Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

Explainable Stacked Ensemble Learning Model for Clinical Heart Disease Risk Prediction

[1]Tony-Alaede, Chisom Sonia, [2]Nwiabu, Nuka Dumle, [3]Emmah, Victor Thomas

[1,2,3]Department of Computer Science, Rivers State University, Port Harcourt, Nigeria
E-mails: [1]chisom.tony-alaede@rsu.edu.ng, [2]nwiabu.nuka@ust.edu.ng, [3]victor.emmah@ust.edu.ng

Abstract: Successive feature-selection stages can simplify clinical prediction models without necessarily improving their performance. This study examines a sequential Mutual Information (MI)–correlation–LASSO procedure within a stacked ensemble using the Framingham Heart Study and UCI Heart Disease datasets. Five configurations were compared: all processed features, MI only, MI plus correlation, the complete sequence, and recursive feature elimination with cross-validation (RFECV). Model development used the training partitions, with performance reported on held-out observations. The complete procedure reduced Framingham from 15 to 14 predictors and UCI from 15 to 13. Framingham ROC-AUC ranged from 0.677 to 0.685 across configurations. On UCI, MI plus correlation achieved the highest F1-score (0.858), while the complete sequence achieved the highest ROC-AUC (0.897) and PR-AUC (0.916). TreeSHAP and full-stack KernelSHAP identified several common leading predictors, alongside retained variables with low attribution magnitudes. The findings show that the benefit of successive selection stages varied by dataset and metric, and that selection status differed from a predictor's contribution to the fitted model.

Keywords: Heart disease prediction; hybrid feature selection; Mutual Information; LASSO; stacked ensemble; SHAP.

I. INTRODUCTION

Cardiovascular disease remains a major global health burden [1], motivating research into predictive models that use routinely recorded clinical information. Machine-learning approaches have been investigated for both long-term coronary heart disease prediction and the identification of existing disease, as represented by the Framingham and UCI datasets, respectively [2]. These tasks differ in their outcomes and predictor characteristics, providing distinct settings in which to examine the performance of a common modelling approach.

Feature selection can reduce the number of model inputs by screening for relevance, redundancy, or contribution within a fitted estimator [3]. Combining these criteria in a sequential procedure offers several opportunities to remove predictors, but the final feature count and model score alone do not reveal the value of each stage. An additional stage may reduce the predictor set without improving discrimination, or improve one performance measure while reducing another. The central question is therefore whether successive stages deliver measurable benefits beyond simpler stopping points.

Feature inclusion also leaves open the question of how a retained predictor influences the final model. This distinction is relevant when selection is followed by a heterogeneous ensemble: the criteria used to retain a variable differ from the way the combined learners use it. Examining both selection decisions and fitted-model explanations can identify predictors that remain in the input set but receive small contributions in the model's output. Such an analysis connects the selection procedure to the behaviour of the resulting predictor without equating model attribution with clinical importance.

This study evaluates sequential Mutual Information (MI), correlation filtering, and LASSO selection within an explainable stacked ensemble on Framingham and UCI data. It compares all processed features, intermediate selection stages, the complete sequence, and recursive feature elimination with cross-validation (RFECV). SHapley Additive exPlanations (SHAP) are used to examine the fitted XGBoost component through TreeSHAP and the complete stack through KernelSHAP. The analysis documents which features each stage removes, how predictive performance changes, and how retained predictors contribute to the fitted models across the two prediction tasks.

II. RELATED WORK

Feature-selection methods include filters, wrappers, and embedded approaches [3]. MI ranks predictors by dependence with the target [4], correlation filtering identifies overlap among predictors, and L1 regularisation can reduce fitted coefficients to

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

zero [5]. Their sequential application combines different selection criteria, but the benefit of each additional stage remains an empirical question.

Recent heart-disease studies have combined feature selection with ensemble learning. Natarajan *et al.* used Firefly optimisation with a stacked ensemble [6], while Sarra et al. reported an accuracy increase from 85.5% to 90.8% after Chi-square selection [7]. Yang and Guan investigated feature optimisation with SMOTE-XGBoost [8]; Ashika and Grace combined Rough Set Theory, stacking, and model ranking [9]; and Dissanayake and Md Johar examined meta-learning-based hybrid feature selection [10]. These approaches report outcomes for their respective selection designs, but do not determine the incremental value of the MI–correlation–LASSO sequence evaluated here.

