IUST
Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

Diagnosis of Coronary Heart Disease via Robust Artificial Neural Network Classifier by Adaptive Synthetic Sampling Approach

Elahe Moradi*(C.A.)

Abstract: With the intricate interplay between clinical and pathological data in coronary heart disease (CHD) diagnosis, there is a growing interest among researchers and healthcare providers in developing more accurate and reliable predictive methods. In this paper, we propose a new method entitled the robust artificial neural network classifier (RANNC) technique for the prediction of CHD. The dataset CHD in this paper has imbalanced data, and in addition, it has some outlier values. The dataset consists of information related to 4240 samples with 16 attributes. Due to the presence of outliers, a robust method has been used to scale the dataset. On the other hand, due to the imbalance of CHD data, three data balancing methods, including Random Over Sampling (ROS), Synthetic Minority Over Sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) approaches, have been applied to the CHD data set. Also, six artificial intelligence algorithms, including LRC, DTC, RFC, KNNC, SVC, and ANN, have been evaluated on the considered dataset with criteria such as precision, accuracy, recall, F1-score, and MCC. The RANNC, leveraging ADASYN to address data imbalance and outliers, significantly improved CHD diagnostic accuracy and the reliability of healthcare predictive models. It outperformed other artificial intelligence methods, achieving precision, accuracy, recall, F1-score, and MCC scores of 95.57%, 96.90%, 99.70%, 97.59%, and 93.42%, respectively.

Keywords: Artificial Neural Network, Robust Classifier, Imbalanced Dataset, Adaptive Synthetic Sampling Approach, Machine Learning.

1 Introduction

ACCORDING to statistics from the World Health Organization (WHO), heart disease is a serious global concern for human health [1]. Numerous factors can contribute to heart disease, such as diabetes, unhealthy eating habits, smoking, obesity, high cholesterol, high blood pressure, and irregular heart rhythms [2]. The World Heart Federation (WHF) stated in 2023 that the death rate from diseases associated with the heart has risen by 60% worldwide over the previous 30 years [3]. Approximately 18 million people die from cardiovascular illnesses each year, according to the WHO [4]. An incorrect diagnosis in the initial stages of cardiac heart disease (CHD) is the primary cause of death for most patients. Understanding disease prediction consequently demands the application of effective disease classification and prediction methods. On the other hand, predicting CHD requires the use of a more accurate model [5]. Machine learning (ML) methods for the prediction of human health diseases are being developed as a result of recent advancements in healthcare technology [6–8]. Developing better ML models has been the focus of numerous researchers. The main objective of machine learning is to create computer code that can access and utilize current data for predicting data in the future [9].

ML algorithms have been widely used in recent decades to use patient electronic medical records as data

Iranian Journal of Electrical & Electronic Engineering, 2024.
Paper first received 27 July 2024 and accepted 01 December 2024.
* The author is with the Department of Electrical and Computer Engineering, Yadegar-e-Imam Khomeini (RAH) Shahre Rey Branch, Islamic Azad University, Tehran, Iran.
E-mail: ElaheMoradi@iau.ac.ir.
Corresponding Author: E. Moradi.

to detect early cardiac problems [10]. [11] utilized a Naïve Bayes algorithm for predicting the diagnosis of heart disease patients. Approximately 500 patients with 11 features were included in the clinical data set used in this study, which was gathered from a top diabetic research facility in Chennai. According to the results, the Naïve Bayes algorithm produced 86.41% accuracy in the shortest amount of time. [12] used the publicly accessible Cleveland heart disease dataset, which is available on the University of California, Irvine (UCI) repository with 14 features of imbalanced data, to propose a number of machine learning (ML) algorithms, including logistic regression (LR), K-nearest neighbor (KNN), support vector machine (SVM), decision tree (DT), and random forest (RF). One-hot encoding for categorical features, data standardization using z-score normalization to normalize the features, and data stratification to split the dataset into training and validation sets to remove the unbalancing effect of disease classes were the techniques employed in this study. [13] presented a machine-learning approach to identify key correlated features in patients' electronic clinical records. It employed various classification algorithms, such as SVM, KNN, DT, RF, and LR, to train two datasets of CHD and failure data from the UCI machine learning repository. The study used the synthetic minority oversampling technique (SMOTE) for dataset balancing. The random forest algorithm outperforms other proposed algorithms. To diagnose heart disease (HD), [14] suggested a machine learning-based diagnostic technique. HD was found using machine learning (ML) prediction models such as ANN, LR, K-NN, SVM, DT, and NB. The features have been selected using standard state-of-the-art feature selection methods, including relief, MRMR, LASSO, and local-learning-based feature selection (LLBFS). The study also suggested a feature selection approach for fast conditional mutual information (FCMIM). To choose the optimal hyperparameters for model selection, the leave-one-subject-out cross-validation (LOSO) technique has been used. Cleveland HD is the dataset that was used to test the suggested approach. By training a DT model with a collection of attributes linked to a high risk of mortality, [15] analyzed ML algorithms that connect patient features with mortality. Because MARKERHF's limitations were generated by two hospitals in San Diego, California, they are biased towards a particular demographic location.

Taking into consideration the issue of data imbalance, [16] explored an intelligence system for the diagnosis of this type of CHD. This study employed the SMOTE approach for unbalanced data, using the K-fold cross-validation model for splitting the data into training and testing. The K-star algorithm was additionally applied to train multiclass classification. The multi-layer perceptron (MLP) neural network approach for the diagnosis of coronary heart disease has been studied by [17]. The Cleveland dataset from the University of California Irvine (UCI) was used for this study. It included 13 features, one of which was the diagnosis's output, that affect the incidence of CHD.

[18] have investigated the convolutional neural networks (CNNs) algorithm for predicting CHD into class imbalanced clinical data. The authors of this study implemented LASSO-based feature weight evaluation and majority voting to identify crucial features using data from the National Health and Nutritional Examination Survey (NHANES). The authors employed several ML algorithms and artificial neural networks (ANNs) for the CHD. To improve performance due to the unbalanced dataset, the SMOTE approach was utilized in this study. The UCI Machine Learning Repository is an open-access resource, and the dataset used in this work is available there [19-20]. Using the South African Heart Disease dataset, the four machine learning techniques (MLP neural networks, SVM, KNN, and logistic regression) were applied and studied [21]. According to the dataset imbalance, K-means SMOTE oversampling techniques were used to solve the problem in the data set, which significantly improved the predictive performance of all models for CHD diagnosis.

