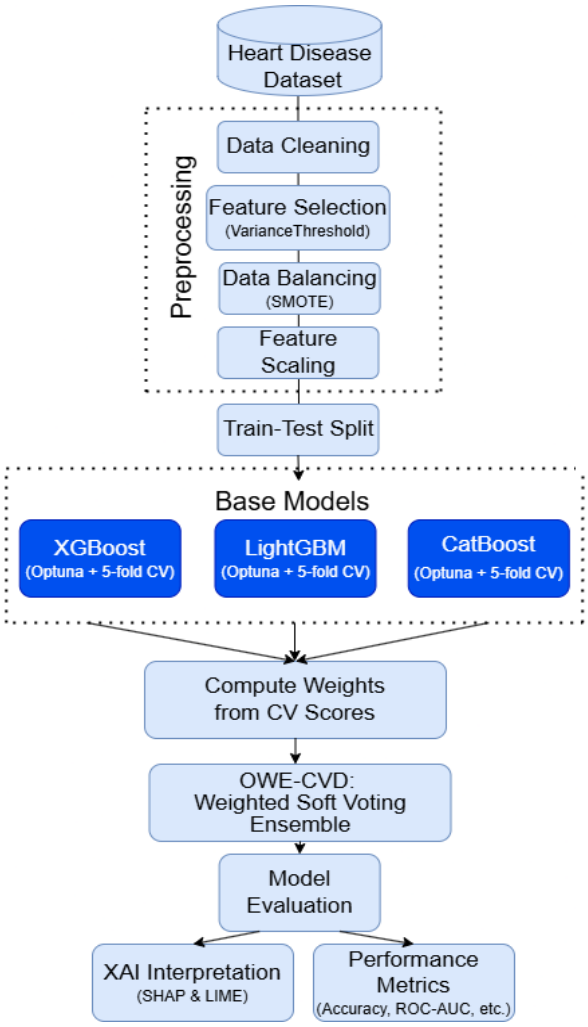

Figure 6. The proposed model (OWE-CVD).

Figure 6 outlines the full process used in this study, covering each step from data collection through model evaluation and interpretation.

4. RESULTS AND DISCUSSION

The performance of the proposed OWE-CVD model was evaluated using several standard classification metrics. These included accuracy, precision, recall, F1-score, confusion matrix, and the area under the receiver operating characteristic curve (ROC AUC).

4.1. Performance Metrics

The weighted soft voting ensemble model achieved an accuracy of 94.44% on the test data. Furthermore, precision, recall, and F1- scores were uniformly balanced across both classes, with a score of 0.94. The consistency of these scores demonstrates the model's ability to accurately classify both heart disease patients and non-heart disease patients.

```
Weighted Voting Model Accuracy on Test Set: 94.44%
Weighted Voting Model ROC AUC on Test Set: 0.9627
Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       126
           1       0.94      0.94      0.94       126

    accuracy                           0.94       252
   macro avg       0.94      0.94      0.94       252
weighted avg       0.94      0.94      0.94       252
```

Figure 7. Classification Report.