Interpretability raises a separate question: a retained predictor may contribute little to the fitted model. SHAP attributes model outputs to individual predictors [11], complementing the inclusion and exclusion decisions made during feature selection. Raman et al. used SHAP with Framingham and Cleveland models [2], while Ganie *et al.* combined ensemble learning and explainable AI across heart-disease datasets [12].

These studies establish a basis for combining feature selection with model explanation, but leave the incremental effects of the MI–correlation–L1 sequence unresolved. The present analysis links the exclusions at each stage to predictive performance and to feature attributions from the fitted base learner and complete stack.

III. METHODOLOGY AND SYSTEM DESIGN

The study used an experimental machine-learning design. The workflow covered dataset preparation, sequential feature selection, stacked ensemble training, ablation testing, and model explanation. The two datasets were processed independently so that the behaviour of the same modelling pipeline could be compared across a long-term coronary heart disease risk task and an existing heart disease classification task.

Figure 1 shows the system architecture. The processed predictors were first passed through Mutual Information, correlation filtering, and LASSO selection. The selected features were then used by Logistic Regression, Random Forest, Support Vector Machine, and XGBoost base learners. Their out-of-fold positive-class probabilities formed the inputs to a LightGBM meta-learner. TreeSHAP explained the fitted XGBoost component, while KernelSHAP explained the complete stacked predictor.

