Ingénierie des Systèmes d'Information
Vol. 31, No. 7, July, 2026, pp. 2315-2324
Journal homepage: http://iieta.org/journals/isi

BioXAI-HeartNet: An Explainable and Adversarially Robust Deep Learning Framework for Reliable Heart Disease Prediction

Iranna Amargol[1*], Dinesha H A[2], Keerthi Devarahatty Shivakumar[3], Suhas G K[4], Pankaja R[5], Sunad Kumara A N[6]

[1] Department of Computer Science and Engineering, Navkis College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi 590018, India
[2] Department of Computer Science and Engineering (Cyber Security), Navkis College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi 590018, India
[3] Department of Electronics and Communication Engineering, Malnad College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi 590018, India
[4] Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi 590018, India
[5] Department of Artificial Intelligence and Machine Learning, BNM Institute of Technology, Bangalore 560070, India
[6] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Mandya 571448, India

Corresponding Author Email: irannaamargol@snpsu.edu.in

https://doi.org/10.18280/isi.310715

Received: 17 April 2026
Revised: 15 June 2026
Accepted: 24 June 2026
Available online: 31 July 2026

Keywords:
expBioXAI-HeartNet, explainable artificial intelligence, adversarial robust deep learning, heart disease prediction, clinical decision support systems, feature optimization, healthcare artificial intelligence

ABSTRACT

Cardiovascular disease (CVD) remains one of the major global health challenges, creating a strong demand for accurate and trustworthy prediction models that can support clinical decision-making. However, many existing deep learning-based diagnostic approaches suffer from limited interpretability and vulnerability to adversarial perturbations, which restrict their practical deployment in healthcare environments. This study proposes BioXAI-HeartNet, an explainable and adversarially robust deep learning framework for heart disease prediction. The proposed framework integrates Particle Swarm Optimization (PSO) for feature selection and hyperparameter optimization, a deep neural network for nonlinear clinical pattern learning, adversarial training based on the Fast Gradient Sign Method (FGSM) for improving model robustness, and explainable artificial intelligence (AI) techniques using SHAP and attention mechanisms for transparent prediction analysis. The effectiveness of the proposed framework is evaluated on three benchmark datasets, including the UCI Heart Disease Dataset, UCI Cleveland Dataset, and MIMIC-IV Dataset. Experimental results demonstrate that BioXAI-HeartNet achieves prediction accuracies of 95.6%, 94.8%, and 93.7%, respectively, with ROC-AUC values exceeding 0.96 across all datasets. Furthermore, the model maintains stable performance under adversarial attacks while providing both global and patient-specific explanations of prediction outcomes. The proposed framework provides a reliable and interpretable solution for developing intelligent clinical decision support systems.

1. INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death globally and requires early and accurate prediction systems for timely clinical action. The increased availability of healthcare data has opened the way for the use of machine learning (ML) and deep learning techniques to predict heart disease. As shown in early efforts using an ensemble of a mixture of traditional and non-traditional methods [1] shown that models can perform more effectively than in previous studies. Additionally, boosting-based ensemble approaches have been applied to both feature selection and classification to improve predictive accuracy and reduce data dimensionality [2]. These examples highlight the continued trend towards using hybrid intelligent systems when predicting CVD risk.

Recent advances in incorporating explainable artificial intelligence (XAI) into predictive frameworks demonstrate how to address the growing issue of model interpretability. For instance, an XAI-enhanced machine-learning heart disease predictive model [3] has improved the transparency of the decision-making process, using XAI with a hybrid ensemble learning framework achieves predictive accuracy while maintaining interpretable results [4]. Also, stacked ensemble learning models [5] have highlighted the importance of using effective feature engineering and XAI in all clinical applications to continue developing improved prediction models. The above approaches are consistent with the overall objective of trustworthy artificial intelligence (AI) [6] and

highlight that explanatory capability is a key prerequisite to deploying AI systems in high-stakes environments.

Many obstacles remain to ensure the reliability and robustness of AI models while these advances were made. Recent research has shown that many ML models including those of XAI are susceptible to adversarial attacks and data disturbances. Study [7] demonstrated that explainable AI models [6] can produce inconsistent and misleading interpretations when faced with corrupt or adversarial inputs; therefore, there are serious concerns about their use in clinical settings. This is particularly true in the healthcare domain where inaccurate predictions may lead to serious consequences. There have also been large-scale studies [8] that demonstrate the utility of AI in cardiovascular risk stratification, and identify the need for robust, reliable systems that generalize across multiple population types.

Integration with new advanced feature learning methodologies, like attention-based models, represents another significant area of research focus. For example, visual-attention-based frameworks have shown improvements in their ability to extract features from medical imaging data [9] and therefore lend themselves well to tabular healthcare data. According to comprehensive reviews on the challenges of predicting heart disease, some key issues that models face include: limited interpretability, vulnerability to attacks from hackers, and an insufficient level of integration with optimization methods [10]. The results from these studies suggest that many existing models have been developed to meet certain criteria e.g., accuracy, interpretability, or robustness but do not work together as a cohesive solution.

In light of this need, the current study presents a new hybrid framework for the prediction of heart disease based on the integrated use of deep learning, biologically inspired optimization methods, explainable AI, and adversarial defences. The proposed hybrid method aims to improve performance and build confidence in the use of predictive models within clinical settings by creating a secure, interpretable, and robust clinical decision support system that can be deployed in real-world healthcare environments.

The major contributions of this research include:

- Novel integration of deep neural networks (DNNs) to create a hybrid method combining both biologically inspired optimization algorithms to provide simultaneous feature selection and hyperparameter tuning.
- Implementation of explainable AI techniques e.g., SHapley Additive exPlanations (SHAP) and attention mechanisms to deliver global and patient-specific interpretability and to build trust with the clinician.
- Use of adversarial training to improve the robustness of models against adversarial or noisy inputs/training and employ attention mechanisms to dynamically assign importance to clinical features to improve predictive performance.
- An extensive evaluation with performance metrics such as accuracy and F1-score on benchmark datasets using several methods including Area under the Receiver Operating Characteristic curve (ROC-AUC).