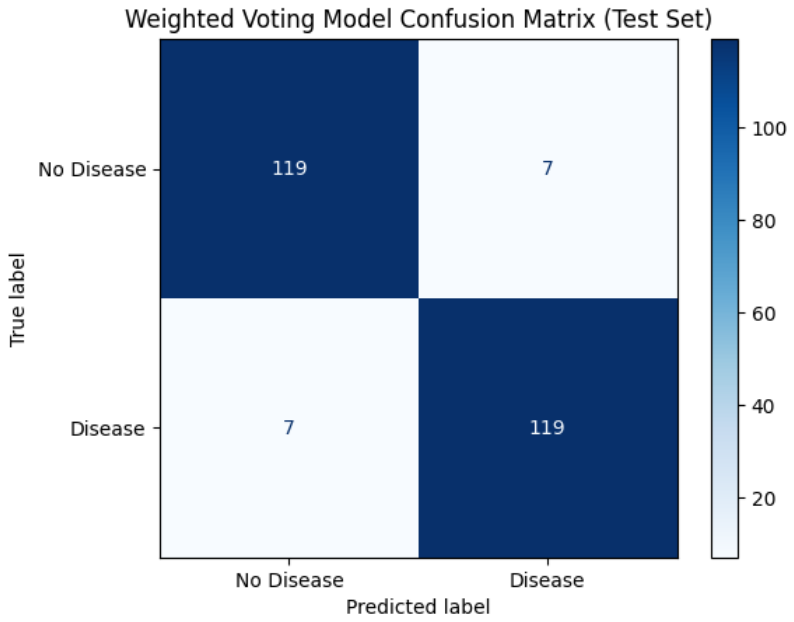

Figure 8. Confusion Matrix.

In figure 8 the confusion matrix illustrates a strong model performance on both classes, with only 14 misclassified samples out of 252 samples. It also illustrates identically high accuracy in the identification of both true positives and true negatives, and indicating balanced sensitivity and specificity.

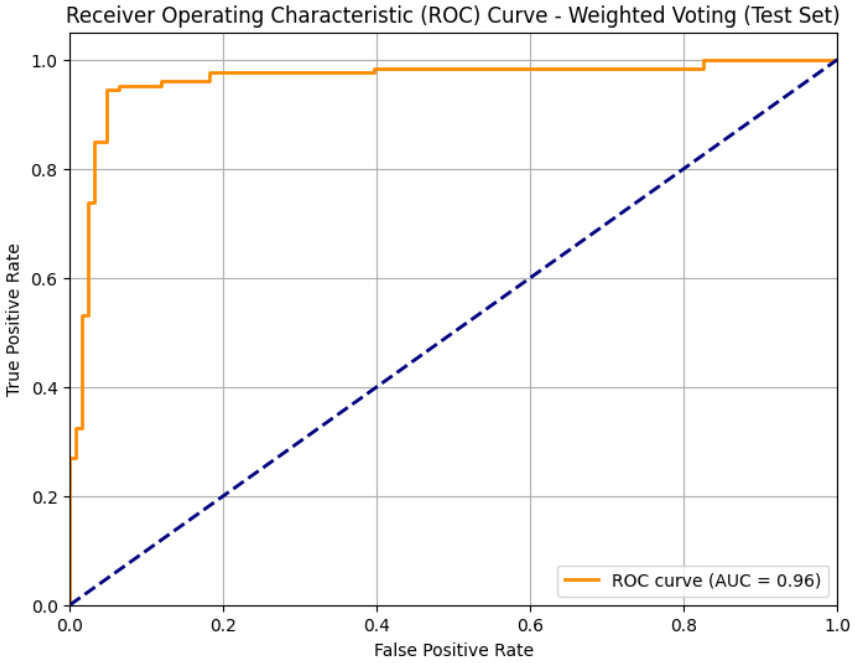

Figure 9. ROC Curve.

The ROC curve in figure 9 illustrates the model's discriminative capacity. The model achieved an AUC score of 0.96, indicating excellent class separation. This is a key indicator of performance in clinical decision contexts where distinguishing between disease presence and absence is critical.

4.2. Explainability Analysis

Model interpretability was assessed using SHAP and LIME explanations:

