



OPEN Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment

Moiz Ur Rehman¹, Shahid Naseem¹, Ateeq Ur Rehman Butt², Tariq Mahmood^{3,6}, Amjad Rehman Khan³, Inayat Khan⁴, Jawad Khan⁵ & Younhyun Jung⁵✉

Worldwide, coronary heart disease (CHD) is a leading cause of mortality, and its early prediction remains a critical challenge in clinical data analysis. Machine learning (ML) offers valuable diagnostic support by leveraging healthcare data to enhance decision-making and prediction accuracy. Although numerous studies have applied ML classifiers for heart disease prediction, their contributions often lack clarity in addressing key challenges. In this paper, we present a comprehensive ML framework that systematically tackles these issues. First, we employ mutual information (MI) for effective feature selection to isolate the most informative predictors. Second, we address the significant class imbalance in the dataset using the Synthetic Minority Oversampling Technique (SMOTE), which substantially improves model training. Third, we propose a novel hybrid model that integrates particle swarm optimization (PSO) with an artificial neural network (ANN) to optimize feature weighting and bias training. Additionally, we conduct a comparative analysis with traditional classifiers, including Logistic Regression and Random Forest, using the National Health and Nutritional Examination Survey dataset. Our results demonstrate that while conventional classifiers achieve an accuracy of 95.8%, the proposed PSO-ANN model attains an enhanced accuracy of up to 97% in predicting CHD. This work clearly defines its contributions by improving feature selection, handling data imbalance, and introducing an innovative hybrid model for superior prediction performance.

Keywords Coronary Heart Disease, Machine Learning, Mutual Information, Particle Swarm Optimization, Artificial Neural Network, Health Informatics, Medical Images, Health Issues

The healthcare sector is one of the largest data producers, which concerns patients, diagnoses, and medical treatments¹. Nevertheless, most of this information is still untapped and cannot provide much valuable information. CHD is currently reported to be the most prevalent killer disease globally, contributing to more than 17 million deaths per year, as estimated by WHO². CVDs encompass a broad classification of diseases that may include coronary artery diseases—a disease affecting the blood vessels of the heart^{3,4}. The formation of the lipoprotein-based plaques in these arteries results in atherosclerosis, thus reducing blood flow and raising the odds of heart attacks and strokes. Sadly, though, four out of five CVD-related deaths can be attributed to these events, and one-third of this total is of people who are below 70 years old⁵. There are several causes of CHD: genetics, age and sex, high cholesterol, obesity, smoking and a sedentary lifestyle⁶. Even though signs of CHD can be detected through MRI, angiography, and x-rays, the equipment is expensive and widely unavailable worldwide^{7,8}. Early ambulatory diagnosis always lowers the mortality rate; however, due to the inadequacy of resources and medical gadgets, many receive treatment way past the appropriate time, hence contributing to high mortality rates⁹. Additionally, risk factors for CHD may differ in male and female patients, which makes diagnosis at an early stage almost impossible^{10,11}.

¹Faculty of Information Sciences, Division of Science & Technology, University of Education, Lahore, Township Campus, 54770 Lahore, Pakistan. ²Department of Computer Science, National Textile University, 37610 Faisalabad, Pakistan. ³Artificial Intelligence and Data Analytics (AIDA) Lab, CCIS, Prince Sultan University, Riyadh 11586, Kingdom of Saudi Arabia. ⁴Department of Computer Science, University of Engineering and Technology Mardan, 23200 Mardan, Pakistan. ⁵School of Computing, Gachon University, 13120 Seongnam, Republic of Korea. ⁶Department of Computer Sciences, University of Education, Vehari Campus, 61161 Vehari, Pakistan. ✉email: younhyun.jung@gachon.ac.kr

Over the past few years, knowledge in the application of ML¹² has grown to be a valuable technique for analyzing extensive data and predicting diseases such as CHD. Manually diagnosing heart illness, however, increases the possibility of incorrect diagnosis, which could fatally postpone critical treatment. To circumvent the issue, numerous scholars are focusing on developing a variety of models to predict cardiovascular diseases in humans. These include mathematical and statistical analysis, but more recently, CHD has been predicted using data extraction and ML techniques. To draw conclusions from the data retrieved using different machine learning models, researchers used preprocessing techniques. The disadvantage of traditional machine learning methods, like Random Forest and K-Nearest Neighbor, is that they perform poorly on large datasets with numerous classes^{13,14}. Non-automated methods of identification are very susceptible to human error¹⁵. As voluminous data handling capacity along with increased accuracy in predictions gains prominent importance in practice, there is a demand for automated methods to be used for identification^{16,17}.

Challenges in previous studies

Despite the advancements in ML-based CHD prediction, several limitations persist in the existing literature¹⁸.

- *Feature Selection*: Traditional machine learning models often treat all features in the dataset equally, leading to suboptimal performance. Many studies fail to address the importance of selecting the most relevant features for accurate predictions¹⁹.
- *Data Imbalance*: Class imbalance within datasets remains a significant hurdle, particularly in medical data. While some studies have used oversampling techniques, such as SMOTE, to address this, they often do not explore its full potential in improving classification results²⁰.
- *Underperformance on Large Datasets*: Conventional machine learning models like Random Forest (RF) and K-Nearest Neighbor (KNN) often underperform when applied to large, complex datasets with multiple classes. These models struggle to scale effectively and deliver robust predictions²¹.
- *Inconsistent Results Across Algorithms*: Many previous studies have presented varying results with different algorithms but lack a comparative analysis to determine which models consistently perform better in predicting CHD²².
- *Lack of Real-World Applicability*: While theoretical models have succeeded in controlled environments, there is often a disconnect between academic research and practical application in real-world clinical settings²³.

Novel contributions

This study addresses key challenges in coronary heart disease prediction by introducing a comprehensive machine-learning framework:

- *Feature Selection*: We employ Mutual Information (MI) to identify and select only the most relevant features, thereby enhancing prediction accuracy.
- *Data Imbalance*: The Synthetic Minority Oversampling Technique (SMOTE) is utilized to improve the class distribution within the dataset, ensuring a balanced learning process.
- *Comparative Analysis*: A detailed comparison of three machine learning models, RF, KNN, and LR, is conducted to assess their effectiveness in predicting CHD.
- *Optimization Technique*: We introduce Particle Swarm Optimization (PSO) in conjunction with Artificial Neural Networks (ANN) for feature weighting and bias training, leading to a significant performance improvement.

Our investigation resolves multiple vital issues by deploying mutual information-based feature selection to identify the most informative features. The state-of-the-art contribution of this study is used SMOTE technique for oversampling the minority class of dataset. To increase the number of minority class instances, SMOTE employs the interpolation method. This is accomplished by creating new bogus instances rather than replicating the minority class. SMOTE can also help to prevent over-fitting when LR, KNN etc; models are being generalized. The development of the human brain like model called PSO-ANN comprises numerous units called neurons. In this model, the input layer receives features as input, the hidden layer computes the first output based on the input connect these layers. This study uses a feed-forward network with two layers to forecast CHD. A feed-forward network's primary job is to calculate every function and utilizes the resultant function to determine the relationship between inputs and outputs. This study employs a range of features, such as gender, marital status, BMI, cholesterol, and glucose levels, to provide a more accurate method of CHD prediction and give hospitals all the necessary recommendations for treating their patients. The incorporation of this united methodology enhances cardiovascular disease prediction accuracy while establishing itself as different from research focusing only on feature selection techniques.

The subsequent section presents the outline of the paper. Section II reviews the related literature on machine learning applications in CHD prediction. Section III outlines the proposed methodology, including data preprocessing, feature selection, and model development. Section IV details the experimental setup and dataset description. Section V presents the results and analysis. and Section VI concludes the study with a discussion of the findings and future directions for research.