The remainder of the paper is organised as follows: Literature review on heart disease prediction, explainable AI, and robustness against adversarial attacks in Section 2; Description of the proposed hybrid framework including bio-inspired optimizations, DNN architecture, and XAI module in Section 3; Experimental setup, datasets, and evaluation metrics of the study are outlined in Section 4; and finally conclusions and recommendations for future research will be drawn in Section 5.

2. RELATED WORK

In recent years, advancements in prediction models for CVD have taken advantage of the potential of ML and deep learning to improve both the accuracy of diagnosis and the ability to make better clinical decisions. For example, a systematic comprehensive review and meta-analysis of ML model implementation on electronic health record (EHR) data by Liu et al. [11] supported the earlier claim that ML models outperform traditional statistical analysis for risk prediction, and therefore help to make better clinical decisions than traditional statistical analysis. Despite this success, the study noted the following challenges: the data used to develop ML models tends to be heterogeneous; the predictions made by ML models are difficult to interpret; and ML models tend to not be generalizable across different patient populations. These challenges have motivated researchers to develop hybrid approaches and also use deep learning methods for more robust performance and better outcomes than originally implemented in the predictive models developed using traditional statistical analysis.

Many types of deep learning models have proven to be advantageous in identifying complex nonlinear relationships in medical data. For example, Sudha and Kumar [12] proposed a hybrid model using a convolutional neural network (CNN) and a long short-term memory (LSTM) network to simultaneously learn both spatial and temporal features to predict heart disease. The hybrid model produced higher rates of accurate classifications than the traditional methods. Rajanna et al. [13] proposed using convolutional neural networks as the basis of a fuzzy inference system in a similar manner to create an improved predictive model for predicting heart disease by providing uncertainty management capabilities in the model. Similarly, many researchers have also started using attention-based deep learning models to create a model that produces more accurate predictions by focusing on which features are clinically relevant for improved overall predictive capability. For example, Xu et al. [14] developed an attention-based model to predict major adverse cardiovascular events, demonstrating that the use of attention mechanisms as part of a predictive model can provide significant improvements in the ability to identify clinically relevant features.

XAI is a new focus area of research that is emerging to address the challenge of understanding and interpreting the output of AI algorithms, including those used in deep learning applications. Literature reviews by Nazir et al. [15] and Tjoa and Guan [16] provide comprehensive summaries of recent research related to XAI approaches within healthcare, such as using feature attribution, saliency maps, and model-agnostic explanations. Moreover, the authors Hassija et al. [17] have presented examples of different approaches for interpreting AI algorithms using explanatory techniques that provide greater transparency in high-stakes situations like medical diagnosis. Foundational works by Holzinger et al. [18] introduced the concept of causability as part of the multi-modal explanation of AI, emphasizing that explanations must be understandable to humans and must align with how clinicians think and operate.

Barredo Arrieta et al. [19] characterized key challenges and opportunities associated with the use of XAI within healthcare,

stressing that AI-based systems must have responsible and trustworthy capabilities. However, incorporating XAI into predictive algorithms may create additional complexity in computing resources as well as providing no inherent guarantee to the robustness of the predictive algorithms. The vulnerability of medical ML systems to adversarial manipulation was highlighted for the first time by Finlayson et al. [20]. This led to concerns about their utility in a clinical environment. Recently, Zahoor and Ghani [21] proposed robust deep learning models with adversarial defence mechanisms and showed they were more resilient than traditional models in the diagnostic process for medical images. However, the use of robustness strategies in heart disease prediction models, especially in conjunction with explainability, is seldom examined.

In addition, bio-inspired algorithmic frameworks have been widely used to improve algorithmic performance by optimizing both feature selection and hyperparameter tuning. A thorough review of Particle Swarm Optimization (PSO) was provided by Shami et al. [22], providing evidence of the effectiveness of PSO to address difficult optimisation problems. Soladoye et al. [23] also illustrated the potential of PSO in solving problems in the health domains such as reducing dimensionality and optimising models. Tawfeek et al. [24] propose a hybrid AI/ ML model for diagnosing CVDs that uses AI to provide personalized diagnoses and risk assessments. Although these hybrid approaches may improve the prediction accuracies of CVDs, they usually do not include all the required components of a complete pipeline, such as interpretability, robustness, and optimization.

Earlier research shows the successes achieved by applying deep learning principles to image classification and cybersecurity, there are still some limitations. The existing methods seem to primarily focus on improving performance without adequately addressing real attacks. The combination of adversarial training and optimized deep learning approaches receives little attention, too, and thus, the developed approaches work efficiently under normal conditions but fail to perform well under adversarial circumstances. Another flaw in the examples analyzed is the inadequacy of comparative studies across various attack conditions. Thus, there is a need for a framework to assess the reliability of proposed approaches across different conflict situations.

3. PROPOSED METHODOLOGY

The proposed BioXAI-HeartNet framework is designed as a hybrid architecture that combines deep learning technology, bio-inspired optimization techniques, XAI, and adversarial robustness to create a system capable of predicting heart disease. The five primary components of the BiXAI-HeartNet framework are data pre-processing, bio-inspired feature selection, optimized DNN modelling, enhancing adversarial robustness, and explainability-based interpretation. The design of this framework is depicted in Figure 1.