In [22], the authors proposed a combination of ML and deep learning (DL) for analysis and prediction of CHD diagnosis. The dataset was selected from the UCI Machine Learning with 13 features as inputs and one feature as a target. The CHD dataset included a few irrelevant features that were eliminated using the isolation forest. To improve the result, the data were additionally normalized. [23] employed two different techniques: logistic regression and decision tree for the CHD dataset. The dataset is publicly accessible on the Kaggle website. The dataset was subjected to the random sample procedure to address the imbalanced data in the percentage of participants with coronary heart disease.

[24] proposed a reliable ensemble strategy that effectively outperformed seven benchmark algorithms. On three CHD datasets—the Mendeley Data Center, the IEEE Data Port dataset, and the Cleveland dataset sourced from the UCI repository—the proposed method achieved more forecast accuracy. The algorithms of ML such as DT, KNN, stochastic gradient descent (SGD), NB, SVM, NB, and LR were employed to classify CHD data [25]. In this study, the authors employed MultiSURF, LASSO, variance threshold, ANOVA, and mutual techniques for feature selection data in the preprocessing step. [26] investigated a neural network ensemble-based, accurate prediction for the diagnosis of

heart disease. This approach combined the posterior probability from various models of processors. The database on CHD was extracted from the machine learning repository at UCI. [27] used genetic algorithms in a hybrid approach to improve neural networks' performance to detect heart disease. This study utilized the resources of the Z-Alizadeh Sani dataset, which included details on 303 people, 216 of whom had cardiac disease. For predicting heart disease, [28] suggested a hybrid optimization method utilizing adaptive stacked residual convolutional neural networks (CNNs). The main objectives of data preprocessing in this work were to address the missing values, standardize the data using the data scaling approach, and apply the random oversampling methodology to deal with unbalanced data. [29] utilized LASSO-CNN, AdaBoost-CNN, and AdaBoost-neural network (NN) for the identification of cardiovascular diseases (CVD). [30] proposed a new hybrid feature selection algorithm and utilized the Nasarian coronary artery disease dataset. In this research, RF, DT, XGBoost, and Gaussian Naïve Bayes (GNB) are employed for the dataset. In addition, the SMOTE technique is used for handling imbalanced data. [31] explored a support vector machine optimization function for the diagnosis of heart disease. In this study, a genetic algorithm (GA) was utilized to choose the more crucial features and improve the performance of the suggested method. Additionally, GA-SVM has been compared with other feature selection techniques such as Chi squares, info gain, relief, and filtered subsets.

From the study of previous research works, it was found that many researchers used different techniques to diagnose heart disease (CHD). The performance of a data-driven model can be increased if a balanced dataset is used for training and testing the model. In the previous study of the techniques used for imbalanced datasets, the researchers often used the SMOTE method.

Furthermore, the prediction accuracy of data preprocessing can be improved by using proper features and scaling the dataset. For this research, the CHD imbalanced dataset is used that is publicly accessible on the Kaggle website.

This study addresses the critical challenges of data imbalance and outliers inherent in CHD diagnosis by introducing a novel approach: a RANNC leveraging the ADASYN technique. This significantly improves both the predictive accuracy and reliability of healthcare predictive models, thereby contributing to more effective clinical decision-making in the field of deep learning. The following are the main contributions to this paper:

I. The given dataset has an unequal distribution of positive and negative classes, which can reduce performance. Therefore, we employed and compared various techniques, such as SMOTE, ROS, and ADASYN, to handle the given imbalanced data.

II. The given dataset has a large number of outliers. In order to improve the predictive performance of artificial intelligence (AI) models, we utilized a robust approach.

III. We employed the six AI classifiers for the CHD dataset in the presence of 4240 samples via tuning hyperparameters and evaluating the metrics such as accuracy, precision, F1-score, recall, and MCC to achieve the highest performance.

The format of this paper is as follows: Section 2 describes the materials and methodology, which include the algorithms that are applied and the Kaggle dataset that is utilized. Section 3 discusses the results, and Section 4 addresses the discussion. Section 5 lists the conclusion.

2 Materials and Methods

The following subsections addressed every background material and study methods.

2.1 Dataset Specifications

A dataset from the Kaggle website is accessible to the public and was used in the present study.[1] The data included people who lived in the Massachusetts town of Framingham. Data on 4240 samples with 16 attributes is included in the dataset. All of the characteristics, which include lifestyle, clinical, and demographic data, have the potential to be risk factors. Table 1 displays the explanation of the attribute. There are 4240 samples in all in the dataset under evaluation. Sixteen qualities define each sample. The information set that was retrieved as a consequence has a total of feature matrices. Table 1 provides a thorough overview of the dataset along with in-depth descriptions of 4240 samples that were obtained from its 16 attributes. Less than 500 samples from the UCI dataset were used in the majority of the datasets in the previous research investigations. Nonetheless, the dataset utilized for this study has more than 4,000 samples.

The train dataset comprises 4240 samples, where 84.81% (3596) of the samples were residences with no CHD, while only 15.19% (644) of the residences have CHD, suggesting that an imbalanced classification exists in the train dataset. Fig. 1 shows the percentage of samples for each type of CHD.

Because most clinical datasets are unbalanced, it is necessary to balance them for algorithms to perform better. There are different methods to balance imbalanced datasets, which we will mention in the next subsection.

[1] https://www.kaggle.com/datasets/palakdoshijain/coronary-heart-disease-prediction-in-ten-years

Table 1 Description of features from CHD dataset.