Literature review

Growing ML-based approaches were used by Alom et al.¹⁸ to predict heart failure stage-I via genetic cardiac diseases and control subjects. They sought to apply their research to use AI as components of people's requirements for psychological functioning, sensory organs, or computational processes. Marco and Farinella¹⁹ looked at how viable computer vision methods for robust healthcare mechanisms incorporate pattern technology, language,

graphics, and robotics into the healthcare provision. The work of Gour et al.²⁰ used KNN, Decision Trees (DT), Support Vector Classifiers (SVC), Logistic Regression (LR), and RF algorithms to guess what would happen in the heart. This showed that supervised machine learning techniques are good at guessing heart disease²⁴. Unsupervised methods, like clustering, use unlabeled data, while supervised methods, like classification and regression, use labeled data. Reinforcement learning, on the other hand, learns from mistakes. Gour et al.²⁰ compared the accuracy of various models and demonstrated that supervised ML could effectively predict heart disease. Marimuthu et al.²¹ presented a method for predicting cardiac illness employing data analytic tools and ML techniques like ANN, DT, fuzzy logic, NB, KNN, and SVM. Some of their work entails reviewing previous studies and performing evaluations of algorithms.

To train and test various approaches, Golande & Pavan²² proposed an architecture that involved data pre-processing steps. Their study supported applying Adaboost to boost the efficiency of all ML approaches and parameter tuning for high accuracy. In predicting heart disease, Sharma & Parmar²³ used the deep learning system on the UCI dataset and showed that deep neural networks played an essential role in classifying or diagnosing cardiac diseases. It showed that Talos hyperparameter optimization outperforms the model tuning compared to other techniques. KNN, RF, SVM, and DT algorithms for adequately forecasting main cardiac illnesses showed that the forecasts produce beneficial and dependable outcomes. Kelwade and Salankar [16] suggested using heart rate series data from the MIT-BIH fibrillation dataset to train a radial basis function artificial neural network that could diagnose eight different types of heart arrhythmias. Their model offered exact estimations on arrhythmias: atrial fibrillation, pre-ventricular premature ventricular contraction, and sinus bradycardia²⁵. The back-propagation method can be used again with Multi-Layer Perceptron (MLP), and Das et al.²⁶ show that it works better than other machine learning models. In another work, Gour et al.²⁰ proposed a hybrid approach to optimizing cardiovascular disease features for decisive feature selection and better heart attack prediction. In their study, Chaki et al.²⁷ made a comparison between different kinds of learning techniques, such as DT, NB, and MLP, and they found MLP to be the most accurate, having an accuracy of 85%. Saboji²⁸ highlighted that the J48 algorithm surpassed traditional LR models in predicting heart disease, demonstrating the proposed model's superior performance. Baccouche et al.²⁹ presented a 2020 study where an ensemble of classification methods, such as BiLSTM and BiGRU with Convolutional Neural Networks (CNN), achieved 91% accuracy in predicting heart disease on the unbalanced Medical Norte Hospital dataset. Also, Javid et al.³⁰ showed that using ensemble models with neural networks (NN) and majority voting made predictions of cardiac illness more accurate³¹.

Miao et al.³² used boosting algorithms to predict CHD and reported that the proposed methodology gave higher accuracy than traditional ML methods in various datasets. They also incorporated a new neural network utilizing feature correlation analysis (FCA) for CHD prediction and proved that the ROC curve area for the NN-FCA was significantly higher than that of the Framingham Risk Score (FRS). Tayefi et al.³³ studied CHD using the decision tree analysis and obtained a high relationship of CHD with the HS-CRP index based on the Gini index. Using the Cox proportional hazards model, Irie et al.³⁴ identified elevated blood creatinine levels as a significant risk factor for congestive heart failure³⁵. Sharma et al.²³ introduced a two-layer CNN optimized by simulated annealing in 2020. They applied Least Absolute Shrinkage and Selection Operator (LASSO) regression to reduce the number of features. They used CNNs generated by simulated annealing to predict CHD, showing that the model outperformed traditional ML algorithms despite using a highly imbalanced dataset. Alotaibi et al.³⁶ Predicting Heart Failure Using Machine Learning and Data from the Cleveland Clinic Foundation Bradley N.e have attained this here as well, where the DT algorithm has given the best result with an accuracy of 93.19%, giving foresight and efficacy in implementing ML algorithms for predicting heart failure earlier, thus ensuring that proactive patient care can be taken accordingly, which is supposed to fetch a higher rate of precision and recall at its core without wasting valuable time (Table 1). Chunduru et al.³⁷ increased the prediction accuracy of MLs and DL for seven diseases, including diabetes, breast cancer, heart disease, kidney disease, liver problems, malaria, and pneumonia, through the Flask web application (the Flask is a very famous Python framework), thereby minimizing misdiagnosis as well. Sevet et al.²⁶ developed an ML framework that uses efficient optimization algorithms for better predicting heart disease and even the likelihood of heart failure. It gets a 70% F-score on Cleveland datasets for heart disease with KNN but only an 88%, and most notably, just a score of 30+ % instead. An F-score of 92.72% is achieved using the flower pollination algorithm for heart disease, which, too, depends on particular population sizes.

For CVD prediction, Hasan and Bao³⁸ compared models, including RF, SVC, kNN, NB, XGBoost, and ANN. The performances indicated that the wrapper approach with XGBoost performed best over other approaches at an accuracy of 73.74%, underscoring the importance of feature selection for achieving accurate models built in CVD predictions; Balasubramaniam et al.³⁹ introduced the GSSA-DMN method developed for heart disease detection, including data preprocessing ReliefF feature selection and optimization via GSSA. This approach can gain performance up to 93.2% accuracy, with 93% sensitivity and 91.5% specificity, which noticeably outperforms traditional ECG methods in effectiveness²⁶. The methodology in²⁶ is a supervised ML-based approach to build an effective prediction model for developing CVD manifested, focusing on a few risk factors. The stacking ensemble model delivered an accuracy of 87.8%, recall of 88.3%, precision, and AUC of 98.2%. On top of individual investigations, kaggle literature also exists concerning performance and capabilities from several machine-learning algorithms that were exercised to predict heart disease. Vishnupriya et al.⁴⁰ have proposed an innovative way of predicting heart disease using hybrid ML and DL mechanisms⁴¹. This hybrid random forest and deep learning approach offered the highest accuracy rate, achieving 92% accuracy. The prediction accuracy is increased by synchronizing a TKinter program comparing current healthcare data and reference distribution⁴². Chang et al.⁴³, techniques like KNN, DT, SVM, CN2, and SGD, which are futuristic approaches, were explored through modern information management practices inserting global standards along with next-generation technologies such as AI/ML/DL/cloud or edge computing. This book introduces

Attributes	Training Data (70%)	Test Data (30%)	Mean	Std
Gender	1.51	1.50	0.49	0.49
Age	48.50	49.03	17.97	18.10
Systolic	124.10	124.05	19.25	19.25
Diastolic	69.91	69.92	13.50	13.73
Platelet Count	252.98	253.08	67.39	67.43
Red-Cell Distribution Width	13.08	13.07	1.31	1.29
Cholesterol	5.07	5.08	1.07	1.09
Creatinine	78.42	79.12	37.47	42.82
Glucose	5.59	5.60	2.06	2.04
LDH	132.17	131.73	32.50	30.65
Uric Acid	321.92	321.25	86.20	85.95
Triglycerides	1.69	1.69	1.28	1.29
HDL	1.37	1.37	0.41	0.41
Glycohemoglobin	5.67	5.68	1.05	1.03
Vigorous Work	1.78	1.77	0.44	0.45
Moderate Work	1.59	1.59	0.51	0.51
Diabetes	1.90	1.90	0.35	0.34
Blood-Rel Diabetes	1.54	1.55	0.49	0.49
Blood-Rel Stroke	1.79	1.79	0.40	0.40

Table 1. Summary statistics of the selected features using mutual information.

new avenues of research and provides a comprehensive survey containing the evolution, norms, standards, and trends in information management. Noroozi et al.⁴⁴ study feature selection in optimizing ML algorithms for heart disease prediction. The study was conducted on the Cleveland Heart Disease dataset, which includes a set of 16 techniques that were tested with seven algorithms. The one using SVM for filtering gets the best accuracy (85.5%) and the other filter with more features takes over all others in ACC, precision, and F-measures.