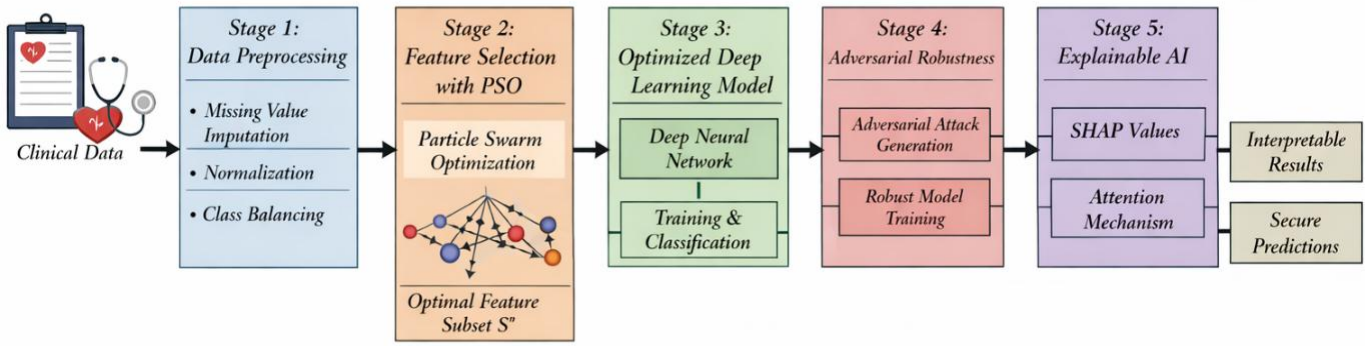

Figure 1. Design diagram of the proposed framework

The unique contribution of BioXAI-HeartNet is not only about combining these techniques but also about using them as a foundation for a collaborative framework under one umbrella, where all components are connected and work together to provide trustworthy heart disease predictions. The framework solves three major challenges in the clinical use of AI: feature redundancy, interpretability gaps, and susceptibility to adversarial perturbations. The process begins with PSO to identify the most relevant features and eliminate unnecessary or overlapping characteristics. The optimal feature mixture is then fed into the DNN, which builds effective models of nonlinear associations among heart disease risk factors. Thus, adversarial training using the Fast Gradient Sign Method (FGSM) can ensure reliability and robustness during model construction. One advantage of such training is the ability to learn from both original and altered samples. In model training, the application of attention mechanisms enables assigning importance to specific features. In this situation, SHAP provides quantitative measures of the contributions of individual clinical variables to both global prediction and prediction for the specific patient. In this regard, it can be said that prediction, optimization, explanation, and robustness are interrelated and cannot be considered independent modules. In a practical context, such a combination means that medical practitioners start receiving predictions that are accurate and at the same time interpretable.

The mathematical formulation for each stage of the processing is described in detail below. The dataset can be expressed using Eq. (1):

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

where, N denotes the total number of samples in the dataset. $x_i \in R^d$ denotes the feature vector for the i-th patient, comprised of d features (e.g., age, cholesterol levels, and blood pressure). $y_i \in \{0,1\}$ denotes the corresponding class label for each patient, indicating whether or not they have heart disease. The input dataset can be represented in matrix form with $X \in R^{N \times d}$, and the labels represented in vector form with $Y \in R^N$.

3.1 Data preprocessing

The preprocessing of the clinical dataset is necessary due to the presence of missing values, noise in the data, and an unequal distribution of class labels across the dataset. There are three stages to preprocessing: The first stage involves using K-nearest neighbors to find a replacement value for each feature that has a missing value according to Eq. (2).

$$x_{ij}^{*} = \frac{1}{k} \sum_{l \in \mathcal{N}_k(i)} x_{lj} \tag{2}$$

where, x_{ij}^{*} denotes the value of the *j*-th feature for the *i*-th patient after imputation has occurred. $N_k(i)$ denotes the k nearest neighbors to the *i*-th patient. The second stage involves scaling all features to be in a common range (scale normalization) according to Eq. (3).

$$x_{ij}^{norm} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3}$$

where, μ_j denotes the mean value of the *j*-th feature across all of its samples and σ_j denotes the standard deviation of the *j*-th feature across all of its samples. The final stage will use an oversampling technique called synthetic minority oversampling to create enough instances of both classes to create an equal representation of both classes in the final processed dataset.

3.2 Bio-inspired feature selection using Particle Swarm Optimization

PSO has been applied as a tool for performing feature subset identification to find the most informative set of features and to reduce the number of dimensions being considered. Each particle represents a candidate solution as shown in Eq. (4).

$$P_i = (p_{i1}, p_{i2}, \ldots, p_{id}) \tag{4}$$

where, $p_{ij} \in \{0,1\}$ indicates if the *j*-th feature has been selected. The velocity update has the following Eq. (5).

$$\begin{aligned} v_{ij}^{t+1} = w v_{ij}^{t} &+ c_1 r_1 \left(pbest_{ij} - p_{ij}^{t}\right) \\ &+ c_2 r_2 \left(gbest_j - p_{ij}^{t}\right) \end{aligned} \tag{5}$$

and the position update can be expressed as Eq. (6).

$$p_{ij}^{t+1} = \begin{cases} 1, & \text{if } sigmoid(v_{ij}^{t+1}) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where, v_{ij}^{t} is the current velocity of the particle *i* at time (*t*), *w* denotes the inertia weight controlling the degree of exploration in the search space, c_1 and c_2 are the cognitive and social coefficients, respectively; and r_1 and $r_2 \sim U(0,1)$ are random values. The p best corresponds to the personal best position of each particle, while g best is the global best position of all particles. The velocity of each particle is transformed to probability via a sigmoid and τ is the threshold for the probability. The fitness function that has been defined for feature set evaluation is given in Eq. (7).

$$F = \alpha \cdot Acc + (1 - \alpha) \cdot \left(1 - \frac{|S|}{d}\right) \tag{7}$$

where, *Acc* is the classification accuracy using selected features, $|S|$ is the total number of selected features, and $\alpha \in [0,1]$ is the ratio between accuracy and feature reduction.