Feature Name	Feature description	Feature range	Feature role
Age	Patient age	[32,70]	Input
Sex	Patient gender	[0,1] 1: Male 0: Female	Input
Education	Education level	[1, 4] 1: High school 2: High school diploma 3: College 4: Degree	Input
Current Smoker	Whether or not the patient is current smoker	[0,1] 1: Participant is a current smoker 0: Participant is non-smoker	Input
Cigs PerDay	The number of cigarettes that the person smoked on average per day	[0,70]	Input
BPMeds	Whether or not the patient was on blood pressure medication	[0,1] 1: on a blood pressure medication 0: not on blood pressure medication	Input
Prevalent Stroke	Whether or not the patient has previously had a stroke	[0,1] 1: has had occurrences of stroke 0: no prevalence of stroke	Input
Prevalent Hyp	Whether or not the patient was hypertensive	[0,1] 1: prevalence of hypertension 0: no prevalence of hypertension	Input
Diabetes	Whether or not the patient had diabetes	[0,1] 1: has diabetes 0: no diabetes	Input
TotChol	Total cholesterol level (mg/dL)	[107,696]	Input
SysBP	Systolic blood pressure (mmHg)	[83.5,295]	Input
DiaBP	Diastolic blood pressure (mmHg)	[48,142]	Input
BMI	Body Mass Index (kg/m²)	[15.54,56.8]	Input
HeartRate	Heart rate in bpm	[44,143]	Input
Glucose	Glucose level (mg/dL)	[40,394]	Input
TenYearCHD	10-year risk of coronary heart disease (CHD)	[0,1] 1: Yes 0: No	Target

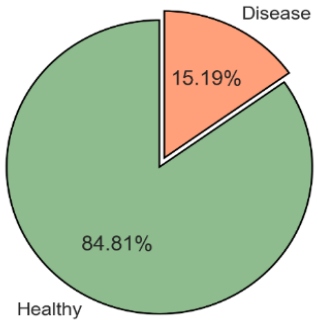

Fig. 1 Imbalanced dataset distribution in CHD.

2.2 Methods

The CHD dataset contains outliers that negatively affect the training of the classifier. Therefore, this increases the training time and undesirable affects the performance of the classifier. On the other hand, the dataset may have some features with high values and be distributed over a wide range, which leads to high training time.

A-Data Scaling

There are various methods for data scaling in the dataset, three common methods Standard Scaling (SS), Normalization Scaling (NS), and Robust Scaling (RS) are given in Table 2. SS and NS methods have been used in most of the previous papers.

It is also worth mentioning that the normalization method is divided into three categories, including Min-Max Scaling (MMS), Max Absolute Scaling (MAS), and Mean Normalization (MN) [32].

Table 2 Data Scaling Methods.

No.	Method	Formulation
1	Standard Scaling (SS)	$x_{scaled} = \frac{x - \mu}{\sigma}$
2	Normalization: Min-Max Scaling (MMS)	$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$
3	Robust Scaling (RS)	$x_{scaled} = \frac{x - Q_1}{Q_3 - Q_1}$

In table 2, x is the original value, and x_{scaled} is the value of data after scaling. Also, μ is the mean and σ is the standard deviation of samples. Further, x_{min} and x_{max} are minimum and maximum feature value, respectively. Finally, Q_1 and Q_3 are the 1st quartile and the third quartile, respectively.

In this paper, due to the presence of outliers in the CHD dataset, the RS method is proposed, and it has the best performance compared to the two data scaling methods in Table 2 [33].

B- Imbalanced dataset

There are also various methods for dealing with unbalanced datasets, such as ADASYN, SMOTE, and ROS. The SMOTE method has been used in most previous research, and the ADASYN method has been used in some of them. In this paper, the mentioned methods are compared with the ROS method. Table 3 shows the number of samples in the real dataset and the balanced dataset with different methods.

Table 3 Imbalanced and balanced datasets of CHD.

	Total samples	Majority of samples	Minority of samples	Ratio of majority to minority samples
Imbalanced dataset	4240	3596	644	5.58
ADASYN method	7263	3667	3596	1.01
SMOTE method	7132	3566	3566	1
ROS method	7192	3596	3596	1

The statistical indices for all 16 features of the CHD, such as the minimum, maximum, mean, and standard deviation of each feature, are depicted in Table 4.

Table 4 Statistical indices of continuous characteristics of the CHD dataset.

Feature Name	Minimum	Maximum	Mean	Std. deviation
Age	32	70	49.58	8.57
TotChol	107	696	236.69	44.59
SysBP	83.50	295	132.35	22.03
DiaBP	48	142.5	82.89	11.91
BMI	15.54	56.8	25.8	4.07
HeartRate	44	143	75.87	12.02
Glucose	40	394	81.96	23.94

The participants' ages range from 32 to 70 years old, with an average age of approximately 50 based on data from Table 4. This suggests that the majority of the participants are older people. The participants are, on average, overweight, as indicated by their average BMI of 25.8 kg/m². The individuals have a maximum BMI of 56.8 kg/m² (obesity) and a minimum BMI of 15.54 kg/m² (underweight). The total cholesterol level is 236.69 mg/dL on average, which is regarded as borderline excessive. Furthermore, 132.35 mmHg is the average systolic blood pressure, which may be a factor in hypertension. The diastolic blood pressure, heart rate, and glucose level averages, however, are all within acceptable limits.

Fig. 2 shows the heatmap of the datasets created from the CHD dataset, which is the source of the research. An effective method for visualizing data in two dimensions that can be used to describe anomalies, patterns, and varying intensities is the heatmap. In addition to showing the associations between features, the heatmap is a crucial tool for comprehending the fundamental interactions that can influence the prediction of CHD.

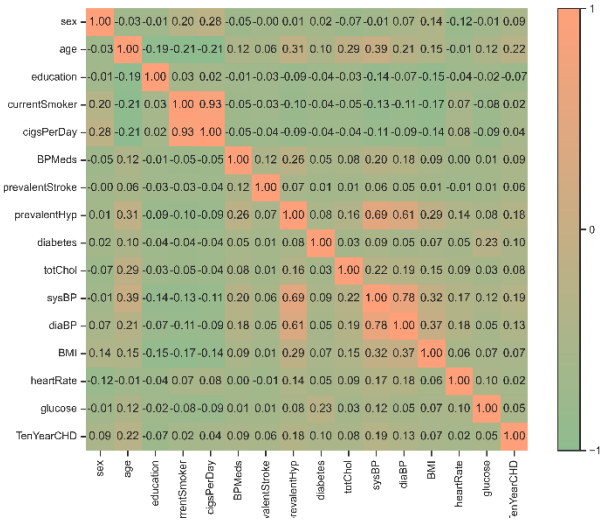

Fig. 2 Correlation heatmap of CHD dataset.