Hong et al.³⁸ developed a CHD prediction model with an improved LightGBM classifier. The model, which achieved an AUC of 97.8% on highly informative Framingham Heart Institute CHD data, translates into cost savings in medical treatment due to early detection and preemptive measures for healthcare across the board, thereby reducing mortality rates and costs associated with disease management. Considering class imbalances, Qin et al.²⁸ proposed a two-layer CNN model to classify CHD clinical data. We use Lasso to determine the weights for features (feature weighting), and enhance the accuracy of predictions with majority voting and randomization per epoch. NHANES Data with Class Imbalance: 77% Accuracy Nandy et al. They aimed to create a more innovative healthcare framework for predicting cardiovascular disease using the Swarm-ANN approach, and in doing so, they generated and evaluated neural nets using novel heuristics for weight adjustments. The prediction achieves an accuracy of approximately 95.78% on a benchmark, surpassing the accuracy of traditional learning. Abdulraheeb et al.²⁸ suggested a way to find CHD better early on that uses bagged decision trees and information about the duplicated misclass instances. In 10-fold CV tests, this method was found to be 97.1% accurate overall.

Proposed methodology

This work introduces a reliable and contactless method for diagnosing CHD using ML classification models and optimizing strategies. The first objective aims at improving diagnostic performance, which is essential in predictive healthcare analytics, given data imbalance and feature redundancy issues. The proposed data preprocessing phase combines recent techniques to select the appropriate feature subsets and optimize the model for better CHD predictability. Figure 1 presents a holistic look at the different parts of the suggested paradigm.

Dataset and preprocessing

The heart disease dataset was tested, obtained from the UCI repository, and normalized for data cleanliness. Before the pre-processing step, missing data were addressed by mean for numerical and mode for categorical variables. Moreover, categorical features were transformed into numeric using the one-hot encoding technique. According to the analysis, features were normalized to enhance consistency and increase the ML algorithms' competency. In detail, in step 1 of the data representation process, the data set was categorized into two categories, normal and diseased, where diseased referred to a group containing CHD and normal to a group that did not. The second step was data cleaning; observations had to be cleaned because of missing and incomplete values. In step 3 of the investigated model, the process of exploratory data analysis was described, during which correlation matrices and data visualization tools were applied to determine specific attributes. The dataset was divided into a training set (70:30%) and a testing set sample in the fourth step. Feature selection was used next to further enhance dimensionality reduction and increase test classification accuracy.

In many practical situations, we have a model for a physical phenomenon, a model that includes several unknown parameters. These parameters need to be determined from the known observations; this determination

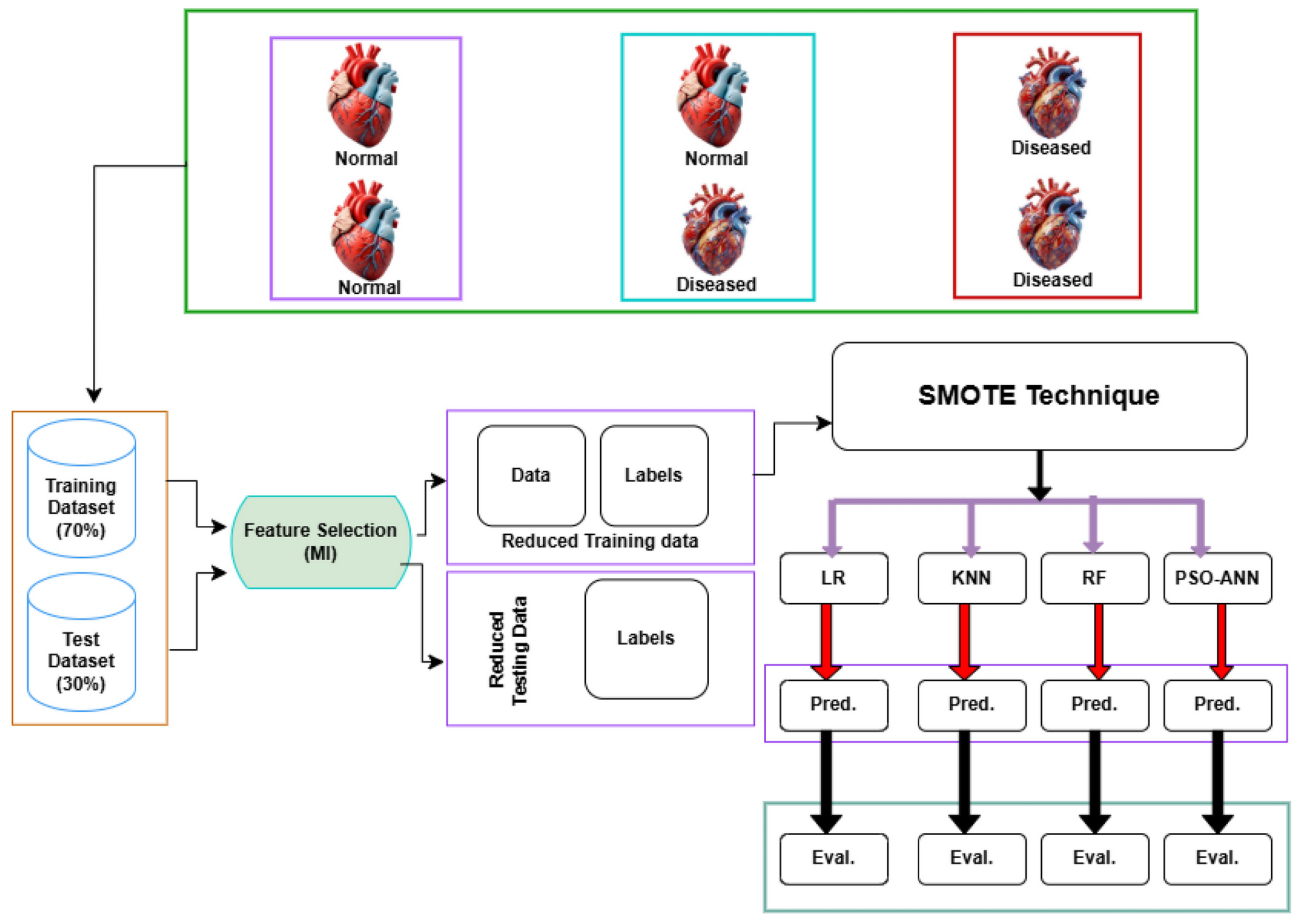


Fig. 1. Workflow of the proposed methodology.

is known as *training the model*. In statistics in general, the more data points we use, the more accurate are the resulting estimates. From this viewpoint, it may seem that the best way to determine the parameters of the model is to use all the available data points in this determination. This is indeed a good idea if we are absolutely certain that our model adequately describes the corresponding phenomenon. We estimate the parameters of the model based on the training set, with $p \cdot N$ elements. Thus, the standard deviation of the corresponding model is proportional to $\frac{1}{\sqrt{p \cdot N}}$. When we gauge the accuracy of the model, we compare the trained model with the data from the testing set. Even if the trained model was exact, because of the measurement errors, we would not get the exact match. Instead, based on $(1 - p) \cdot N$ measurements, we would get the standard deviation proportional to $\frac{1}{\sqrt{(1-p) \cdot N}}$. We want to estimate the difference d_k between the trained model and the testing data. It is reasonable to assume that, in general, the errors corresponding to the training set and to the testing set are independent: we may get positive correlation in some cases, negative correlation in others, so, on average, the correlation is 0. We first train our model on the training set, and then we use the data from the testing set to gauge the accuracy of the resulting model. Empirical studies show that the best results are obtained if we use 20–30% of the data for testing, and the remaining 70–80% of the data for training.

- *Valid* – in the sense that they do not overestimate the accuracy (i.e., do not underestimate the approximation error),
- *More accurate among the valid estimates* – i.e., their overestimation of the approximation error is the smallest possible.