3.3 Optimized deep neural network

The subset of selected features is used as the input to a DNN for classification. The equation for forward propagation through the DNN can be defined as Eq. (8).

$$h^{l} = \sigma\left(W^{l} h^{(l-1)} + b^{l}\right) \tag{8}$$

where, h^{l} is the output from *l*-th hidden layer, W^{l} is the weight matrix at layer *l*, b^{l} is the bias vector at layer *l*, and $\sigma(\cdot)$ is the activation function (ReLU). The output layer has a sigmoid activation function is given as Eq. (9).

$$\hat{y}_i = \frac{1}{1 + e^{-z_i}} \tag{9}$$

where, $\hat{y}_i$ is the probability of heart disease prediction for sample *i* and z_i is the linear output of the network's final layer. The classification loss function will be stated using binary cross-entropy as shown in Eq. (10).

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{10}$$

3.4 Adversarial robustness using Fast Gradient Sign Method

The FGSM is used to create adversarial samples, thereby providing a method to increase model robustness. For example, to create an adversarial sample of each sample x_i from the training data set, the model will take the following steps shown in Eq. (11).

$$x_i^{adv} = x_i + \epsilon \cdot sign\left(\nabla_{x_i} L_{cls}\right) \tag{11}$$

where, ϵ defines the amount of perturbation (i.e., attack strength), and $\nabla(x_i)L_{cls}$ is the gradient of the classification loss function with respect to the input features. The total loss function combines both classification and adversarial loss as shown in Eq. (12).

$$L_{total} = L_{cls} + \lambda L_{adv} \tag{12}$$

where, L_{adv} represents the amount of relative weight assigned to the contribution made from enhancing the model's robustness λ to adversarial examples.

3.5 Explainability using SHapley Additive exPlanations and attention

Using SHAP and Attention for Explainability will become a function of SHAP values and attention weighted methods based on results from training to understand how each sample contributes globally (dataset level) and locally (individual patient level). The SHAP value for feature *j* is defined as Eq. (13).

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (13)$$

where, F is the full set of features, and S represents all the possible subsets of features. The value ϕ_j applications of the feature for predictions. Additionally, an attention mechanism will apply a corresponding weight on each feature using Eq. (14).

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^{d} \exp(e_k)} \quad (14)$$

where, α_j is the attention weight for feature j based on the training process and e_j corresponds to the learned importance or relevance for feature j. Mechanisms are provided for interpretability of both total loss functions calculated at the dataset level and for each data point at the patient level.

4. RESULTS AND DISCUSSIONS

This section provides the experimental setup and result discussion of the proposed model. Python programming is used to implement the deep learning framework via TensorFlow and Scikit-learn libraries. The data set consists of training and test sets that have been divided at a 70:30 split, the model can be tested with new and unseen data. In addition to the 70:30 data split five-fold cross-validation is implemented during training to produce a more accurate estimate of the overall model's predictability. The DNN is trained using an Adam optimizer with a starting learning rate of 0.001 is updated based on validation performance throughout the training process. The bio-inspired optimization process is based on PSO and is assigned a population of 30 particles to operate on and a maximum iteration limit of 50. The inertia weight for PSO is set to 0.7, this parameter balanced exploratory and exploitative behaviour. Both cognitive and social coefficient values for the particles are set equally at 1.5. The fitness function is developed to include both the classification accuracy of each trained model and the number of features included in each feature set.

4.1 Datasets

To determine the effectiveness of the proposed BioXAI-HeartNet framework, experiments are carried out on three benchmark cardiovascular datasets to determine the model's generalization and robustness across different datasets. The UCI Heart Disease Dataset is primarily used for the experiments, which contains clinical data such as age, gender, and the type of chest pain as well as features such as resting blood pressure, cholesterol, fasting blood glucose levels, and ECG results. This dataset also includes binary and multi-class labels where the presence, and severity of heart disease is indicated. In addition, the MIMIC-IV Dataset will be used to evaluate the model in a real clinical environment. Furthermore, to assess the robustness of the proposed framework against different data distributions, experiments will be conducted using a processed version of the UCI Cleveland Dataset that is commonly used as a benchmark for heart disease prediction models.

The datasets that are used will be pre-processed with missing value imputation, normalization of the input features to create balance within each class, and create consistent results for all experiments. By using multiple datasets, comprehensive validation will be done on the ability of the model to predict outcomes as well as to generalize its predictive performance across different datasets. The method used to implement against adversarial attacks is based on the FGSM attacks models by creating input perturbations level of 0.01 The model is trained using both the original input as well as inputs that are subjected to FGSM through this training and the model improved its robustness to input perturbations. Additionally, dropout regularization with a rate of 0.3–0.5 is used during training to avoid overfitting issues within the model. Furthermore, batch normalization is used to enhance the stability of the model during training.

This DNN model has an architecture consisting of 3 hidden layers having sizes of 128, 64 and 32 neurons respectively. The activation function used in the hidden layers is the rectified linear unit but the output layer uses a sigmoid activation for binary classification. The batch size is 32 and the model trained for 100 epochs. To prevent the model from overfitting, an early stopping method based on validation loss is used. Using a PSO-based feature selection method dynamically selects the best subset of features from the feature space. The method reduces the feature space between 20%–40% without losing accuracy in terms of making accurate predictions. Regularization coefficient (λ) is used to regulate adversarial training so that the balance between accuracy of classification and robustness is achieved; this coefficient is set to 0.5 based on empirical data.