C- Splitting the CHD dataset

The CHD dataset is divided arbitrarily into two subsets. In machine learning, the 80:20 split, which allocates 80% for training and 20% for testing, is commonly utilized to maximize learning while guaranteeing robust assessment, in accordance with accepted practices [34-35]. In our study, we divided the CHD dataset 80:20 so that numerous algorithms could be trained and tested. A five-fold cross-validation approach was also used in conjunction with this split to improve the assessment of the proposed methods performance. The proposed methods were trained on four of the five equal folds created by splitting the training subset, and the remaining fold was used for validation. Repeating this procedure over all folds reduced the chance of overfitting and increased the reliability of performance metrics.

D- Artificial intelligence classifiers

After preprocessing the CHD dataset, this study utilized various AI classifiers including Logistic Regression Classifier (LRC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), K-Nearest Neighbors Classifier (KNNC), Support Vector Classifier (SVC), and Artificial Neural Network (ANN).

- LRC

One supervised machine learning approach for solving

classification problems and predicting probability-based target variables is the logistic regression classifier (LRC). When performing binary classification tasks, the objective is to predict the value of the dependent variable when it takes one of two possible values: 0 for the negative class and 1 for the positive class [36]. LRC is often used for these types of tasks. Three or more ordinal variables, or three or more ordered variables in the target variable, are characteristics of multinomial target variables [13]. In this research, the LRC is implemented using a solver set to 'lbfgs'.

- DTC

DTC is employed to solve classification problems. There are four types of nodes in it: root, inner, branch, and leaf. The structure of the data is a tree, with the root node signifying the entire data set, the inner nodes denoting its features, the branches signifying the decision-making region, and the leaf nodes signifying the final result. The features chosen from the dataset are used to make decisions [37]. The algorithm begins at the root node when predicting attributes from the dataset. After comparing the value of the root attribute with the value of the feature in the dataset, the algorithm advances to the next node. The procedure proceeds to the subsequent node, where the features of that node are contrasted with those of the subsequent nodes. This process continues until the leaf node is reached [13]. In this paper, the DTC is employed with the criterion for splitting set to 'Gini'.

- RFC

The random forest method uses many decision trees (DT) to classify data. The RFC considers the outcomes of every tree to produce an accurate forecast, and it ultimately decides which results receive the greatest votes [38]. RFC employs ensemble learning to address problems by combining many classifiers to enhance algorithmic performance. Multiple classifiers for DT are included in the method. To increase the prediction accuracy, each DT uses a subset of the data; the average is then calculated. The RFC employs a majority vote to decide how to proceed with predictions based on each tree's predictions rather than just one [13]. In this paper, the RFC is implemented with 100 trees in the forest and using 'Gini' as the criterion for measuring the quality of a split.

- KNNC

For new input instances, KNNC is used to predict the class label. To make predictions, this algorithm compares the new input to the input samples from the training set. When new input is the same as samples that are already in the training set, KNNC's performance is suboptimal [39]. The dataset is trained by KNNC and then stored in memory. When new data points are being tested for classification, the algorithm finds the most similar class based on K value and the nearby one based on the Euclidean distance. This is done by comparing the state similarity of the new data point and the stored data set [13]. In this work, the KNNC is used with k set of 5.

- SVC

Because of the SVC technique's exceptional machine-based classification performance, it is mainly employed for classification tasks. There are numerous applications for this method that are extensively utilized. This model is comparable to neural networks in that it aims to fine-tune a set of parameters, enabling the establishment of boundaries in a dimensional space and the approximation of functions or distinct patterns in various regions of the feature space. The training procedure used to change the settings accounts for the discrepancy. In contrast, SVC bases its training on the maximization of the margin between the instances of the two classes and the hyperplane (this model was originally meant to address issues of classifying two classes, nevertheless there are adaptations for multiclass and regression issues as well). Both linear and nonlinear data can be used with this approach. The procedure determines a hyperplane with the maximum margin as the greatest distance between data points of two classes when the data are linear. The method can categorize the test dataset with high confidence due to the maximum margin. The decision boundary that divides the class data is called a hyperplane. The data points that develop in proximity to the hyperplane are known as support vectors. To optimize the margin under support vectors, the distance is increased. Therefore, when these support vectors are eliminated, the hyperplanes alter. As a result, these ideas create an SVC. The original coordinate region is transformed (NN) using classification inputs to determine the most closely matching class among numerous into a separable space for nonlinear data [13, 40-41]. In this study, the SVC is utilized with the default radial basis function (RBF) kernel, and the default value for the regularization parameter C is set to 1.0.

- ANN

In ANN, a neural network has options for an observation. Although both numerical and categorical input characteristics (independent variables) are allowed, a category-dependent feature is required. An input layer, a hidden layer, and an output layer appear as the three layers of the NN classifier. Input values utilized in the network's training phase are given to the neural network's input layer. For the class that is already known, the NN's output is determined. By taking into consideration the difference between the anticipated and observed class values, the weight is reevaluated [42]. In this research, the ANNC Classifier consists of a sequential model with three hidden layers: the first hidden layer has 400 neurons, the second hidden layer

has 300 neurons, and the third hidden layer has 200 neurons. Also, the output layer consists of 2 neurons, suitable for binary classification tasks.

E- Data Balancing Techniques

In this subsection, a detailed explanation of the three data balancing techniques employed in this study is provided: Random Over Sampling (ROS), Synthetic Minority Over Sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN).

- ROS

A simple technique such as replication of occurrences from the minority class to even out the dataset imbalance would be an approach called ROS. On one hand, this method is easy to implement and helps overcome instances of the problem; however, such could be viewed as being too artificial since the repeated instances tend to evoke memory for overfitting. One inherent demerit of random oversampling, besides this, relates generally to the generation of instances that do not bring any new information into the data set. Therefore, ROS might not be that effective in improving the performance of models. Despite these limitations, ROS serves as a foundational method that can be beneficial when combined with other techniques or used in scenarios where computational simplicity is desired [43].