Feature selection and engineering

Performance optimization was a key target, and a unique feature engineering methodology was followed to achieve that. First, we removed all the features that can be considered noise or are similar to other features through different statistical procedures, including variance thresholding. Subsequently, the use of the MI to order features concerning their importance in the output class was carried out. The MI metric quantifies the mutual dependence between two variables as represented in Eq. 1 and is calculated as follows:

$$I(f_i; c_j) = \log \left(\frac{P_r(f_i, c_j)}{P_r(f_i) P_r(c_j)} \right) \quad (1)$$

If a feature f_i is independent of the class c_j , then $I(f_i; c_j) = 0$. MI helps identify the most informative features, minimizing computational complexity while retaining key predictive attributes. The following methods are used to calculate the total feature score among all terms:

$$I_{\text{avg}}(f_i; C) = \sum_{j=1}^n P_r(c_j) I(f_i; c_j) \quad (2)$$

$$I_{\text{max}}(f_i; C) = \max_{j=1, \dots, n} \{I(f_i; c_j)\} \quad (3)$$

PSO was introduced as an approach to optimization by Kennedy et al. in 1995⁴⁵, and was inspired by the social interactions of fish schools and bird flocks. In this, the n -dimensional problem space contains a large number of particles. Any particle can define the loss function, which can be optimized by estimating it at its present position. Next, each particle's location is updated by comparing it to the global average best position ($gBest$) or its particular best position ($pBest$). Let p be the particle's position and v be the particle's velocity. The particle update rule for each particle is thus given as:

$$v = v + c_1 \text{rand} (p_{\text{Best}} - p) + c_2 \text{rand} (g_{\text{Best}} - p) \quad (4)$$

$$p = p + v \quad (5)$$

Furthermore, principal component analysis was also performed to reduce the dimensions of the dataset as far as possible while retaining the most variance. It minimizes overfitting and enhances a model's generalization ability by considering only the most important features.

Handling data imbalance with SMOTE

In this study, the SMOTE (Synthetic Minority Oversampling Technique) was used to predict patient survival, and the datasets were equalized. The highest-rated features chosen by PSO were used to train machine learning systems. In predicting the survival of people with cardiac issues, MI has been proven to perform well in experiments, with a high accuracy, using SMOTE. The proposed methodology predicts heart attacks using several types of machine learning classifiers on a UCI dataset without the use of feature engineering. To manage the imbalanced data, the proposed system employs a SMOTE. From the experimental results, it is found that SMOTE-based artificial neural networks performed the best among all machine learning classifiers. In an imbalanced dataset, the classification models tend to provide poorer results for the minority class than the majority class. Since SMOTE synthesizes innovative points belonging to the minority class by locating the closest neighbors among samples, it performs well in generating synthetic instances that cover a broader area in terms of features. This helps to make sure that both of the classes play the same role in the process of classification. SMOTE operates as follows:

- Select a sample from the minority class.
- Find its k -nearest neighbors.
- Interpolate new samples between the selected sample and its neighbors.

This technique significantly improves the classifier's ability to detect minority class instances (i.e., "diseased" cases) and prevents overfitting.

Machine learning algorithms

Following data preparation, various machine learning algorithms were applied to evaluate the performance of different classification techniques in predicting CHD. We worked with many models, including KNN, Support Vector Machine (SVM), LR, RF, and Naïve Bayes (NB). Each algorithm's performance was refined through hyperparameter optimization approaches.

K-nearest neighbors (KNN)

KNN is a non-parametric, supervised ML algorithm that classifies a data point based on the majority class of its k -nearest neighbors. The algorithm was tuned by optimizing the value of k , which controls the number of neighbors considered during classification. Cross-validation was employed to select the optimal k , minimizing classification error.

Support vector machine (SVM)

SVM constructs a hyperplane that best separates the data into two classes. This study used a radial basis function (RBF) kernel to allow for non-linear classification. The hyperparameters C (regularization) and γ (kernel coefficient) were optimized using a grid search to maximize accuracy.

Logistic regression (LR)

LR is a simple yet effective method for binary classification. This study applied regularized logistic regression to prevent overfitting, with L_2 regularization used to penalize large coefficient values. The model was optimized using a stochastic gradient descent approach.

Random forest (RF)

RF is an ensemble method that constructs multiple decision trees and aggregates their results to improve accuracy. To mitigate overfitting, the number of trees in the forest ($n_{estimators}$) and the maximum depth of each tree was optimized. The model was further enhanced by implementing bootstrap sampling to increase its robustness.

Naïve Bayes (NB)

The Naïve Bayes classifier assumes conditional independence between features and applies Bayes' theorem to predict a class's probability. This approach is efficient for high-dimensional data and was implemented with Gaussian assumptions for continuous features.

Optimization via PSO-ANN

The PSO algorithm was applied to optimize the weights of an ANN. The social behavior of bird flocks and fish schools inspires PSO. It works by iteratively updating a swarm of candidate solutions (particles) to find the optimal set of weights for the ANN.

The ANN architecture used in this study consisted of three layers: an input layer, a hidden layer, and an output layer. The optimization process involved the following steps:

- Initialize a swarm of particles, each representing a possible set of ANN weights.
- Evaluate the performance of each particle using a loss function.
- Update the particle velocities and positions based on their personal best and global best positions.
- Repeat until convergence and use the optimized weights to train the ANN.

PSO-ANN was designed to balance the model complexity and prediction accuracy trade-offs, ensuring robust performance even with noisy or incomplete data. The swarm size was set to 30 particles, and the ANN contained five neurons in the hidden layer. A learning rate of 0.001 was used to train the model, and convergence was achieved after 100 iterations.

Cross-validation and performance metrics

To ensure the generalization of the models, we employed 10-fold cross-validation. This method splits the dataset into ten equal parts, using nine parts for training and one for testing in each iteration. The average performance across all folds was used to evaluate the final model.

The following performance metrics were used:

- Accuracy: Proportion of correctly classified instances.
- Precision: Ratio of true positives to the sum of true and false positives.
- Recall (Sensitivity): Ratio of true positives to the sum of true positives and false negatives.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC Score: Area under the Receiver Operating Characteristic curve, measuring the trade-off between true and false positive rates.

Experimental settings

Dataset

The National Health & Nutritional Examination Survey (NHANES) data covering the years 1999–2000 to 2015–2016 is examined in this research. The dataset combines demographic, examination, and laboratory data from questionnaires. It contains 51 attributes and approximately 37,000 instances. There is a significant class imbalance, with about 35,500 cases of non-coronary cardiovascular disease and 1,500 cases of coronary heart disease (CHD). The dataset features include mean cell volume, gender, age, systolic blood pressure, weight, height, BMI, and hemoglobin levels. Preprocessing is crucial for machine learning algorithms. During preprocessing, variables such as yearly family income and the ratio of family income to poverty were removed. Feature selection was applied to 38 remaining features, where those with the highest mutual information regarding the output class (CHD) were selected. Table 1 summarizes the selected traits, and Fig. 2 illustrates their influence.

Baseline models

The proposed model's performance was compared with various machine learning algorithms. These include:

- Logistic Regression
- Random Forest with 50 trees
- k-Nearest Neighbor with $k = 3$ and Euclidean distance as the distance function

Results

Evaluation metrics

The main contribution of the proposed method is to build a predictive model that can assist health workers in the timely detection of Coronary Heart Disease (CHD) and ultimately reduce mortality. This study performs a

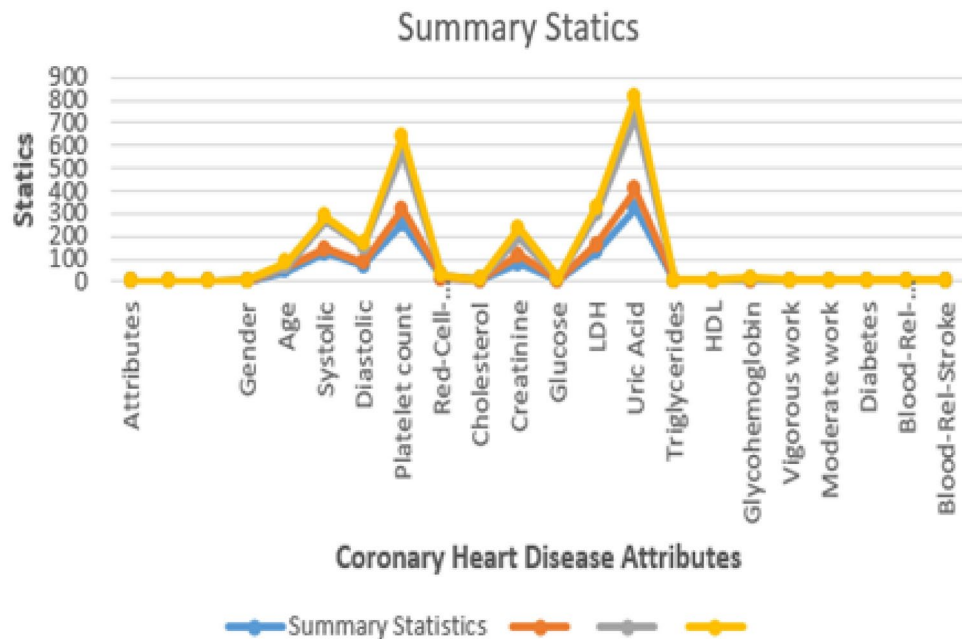


Fig. 2. Coronary heart disease affected attributes.