4.2 Evaluation metrics

To effectively evaluate how well the design does its job and how efficiently it accomplishes its task, many different ways to look at performance will be considered. The most basic of these measures is accuracy a measure that tells you the proportion of correct predictions made by the Design. Other metrics include precision (a measure of the proportion of positive examples identified as such) and recall (a measure of the proportion of actual positive examples that are identified as such). The F1-score, which provides an average of both precision and recall, serves as a good example of an overall performance metric. Finally, the ROC-AUC, a measure of how well the Design performs, provides a great measure of performance at varying thresholds. In equations, the above metrics can be expressed mathematically as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

$$Recall = \frac{TP}{TP + FN} \quad (17)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (18)$$

where, TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

Table 1. Clinical features selected by Particle Swarm Optimization (PSO)

Feature	Selected
Age	Yes
Sex	Yes
Chest Pain Type (CP)	Yes
Resting Blood Pressure	Yes
Cholesterol	Yes
Fasting Blood Sugar	Yes
Resting ECG Results	Yes
Maximum Heart Rate Achieved (Thalach)	Yes
Exercise-Induced Angina	Yes
ST Depression (Oldpeak)	Yes
Slope of ST Segment	Yes
Number of Major Vessels (CA)	Yes
Thalassemia (Thal)	Yes
Other Redundant Attributes	No

PSO is applied in order to optimize the efficiency of the model as well as to decrease the redundancy of the features. PSO allowed for picking the optimal clinical features right from the entire set of features by maximizing the classification performance of the selected features and minimizing the dimensionality of the solution according to the fitness function introduced in Eq. (7). The performed optimization is able to provide a reduction in the feature space by about 30–40% without any significant loss of the quality of prediction. Table 1 shows the selected features by PSO.

The selected features have good agreement with the known features related to CVD. It can be seen that four variables are found to be consistently important according to the assigned importance scores during both optimization and SHAP analysis. For the variables identified through the optimization process, the most common attributes are those known as the most important ones for the cardiac disease and coronary artery disease. The PSO algorithm helped to remove irrelevant or weak features while preserving the relevant clinical information. The dimensionality reduction improved the computational efficiency of the model as well as increased the chances of overfitting. Besides, the coincidence between the variables selected by PSO and the ones that are important according to the SHAP explanatory module serves as an additional proof of the reliability of the proposed feature selection method.

4.3 Performance evaluation

Five different evaluative measurements are used to assess the performance of the new BioXAI-HeartNet framework on three different benchmark datasets (the UCI Heart Disease Dataset, UCI Cleveland Dataset, and MIMIC-IV Dataset). These evaluative measurements include accuracy, precision, recall, F1-score, and ROC-AUC curve.

The model produced the best overall accuracy of 95.6% on the UCI Heart Disease Dataset as illustrated in Table 2 and Figure 2. This indicates that the new model has the ability to accurately learn patterns from structured data. Additionally, the UCI Cleveland Dataset produced similar results, as the new model indicated a 94.8% accuracy rate. Therefore, it can be inferred that the model has the ability to generalize across multiple benchmark datasets. Furthermore, the UCI Cleveland and MIMIC-IV datasets produced accuracies of 94.8% and 93.7%, respectively. Thus, supporting the conclusion that the new model has demonstrated to be robust and effective in learning from heterogeneous data. Each of the evaluated datasets also produced very high precision and recall scores, that indicates that the new model will reduce the number of false positive and false negative predictions. Each of the evaluated datasets produced high F1-scores, which indicate that the new model demonstrates a balanced predictive performance between false positives and false negatives. In addition, each of the evaluated datasets produced a ROC-AUC value of greater than 0.96. This implies a high level of discriminatory ability provided by the new model, thus making it an ideal and reliable candidate for providing clinical decision support.

Table 2. Performance of proposed model on different datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
UCI Heart Disease	95.6	95.1	94.8	94.9	0.978
UCI Cleveland	94.8	94.2	93.9	94.0	0.972
MIMIC-IV	93.7	93.1	92.6	92.8	0.965

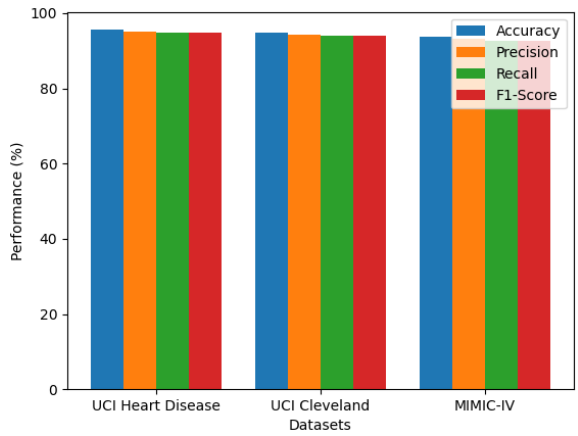

Figure 2. Performance of proposed model on different datasets

The performance of the new BioXAI-HeartNet model under FGSM-based adversarial attacks is evaluated in Table 3 using three datasets: UCI Heart Disease Dataset, UCI Cleveland Dataset, and MIMIC-IV dataset. While there is a slight drop in accuracy from unperturbed to perturbed inputs, the model performed well overall with all data sets reflecting over 86% accuracy with respect to their adversarial perturbed examples. All robustness scores are above or equal to 0.90; therefore, the model appears robust against input perturbations. Thus, these findings suggest that the adversarial training procedure successfully improves the model's robustness and applicability for secure clinical use cases with a critical need for robustness.

The performance variations observed across the three benchmark datasets can be attributed to differences in their underlying characteristics. Factors such as sample size, class balance, feature distribution, and data complexity significantly influence model learning and generalization capabilities. Datasets containing larger numbers of training samples and well-separated feature representations tend to produce higher classification accuracy and more stable robustness performance. In contrast, datasets with limited samples, overlapping class boundaries, or greater feature variability

pose more challenging learning conditions, potentially resulting in lower performance. Additionally, differences in data distributions can affect the model's sensitivity to adversarial perturbations, leading to varying robustness levels across datasets. These findings highlight the importance of evaluating the proposed framework on multiple datasets to ensure its effectiveness under diverse data characteristics and real-world conditions.