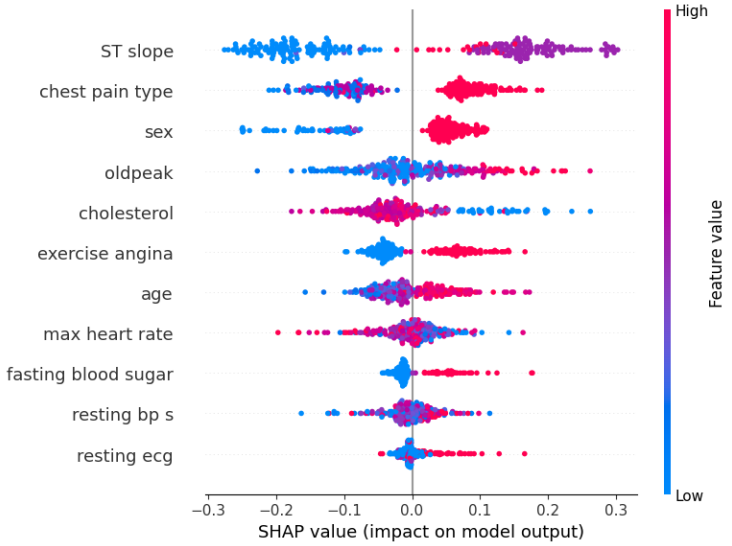

Figure 10. SHAP Summary Plot.

Figure 10 shows the SHAP summary plot which identified ST slope, chest pain type, and sex as the most influential features. High values of ST slope and chest pain were positively correlated with disease presence, which aligns with clinical findings.

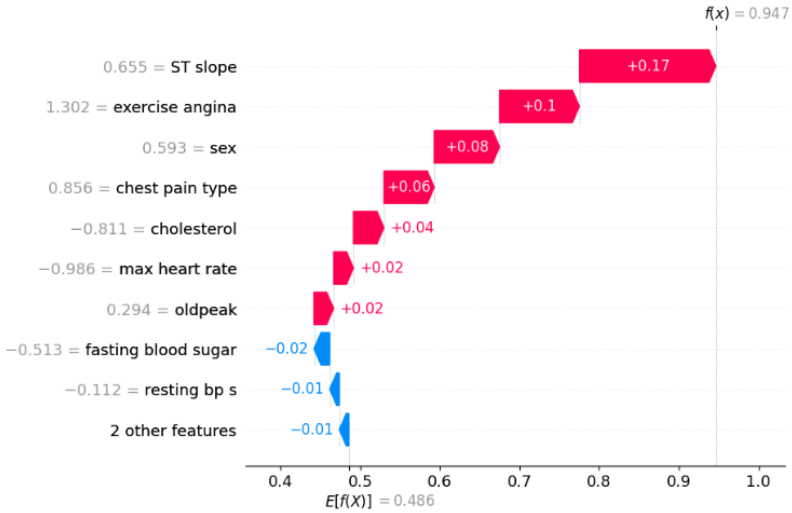

Figure 11. SHAP Waterfall Plot.

Figure 13 shows the contributions of individual features for a single prediction. ST slope, exercise angina, and oldpeak contributed positively to the heart disease prediction in this case.

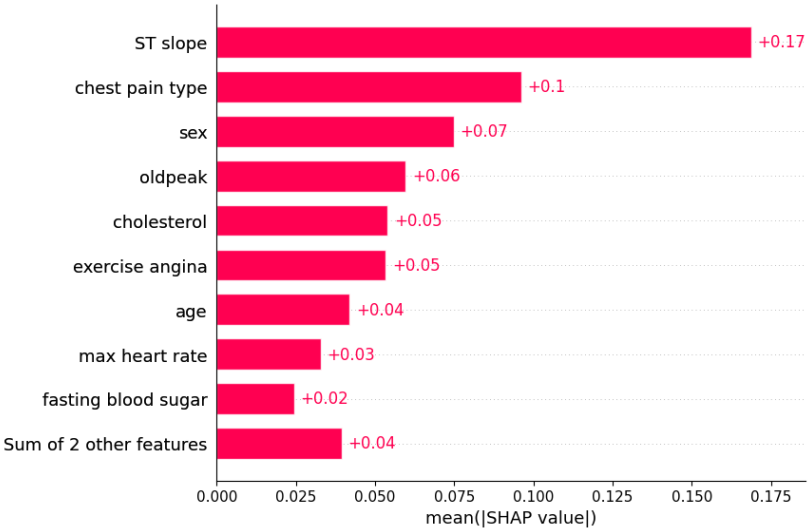

Figure 12. SHAP Waterfall Plot.

SHAP Feature Importance Plot (Figure 14): Ranked features by average absolute SHAP value, with ST slope leading as the most impactful feature, followed by chest pain type, sex, and oldpeak.

