International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 10 October-2026 | Impact Factor: 3.5



A Trustworthy Hybrid AI Framework for Heart Disease Prediction Using Clinical Features and Explainable Deep Learning

Md. Aamir Sohail
Assistant Professor, Department Of Computer Science and Engineering, GMR Institute Of Technology(GMRIT)- Deemed to be University, Rajam, Viziangaram, Andhra Pradesh, India.

How to Cite this Article:

Sohail, M. A. (2026). A Trustworthy Hybrid AI Framework for Heart Disease Prediction Using Clinical Features and Explainable Deep Learning. International Journal of Creative and Open Research in Engineering and Management, 02(10), 1-9. https://doi.org/10.55041/ijcope.v2i10.207

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i10.207

Abstract: Heart disease remains one of the major causes of morbidity and mortality worldwide. Early identification of individuals at risk can support timely clinical assessment and appropriate preventive interventions. Machine learning (ML) and deep learning (DL) techniques have demonstrated potential for predicting heart disease using structured clinical data. However, conventional predictive models may lack interpretability, provide poorly calibrated risk estimates, and exhibit inconsistent performance when applied to different patient populations. This paper proposes a trustworthy hybrid artificial intelligence framework that combines conventional machine learning algorithms, a multilayer perceptron (MLP), probability-level fusion, and explainable artificial intelligence (XAI). The framework processes structured clinical features, including age, sex, chest pain characteristics, resting blood pressure, serum cholesterol, fasting blood sugar, electrocardiographic findings, maximum heart rate, exercise-induced angina, and other relevant clinical indicators. A preprocessing pipeline performs data validation, missing-value imputation, categorical encoding, feature scaling, and class-distribution assessment. Random Forest and Gradient Boosting models capture nonlinear relationships and feature interactions, while the MLP learns complex representations from the clinical variables. Their predicted probabilities are combined through a validation-based fusion strategy. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are incorporated to explain global feature importance and individual predictions.

Keywords: Heart disease prediction, hybrid artificial intelligence, machine learning, deep learning, explainable AI, SHAP, LIME, clinical decision support, model calibration, trustworthy AI.

2.Introduction

Cardiovascular diseases represent a substantial global public health challenge. Early identification of patients who may have heart disease can help clinicians prioritize further examination, diagnostic testing, and preventive care. Traditional clinical assessment involves interpreting symptoms, patient history, physiological measurements, laboratory results, and other diagnostic evidence. Although these methods remain essential, evaluating multiple clinical variables simultaneously can be challenging, particularly when relationships among variables are nonlinear. Artificial intelligence has emerged as a useful approach for analyzing clinical data and identifying patterns associated with disease. Machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting have been used for medical classification tasks. Deep learning models, including multilayer perceptrons, can additionally learn complex relationships among numerical and categorical clinical variables. Despite these advantages, achieving high predictive performance alone is insufficient for clinical applications. Healthcare professionals need to understand why a model predicts an increased risk for a particular patient. They also need evidence that its predicted probabilities are reliable and that performance does not deteriorate substantially across relevant patient subgroups.



Three major challenges motivate this research.

First, individual algorithms have different strengths and limitations. Linear models offer relatively straightforward interpretation but may not capture complex nonlinear interactions. Tree-based models can learn nonlinear relationships but may produce explanations that are difficult to communicate consistently. Neural networks can learn complex representations but are often regarded as less transparent.

Second, predictive models may provide inaccurate probability estimates even when their classification accuracy appears satisfactory. For example, a model that assigns a predicted probability of 0.80 should ideally demonstrate approximately 80% positive outcomes among comparable predictions in a sufficiently large, representative evaluation sample. This property is known as calibration.

Third, explanations must be meaningful and carefully interpreted. Feature importance does not establish causation, and an explanation generated by an AI model does not independently confirm a medical diagnosis.

The major contributions are:

1. A hybrid prediction architecture combining tree-based machine learning and a neural-network model.
2. A probability-fusion mechanism designed to combine complementary predictive outputs.
3. A global and local explanation strategy using SHAP and LIME.
4. An evaluation protocol covering discrimination, sensitivity, specificity, calibration, and subgroup performance.
5. A reproducibility and responsible-use framework for evaluating clinical prediction models.