Table 3. Robustness performance under adversarial attacks

Dataset	Clean Accuracy (%)	Adversarial Accuracy (%)	Robustness Score
UCI Heart Disease	95.6	89.4	0.93
UCI Cleveland	94.8	88.1	0.92
MIMIC-IV	93.7	86.5	0.92

Figure 3 gives the overall interpretability of the constructed BioXAI-HeartNet system, we evaluated the feature dynamics across the dataset from the perspective of the mean absolute Shapley values. The resultant global attribution matrix showed that the Chest Pain Type (CP) is the most influential feature in the diagnostic classification with a mean SHAP value of 0.38, followed by key physiological variables such as ST Depression (Oldpeak) with a value of 0.31, Maximum Heart Rate Achieved (Thalach) with a value of 0.27, and the Number of Major Vessels (CA) with a value of 0.24. Other demographic and metabolic factors that had relatively small impacts on the decisions made by the system, including Age (0.16) and Serum Cholesterol (0.12), provided a minor baseline modulation. Overall, this ranking reveals the mathematical consistency of the whole system, proving that it is aligned with those aspects indicated by our PSO feature selection method as well as the corresponding scientific knowledge on CVDs.

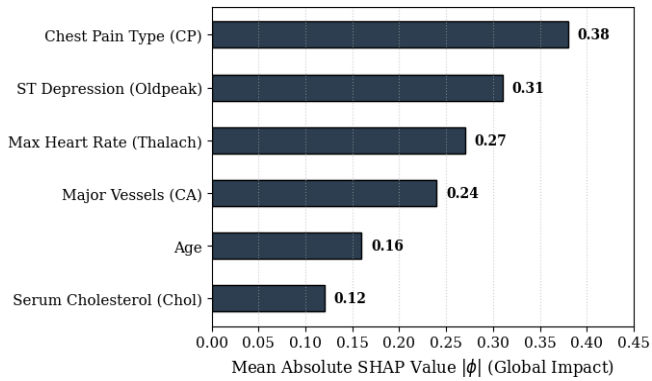

Figure 3. Global feature importance attribution via SHapley Additive exPlanations (SHAP) matrix

The local explanatory power can be achieved for individual clinical decision support, we explored a case study at the patient level such that local additive feature contributions can be illustrated using a SHAP waterfall plot shown in Figure 4. With a cohort random prediction baseline of 0.31, the actual diagnostic probability vector is substantially increased due to specific clinical profile variables in red such as being asymptomatic (CP Type 4), presence of extreme ECG changes (Oldpeak 3.2 mm), bradycardic heart rate pattern (Thalach 112), and hyperlipidemia (Chol 294). This risk increase exceeds the threshold of 0.50 and then slightly reduced by protective clinical features (in blue) such as a normal blood glucose value (FBS) and stable ECG readings (RestECG), resulting in an actual diagnosis risk of 0.84. This patient-centered approach eliminates the classic danger of untransparent "black-box" solutions and provides the healthcare professionals with reliable straightforward scientific knowledge for diagnosis validation and communication.

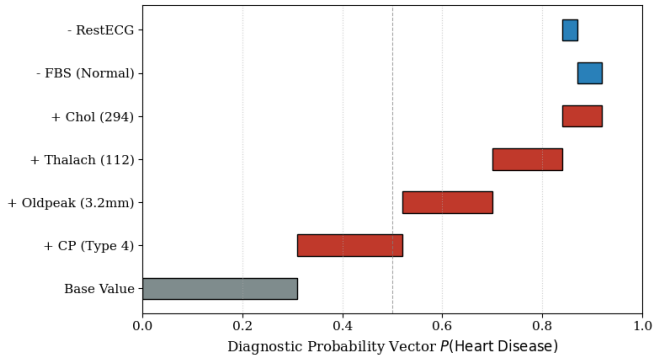

Figure 4. Local diagnostics attribution graph

Table 4 shows how each part of the BioXAI-HeartNet framework contributes to its performance by showing the results of an ablation study. The baseline deep learning model has a 91.3 percent accuracy and increases to 93.9 percent with the addition of PSO feature selection, showing that PSO to get the optimal features is important. Adding adversarial training further increases performance to 94.8 percent, providing additional robustness and generalizability. Adding the XAI module keeps performance comparable (94.5%) while allowing for interpretation of results. The total accuracy for the full BioXAI-HeartNet model is 95.6 and an F1-score of 94.9. Overall, this suggests that including all three components together provides the best overall performance.

Table 4. Ablation study of proposed framework

Model Variant	Accuracy (%)	F1-Score (%)
DL Only	91.3	90.8
DL + PSO	93.9	93.4
DL + PSO + Adversarial Training	94.8	94.3
DL + PSO + XAI	94.2	93.8
Full Model (BioXAI-HeartNet)	95.6	94.9

Note: DL = deep learning, PSO = Particle Swarm Optimization, XAI = explainable artificial intelligence (XAI).

In order to assess more deeply the reliability of the proposed BioXAI-HeartNet framework, adversarial examples are created using FGSM with perturbation levels of 0.01. It happened actually as it is expected that the accuracy of the model dropped as the perturbation changes increased. The perturbations of small values did not cause significant degradation in performance indicating that the features learned by both networks are consistent. In case of introduction of big perturbations, there is some decline in the accuracy of the model. The classification proved to be acceptable in all cases.

The stable performance is explained by the adversarial training approach applied by this framework. During training the model is provided with samples of both clean data and perturbed data. As a result of this, this model is able to find the

decision boundaries less sensitive to smaller perturbations produced by variance of input data.

Moreover, the PSO-based feature selection method allows for removing attributes that do not have significant value for the formation of discriminative features.

