Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Effective Heart Disease Prediction Model Using Deep Learning-based Dimensionality Reduction on Imbalanced Data

S. Kabirirad [1], V. Afshin [2], S. H. Zahiri [2,*]

[1]*Department of Computer Science, Faculty of Computer and Industrial Engineering, Birjand University of Technology, Birjand, Iran.*

[2]*Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.*

Article Info

Article History:
Received 25 June 2024
Reviewed 08 August 2024
Revised 03 October 2024
Accepted 15 October 2024

Keywords:
Dimensionality reduction
Imbalanced data
Heart disease prediction
Autoencoder
PCA
Information bottleneck
Feature extraction

*Corresponding Author's Email Address: *hzahiri@birjand.ac.ir*

Abstract

Background and Objectives: When dealing with high-volume and high-dimensional datasets, the distribution of samples becomes sparse, and issues such as feature redundancy or irrelevance arise. Dimensionality reduction techniques aim to incorporate correlation between features and map the original features into a lower dimensional space. This usually reduces the computational burden and increases performance. In this paper, we study the problem of predicting heart disease in a situation where the dataset is large and (or) the proportion of instances belonging to one class compared to others is significantly low.

Methods: We investigated the prominent dimensionality reduction techniques, including Principal Component Analysis (PCA), Information Bottleneck (IB), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) and Variational Autoencoder (VAE) on popular classification algorithms. To have adequate samples in all classes to properly feed the classifier, an efficient data balancing technique is used to compensate for fewer positives than negatives. Among all data balancing methods, a SMOTE-based method is selected, which generates new samples at the boundary of the samples distribution and avoids the synthesis of noise and redundant data.

Results: We used UCI and Kaggle datasets to simulate and evaluate the model. The experimental results show that VAE-based method outperforms other dimensionality reduction algorithms in the performance measures. The proposed hybrid method improves accuracy to 97.7% and sensitivity to 99.4%. Also, a feature importance analysis is provided to show insights into the factors driving the predictions and help understand the underlying mechanisms of heart disease.

Conclusion: Finally, it can be concluded that the combination of VAE with oversampling algorithms can significantly enhance system performance as well as computational time.

Introduction

Heart disease or cardiovascular disease is one of the leading causes of death in humans and its early diagnosis is quite challenging. Many studies are performed to improve the early detection of heart disease and reduce mortality. These studies aim to develop computer-aided diagnostic systems using emerging technologies. These systems predict heart disease based on data classification

algorithms; thus, the application of efficient algorithms plays an essential role in their accuracy. Many researchers have employed machine learning algorithms to construct diverse models and have attained remarkable accomplishments [1], [2]. To incorporate correlation between features, dimensionality reduction methods can be used. These methods can map the initial features into a space with fewer dimensions and extract effective features to feed the classification models. Many researchers have emphasized that feature reduction can improve performance and lead to faster processing per record.

Recently, AEs have excelled in unsupervised machine learning works for denoising data, compression and feature reduction. These networks can represent features in complex and large datasets with exceptional performance [3]. AEs can be considered as feedforward networks that their hidden layers have fewer neurons than the input and output layers. An AE is an encoder-decoder pair that generates an encoded representation and then reconstructs the input with encoded knowledge.

Usually, most dataset instances are normal and only a small percentage of them are related to abnormal or patient cases, as a result, the lack of patient instances may cause the model to not be properly fed and fully trained to recognize patients. Therefore, we use a data balancing phase to compensate for fewer patient instances than normal ones.

The Synthetic Minority Oversampling Technique (SMOTE) has promising results in addressing imbalanced data [4]. However, SMOTE has limitations, as it can generate noise and redundant data that do not significantly enhance the performance parameters. To overcome these limitations, improved versions of SMOTE, such as the Borderline Synthetic Minority Oversampling Technique (BSM) are proposed [5]. This technique focuses on generating samples at the boundary of the sample distribution to avoid the synthesis of noise and redundant samples.

When the training data has a large volume or high sample dimensions, there are problems such as the feature redundancy or feature irrelevance. In such a situation, SMOTE-based sampling methods lead to failure. Therefore, dimensional reduction methods can be helpful to implement sampling methods in low-dimensional space. The traditional dimension reduction method creates a great deal of redundancy in the feature space and the distribution of samples between the categories is mixed. This is a challenge for data synthesis with edge samples.