- SMOTE

SMOTE is a well-known method of oversampling that synthesizes a new example by interpolating between two samples from the minority class. Specifically, SMOTE chooses a minority instance and finds its k-nearest neighbors; new instances will then be created from the selected instance by following the line segments connecting it to the k-neighbors. Overall, SMOTE increases the number of samples in the minority class but also populates the feature space with different synthetic examples of numerous ones than other oversampling techniques such as ROS. SMOTE thereby increases the robustness of the model while reducing overfitting tendencies associated with simpler methods. Thus, it has been validated that SMOTE improves classification performance for multiple applications [19, 43].

- ADASYN

ADASYN is an advanced oversampling method that generates synthetic data points for the minority class. Unlike traditional approaches, ADASYN adapts the number of synthetic instances generated based on the local density of minority class samples. This means that regions, where the minority class is less dense, will have more synthetic samples created, efficiently balancing the dataset while maintaining its underlying structure. This adaptive nature supports improving the classifier's performance by enhancing its ability to learn from underrepresented areas of the feature space [43].

F- Evaluation indices

Performance metrics can be employed to assess the efficacy of AI methodologies. There are various evaluation metrics in AI classifiers, such as accuracy, precision, F1-score, and recall. One of the other criteria that is effective in binary classifications but has received less attention from researchers regarding the CHD dataset is the Matthews Correlation Coefficient (MCC). In this study, we apply indices such as accuracy, precision, F1-score, recall, and MCC to evaluate the performance of robust artificial intelligence methods on the CHD dataset. To determine the index formula, it is necessary to define the main components of the confusion matrix as follows:

- True Positive (TP):

A positive test result indicates that the patient appears to have CHD.

- True Negative (TN):

Test results are negative even though the patient is diseased.

- False Positive (FP):

Despite testing positive, the patient is negative for CHD.

- False Negative (FN):

The patient does not have the disease, according to negative test results.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1-score = \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Although they are widely used, F1-score and accuracy might produce inflated, overoptimistic results, particularly when used with imbalanced datasets [34].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (5)$$

Its range is [−1, +1], with the extreme values being obtained in the cases of perfect misclassification (-1) and perfect classification (+1), respectively. The expected value for the coin-tossing classifier is MCC = 0 [34].

G- Hyperparameter Tuning

Hyperparameter tuning of the six proposed artificial intelligence classifiers has been performed with the systematic trial-and-error method of fine-tuning hyperparameters iteratively to discover configurations yielding optimal performance. Model evaluations in performance were carried out with metrics of precision, accuracy, recall, F1-score, and MCC ensuring refined and effective selection classifiers.

3 Results

We employed the imbalanced dataset of CHD, including all of its properties, and utilized three methods

of balanced dataset due to the best performance of our proposed method, including ADASYN, SMOTE, and ROS. For data scaling of the dataset, we used RS and applied six distinct AI algorithms, namely LRC, DTC, RFC, KNNC, SVC, and ANN, to it. The accuracy, precision, recall, F1-score, and MCC of each of the six algorithms with each of its 15 features and three methods of balancing the dataset are shown in Figures 3 to 7, respectively. According to the findings of the research shown in these figures, the ADASYN method seems to be the most efficient method since it achieves the most significant proportion of metric predictions. Also, according to these figures, the ANN algorithms with each of the three balanced methods seem to be the most efficient classification method since they achieve the most significant proportion of metric predictions.

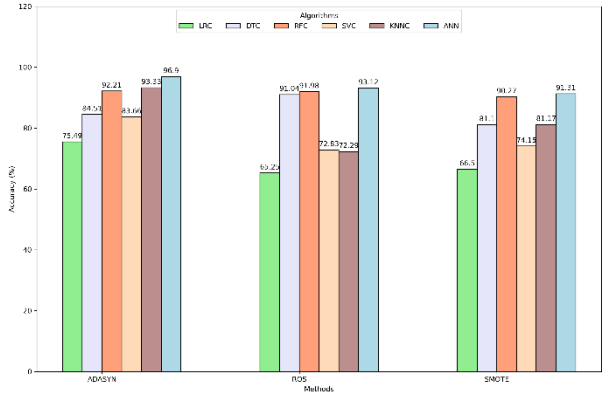

Fig. 3 Comparison of the RAI methods on the CHD dataset using the accuracy criteria.

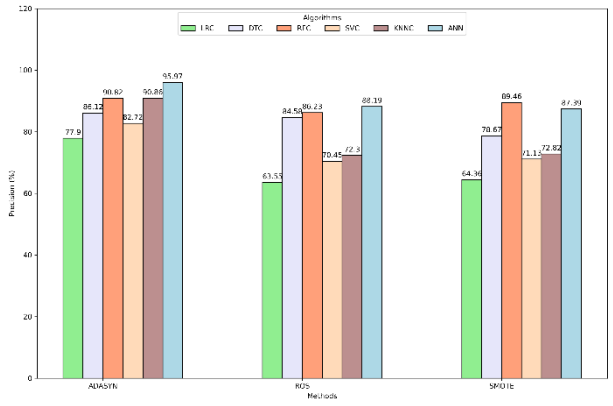

Fig. 4 Comparison of the RAI methods on the CHD dataset using the precision criteria.

To implement model codes for RAI algorithms, we used Jupyter Notebook, which is based on Python. To increase the reproducibility of the results, the RANN model details and the range of the hyperparameters used for optimizing the model are mentioned in Table 5. All analyses are performed in Python and its frameworks, such as Numpy, Matplotlib, Pandas, Seaborn, Scikit-Learn, Keras, and TensorFlow.

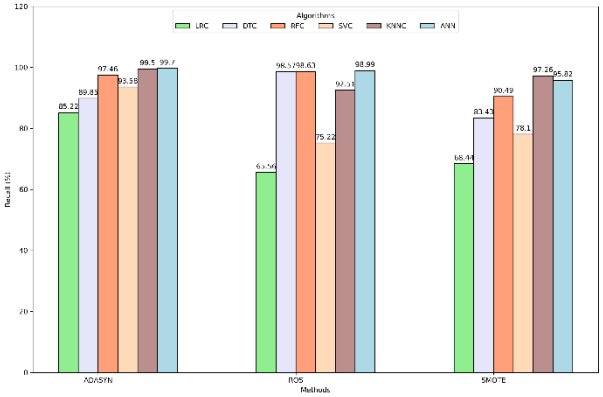

Fig. 5 Comparison of the RAI methods on the CHD dataset using the recall criteria.