Also, the role of the attention mechanism is to aid in the ranking of pertinent features during the training phase thus eliminating inferior inputs that do not contribute any diagnostic value during medical practice. Overall, this proves that the incorporation of feature optimization, adversarial training as well as attention-enabled learning makes it possible for BioXAI-HeartNet to remain stable in its prediction even when faced with adversarial attacks. This is particularly important in the healthcare field since it plays a vital role in the provision of viable medical decision support systems to its practitioners.

The framework known as BioXAI-HeartNet has been compared against the current best performing prediction methods in two datasets: UCI Cleveland Dataset and UCI Heart Disease Dataset (see Table 5). Ultimately, this model performed better than the other methodologies on either dataset with an accuracy of 95.6%. The performance also exceeded other methodologies including: Two-Level Boosting and CNN-LSTM to name a few, for Cleveland; and fuzzy logic, decision tree based and Genetic Algorithm – Support Vector Machine (GA-SVM) for UCI. In terms of improvement performance over other methodologies, the BioXAI-HeartNet model shows to have improved performance between approximately 2% to 17% which further confirms the effectiveness of unifying bio-inspired optimization, deep learning, adversarial robustness and explainable AI into one overall framework. Therefore, these findings demonstrate that the BioXAI-HeartNet model is the more optimal and reliable choice to predict heart disease compared to all of the existing methodologies.

Table 5. Comparative analysis of the proposed model with state-of-the-art methods

Methodologies	Dataset	Accuracy (%)
Two-Level Boosting [2]	Cleavland	93.44
CNN-LSTM Hybrid [12]	Cleavland	89
Fuzzy-CNN Model [13]	Cleavland	91.1
BioXAI-HeartNet (Proposed)	**Cleavland**	**95.6**
Fuzzy rules + DT [25]	UCI	88
Hybrid DT [26]	UCI	78
Rule based fuzzy logic (RBFL) [27]	UCI	78
GA-SVM [28]	UCI	90
DCNN [29]	UCI	91.7
BioXAI-HeartNet (Proposed)	**UCI**	**95.6**

Note: CNN = convolutional neural network, LSTM = long short-term memory, GA-SVM = Genetic Algorithm – Support Vector Machine.

Table 5 shows that the baseline results are selected because they compare their methodologies against the same benchmark datasets used in the current analysis and present similar performance indicators. Although some variations in the experimental design may exist, they are not significant enough to invalidate the comparison. The proposed model is tested against adversarial attacks and achieves strong results, demonstrating its reliability in clinical environments. The integrated XAI methods provide clear predictions. The framework supports the secure deployment of this model in healthcare.

Despite the good results achieved by the developed BioXAI-HeartNet system, it has some drawbacks. First, even though the research used three well-known datasets, the overall sample size is relatively small compared to other healthcare databases. This can negatively affect the system's generalizability to other patient populations. Second, only retrospective data are used for the research. Although using multiple databases helps determine whether the system can be generalized, the evidence would be more robust if several medical centers from different regions of the globe participated in verifying the BioXAI-HeartNet system. Third, the tests are conducted only under FGSM attacks; therefore, further research should address other attack methods, such as PGD, BIM, and others.

In practice, the deployment of AI in clinical decision support systems requires validation, regulatory compliance, and seamless integration with EHR, resulting in a clinician-centered assessment of explainability outputs. All these aspects have to be considered in future research. Thus, future studies will inevitably involve multicenter, larger-scale studies, the use of various types of healthcare data, the further implementation of techniques for privacy preservation in healthcare analysis, and the improvement of current techniques for protecting AI systems against adversarial attacks.

5. CONCLUSION

A unique hybrid structure called BioXAI-HeartNet has been developed, which uses deep learning, bio-inspired optimization, explainable AI, and adversarial robustness to predict heart disease. The datasets used for this research study are UCI Heart Disease Dataset, UCI Cleveland Dataset, and MIMIC-IV achieved the accuracy of BioXAI-HeartNet is 95.6%, 94.8%, and 93.7%, accordingly to each dataset. Also, the ROC-AUC values on each of the datasets exceeded 0.96. The robustness of BioXAI-HeartNet is 85% or more on each dataset indicating only slight degradation in performance with the application of adversarial conditions. Therefore, only slight degradation occurred in the applicants' performance when the adversarial conditions are used. By utilizing a PSO-based feature selection process, the efficiency of the entire process is improved. Therefore, applicants have improved trust and confidence in the clinical data and the results. Finally, through the use of XAI techniques, the interpretability of clinical and diagnostic results is achieved. In future work, it can expand the framework to include multi-modal data sources such as ECG signals and medical imaging, as well as to incorporate federated learning for distributed healthcare systems, maintaining individual patient privacy. It can also explore the real-time deployment and validation of the framework in clinical settings to further increase the practical utility of the proposed framework.

REFERENCES

[1] Riyaz, L., Butt, M.A., Zaman, M. (2022). A novel ensemble deep learning model for coronary heart disease prediction. Revue d'Intelligence Artificielle, 36(5): 825-832. https://doi.org/10.18280/ria.360602

[2] Dissanayake, K., Johar, M.G.M. (2023). Two-level

boosting classifiers ensemble based on feature selection for heart disease prediction. Indonesian Journal of Electrical Engineering and Computer Science, 32(1): 381-391. https://doi.org/10.11591/ijeecs.v32.i1.pp381-391

[3] El-Sofany, H., Bouallegue, B., El-Latif, Y.M. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14: 23277. https://doi.org/10.1038/s41598-024-74656-2

[4] Shah, P., Shukla, M., Dholakia, N.H., Gupta, H. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. Scientific Reports, 15: 17927. https://doi.org/10.1038/s41598-025-01650-7

[5] Ghose, P., Oliullah, K., Mahbub, M.K., Biswas, M., Uddin, K.N., Jamil, H.M. (2025). Explainable AI assisted heart disease diagnosis through effective feature engineering and stacked ensemble learning. Expert Systems with Applications, 265: 125928. https://doi.org/10.1016/j.eswa.2024.125928