3. Related Work

3.1 Traditional machine learning for heart disease prediction

Traditional machine learning has been widely investigated for classification using structured healthcare records. Logistic Regression estimates the relationship between clinical variables and the log-odds of a binary outcome. Its coefficients can be interpreted under the assumptions of the fitted model, making it a useful baseline. Decision trees provide rule-based predictions that can be communicated through branching conditions. Random Forest extends this approach by combining predictions from multiple decision trees trained using randomized samples and feature subsets. Gradient Boosting builds a sequence of weak learners to reduce errors made by previous learners.

These algorithms are suitable for structured clinical datasets because they can capture relationships among patient characteristics and disease labels. However, performance depends on data quality, feature selection, hyperparameter configuration, sample size, and the evaluation procedure.

3.2 Deep learning in clinical prediction

Deep learning models use multiple computational layers to learn representations from data. For structured clinical data, an MLP can learn nonlinear combinations of clinical variables through interconnected neurons and nonlinear activation functions.

An MLP typically contains an input layer, one or more hidden layers, and an output layer. The input layer receives encoded clinical variables, while hidden layers learn intermediate representations. A sigmoid output unit can estimate the probability of a binary outcome.

Nevertheless, neural networks may overfit small datasets, and their outputs can be difficult to explain without additional techniques. Regularization, dropout, early stopping, and careful validation can help control overfitting.



3.3 Explainable artificial intelligence

Explainable AI aims to provide information about how a model generates its predictions.

SHAP uses Shapley-value-based attribution to quantify the contribution of features to a prediction relative to a specified baseline. Global SHAP summaries can reveal which variables contribute most strongly to predictions across a dataset. Local explanations can illustrate how individual features contribute to a particular model output. LIME approximates a model locally using an interpretable surrogate model fitted around an instance of interest. It can produce understandable explanations for individual predictions, although the results may depend on the perturbation strategy, feature representation, and sampling procedure.

These techniques improve transparency but do not guarantee that a model is clinically valid, unbiased, or causally correct.

3.4 Research gap

Many predictive studies emphasize model accuracy without comprehensively evaluating calibration, explanation stability, subgroup performance, and reproducibility. In addition, explanations are sometimes reported separately from the predictive pipeline rather than being treated as an integral component of the system.

The proposed framework addresses this gap by combining hybrid prediction with explainability and a broader evaluation protocol. Its contribution is a unified methodology rather than an assumption that model fusion automatically produces superior results.

4. Problem Statement and Research Objectives

4.1 Problem statement

Heart disease is one of the major causes of mortality worldwide, and its early prediction is essential for timely diagnosis and effective treatment. Traditional heart disease prediction methods often rely on individual machine learning algorithms and clinical features, which may not adequately capture complex relationships among patient attributes. Furthermore, many existing AI-based models operate as black-box systems, providing limited explanations for their predictions and reducing transparency and clinical trust.

Therefore, this research proposes a trustworthy hybrid AI framework for heart disease prediction using clinical features and explainable deep learning. The framework integrates multiple machine learning algorithms with a Multilayer Perceptron (MLP) through probability-level fusion to improve predictive performance. SHAP and LIME are incorporated to explain model predictions and identify influential clinical features. The proposed framework also considers probability calibration, uncertainty estimation, and subgroup performance evaluation to enhance reliability, interpretability, and fairness. The ultimate objective is to develop an accurate, transparent, and trustworthy decision-support framework that assists healthcare professionals in identifying patients who may be at risk of heart disease.

4.2 Research objectives

The proposed study has the following objectives:

- Develop a preprocessing pipeline for heterogeneous clinical features.
- Train conventional machine learning and deep learning models.
- Combine model probabilities using validation-based fusion.
- Generate global and patient-level explanations using SHAP and LIME.
- Evaluate discrimination, classification performance, and probability calibration.
- Investigate performance differences across relevant patient subgroups.



- Establish a reproducible evaluation procedure that avoids information leakage.

5. Proposed Methodology

5.1 System architecture

The proposed framework consists of six major stages:

1. Clinical data acquisition and validation.
2. Data preprocessing and feature preparation.
3. Individual machine learning and deep learning model training.
4. Probability-level hybrid fusion.
5. Explainability and uncertainty assessment.
6. Performance evaluation and reporting.