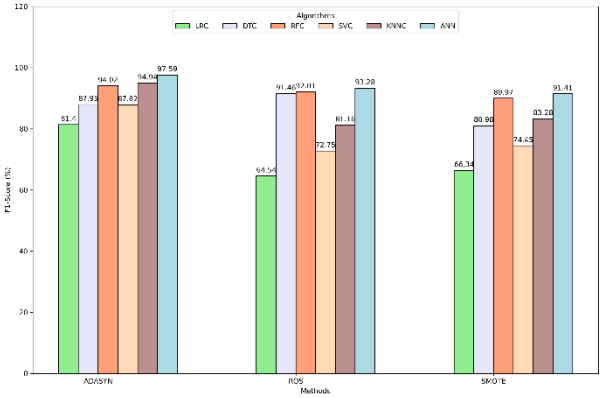

Fig. 6 Comparison of the RAI methods on the CHD dataset using the F1-score criteria.

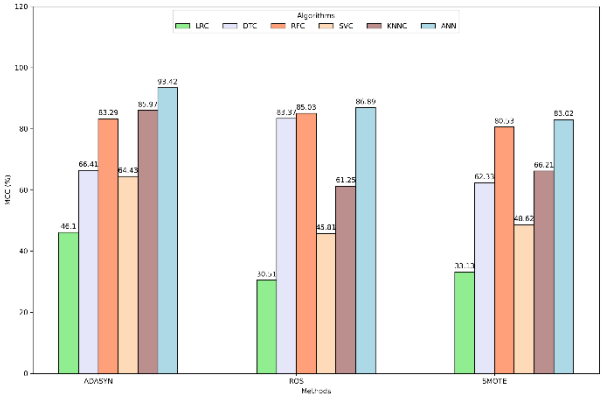

Fig. 7 Comparison of the RAI methods on the CHD dataset using the MCC criteria.

Table 5 Hyperparameters of RANN algorithm.

	ADASYN	ROS	SMOTE
Activation function	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	RMSprop
Learning rate	0.001	0.001	0.001
Epochs	400	400	400
Batch size	64	128	64

4 Discussion

According to the results obtained in the figures of the previous section, it can be seen that the best performance of the proposed algorithms after the ANN algorithm, the RFC, KNNC, DTC, SVC, and LRC algorithms are respectively based on the evaluation criteria.

Also, according to the obtained results, it is clear that the ADASYN method has performed better in balancing the CHD dataset than the other two methods, SMOTE and ROS. As mentioned, one of the important indicators in binary calcification datasets is the MCC criterion. Therefore, six of the proposed algorithms with three data balancing methods using the MCC criterion can be seen in Fig. 8. The value of the MCC index for the robust ADASYN method is the best, and it is 93.42%, 85.97%, 83.29%, 66.43%, 66.41%, and 46.10% in algorithms ANN, KNNC, RFC, DTC, SVC, and LRC, respectively. The value of the MCC index for the robust ROS method is 86.89%, 85.37%, 85.03%, 61.25%, 45.81%, and 30.51% in algorithms ANN, DTC, RFC, KNNC, SVC, and LRC, respectively. The value of the MCC index for the robust SMOTE method is 83.02%, 80.53%, 66.21%, 62.33%, 48.62%, and 33.13% in algorithms ANN, RFC, KNNC, DTC, SVC, and LRC, respectively.

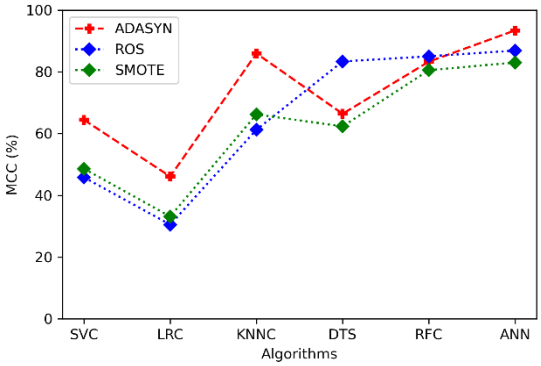

Fig. 8 Comparison of MCC criteria after implementing ADASYN, ROS, SMOTE techniques on CHD datasets.

The accuracy and loss curves of the ANN algorithm with the SMOTE method and RS data scaling can be seen in Figures 9 and 10, and the accuracy and loss values are equal to 91.31 and 0.38.

The accuracy and loss curves of the ANN algorithm with the ROS method and RS data scaling can be seen in Figures 11 and 12, and the accuracy and loss values are equal to 93.12 and 0.37.

It is worth noting that among the proposed methods, the best performance with the investigated indicators belongs to the ANN algorithm with the ADASYN method and RS data scaling. Therefore, the accuracy and loss curves of the robust proposed method can be seen in Figures 13 and 14, and the accuracy and loss values are equal to 96.90 and 0.13.

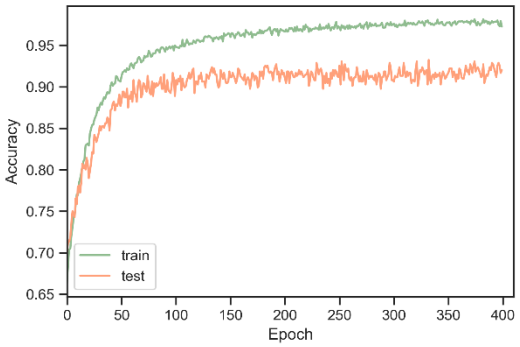

Fig. 9 Accuracy of ANN algorithm on CHD dataset after implementation of SMOTE.

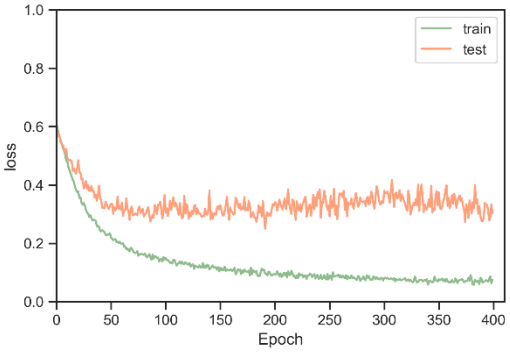

Fig. 10 Loss of ANN algorithm on CHD dataset after implementation of SMOTE.