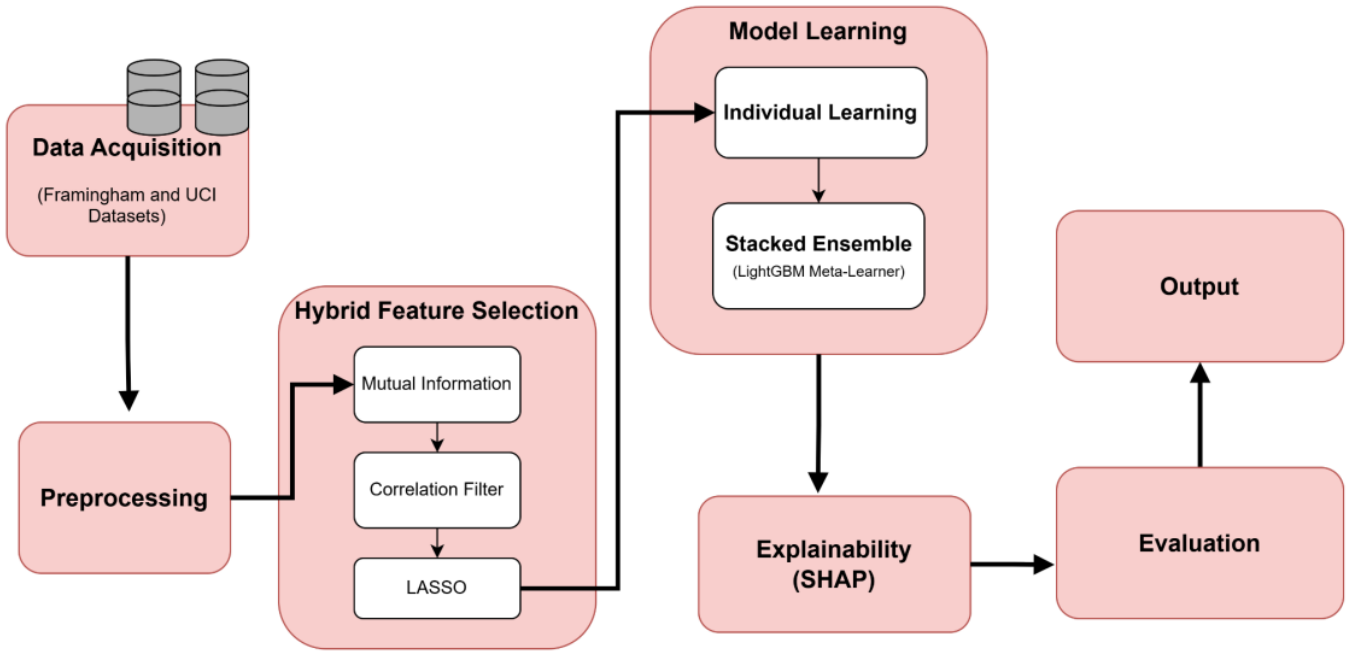


Figure 1: System Architecture of the Explainable Stacked Ensemble

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

Datasets and Preprocessing

The experiments used the Framingham Heart Study dataset and the UCI Heart Disease dataset. Framingham dataset contained 4,240 records, 15 predictors, and the binary ten-year coronary heart disease outcome, TenYearCHD [13]. The UCI dataset contained 920 records from Cleveland, Hungary, Switzerland, and VA Long Beach [14]. Its target was recoded as absence of disease for num = 0 and presence of disease for num > 0. Positive cases represented 15.19% of Framingham and 55.33% of UCI.

Missing numerical values were imputed with the median, while categorical values were imputed with the mode. In UCI, zero values for resting blood pressure and cholesterol were treated as missing. The identifier and source-cohort columns were removed, together with ca and thal because of high missing-value rates. One-hot encoding was applied to categorical UCI predictors, giving 15 processed columns. Framingham retained 15 numeric predictor columns after imputation. Each dataset was split into 80% training and 20% test partitions using stratification and random seed 42.

Sequential Feature Selection

The feature-selection stage combined filter and embedded methods. Mutual Information ranked predictors by their dependence on the target. For variables X and Y, Mutual Information is given as:

$$I(Y) = \sum_{x} \sum_{y} p(y)\ log\ log \left[\frac{p(y)}{p(x)p(y)}\right] \quad (1)$$

Where $p(x, y)$ is the joint probability, while $p(x)$ and $p(y)$ are marginal probabilities. After Mutual Information ranking, correlation filtering removed the lower-ranked member of highly correlated predictor pairs. LASSO selection was then applied through L1-regularised Logistic Regression on standardised predictors. Predictors with non-zero coefficients were retained, and RFECV with Logistic Regression was used as an additional wrapper comparison.

The search considered MI top-k values from 4 to 15 and correlation thresholds from 0.60 to 0.95. Candidate feature subsets were evaluated using training-stage cross-validation with a score that combined ROC-AUC and average precision. The selected settings were MI top-k = 14 and correlation threshold = 0.80 for Framingham, and MI top-k = 15 and correlation threshold = 0.70 for UCI.

Stacked Ensemble Training

The stacked model used four base learners: Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. Their out-of-fold probability outputs were used to train a LightGBM meta-learner. The stack used five-fold stratified cross-validation to generate meta-features and did not pass the original predictors directly to the meta-learner.

No Resampling, Random Oversampling, SMOTE, BorderlineSMOTE, and SMOTETomek were compared during training. No Resampling was selected for Framingham, while Random Oversampling was selected for UCI. Base-learner hyperparameters were searched with five-fold cross-validation. The ablation compared five feature configurations: all processed features, MI only, MI plus correlation, MI plus correlation plus LASSO, and RFECV.

Evaluation and Explainability

The held-out test partitions were used for final evaluation. Accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC were computed for each ablation setting. Classification thresholds were selected from training-stage out-of-fold probabilities by maximising F1-score, with recall used to break ties.

TreeSHAP was applied to the fitted XGBoost base learner using sampled test observations. KernelSHAP used sampled training observations as background and explained the complete stacked predictor through its positive-class probability output. Global importance was summarised by mean absolute SHAP values, and the two explanation methods were compared by their leading-feature rankings.

IV. RESULTS AND DISCUSSION

This section presents the main feature-selection, prediction, and explanation results. The emphasis is on the effect of each selection stage on the two datasets and on whether retained predictors made meaningful contributions to the fitted stacked model.

Feature-Selection Behaviour

The sequential selection procedure behaved differently on the two datasets. For Framingham, Mutual Information removed only the male indicator, while correlation filtering and LASSO retained the remaining 14 predictors. For UCI, Mutual

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

Information retained all 15 processed predictors, correlation filtering removed slope_flat, and LASSO further removed trestbps. RFECV retained all processed predictors in both datasets, which suggests that the wrapper criterion did not find a strong basis for additional reduction.