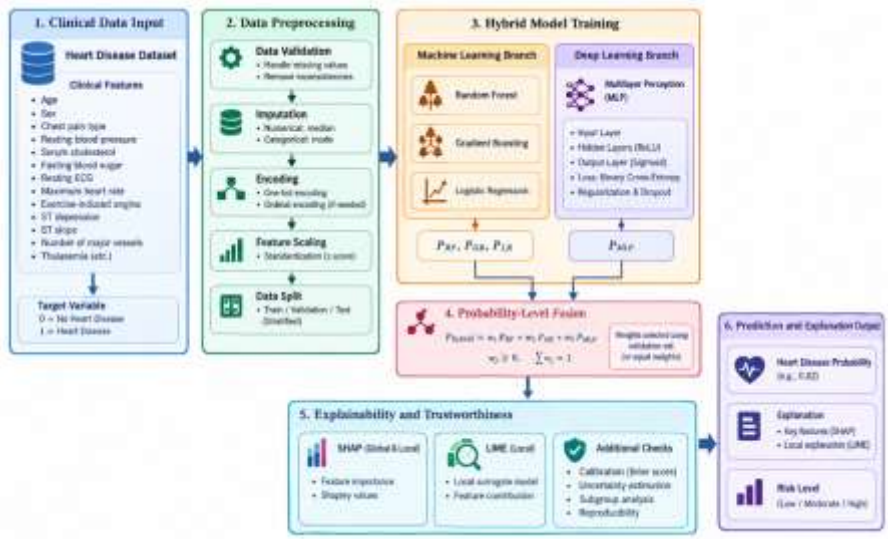

Figure 1. Proposed Hybrid AI Framework for Heart Disease Prediction Using Clinical Features and Explainable Deep Learning

5.2 Dataset description

The UCI Heart Disease dataset is a potential benchmark for evaluating the proposed framework. It contains clinical attributes used in heart disease classification research. The Cleveland subset is commonly used in experiments involving approximately 300 records and 14 commonly selected attributes, including a target variable. The precise record count, feature definitions, missing values, and label transformation must be verified against the exact dataset version used.

Potential predictor variables include:

Feature	Description
Age	Patient age in years
Sex	Recorded sex category
Chest pain type	Encoded chest pain classification
Resting blood pressure	Resting systolic blood pressure
Serum cholesterol	Recorded cholesterol measurement
Fasting blood sugar	Indicator based on a specified threshold
Resting ECG	Resting electrocardiographic category
Maximum heart rate	Maximum heart rate achieved
Exercise-induced angina	Indicator of angina during exercise
ST depression	Exercise-related ST-segment depression
ST slope	Slope of the ST segment



Number of major vessels	Number of major vessels represented in the dataset
Thalassemia-related field	Dataset-specific thalassemia category

Feature definitions and encodings must follow the selected dataset's documentation. These attributes should not automatically be interpreted as independently validated clinical risk factors in every population.

5.3 Data preprocessing

Preprocessing is essential because clinical datasets may contain missing values, inconsistent encodings, different measurement scales, and class imbalance.

Missing-value handling: Numerical features can be imputed using training-set medians, while categorical features can be imputed using training-set modes. More advanced imputation methods may be investigated when justified by the dataset.

Categorical encoding: Nominal variables can be one-hot encoded, while ordinal encoding should be used only when the category ordering is meaningful and supported by the data definition.

Feature scaling: Standardization can be applied to numerical inputs for Logistic Regression and the MLP:

Data partitioning: Training, validation, and test partitions should be separated before fitting preprocessing transformations. A stratified split can help preserve class proportions. For small datasets, repeated stratified cross-validation may provide a more stable estimate of performance.

All preprocessing operations must be fitted within the training fold to prevent data leakage.

5.4 Machine learning branch

The machine learning branch uses Random Forest and Gradient Boosting models.

Random Forest builds multiple decision trees and combines their predictions. It can model nonlinear relationships and interactions among clinical attributes without requiring the same degree of scaling as many linear or neural models.

Gradient Boosting trains sequential learners, where later learners attempt to reduce the errors of earlier learners. This approach can capture complex patterns but may be sensitive to hyperparameters and small datasets.

Logistic Regression is included as a baseline to determine whether the additional model complexity produces a meaningful improvement.

Hyperparameters should be selected using training and validation data. The final test set must not be used for model selection.