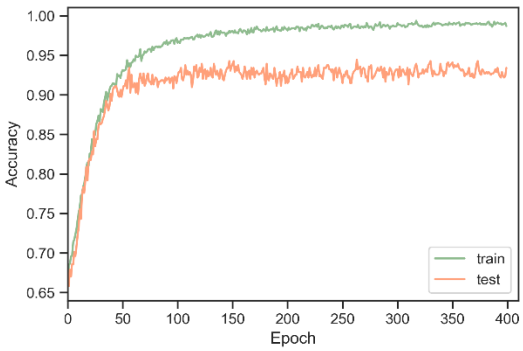

Fig. 11 Accuracy of ANN algorithm on CHD dataset after implementation of ROS.

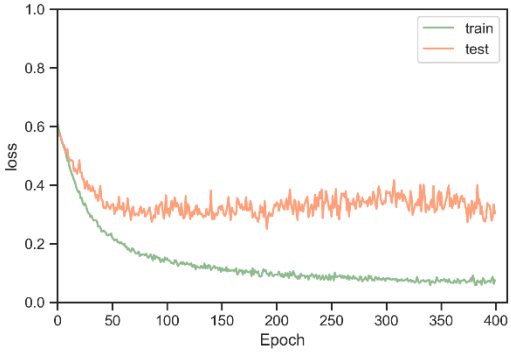

Fig. 12 Loss of ANN algorithm on CHD dataset after implementation of ROS.

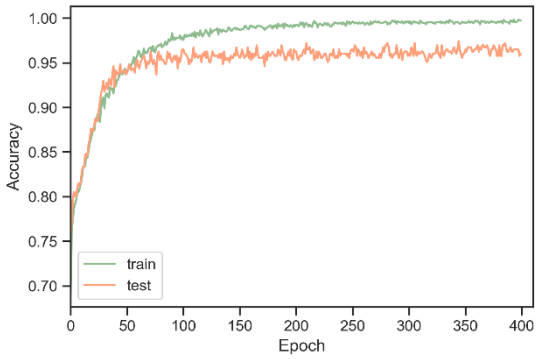

Fig. 13 Accuracy of ANN algorithm on CHD dataset after implementation of ADASYN.

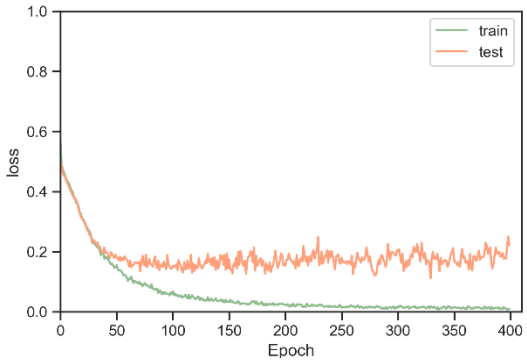

Fig. 14 Loss of ANN algorithm on CHD dataset after implementation of ADASYN.

The findings and simulations studied showed that the robustness of the proposed technique is attributed to several aspects. A robust approach was employed to handle outliers in the CHD dataset, which is critical for maintaining the veracity of the model's predictions. Furthermore, the subject of imbalanced data was explicitly addressed using the ADASYN technique. This approach was highlighted as the most effective for balancing the dataset, contributing to the robustness of the ANN by ensuring that the model is not biased towards the majority class. These strategies collectively improved the proposed approach's ability to generalize well across variations in the dataset, making it robust against potential biases and errors introduced by outliers and class imbalances.

The outcomes show the efficiency of the RANNC in enhancing diagnostic accuracy and its potential to transform healthcare systems. Its robust handling of data imbalance and outliers provides a reliable foundation for developing advanced diagnostic tools, leading to improved clinical decision-making and patient outcomes.

5 Conclusion and future research

Define This research paper aimed to classify the binary of the CHD dataset effectively using robust artificial intelligence (RAI) based on different balanced methods. Preprocessing is done on the dataset, and due to the presence of outliers in the dataset, we used the robust scaling approach for data scaling. Further, we investigated three methods, such as ROS, SMOTE, and ADASYN, to overcome imbalanced classes. Finally, six algorithms of RAI were evaluated against a publicly available dataset to ensure that the model is reliable. Also, in this paper, in addition to common indicators such as precision, accuracy, recall, and F1-score, we have used a special binary classification index called MCC. The RANN model, after implementing the ADASYN approach, performed better in classification than other models like LRC, DTC, RFC, KNNC, and SVC in terms of disease predictions. The RANNC not only enhances predictive accuracy for the CHD but also pointedly improves healthcare systems by effectively managing data imbalance and outliers. This study can be further extended by exploring the application of IoT in real-time sample testing.

Conflict of Interest

The author declares no conflict of interest.

Author Contributions

Elahe Moradi performed the simulation, assessed the outcomes, and composed the ultimate draft of the.

Funding

No funding was received for conducting this study.

Informed Consent Statement

Not applicable.

References

[1] World Health Statistics. Cardiovascular Diseases, Key Facts. 2021. Available online: https://www.who.int/news-room/factsheets/ detail /cardiovascular-diseases-(accessed on 10 December 2022).

[2] R. P. Choudhury, N. Akbar, “Beyond Diabetes: A Relationship between Cardiovascular Outcomes and Glycaemia Index”, *Cardiovascular Research*,117, 97–98, 2021.

[3] World Heart Federation, Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report – World Heart Federation. https://world-heartfederation, 9 August, 2023.

[4] World Health Organization: WHO, Cardiovascular diseases. https://www.who.int/health topics/ cardiovascular diseases, 11 June, 2019.

[5] N. Chandrasekhar, S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization”, *Processes*, 11(4), 2023.

[6] S. Mohan, C. Thirumalai, G. Srivastava, “Effective

Heart Disease Prediction Using Hybrid Machine Learning Techniques". *IEEE Access*, 7, 81542 – 81554, 2019.

[7] K. R. Chowdary, P. Bhargav, P.; N. Nikhil, K. Varun, D. Jayanthi, "Early Heart Disease Prediction Using Ensemble Learning Techniques", International Conference on Electronic Circuits and Signaling Technologies, *J. Phys. Conf. Ser*, 2325, 2022.

[8] J. Liu, X. Dong, H. Zhao, Y. Tian, "Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion", *Processes*, 10 (4), 749, 2022.