Table 1: Feature-Selection Summary Across Datasets

Dataset	Initial features	After MI	After correlation	After LASSO	Main exclusions
Framingham	15	14	14	14	male
UCI Heart Disease	15	15	14	13	slope_flat, trestbps

The final Framingham subset retained age, cigsPerDay, sysBP, glucose, totChol, prevalentStroke, diaBP, BPMeds, prevalentHyp, diabetes, heartRate, BMI, currentSmoker, and education. The UCI subset retained cp_atypical angina, sex, exang, oldpeak, cp_non-anginal, age, thalch, chol, cp_typical angina, slope_upsloping, fbs, restecg_normal, and restecg_st-t abnormality. The smaller reduction in Framingham shows that successive feature-selection stages do not automatically remove many variables, especially when the available predictors contain limited redundancy under the selected thresholds.

Predictive Performance

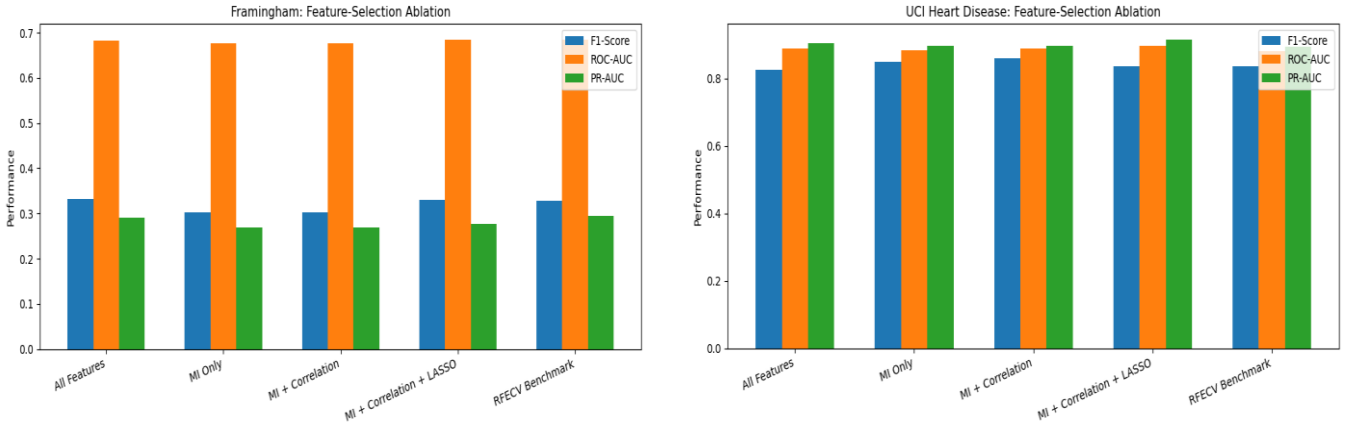
Table 2 gives the held-out ablation results. On Framingham, the ROC-AUC values were close across all configurations, ranging from 0.677 to 0.685. The complete MI plus correlation plus LASSO sequence produced the highest ROC-AUC, but the gain over all features was small. Its F1-score was also nearly identical to the all-feature configuration. This indicates that the hybrid selection sequence simplified the feature set without producing a clear performance advantage on the Framingham task.

Table 2: Feature-Selection Ablation on the Held-Out Test Sets

Dataset	Configuration	Features	F1	Recall	ROC-AUC	PR-AUC
Framingham	All features	15	0.332	0.496	0.683	0.290
Framingham	MI only	14	0.302	0.411	0.677	0.269
Framingham	MI + Corr.	14	0.302	0.411	0.677	0.269
Framingham	MI + Corr. + LASSO	14	0.331	0.496	0.685	0.276
Framingham	RFECV	15	0.329	0.442	0.684	0.294
UCI	All features	15	0.827	0.912	0.888	0.905
UCI	MI only	15	0.848	0.931	0.884	0.897
UCI	MI + Corr.	14	0.858	0.980	0.889	0.897
UCI	MI + Corr. + LASSO	13	0.836	0.922	0.897	0.916
UCI	RFECV	15	0.836	0.922	0.882	0.894

On UCI, the selection stages created a clearer trade-off. MI plus correlation gave the highest F1-score and recall, while the full MI plus correlation plus LASSO sequence gave the highest ROC-AUC and PR-AUC with only 13 features. Figure 2 summarises these ablation patterns for both datasets. The result shows that feature reduction can improve ranking quality without necessarily improving the threshold-dependent F1-score.