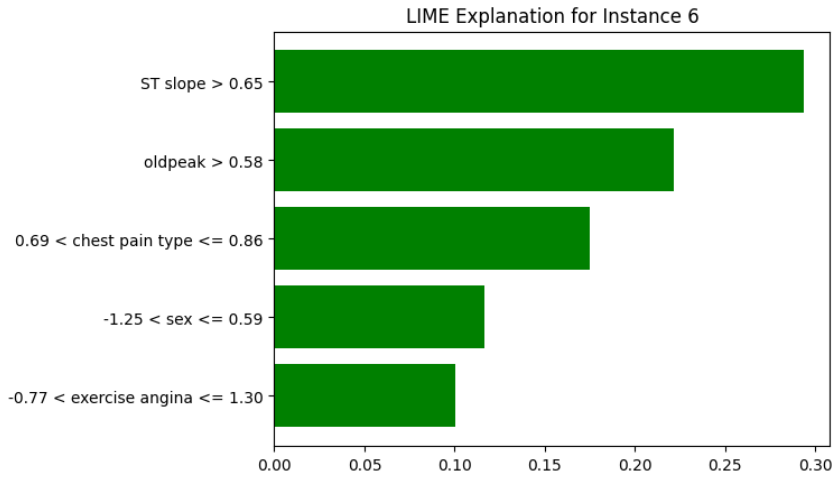

Figure 13. LIME Explanation for Instance 6.

In figure 13 ST slope and oldpeak were the strongest positive contributors to the disease prediction. These matched the findings from SHAP.

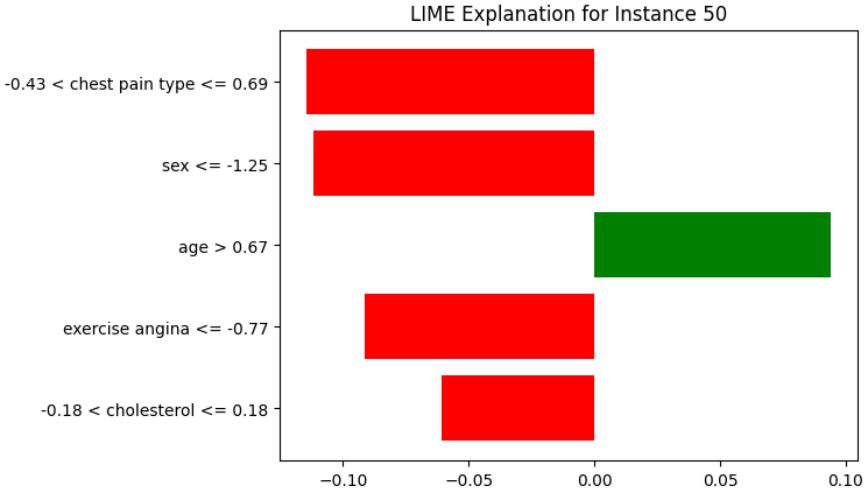

Figure 14. LIME Explanation for Instance 6.

The model predicted a low risk of heart disease in figure 14, driven by negative contributions from cholesterol, chest pain type, and exercise angina, further validating the model's interpretability.

4.3. Comparative Evaluation

Table 3. Result Comparison Table.

Reference	Methodology	Accuracy
OWE-CVD	Weighted Voting Ensemble (XGB+LGBM+CatBoost, Tuned)	94.44
Sadr et al. [5]	CNN + LSTM + KNN + XGBoost	90.87
Rezk et al. [26]	Voting Ensemble (LightBoost + XGBoost)	96.22
Tiwari et al. [17]	Stacking Ensemble (ExtraTrees, Random Forest, XGBoost)	92.34
Sultan et al. [22]	NCDG Model (stacking model)	91.00
Mohapatra et al. [16]	Stacking Classifiers Model	91.8
Zhang et al. [20]	Neural Networks Arbitration Model (Feedforward MLP + SVM)	87.5

Table III provides the classification accuracy results that show the proposed OWE-CVD model provides competitive performance compared to current cardiovascular disease (CVD) prediction methods. Rezk et al. [26] had the highest accuracy of 96.22%, using an ensemble of LightBoost combined with XGBoost. However, they did not offer built-in interpretability, which is an important and growing reporting element for models used in clinical settings.