comparative analysis of three supervised classification machine learning algorithms-mutual information (MI), PSO, and SMOTE-in predicting the case-control status of the patient. These algorithms are employed to select the more relevant ones from the available features. The performance of the employed models is compared using sensitivity, specificity, accuracy, etc. Several evaluation metrics were used to quantify the classifiers' performance. The selection of metrics depends on the experiment's goal and the field in which it is conducted. For instance, sensitivity and specificity are common in medicine, while accuracy, precision, and recall are frequently used in computer science.

This research uses accuracy and Normalized Mutual Information (NMI) as the performance measuring parameters of classifiers. Accuracy is the ratio of correct predictions to the number of cases examined. NMI is based on the mutual information of variables, which measures the average amount of information shared between different features. It depends on the mutual information (MI) and the entropy labeled as $H(C)$ and $H(S)$.

The entropy across the class variable C and the set of selected features S is given by $H(C)$ and $H(S)$, respectively. Accuracy and NMI are computed using equations (6) and (7) as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$\text{NMI}(C, S) = \frac{MI(C, S)}{\text{mean}(H(C), H(S))} \quad (7)$$

TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

Feature selection results

Not all features in the dataset contribute equally to preventing coronary artery disease (CHD). The mutual information (MI) technique was used to identify the most essential features, enhancing the accuracy of machine learning models. Table 2 presents the MI scores for various features, while Fig. 3 shows the impact of different machine-learning techniques.

SMOTE results

The class imbalance in the dataset was addressed using the SMOTE. Table 3 shows the count of each label before and after applying SMOTE, and Fig. 4 illustrates the impact of SMOTE on the dataset.

Classification results

This section presents the results of the various machine-learning algorithms applied to the dataset. Table 4 compares the classifiers' results without feature selection, while Table 5 shows the results with feature selection. Figure 5 shows the results of the SMOTE features selection, and Table 4 shows the results overall. Table 5 shows that, compared to traditional ML modes, PSO-ANN achieves relatively high accuracy, and Fig. 6 shows the results. You can see how the suggested model stacks up against the 2-layer CNN model in Table 6. Both models were tested on the same dataset. In Fig. 6, we have defined the accuracy of different machine learning algorithms and our proposed algorithm after applying the SMOTE technique for feature selection in the blue

SR. NO.	Feature Description	MI Score
0	SEQN	0.0003
1	Gender	0.0082
2	Age	0.0302
3	X60-sec-pulse	0.0043
4	Systolic	0.0019
5	Diastolic	0.0006
6	Body-Mass-Index	0.0000
7	White-Blood-Cells	0.0000
8	Basophils	0.0011
9	Red-Blood-Cells	0.0006
10	Hemoglobin	0.0000
11	Mean-cell-Hemoglobin	0.0000
12	Platelet-count	0.0050
13	Mean-Platelet-Vol	0.0000
14	Red-Cell-Distribution-Width	0.0024
15	Albumin	0.0024
16	ALP	0.0017
17	AST	0.0007
18	ALT	0.0010
19	Cholesterol	0.0044
20	Creatinine	0.0124
21	Glucose	0.0071
22	GGT	0.0022
23	Iron	0.00002
24	LDH	0.0033
25	Phosphorus	0.0000
26	Bilirubin	0.0007
27	Protein	0.0010
28	Uric Acid	0.0028
29	Triglycerides	0.0021
30	HDL	0.0029
31	Glycohemoglobin	0.0113
32	Vigorous Work	0.0057
33	Moderate Work	0.0039
34	Diabetes	0.0141
35	Blood-Rel-Diabetes	0.0043
36	Blood-Rel-Stroke	0.0062

Table 2. Features and their respective MI scores.

line. The orange line shows the comparison without feature selection, which is at the zero-level line because in this diagram, our main focus is on feature selection. When we use different machine learning algorithms, the difference in accuracy becomes very low, and the graph lines reach near a straight line. Figure 7 shows the training loss curves for various models.

Comparison with CNN

Table 6 compares the prediction accuracy of the proposed PSO-ANN model with a 2-layer CNN model. Both models were evaluated on the same dataset.

Discussion

The enhanced performance of the proposed model for predicting coronary heart disease (CHD) can be attributed to a few key methodological choices that collectively optimize the model's accuracy and reliability. The combination of feature selection, data balancing, and the integration of PSO-ANN has played a critical role in achieving superior results compared to traditional approaches, as evidenced by the results in Section.

One of the significant factors contributing to the model's accuracy is the use of the Mutual Information (MI) feature selection technique. By prioritizing features with higher MI scores (refer to Table 2), the model is able to focus on variables that have the most predictive value while ignoring less relevant data points. This reduction in input dimensionality effectively minimizes noise, which can otherwise lead to overfitting or inaccurate predictions. The MI-based selection also offers a way to handle redundant data, allowing the classifiers to be

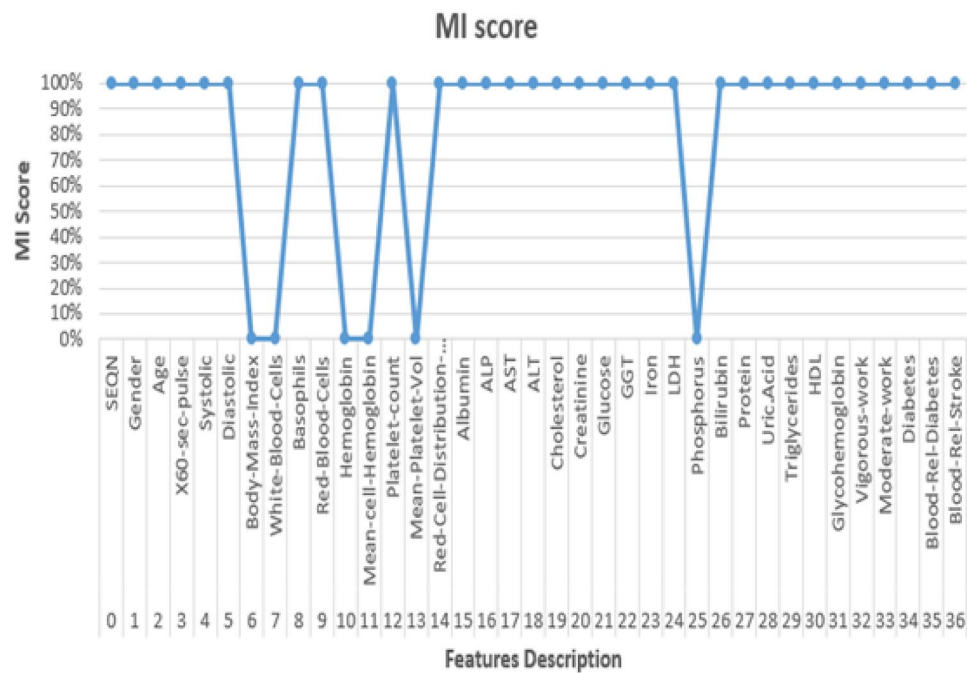


Fig. 3. MI score of different machine learning algorithms.

Counts of Labels	Before SMOTE	After SMOTE
"0"	24,915	24,915
"1"	1,040	24,915

Table 3. Counts of each label in the dataset.