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com



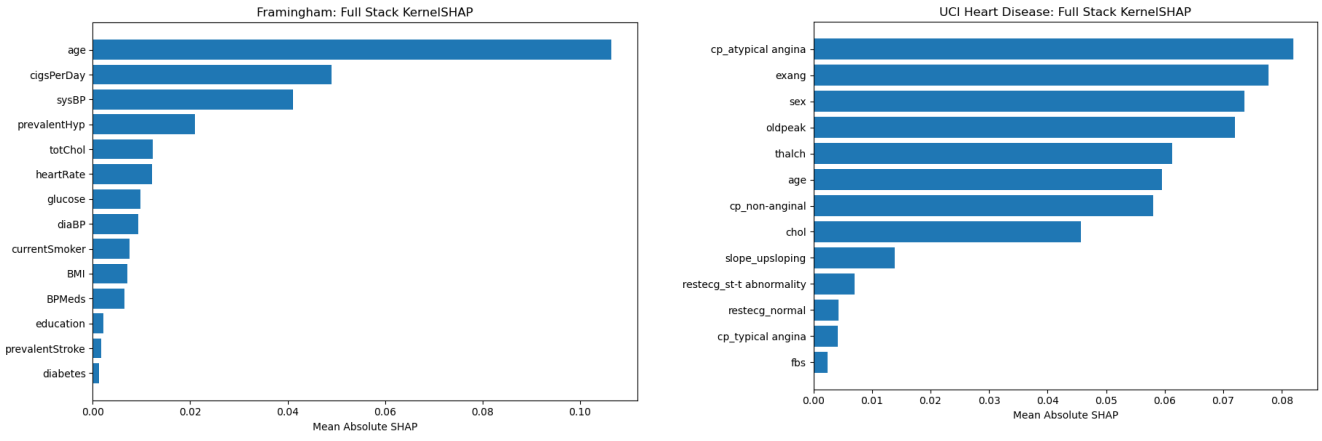
(a) **(b)**

Figure 2: Feature-selection ablation: (a) Framingham; (b) UCI

The separately tuned final model recorded Framingham ROC-AUC of 0.687 and PR-AUC of 0.279 at threshold 0.56, and UCI ROC-AUC of 0.898 and PR-AUC of 0.917 at threshold 0.34. These values were close to the complete-hybrid ablation rows, supporting the same interpretation: the selection sequence was more useful for compactness and ranking behaviour than for uniform improvement across all metrics.

Explainability and Interpretation

The explanation results showed that being selected did not mean that every retained predictor contributed equally to the fitted model. In Framingham, age had the strongest TreeSHAP contribution, followed by smoking intensity and systolic blood pressure. Diabetes and previous stroke were retained by the selection process but had low attribution values. For UCI, the leading contributors were atypical chest pain, exercise-induced angina, oldpeak, sex, and cholesterol. The full-stack KernelSHAP results in Figure 3 followed the same broad pattern, with age, cigsPerDay, and sysBP leading in Framingham, and atypical chest pain, exercise-induced angina, sex, and oldpeak leading in UCI.



(a) **(b)**

Figure 3: Full-stack KernelSHAP rankings: (a) Framingham; (b) UCI

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

Table 3: Retained Predictors with High and Low Explanation Evidence

Dataset	Predictor group	Explanation evidence
Framingham	Age, cigarettes/day, systolic BP	Highest TreeSHAP contributions; higher values generally increased the model output
UCI	Atypical chest pain, exercise-induced angina, oldpeak, sex, cholesterol	Main contributors in the UCI fitted model
Framingham	Diabetes, previous stroke	Retained by selection but low in the fitted-model explanation

These findings add an interpretation layer to the feature-selection results. Selection identified variables that passed the staged screening criteria, but SHAP showed how strongly the fitted model used those variables after training. This distinction is important because a clinical predictor may remain in the selected subset even when its marginal contribution to the trained ensemble is small.