In this paper, a hybrid system is proposed that uses dimensionality reduction techniques namely, PCA, Information bottleneck (IB), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) and variational AE (VAE) to incorporate the correlation between features and extract the most essential features. Then, new samples are synthesized using BSM, especially at the boundary of the sample distribution. Finally, the combined samples are applied to train classification algorithms, including MLP, SVM and Logistic regression (LR) algorithms. We analyze the impact of dimensionality reduction and data balancing techniques on the performance of the classification algorithms. The experimental results show that VAE outperforms PCA and IB, besides, PCA has better computational time than VAE and IB. Also, data augmentation improves performance metrics. It can be concluded that the use of deep learning methods increases performance and efficiency, especially in large data sets. The results show that the proposed model using AE-based dimensionality reduction and BSM oversampling methods provides better performance, accuracy of 97.7% and sensitivity of 99.4%. The main contributions of the paper are as follows:

- Investigating the impact of applying three dimensionality reduction methods, PCA, IB, t-SNE, UMAP and VAE, on several classification algorithms using performance measures (accuracy, sensitivity, F1-score, precision, ROC- AUC score).
- Applying an improved SMOTE algorithm, BSM, after dimensionality reduction. This has a significant effect on the performance in two ways: First, essential features are restored and the synthesized data is generated based on these features. Second, after reducing the dimension, the problem of synthesis of noise data is solved.
- Proving the higher performance of VAE rather than the other dimensionality reduction techniques.
- Studying the effect of dimensionality reduction on computational time of large datasets.
- Proposing a hybrid model based on VAE and BSM with high accuracy and sensitivity 97.7% and 99.4%.

The rest of the paper is organized as follows: Section "Methodology" reviews the building blocks of the proposed model including dimensionality reduction techniques, oversampling methods and machine learning algorithms. The proposed model is described in Section "Architecture of the Model". Section "Experimental Results" shows experimental results and performance analysis. Finally, the paper concludes in Section "Conclusion".

Related Work

Bhatt et al. [6] examined the efficacy of several machine learning algorithms in predicting heart disease. They proposed a k-mode clustering algorithm that utilizes random forest, decision tree, multilayer perceptron, and XGBoost. Khan et al. [7] presented a hybrid machine learning method and performed experimental analysis.

Hassan et al. [8] proposed a system with combining a pre-trained Deep Neural Network (DNN) for feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and Logistic Regression (LR) for classification. The system demonstrated accuracy rates of 91.79% and 93.33% on the Cleveland dataset. In [9], authors developed a system based on machine learning and feature selection algorithms to achieve acceptable results. In [10], a system was developed that combines ensemble deep learning and feature fusion methods. This system utilized two algorithms, information gain and conditional probability, to reduce the number of features and assign specific weights to heart disease features. Following this, an ensemble deep learning classifier was trained to forecast heart disease in patients.

PCA is a common statistical technique that has found applications for finding patterns in high-dimensional data.

Results of recent research demonstrate that utilization of deep learning methods enhances the accuracy of predictions. For example, in [11], some machine learning techniques, including logistic regression (LR), SVM, deep neural network, decision tree, Nave Bayes, random forest and k-nearest neighbor are investigated and it concluded that DNN had the best performance with 98.15% accuracy and 98.68% sensitivity. Deep learning has been used successfully in various fields, especially in image analysis, visualization and working with large volumes of data. It is an evolving technique that is capable of representation of multi-level records [12]. DNN is a complex neural network with several hidden layers between the input and output layers. The input data is converted to nonlinear or activation functions to generate classes. In [2], a hybrid DNN is proposed to utilize convolutional neural network (CNN) and long short-term memory jointly. This method can predict heart disease with an accuracy of 93.7%.

For example, in [13], the authors used various feature selection techniques to forecast heart disease. In particular, they employed an SVM classifier for forward feature extraction, along with back-elimination feature selection. Their results demonstrated a reduction in the number of input variables, leading to an improvement of accuracy up to 85%. In another paper, Shao et al. [14] proposed a rough set strategies and multivariate adaptive regression splines to optimize the number of descriptive features and achieve an accuracy of 82.14%.