In comparison to the ensemble-based models from Tiwari et al. [17] (92.34%), Sultan et al. [22] (91.00%), and Mohapatra et al. [16] (91.80%), OWE-CVD achieved a slightly higher accuracy (94.44%), whilst employing XAI (Explainable AI) techniques such as SHAP and LIME. These tools enabled both global and local explanations for the model, giving indications of feature contribution at a generalized level and thus enabling users to develop trust and understanding.

Comparing to deep learning-based models such as those from Sadr et al. [5] (90.874%) and Zhanget al. [20] (87.50%), which provided high accuracy based on applying strong models, but with no ability to provide interpretability due to their black-box nature.

OWE-CVD improves on existing models by utilizing both accuracy and transparency. The classification accuracy benefits from hyperparameter tuning via Optuna and cross-validated weighted risk adjustment, improving reliability across datasets.

In the context of cardiovascular disease diagnostics, it is just as important to understand how predictions were derived and not simply that predictions were made. OWE-CVD was designed with this approach to ensure that interpretability was inherently built into its operation.

4.4. Discussion and Clinical Implications

The results suggest that OWE-CVD achieves high classification accuracy while having acceptable precision and recall, which is especially useful in minimizing false negatives and false positives in the healthcare context. The incorporated ensemble learning enhances predictive power, while Optuna tuning and SMOTE rebalancing improve model performance reliability. The SHAP and LIME model explainability provide interpretability that allows clinicians to trust the decision made by the model.

This model shows great promise in being integrated into clinical workflows or decision support systems to detect cardiovascular disease. Future work should investigate predictions on larger, multi-institutional datasets, and also implement it in real-time in electronic health record systems.

5. CONCLUSION

This study presented OWE-CVD, an optimized weighted ensemble model utilizing XGBoost, LightGBM, and CatBoost through a soft voting process. The design process included hyperparameter optimization with Optuna and was balanced with SMOTE because of the class imbalance in clinical datasets. OWE-CVD demonstrated strong predictive performance, achieving 94.44% accuracy and an AUC of 0.9627 on the test set. These results speak to the model's robustness, generalizability, and ability to accurately detect heart disease across a balanced population of samples.

In addition to the predictive ability, one prominent feature of the proposed design is the explainability. By including SHAP and LIME in the model, OWE-CVD is interpretable and provides explanations about feature importance globally and for each individual patient. Clinically relevant features (e.g., ST slope, chest pain type, oldpeak) ranked as the most important features, supporting—without bias—the context of already known cardiovascular risk factors. This dual-explaining facilitate clinical decision-making and verification of what the model predicted, in an easy-to-understand medical context.

OWE-CVD was also compared against previously published models using more traditional stacking classifiers, hybrid neural models, and ensemble models of a variety of types. A few of these models achieve similar accuracy ratings or slightly higher ratings but couldn't replicate the general, clear, uncomplicated decision logic that OWE-CVD provides. These results suggest that OWE-CVD is more suitable for integration into clinical workflows, where explainability is as important as predictive accuracy.

Future work may focus on conducting validation of OWE-CVD against multi-institutional datasets and real-time datasets, in particular, datasets within electronic health record systems. Moreover, future prospective trials with clinical experts would support insight into usability and trustworthiness of the model prior to deployment. Overall, this study contributes to the development of clinically applicable, explainable AI for cardiovascular disease prediction and promotes a balanced, socially responsible approach to model deployment beyond technical performance.

ACKNOWLEDGEMENT

The researchers would like to acknowledge Deanship of Scientific Research, Taif University for funding this work and support.