5.5 Deep learning branch

The deep learning component uses a multilayer perceptron. For a binary classification problem, the network can be represented as

$$h^{(1)} = \mathrm{ReLU}(W^{(1)}x + b^{(1)})$$

$$h^{(2)} = \mathrm{ReLU}(W^{(2)}h^{(1)} + b^{(2)})$$

$$p_{\mathrm{MLP}} = \sigma(W^{(3)}h^{(2)} + b^{(3)})$$



Regularization, dropout, and early stopping may be used to reduce overfitting. The architecture and training parameters should be reported explicitly in the final manuscript.

5.6 Hybrid probability fusion

The proposed hybrid model combines the outputs of the individual models rather than simply combining their binary class labels.

The fusion weights should be selected using validation predictions, preferably out-of-fold predictions when the dataset is small. Equal weights provide a simple baseline. Validation-based weights can be investigated to determine whether they improve generalization.

The threshold should be selected according to the intended use and clinical consequences of false-positive and false-negative errors. A threshold of 0.5 should not be assumed to be optimal for every application.

5.7 Explainability using SHAP

SHAP is used to estimate how individual features contribute to model predictions relative to a specified reference.

5.8 Explainability using LIME

LIME generates a local approximation of the prediction model around a selected patient record. It perturbs the input, queries the original model, and fits a simpler interpretable model to approximate its local behavior.

For example, an explanation report might state that a patient's recorded age, exercise-related measurements, and chest pain category contributed to the model's prediction. The actual direction and magnitude of each contribution must come from the fitted model; they must not be invented for the paper.

5.9 Calibration and uncertainty

A trustworthy prediction system should evaluate whether estimated probabilities correspond to observed outcomes.

Lower Brier scores indicate better overall probabilistic accuracy, although comparisons should consider class prevalence and evaluation context.

Calibration plots can compare average predicted probabilities with observed positive-class frequencies. Platt scaling or isotonic regression may be investigated when sufficient validation data are available. Calibration methods must be fitted without using the final test labels.

Uncertainty can be investigated using variation across cross-validation folds, bootstrap confidence intervals, model ensembles, or other suitable methods. A prediction with substantial uncertainty should be flagged for further review rather than presented as definitive.

6. Experimental Setup

6.1 Evaluation design

The experiments should compare Logistic Regression, Random Forest, Gradient Boosting, the MLP, and the proposed hybrid model using identical data partitions. Hyperparameter tuning must occur within the training and validation procedure.



For a small benchmark dataset, repeated stratified cross-validation can reduce dependence on a single split. If a separate test set is used, it should remain untouched until model selection is complete.

The following metrics should be reported:

- Accuracy
- Precision
- Recall or sensitivity
- Specificity
- F1-score
- ROC-AUC
- Brier score
- Calibration performance
- Confidence intervals, where feasible

7. Illustrative Results and Discussion

Important: The following values are illustrative examples for demonstrating the result-reporting format. They are not outputs from executed experiments and must not be reported as measured research findings.

7.1 Illustrative comparison of models

Model	Accuracy	ROC-AUC
Logistic Regression	0.80	0.84
Random Forest	0.83	0.88
Gradient Boosting	0.82	0.87
MLP	0.81	0.85
Hybrid fusion	0.85	0.90

In this hypothetical comparison, the hybrid model has the highest accuracy and ROC-AUC. Such a result would suggest that combining models with complementary predictive behavior may improve discrimination. However, this conclusion can only be supported after running the experiments on a documented dataset with an appropriate validation strategy. A genuine evaluation should also report variation across repeated runs, confidence intervals, and whether the observed differences are statistically meaningful.

7.2 Illustrative confusion matrix

The following example assumes 60 evaluated records, comprising 30 positive and 30 negative cases.

Actual class	Predicted positive	Predicted negative
Positive	27 (TP)	3 (FN)
Negative	6 (FP)	24 (TN)

The corresponding illustrative metrics are:

- Accuracy: 51/60 = 85%
- Sensitivity: 27/30 = 90%
- Specificity: 24/30 = 80%
- Precision: 27/33 = 81.82 %
- F1-score: **85.71%**



These calculations are mathematically consistent with the hypothetical confusion matrix. They do not establish the real-world performance of the proposed framework.

7.3 Explainability analysis

The explainability component should generate a global feature-importance summary and local explanations for selected test records. Global SHAP values can be used to rank features according to their average contribution magnitude. LIME can be used to generate a local approximation for an individual prediction.

7.4 Calibration and subgroup evaluation

The hybrid model should be assessed for probability calibration using reliability diagrams and Brier scores. Calibration should be compared before and after any calibration procedure.