[9] A.G. Devi, S.R. borra, "A Method of Cardiovascular Disease Prediction Using Machine Learning", *Int. J. Eng. Res. Technol*, 9 (5), 243–246, 2021.

[10] M. Ganesan, N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models", In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5), March 2019.

[11] K. Vembandasamy, R. Sasipriya, E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", *Int. J. Innov. Sci. Eng. Technol*, 2(9), 441–444, 2015.

[12] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis", *IEEE Access*, 8, pp. 14659-14674, 2020.

[13] E.M. Senan, I. Abunadi, M.E. Jadhav, S.M. Fati, "Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms", *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, 8, 107562-107582, 2020.

[15] E. D. Adler, A. A. Voors, L. Klein et al., "Improving risk prediction in heart failure using machine learning," *European Journal of Heart Failure*, vol. 22, no. 1, pp. 139–147, 2020.

[16] W. Wihharto, MCom, H. Kusnanto, DRPh, H. Heranto, DrEng, "Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm", *Health Informatics Research*, 22(1), 2016.

[17] H. Wiharto, H. Kusnanto, H. Herianto, "Hybrid System of Tiered Multivariate Analysis and Artificial Neural Network for Coronary Heart Disease Diagnosis", *International Journal of Electrical and Computer Engineering*, 7(2), 1023-1031, 2017.

[18] A. Dutta, T. Batabyal, M. Basu, S.T. Acton, "An efficient convolutional neural network for coronary heart disease prediction", *Expert Systems with Applications*, 159, 2020.

[19] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction", *Scientific Programming*, 1-12, 2021.

[20] N. M. Lutimath, H. V. Ramachandra, S. Raghav, and N. Sharma, "Prediction of Heart Disease Using Genetic Algorithm," Proceedings of Second Doctoral Symposium on Computational Intelligence. Advances in Intelligent Systems and Computing, vol 1374. Springer, Singapore, September 2021.

[21] H. Khadir, N. M. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction", *International Journal of Advanced Computer Science and Applications*, 12(5), 2021.

[22] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*, 2021.

[23] A.N.M. Zamhari, N.F.M. Fazli, Z.A. Rahim, B.M. Osman, S. Sapri, Z. Derasit, "HANDLING IMBALANCE DATA IN MODELLING CORONARY HEART DISEASE", *International Journal of Social Science Research*, 4(4), 2022.

[24] B.P. Doppala, D. Bhattacharyya, M. Janarthanan, N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques", *Journal of Healthcare Engineering*, 1-13, 2022.

[25] F. Ullah, X. Chen, K. Rajab, M.S. Reshan, A. Shaikh, M.A. Hassan, M. Rizwan, M. Davidekova, "An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Predication", *Computational Intelligence and Neuroscience*, 1-12, 2022.

[26] R. Das, I. Turkoglu, A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", *Expert Systems with Applications*, 36, 7675-7680, 2009.

[27] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Roshanzamir, A.A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm", *Computer Methods and Programs in Biomedicine*, 141, 19-26, 2017.

[28] R. Srinivas, R.K. Bagadi, T.R. Reddy, N. Praveen, G. Aparanjini, "An efficient hybrid optimization algorithm for detecting heart disease using adaptive stacked residual convolutional neural networks", *Biomedical Signal Processing and Control*, 87, 2024.
[29] A. Najafi, A. Nemati, M. Ashrafzadeh, S.H. Zolfani, "Multiple-criteria decision making, feature selection, and deep learning: A golden triangle for heart disease identification", *Engineering Applications of Artificial Intelligence*, 125, 2023.
[30] E. Nasarian, M. Abdar, "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach", *Pattern Recognition Letters*, 133, 33-40, 2020.
[31] C.B. Gokulnath, S.P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease", *Cluster Computing*, 2019.
[32] M.M. Ahsan, M.A.P. Mahmud, P.K. Saha, K.D. Gupta, Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance", *Technologies*, 9, 2021.
[33] E. Moradi, "A Data-Driven based Robust Multilayer Perceptron Approach for Fault Diagnosis of Power Transformers" 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP).
[34] D. Chicco, G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification", *BioData Mining*, 16(4), 2023.
[35] D. Chicco, G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification", *BMC Genomics*, 21, 2020.
[36] P. Ranganathan, C.S. Pramesh, R. Aggarwal, "Common pitfalls in statistical analysis: Logistic regression", *Perspectives in Clinical Research*, 2017.
[37] H.H. Patel, P. Prajapati, "Study and analysis of decision tree-based classification algorithms", *International Journal of Computer Sciences and Engineering*, 6(10), 74-78, 2018.
[38] S.J. Rigatti, "Random Forest", *Journal of Insurance Medicine*, 47(1), 31-39, 2017.
[39] H.A. Alfeilat, A.B. Hassanat, O. Lasassmeh, A.S. Tarawneh, M.B. Alhasanat, H.S. Eyal Salman, V.S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review", *Big data*, 7(4), 221-248, 2019.
[40] M. Nilashi, H. Ahmadi, A.A. Manaf, T.A. Rashid, "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates", *International Journal of Fuzzy Systems*, 22, 1376-1388, 2020.
[41] M. Nilashi, H. Ahmadi, A.A. Mand, T.A. Rashid, S. Samad, L. Shahmoradi, N. Aljojo, E. Akbari, "Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates", *Int. J. Fuzzy Syst*, 2020.
[42] S. Bianco, R. Cadene, L. Celona, P. Napoletano, "Benchmark analysis of representative deep neural network architectures", *IEEE Access*, 6, 64270-64277, 2018.
[43] V. W. De Vargas, J. A. S. Aranda, R. D. S. Costa, P. R. Da Silva Pereira, and J. L. V. Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowledge and Information Systems*, vol. 65, no. 1, pp. 31–57, Nov. 2022.

Elahe Moradi received her PhD degree in Electrical Engineering-Control from Islamic Azad University, Science and Research Branch, in the year 2017. At present, Elahe Moradi is an Assistant Professor of electrical and computer engineering department at Islamic Azad University, Yadegar-e-Imam Khomeini (RAH) Shahr-e-Rey Branch.
Her main areas of research are optimal control, robust control, fault detection and diagnosis, neural networks, machine learning, and deep learning.