REFERENCES

[1] “Cardiovascular Diseases Ministry Of Health Saudi Arabia,” Ministry Of Health Saudi Arabia. Accessed: Apr. 30, 2025. [Online]. Available: https://www.moh.gov.sa/en/Pages/Default.aspx

[2] “Cardiovascular diseases (CVDs).” Accessed: Apr. 30, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nat. Med., vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[4] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," N. Engl. J. Med., vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[5] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models," Eur. J. Med. Res., vol. 29, no. 1, p. 455, Sep. 2024, doi: 10.1186/s40001-024-02044-7.

[6] G. M. Rao, D. Ramesh, V. Sharma, A. Sinha, Md. M. Hassan, and A. H. Gandomi, "AttGRU-HMSI: enhancing heart disease diagnosis using hybrid deep learning approach," Sci. Rep., vol. 14, no. 1, p. 7833, Apr. 2024, doi: 10.1038/s41598-024-56931-4.

[7] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, 0 ed. Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

[8] S. A. J. Zaidi, A. Ghafoor, J. Kim, Z. Abbas, and S. W. Lee, "HeartEnsembleNet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction," Healthcare, vol. 13, no. 5, p. 507, Feb. 2025, doi: 10.3390/healthcare13050507.

[9] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017, arXiv. doi: 10.48550/ARXIV.1702.08608.

[10] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[11] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Apr. 23, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[13] P. Domingos, "A few useful things to know about machine learning," Commun. ACM, vol. 55, no. 10, pp. 78–87, Oct. 2012, doi: 10.1145/2347736.2347755.

[14] S. S. Yadav, S. M. Jadhav, S. Nagrale, and N. Patil, "Application of Machine Learning for the Detection of Heart Disease," in 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India: IEEE, Mar. 2020, pp. 165–172. doi: 10.1109/ICIMIA48430.2020.9074954.

[15] S. Patidar, D. Kumar, and D. Rukwal, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction," in Advances in Transdisciplinary Engineering, R. M. Singari and P. K. Kankar, Eds., IOS Press, 2022. doi: 10.3233/ATDE220723.

[16] S. Mohapatra et al., "A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease," Healthc. Anal., vol. 3, p. 100133, Nov. 2023, doi: 10.1016/j.health.2022.100133.

[17] 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.compbiomed.2022.105624.

[18] 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.

[19] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, and H.-L. Jiang, "Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure," J. Clin. Med., vol. 11, no. 21, p. 6460, Oct. 2022, doi: 10.3390/jcm11216460.

[20] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," IEEE Access, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[21] Y. Wu, Z. Xia, Z. Feng, M. Huang, H. Liu, and Y. Zhang, "Forecasting Heart Disease Risk with a Stacking-Based Ensemble Machine Learning Method," Electronics, vol. 13, no. 20, p. 3996, Oct. 2024, doi: 10.3390/electronics13203996.

[22] S. Q. Sultan, N. Javaid, N. Alrajeh, and M. Aslam, "Machine Learning-Based Stacking Ensemble Model for Prediction of Heart Disease with Explainable AI and K-Fold Cross-Validation: A Symmetric Approach," Symmetry, vol. 17, no. 2, p. 185, Jan. 2025, doi: 10.3390/sym17020185.

[23] E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," Jun. 30, 2020, arXiv: arXiv:2002.11097. doi: 10.48550/arXiv.2002.11097.

[24] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable AI meets Healthcare: A Study on Heart Disease Dataset," Nov. 06, 2020, arXiv: arXiv:2011.03195. doi: 10.48550/arXiv.2011.03195.

[25] U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in Healthcare," in 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland: IEEE, Jun. 2020, pp. 1–2. doi: 10.1109/CyberSA49311.2020.9139655.

[26] N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach," Bioengineering, vol. 11, no. 10, p. 1016, Oct. 2024, doi: 10.3390/bioengineering11101016.

[27] M. Siddhartha, "Heart Disease Dataset (Comprehensive)." IEEE DataPort, Nov. 05, 2020. doi: 10.21227/DZ4T-CM36.