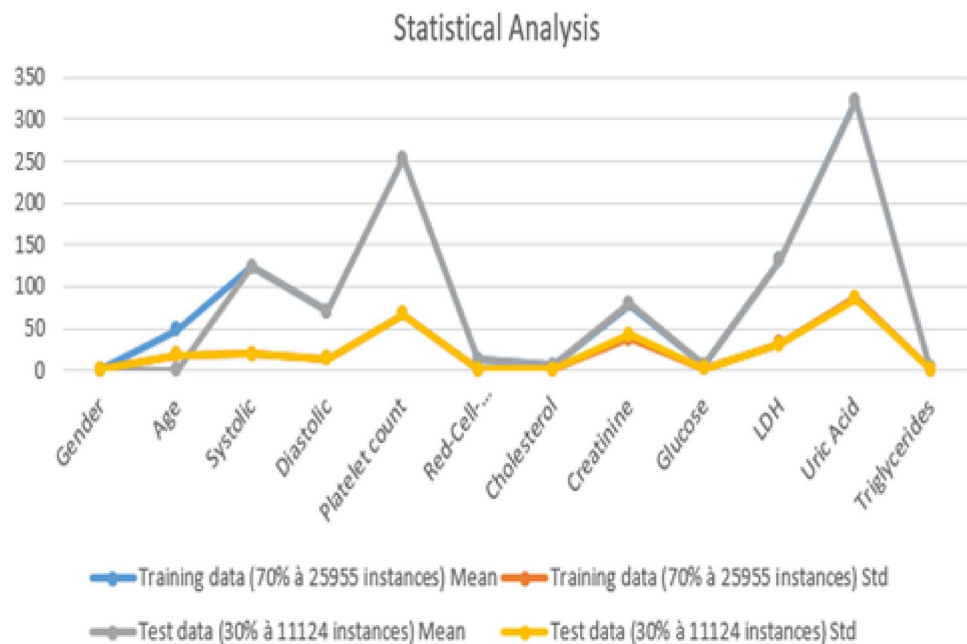


Fig. 4. Statistical analysis of training and testing data after applying SMOTE.

Models	Before SMOTE ACC.	Before SMOTE NMI	After SMOTE ACC.	After SMOTE NMI
Logistic Regression	96.0	0.0003	85.0	0.00009
KNN	95.7	0.0003	78.2	0.00003
Random Forest	96.0	0.0001	92.8	0.00002
PSO-ANN	96.1	0.00028	96.1	0.00028

Table 4. Comparison of machine learning algorithms without feature selection.

Models	Before SMOTE ACC.	Before SMOTE NMI	After SMOTE ACC.	After SMOTE NMI
Logistic Regression	95.8	0.013	83.6	0.013
KNN	95.1	0.015	84.0	0.015
Random Forest	95.8	0.0001	92.1	0.0001
PSO-ANN	96.1	0.00028	96.1	0.00028

Table 5. Comparison of machine learning algorithms with feature selection.

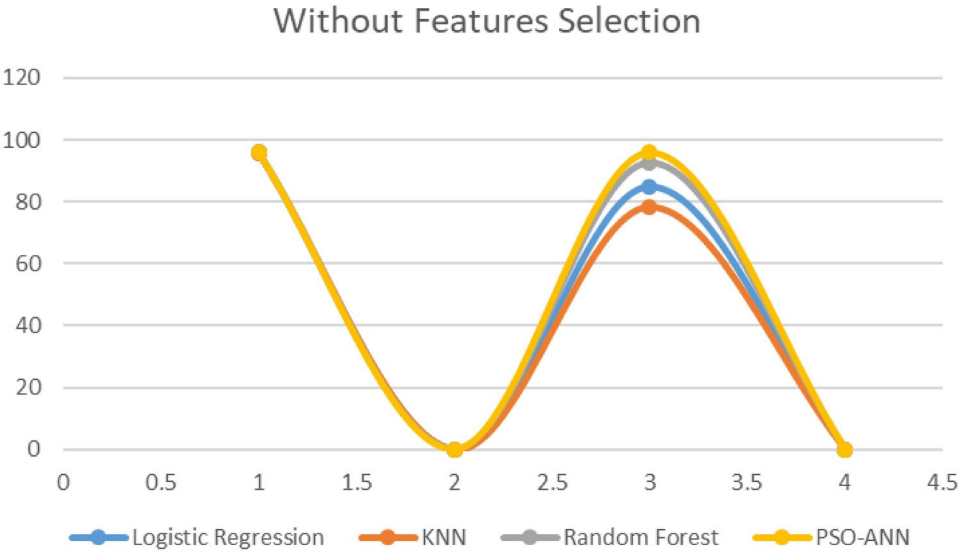


Fig. 5. Without SMOTE feature selection through ML algorithms.

more streamlined and computationally efficient. Given that not all features in the NHANES dataset contribute equally to predicting CHD risk, MI-based selection ensures the model only considers data points that genuinely impact outcomes. Another crucial component of the model's success is the SMOTE, used to address the class imbalance within the dataset. The NHANES dataset has an inherent imbalance between CHD-positive and CHD-negative samples (as shown in Table 3), which could skew model predictions in favor of the majority class if left untreated. SMOTE mitigates this by generating synthetic samples for the minority class, balancing the distribution of CHD-positive and CHD-negative instances. This balanced dataset allows the classifiers to learn patterns from both classes equally, enhancing the model's sensitivity to CHD cases without compromising on specificity. The impact of SMOTE on classification performance is evident from the improved accuracy and Normalized Mutual Information (NMI) scores, especially for models like Logistic Regression and KNN after applying SMOTE (refer to Tables 4 and 5). By effectively handling class imbalance, SMOTE contributes to the robustness and fairness of the prediction model, making it more applicable for real-world clinical use.

Particle Swarm Optimization with Artificial Neural Networks (PSO-ANN) is another innovative element in the proposed approach. PSO-ANN combines the optimization power of PSO with the learning capacity of ANN, allowing the model to achieve optimal weight initialization and faster convergence during training. PSO aids in fine-tuning the ANN model by iteratively searching for the best parameter values, which reduces the risk of getting stuck in local minima. The PSO-ANN model, as shown in the results (Table 6), significantly outperformed other classifiers, including CNN, by providing both high accuracy and computational efficiency. This is a notable achievement since CNNs typically require substantial training data and computational resources, which may not always be feasible in healthcare settings. By contrast, PSO-ANN achieves high accuracy (96.1%) with relatively moderate data requirements and fewer computational costs, making it a practical option for early CHD

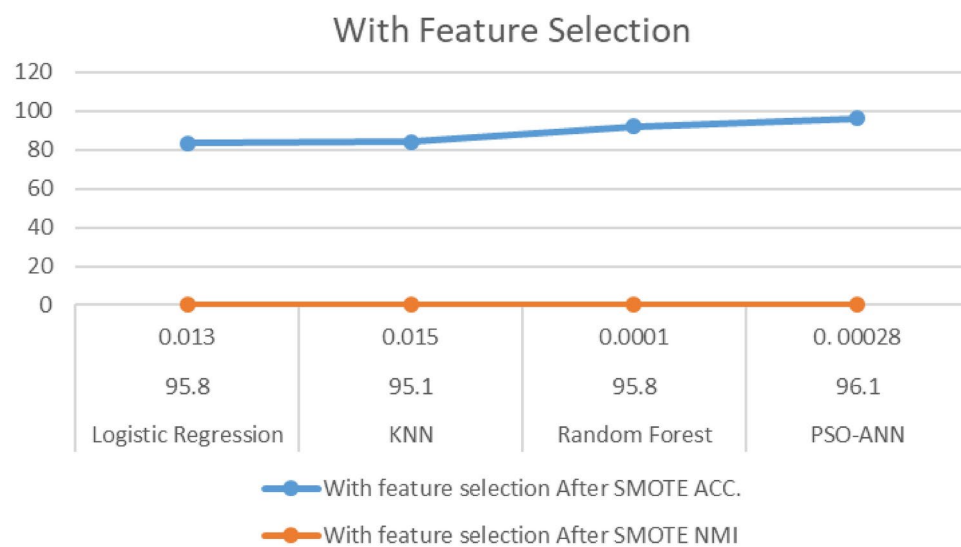


Fig. 6. With SMOTE feature selection through ML algorithms.