Subgroup evaluation should be conducted where sample sizes permit meaningful assessment. Potential groupings include age categories and recorded sex.

8. Limitations and Ethical Considerations

The proposed framework has several limitations.

First, a small benchmark dataset may not represent the demographic, clinical, and geographic diversity of real-world patient populations. Consequently, strong benchmark performance does not establish generalizability to other hospitals or healthcare systems.

Second, probability fusion introduces additional design choices, including weight selection and threshold determination. These choices must be made using training and validation data to avoid optimistic estimates.

Third, SHAP and LIME provide model-based explanations rather than causal evidence. Explanations may be affected by correlated features, reference distributions, perturbation methods, and model instability.

Fourth, missing values, measurement errors, class imbalance, and changes in clinical practice can affect prediction quality. External validation and monitoring are required before considering deployment.

9. Conclusion and Future Work

This paper presents a trustworthy hybrid AI framework for heart disease prediction using structured clinical features, conventional machine learning, deep learning, probability fusion, and explainable AI. The framework integrates predictive modeling with SHAP- and LIME-based interpretation and recommends calibration, uncertainty assessment, subgroup evaluation, and reproducibility checks as essential components of the evaluation process.

The main objective is to move beyond a prediction-only approach toward a system whose outputs can be examined, evaluated, and communicated responsibly. Although hybrid fusion may improve predictive performance, its effectiveness must be demonstrated through carefully designed experiments rather than assumed from the architecture alone. Future work will include implementation on a documented clinical dataset, nested or repeated stratified cross-validation, independent external validation, statistical comparison of models, calibration analysis, explanation-stability assessment, and evaluation of performance across relevant patient groups. Further development may also investigate privacy-preserving training and prospective clinical evaluation. The framework therefore provides a methodological foundation for developing interpretable and carefully evaluated AI-based heart disease prediction systems.



References

1. "HXAI-ML: A hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection," Results in Engineering, vol. 25, article 104370, 2025. https://doi.org/10.1016/j.rineng.2025.104370.
2. Mridha, K., Kuri, A. C., Saha, T., and Shukla, M., "SmartHeart: A conceptual framework for explainable machine learning in cardiovascular risk prediction," Computers in Biology and Medicine, vol. 198, article 111221, 2025. https://doi.org/10.1016/j.compbiomed.2025.111221.
3. Walia, S., Walia, S., Bhardwaj, A., et al., "Explainable soft-voting classifier for heart disease prediction using SHAP and LIME," Discover Computing, vol. 29, article 83, 2026. https://doi.org/10.1007/s10791-026-09972-4.
4. Detrano, R., Janosi, A., Steinbrunn, W., et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310,.
5. Breiman, L., "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.
6. Friedman, J. H., "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
7. Lundberg, S. M., and Lee, S.-I., "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
8. Ribeiro, M. T., Singh, S., and Guestrin, C., "'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, pp. 1135–1144, 2016.
9. Collins, G. S., Moons, K. G. M., Dhiman, P., et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, 2024.
10. Chen, T., and Guestrin, C., "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
11. Niculescu-Mizil, A., and Caruana, R., "Predicting good probabilities with supervised learning," in Proceedings of the 22nd International Conference on Machine Learning, pp. 625–632, 2005.
12. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., "On calibration of modern neural networks," in Proceedings of the 34th International Conference on Machine Learning, PMLR, vol. 70, pp. 1321–1330, 2017.
13. Steyerberg, E. W., Vickers, A. J., Cook, N. R., et al., "Assessing the performance of prediction models: A framework for traditional and novel measures," Epidemiology, vol. 21, no. 1, pp. 128–138, 2010.
14. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W., "Calibration: The Achilles heel of predictive analytics," BMC Medicine, vol. 17, article 230, 2019.
15. Wolpert, D. H., "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
16. Dietterich, T. G., "Ensemble methods in machine learning," in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, pp. 1–15, 2000.
17. Saito, T., and Rehmsmeier, M., "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.
18. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
19. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D., "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, article 195, 2019.
20. Topol, E. J., "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.
21. Rudin, C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.
21. Ghassemi, M., Oakden-Rayner, L., and Beam, A. L., "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, 2021.
22. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M., "Ethical machine learning in health care," Annual Review of Biomedical Data Science, vol. 4, pp. 123–144, 2021.