[6] Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99: 101805. https://doi.org/10.1016/j.inffus.2023.101805

[7] Repetto, S., Maljkovic, I., Lotto, M., Cinà, A.E., Vascon, S., Roli, F. (2026). Evaluating the robustness of explainable AI in medical image recognition under natural and adversarial data corruption. Machine Learning, 115: 4. https://doi.org/10.1007/s10994-025-06919-6

[8] Dhingra, L.S., Aminorroaya, A., Sangha, V., et al. (2025). Heart failure risk stratification using artificial intelligence applied to electrocardiogram images: A multinational study. European Heart Journal, 46(11): 1044-1053. https://doi.org/10.1093/eurheartj/ehae914

[9] Tomar, N., Chandel, S., Bhatnagar, G. (2024). A visual attention-based algorithm for brain tumor detection using an on-center saliency map and a superpixel-based framework. Healthcare Analytics, 5: 100323. https://doi.org/10.1016/j.health.2024.100323

[10] Kumar, R., Garg, S., Kaur, R., et al. (2025). A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions. Frontiers in Artificial Intelligence, 8: 1583459. https://doi.org/10.3389/frai.2025.1583459

[11] Liu, T.Y., Krentz, A., Lu, L., Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1): 7-22. https://doi.org/10.1093/ehjdh/ztae080

[12] Sudha, V.K., Kumar, D. (2023). Hybrid CNN and LSTM network for heart disease prediction. SN Computer Science, 4: 172. https://doi.org/10.1007/s42979-022-01598-9

[13] Rajanna, S., Jayaramaiah, C., Sridhar, R., Chandrappa, P.H., Venkatesh, R.T. (2023). Fuzzy inference with enhanced convolutional neural network based classification framework for predicting heart attack using sensor data. Revue d'Intelligence Artificielle, 37(1): 93-99. https://doi.org/10.18280/ria.370112

[14] Xu, Z.Y., Xu, X., Zhu, X.M., Niu, K., Dong, J., He, Z.Q. (2024). Attention-based deep learning model for prediction of major adverse cardiovascular events in peritoneal dialysis patients. IEEE Journal of Biomedical and Health Informatics, 28(2): 1101-1109. https://doi.org/10.1109/JBHI.2023.3338729

[15] Nazir, S., Dickson, D.M., Akram, M.U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Computers in Biology and Medicine, 156: 106668. https://doi.org/10.1016/j.compbiomed.2023.106668

[16] Tjoa, E., Guan, C. (2021). A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11): 4793-4813. https://doi.org/10.1109/tnnls.2020.3027314

[17] Hassija, V., Chamola, V., Mahapatra, A., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation, 16: 45-74. https://doi.org/10.1007/s12559-023-10179-8

[18] Holzinger, A., Malle, B., Saranti, A., Pfeifer, B. (2021). Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Information Fusion, 71: 28-37. https://doi.org/10.1016/j.inffus.2021.01.008

[19] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58: 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[20] Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S. (2019). Adversarial attacks on medical machine learning: Emerging vulnerabilities demand new conversations. Science, 363(6433): 1287-1289. https://doi.org/10.1126/science.aaw4399

[21] Zahoor, K., Ghani, U. (2025). Robust deep learning model for adversarial defense in medical image diagnosis. In 2025 International Conference on Emerging Technologies in Electronics, Computing, and Communication (ICETECC), Jamshoro, Pakistan, pp. 1-6. https://doi.org/10.1109/ICETECC65365.2025.11070282

[22] Shami, T.M., El-Saleh, A.A., Alswaitti, M., Al-Tashi, Q., Summakieh, M.A., Mirjalili, S. (2022). Particle swarm optimization: A comprehensive survey. IEEE Access, 10: 10031-10061. https://doi.org/10.1109/access.2022.3142859

[23] Soladoye, A.A., Joseph, T.S., Sobowale, A.A., Bello, O.J., Analikwu, C.C., Olawade, D.B. (2026). Particle swarm optimization in medical applications: A systematic review of data dimension reduction and algorithmic configuration. Intelligent Hospital, 2(2): 100052. https://doi.org/10.1016/j.inhs.2026.100052

[24] Tawfeek, M.A., Alrashdi, I., Alruwaili, M., Allahem, H. (2025). Cardiovascular disease detection: A hybrid machine learning-AI framework for personalized diagnosis and risk assessment. PLoS One, 20(10): e0335421. https://doi.org/10.1371/journal.pone.0335421

[25] Pathak, A.K., Valan, J.A. (2019). A predictive model for heart disease diagnosis using fuzzy logic and decision tree. In Smart Computing Paradigms: New Progresses and Challenges. Advances in Intelligent Systems and Computing, pp. 131-140. https://doi.org/10.1007/978-981-13-9680-9_10

[26] Maji, S., Arora, S. (2019). Decision tree algorithms for

prediction of heart disease. In Information and Communication Technology for Competitive Strategies. Lecture Notes in Networks and Systems, Springer, Singapore, pp. 447-454. https://doi.org/10.1007/978-981-13-0586-3_45

[27] Reddy, G.T., Khare, N. (2017). An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model. Journal of Circuits, Systems and Computers, 26(4): 1750061. https://doi.org/10.1142/S021812661750061X

[28] Murad, S.H., Tayfor, N.B., Mahmood, N.H., Arman, L. (2025). Hybrid genetic algorithms-driven optimization of machine learning models for heart disease prediction. MethodsX, 15: 103510. https://doi.org/10.1016/j.mex.2025.103510

[29] Arooj, S., Rehman, S.U., Imran, A., Almuhaimeed, A., Alzahrani, A.K., Alzahrani, A. (2022). A deep convolutional neural network for the early detection of heart disease. Biomedicines, 10(11): 2796. https://doi.org/10.3390/biomedicines10112796