Models	Accuracy
2-layer CNN	81.78
Proposed Model	96.1

Table 6. Comparison of prediction accuracies of the proposed model with CNN.

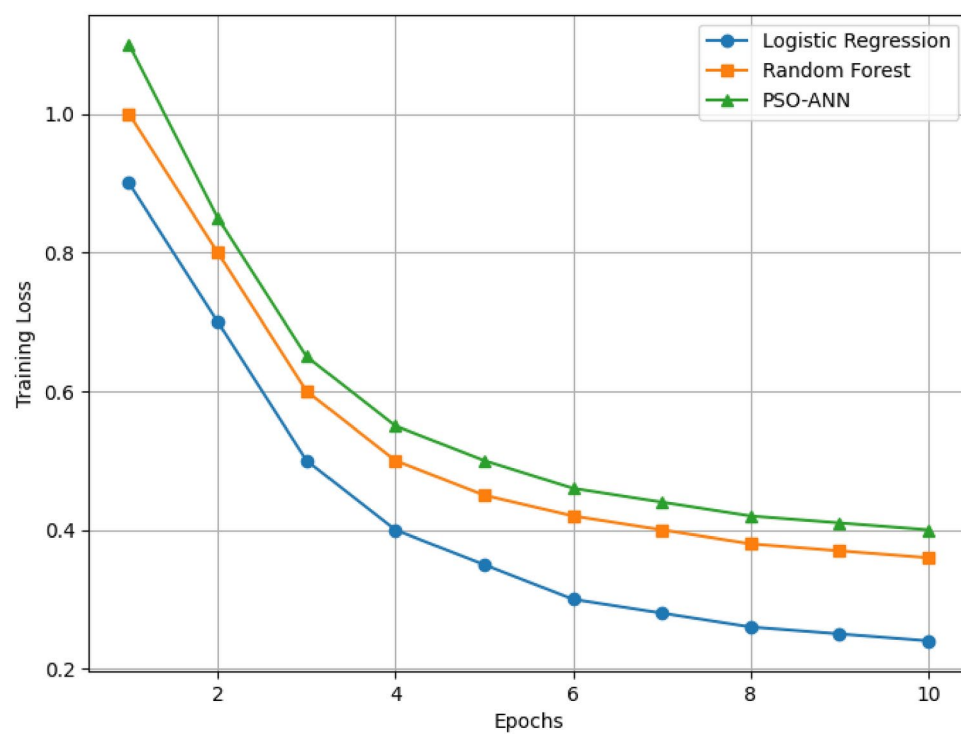


Fig. 7. Training Loss Curves for Various Models.

prediction. In addition to these core components, the choice of evaluation metrics-accuracy and NMI-provides a balanced view of the model's predictive capabilities. Accuracy alone may not suffice, especially when dealing with imbalanced data, as it may not reflect the classifier's performance on the minority class. NMI measures the shared information between predicted and actual labels. It adds another level of clarity to the results by proving that certain features are relevant and that different data attributes are linked to a higher risk of coronary heart disease. It is clear from the high NMI scores in Tables 4 and 5 that the feature selection process did a good job of matching the model to the patterns in the dataset.

The comparison with the 2-layer CNN model further underscores the proposed model's strengths. While CNNs have gained popularity for their high-dimensional feature extraction capabilities, they may not always be the best choice for structured medical datasets, such as those with tabular data. The proposed PSO-ANN model outperforms CNN in terms of accuracy, indicating that the feature selection and optimization techniques employed are more suitable for this dataset. Furthermore, CNNs often require considerable tuning and more extensive datasets to achieve optimal performance, which may not be practical in resource-limited settings. The results from Table 6 highlight the effectiveness of PSO-ANN as a more adaptable solution, capable of delivering high accuracy with fewer demands on computational and data resources.

Overall, the proposed model's improved performance can be attributed to the careful integration of advanced techniques tailored to the dataset's characteristics. By balancing data, selecting the most predictive features, and employing an optimized neural network, this approach addresses many of the limitations encountered in traditional models. Moving forward, incorporating additional data sources, such as genetic and lifestyle factors, could further enhance the model's predictive capacity and clinical relevance. Using more advanced deep learning methods, like long short-term memory (LSTM) networks or ensemble learning, may also give us more information, especially when looking at patterns over time or combining data from different sources. This multipronged approach to CHD prediction offers a promising pathway for more accurate and early diagnosis, potentially contributing to improved patient outcomes and reduced healthcare costs.

Conclusion

The effectiveness of machine learning (ML) algorithms for predicting coronary heart disease (CHD) is examined in this study. CHD prediction was performed using both test and training datasets, with 30% of the data allocated for model evaluation and 70% for training. Using the Mutual Information feature selection technique significantly enhanced the models' prediction accuracy by reducing irrelevant input variables. One way that the class imbalance in the NHANES dataset was fixed was by using the synthetic minority oversampling technique (SMOTE). This helped the classification models work better. By accurately predicting the likelihood of CHD, this study has the potential to aid doctors and hospitals in initiating timely treatment, ultimately reducing the risk of fatalities associated with misdiagnosis or delayed intervention. The study used MATLAB to run the machine learning techniques. The results show that the proposed model, which combines feature selection and SMOTE for class imbalance, was more accurate than traditional machine learning techniques and the convolutional neural network (CNN) model. Cardiovascular diseases claim thousands of lives each year, with many of these deaths attributed to misdiagnosis or underdiagnosis of heart conditions. This study uses selected features and well-known ML methods to predict CHD from real-world datasets. The results demonstrate that the proposed model significantly improves the accuracy of CHD prediction, offering more reliable forecasts than other traditional approaches. Despite the promising results and potential uses of the proposed ML-based PSO-ANN for coronary heart disease prediction, there are several limitations to consider. The performance and reliability of ML models depend on the quality and availability of testing and training datasets, and we employed CHD and private datasets in our study, which may face limitations in availability, representativeness, and data quality. SMOTE generates synthetic minority class samples to overcome class imbalance, but its effectiveness depends on the dataset and situation, and class imbalance can cause models to perform well for the majority group but poorly for the minority class, which is often the main concern. To address this issue, it's essential to discuss and compare different approaches for handling class imbalances along with SMOTE, such as ensemble methods or algorithm selection. Future research will encompass the examination of diverse optimizers for artificial neural networks (ANN) alongside implementing advanced deep learning methodologies. Models like recurrent neural networks (RNN), long short-term memory (LSTM), and other deep architectures will help us better predict coronary heart disease and make the diagnosis more reliable. This investigation may facilitate more thorough data analysis and enhanced model efficacy, hence enabling sophisticated applications in clinical environments.

Data availability

The data used and analyzed during the current study is available from the corresponding author upon reasonable request.

Received: 27 November 2024; Accepted: 28 March 2025

Published online: 17 April 2025

References

1. Al-Ssulami, A. M., Alsorori, R. S., Azmi, A. M. & Aboalsamh, H. Improving coronary heart disease prediction through machine learning and an innovative data augmentation technique. *Cogn. Comput.* **15**, 1687–1702 (2023).
2. Nandy, S. et al. An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Comput. Appl.* **35**, 14723–14737 (2023).
3. Yang, H., Chen, Z., Yang, H. & Tian, M. Predicting coronary heart disease using an improved lightgbm model: Performance analysis and comparison. *IEEE Access* **11**, 23366–23380 (2023).