The results also agree with prior heart-disease studies in showing that feature selection can help, but that its benefit depends on the dataset and modelling setting. Sarra *et al.* reported improvement after Chi-square selection [7], and Natarajan *et al.* reported strong classification with Firefly-based feature reduction [6]. In the present study, Framingham showed limited gain from additional selection, while UCI showed stronger benefits in ranking measures. Raman *et al.* also observed stronger discrimination on Cleveland/UCI than Framingham [2], which is consistent with the direction of the performance difference observed here.

The study is limited by its use of two public datasets, fixed hold-out partitions, and no external clinical validation. The ablation results are point estimates, and the differences between selection stages were not tested with paired statistical procedures. SHAP comparisons were also based on sampled explanation sets, so they should be interpreted as model-behaviour evidence rather than clinical causality. Within these limits, the results show that sequential feature selection can reduce predictor count and clarify model behaviour, but its predictive benefit is dataset- and metric-dependent.

V. CONCLUSION

Sequential MI–correlation–LASSO selection reduced Framingham from 15 to 14 predictors and UCI from 15 to 13, but its predictive benefit varied by dataset and metric. Framingham showed little numerical improvement, while UCI achieved its highest F1 after correlation filtering and its highest ROC-AUC and PR-AUC after the complete sequence. TreeSHAP and full-stack KernelSHAP shared several leading predictors and showed that retained variables could have markedly different attribution magnitudes. Successive selection stages therefore produced different trade-offs between feature count and prediction, while model explanations distinguished retained predictors by their contributions to fitted outputs. These findings are limited to the evaluated datasets and model configurations.

DATA AVAILABILITY

The datasets used in this study are available from the Kaggle repositories **Framingham Heart Study Dataset** [13] and **Heart Disease Data** [14]. The versions recorded for this study were accessed on 5 July 2026. No new patient data were collected.

REFERENCES

[1] B. Chong *et al.*, "Global burden of cardiovascular diseases: Projections from 2025 to 2050," *European Journal of Preventive Cardiology*, vol. 32, no. 11, pp. 1001–1015, 2025.

[2] S. Raman *et al.*, "Machine learning for coronary heart disease prediction: Comparative analysis of Framingham and Cleveland subset of the UCI dataset with SHAP-based interpretability," *Epidemiologia*, vol. 7, p. 75, 2026.

[3] I.Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] T. M. Cover and J. A. Thomas, Elements of Information Theory, *2nd ed. Wiley*, 2006.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] K. Natarajan *et al.*, "Efficient heart disease classification through stacked ensemble with optimized firefly feature

Journal of Artificial Intelligence and Emerging Technologies
Volume 3, Issue 9, pp 29-35, September-2026
ISSN (online): 3049-1967
DOI: https://doi.org/10.47001/JAIET/2026.309004
Available Online at https://www.jaiet.com

selection," *International Journal of Computational Intelligence Systems*, vol. 17, p. 174, 2024.

[7] R. R. Sarra *et al.*, "Enhanced stacked ensemble-based heart disease prediction with chi-square feature selection method," *Journal of Robotics and Control*, vol. 5, no. 6, pp. 1753–1763, 2024.

[8] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and SMOTE-XGBoost algorithm," *Information*, vol. 13, no. 10, p. 475, 2022.

[9] T. Ashika and H. G. Grace, "Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: An optimized RST-ML approach," *Frontiers in Digital Health*, vol. 7, 1609308, 2025.

[10] K. Dissanayake and M. G. Md Johar, "Heart disease diagnostics using meta-learning-based hybrid feature selection," *Applied Computational Intelligence and Soft Computing*, vol. 2024, 8800497, 2024.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *in Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[12] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, vol. 15, 13912, 2025.

[13] Aasheesh, "Framingham Heart Study Dataset," *Kaggle, n.d.* [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset. Accessed: July 5, 2026.

[14] Karimsony, "Heart Disease Data," *Kaggle, n.d.* [Online]. Available: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data. Accessed: July 5, 2026.

[15] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

Citation of this Article:

Tony-Alaede, Chisom Sonia; Nwiabu, Nuka Dumle; & Emmah, Victor Thomas. (2026). Explainable Stacked Ensemble Learning Model for Clinical Heart Disease Risk Prediction. *Journal of Artificial Intelligence and Emerging Technologies (JAIET).* 3(9), 29-35. Article DOI: https://doi.org/10.47001/JAIET/2026.309004

***** End of the Article *****