4. Ashraf, R. *et al.* An efficient technique for skin cancer classification using deep learning. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 1–5 (IEEE, 2020).
5. World Health Organization. Cardiovascular diseases (cvds). <https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab1>. Accessed: 2022-01-10 (2022).
6. Sabahi, F. Analyzing causality in uncertain domains using extended fuzzy logic (fle) applied to coronary heart disease diagnosis. *Scientia Iranica* (2024).
7. Lim, W. *et al.* Type 2 diabetes in patients with end-stage kidney disease: influence on cardiovascular disease. *Med. J. Aust.* **209**, 440–446 (2018).
8. Khan, I. & Siddique, M. Z. Ateeq ur rehman butt, azhar imran mudassir, muhammad azeem qadir, and sundus munir. 2021. "towards skin cancer classification using machine learning and deep learning algorithms: A comparison". *Int. J. Innov. Sci. Technol.* **3**, 110–118.
9. Butt, A. U. R. *et al.* An optimized role-based access control using trust mechanism in e-health cloud environment. *IEEE Access* **11**, 138813–138826 (2023).
10. Benjamin, M. F. *et al.* Heart disease and stroke statistics. *Circulation* **138**, 520–528 (2019).
11. Khan, F. A. *et al.* Burnt human skin segmentation and depth classification using deep convolutional neural network (dcnn). *J. Med. Imaging Health Inform.* **10**, 2421–2429 (2020).
12. Ravindhar, N., Anand, H. & Ragavendran, G. Intelligent diagnosis of cardiac disease prediction using machine learning. *Int. J. Innov. Technol. Explor. Eng.* **8**, 1417–1421 (2019).
13. Komilovich, E. B. & Khalimovich, M. N. Dependencies in the clinic and diagnosis of coronary heart disease and arterial hypertension. *J. New Century Innov.* **46**, 61–69 (2024).
14. Khan, F. A. *et al.* Computer-aided diagnosis for burnt skin images using deep convolutional neural network. *Multim. Tools Appl.* **79**, 34545–34568 (2020).
15. Butt, A.U.R., Ahmad, W., Ashraf, R., Asif, M. & Cheema, S.A. Computer aided diagnosis (cad) for segmentation and classification of burnt human skin. In *2019 international conference on electrical, communication, and computer engineering (ICECCE)*, 1–5 (IEEE, 2019).
16. Hussain, S., Mushtaq, M. & Halim, Z. Multi-view document clustering via ensemble method. *J. Intell. Inf. Syst.* **43**, 81–99 (2014).
17. Khan, M. I. *et al.* Activity detection of elderly people using smartphone accelerometer and machine learning methods. *Int. J. Innov. Sci. Technol.* **3**, 186–197 (2021).
18. Alom, Z. *et al.* Early stage detection of heart failure using machine learning techniques. In *International Conference on Big Data, IoT, and Machine Learning* (Bangladesh, 2021).
19. Marco, L. & Farinella, G. *Computer Vision for Assistive Healthcare* (Academic Press, 2018).
20. Gour, S., Panwar, P., Dwivedi, D. & Mali, C. A machine learning approach for heart attack prediction. *Intell. Sustain. Syst.* **12**, 741–747 (2022).
21. Marimuthu, M., Abinaya, M., Hariesh, K., Madhankummar, K. & Pavithra, V. An aalto review on heart disease prediction using machine learning and data analytics approach. *Int. J. Comput. Appl.* **181**, 975–1007 (2018).
22. Golande, A. & Kumar, T. P. Heart disease prediction using effective machine learning techniques. *Int. J. Recent Technol. Eng.* **10**, 944–950 (2019).
23. Sharma, S. & Parmar, M. Heart diseases prediction using deep learning neural network model. *Int. J. Innov. Technol. Explor. Eng.* **15**, 124–137 (2020).
24. Khabibovna, Y. S. & Alisherovna, K. M. Stress testing in patients with coronary heart disease. *J. New Century Innov.* **45**, 28–33 (2024).
25. Rehman, A., Mahmood, T. & Saba, T. Robust kidney carcinoma prognosis and characterization using swin-vit and deeplabv3+ with multi-model transfer learning. *Appl. Soft Comput.* **170**, 112518 (2025).
26. Das, P., Turkoglu, I. & Sengur, A. Effective diagnosis of heart disease through neural networks ensembles. *Expert Syst. Appl.* **36**, 7675–7680 (2009).
27. Chaki, D., Das, A. & Zaber, M. A comparison of three discrete methods for classification of heart disease data. *Bangladesh J. Sci. Ind. Res.* **50**, 293–296 (2015).
28. Saboji, R. & Ramesh, P. A scalable solution for heart disease prediction using classification mining technique. In *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)* (Chennai, India, 2017).
29. Baccouche, A., Garcia, B., Olea, C. & Elmaghraby, A. Ensemble deep learning models for heart disease classification: A case study from mexico. *Informatics* **11**, 1–28 (2020).
30. Javid, I., Alsaedi, Z. & Ghazali, R. Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method. *Int. J. Adv. Comput. Appl.* **11**, 540–551 (2020).
31. Jabbar, A. *et al.* Deep transfer learning-based automated diabetic retinopathy detection using retinal fundus images in remote areas. *Int. J. Computat. Intell. Syst.* **17**, 135 (2024).
32. Mahmud, I. *et al.* Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel. *Diagnostics* **13**, 2540 (2023).
33. Tayefi, M. *et al.* hs-crp is strongly associated with coronary heart disease (chd): A data mining approach using decision tree algorithm. *Computational Methods and Programs in Biomedicine* 105–109 (2017).
34. Irie, F. The relationships of proteinuria, serum creatinine, glomerular filtration rate with cardiovascular disease mortality in the japanese general population. *Kidney Int. J.* **69**, 1264–1271 (2006).
35. Anjum, S. *et al.* Lung cancer classification in histopathology images using multiresolution efficient nets. *Comput. Intell. Neurosci.* **2023**, 7282944 (2023).
36. Alotaibi, F. S. Implementation of machine learning model to predict heart failure disease. *Int. J. Adv. Comput. Sci. Appl.* **10**, 261–268 (2019).
37. Chunduru, A., Kishore, A. R., Sasapu, B. K. & Seepana, K. Multi chronic disease prediction system using cnn and random forest. *Social Netw. Comput. Sci.* **5**, 157 (2024).
38. Hasan, N. & Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* **11**, 49–62 (2021).
39. Balasubramaniam, S., Joe, C. V., Manthiramoorthy, C. & Kumar, K. S. Relief based feature selection and gradient squirrel search algorithm enabled deep maxout network for detection of heart disease. *Biomedical Signal Processing and Control* **87** (2024).
40. Vishnupriya, G., Pradeep, T., Mehanethra, M. & Mitra, V. J. A. Prediction of pre-cardiac disease using ml & dl techniques in t-kinter. *J. Surv. Fish. Sci.* **10**, 1874–1883 (2023).
41. Butt, A. U. R. *et al.* Beyond boundaries: A novel ensemble approach for breast cancer detection in ultrasound imaging using deep learning. In *2024 21st International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 315–320 (IEEE, 2024).
42. Khan, R. U. *et al.* Blockchain-based trusted tracking smart sensing network to prevent the spread of infectious diseases. *IRBM* **45**, 100829 (2024).
43. Chang, V., Bhavani, V.R., Xu, A.Q. & Hossain, M. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics* **2** (2022).
44. Noroozi, Z., Orooji, A. & Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci. Rep.* **13**, 22588 (2023).

45. Kelwade, P. & Salankar, R. Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series. In *IEEE First International Conference on Control, Measurement and Instrumentation* (2016).

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00554526), the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project Name: Cultural Technology Specialist Training and Project for Metaverse Game, Project Number: RS-2023-00227648), and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (grant number: HI22C1651). The authors also would like to thank the Artificial Intelligence and Data Analytics (AIDA) Lab, College of Computer & Information Sciences, Prince Sultan University Riyadh, Kingdom of Saudi Arabia, for support.

Author contributions

Conceptualization, M.R, S.N., A.B., T.M., A.R., I.K., J.K., Y.J.; methodology, T.M., A.B; M.R; software, A.B; validation, A.B; formal analysis, T.M, S.N. A.R.; investigation, A.R., I.K.; resources, I.K. A.R., T.M. M.R.; data curation, T.M, A.R, A.B, S.N.; writing- original draft preparation, T.M, A.B., A.R, M.R.; writing-review and editing, T.M.; visualization, S.A., A.R; supervision, A.R, Y.J.; project administration, A.R; funding acquisition, J.K, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Gachon University, Republic of Korea under grant no. RS-2023-00227648.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025