In recent research, applying newer feature selection algorithms such as fuzzy-based systems or DNN has significantly improved performance metrics. In [15], a hybrid model based on Fuzzy C-means and ANN along with PCA was proposed. PCA was used to select important features of the dataset. The extracted data from PCA was clustered using fuzzy C-means and finally, ANN was applied to predict cardiovascular disease. Its simulation results showed the effectiveness of the method with an accuracy value 99.55%, however, the precision is 33.27% and it requires a significant improvement.

In [16], the authors proposed a two-phase method in which the first phase involved sparse AE training to learn the best representation of training data. The second phase utilized ANN to predict health status based on trained records. Its experimental results showed that the model's accuracy is 90%, which shows a better performance than some traditional machine learning and neural network approaches. Authors of [17] proposed a system based on two deep neural networks that consist of one PCA and four deep learning models, including two variational AE and two DNN models. Ebiaredoh-Mienye et al. [18] proposed a model consisting of feature selection and classification phases that integrate an improved sparse AE and Softmax regression. They showed that the model has a robust feature learning algorithm and a high-performance classification.

Methodology

In this subsection, we briefly review the algorithms used in the proposed model, including data balancing techniques, dimensionality reduction methods and machine learning algorithms.

A. Data Balancing

The classification of imbalanced datasets is a challenging issue. When imbalanced data appear in the classification, the problem of overfitting arises and the result will be biased toward the majority class. In such situations, the data should be balanced either by oversampling or undersampling techniques to improve the performance. An oversampling technique increases the number of samples of the minority class, such as ADYSAN, SMOTE, SMOTE-TOMKE, etc. An undersampling technique reduces the number of samples of the majority class, like Dense Nearest Neighbor, Edited Nearest Neighbor and so on. We can apply a hybrid method such that oversampling is applied to the minority class to improve the model detection for the minority class samples, and undersampling is performed on the majority class to reduce the bias in the majority class samples.

SMOTE is an extended method that builds upon the random oversampling algorithm. Initially, it computes the Euclidean distance between K neighboring samples belonging to the same category surrounding each sample x_i in the minority class. Subsequently, a neighbor is randomly chosen, and a synthetic sample is created with a probability that falls on the line connecting the sample and its neighbor. The formula for synthesis can be represented as below:

$$x_{new} = x_i + r * (\hat{x_i} - x_i) \quad (1)$$

where r is a random number between [0,1], x_i is a sample

to be oversampled and $\hat{x}_i$ is a random neighbor sample.

The SMOTE method is one of the widely used methods for data synthesis, for which many improvements have been proposed so far. But most of them are either complicated or only focus on one of SMOTE's weakness. Among the proposed methods, the BSM approach is an approach that, in addition to removing noise data and detecting main features, also considers border data.

BSM categorizes the samples into safe, noisy and dangerous samples, where the dangerous samples are those that are on the boundary of the distribution. By generating synthetic data for these samples, the ability of the method to identify patient samples is significantly increased. Applying BSM can help the method to predict heart disease to a great extent, especially when minority samples are difficult to detect.

B. Dimensionality Reduction

Dimensionality reduction is a preprocessing step that reduces high-dimensional data to a controllable size while retaining the original information intact. It is a common step used for pattern recognition, classification applications and compression schemes. The dimensionality reduction has been effective in multiple aspects: first, the reduced representation combines different features of the records. Second, reducing the dimension speeds up the execution of the algorithm and improves the performance of the system in some cases. In this paper, several common dimensionality reduction methods have been used: PCA, IB and AE. AE-based model is shown to provide better performance while PCA-based model improves speed compared to IB and AE.

I) Principal Component Analysis (PCA)

PCA is a linear transformation that reduces the dimensionality of the input data, keeping its most significant parts. To achieve this, one must calculate the eigenvalues and eigenvectors of the data covariance matrix, then arrange the eigenvectors based on the eigenvalues in a descending manner and ultimately project the original data onto the directions of the eigenvectors. This method is suitable for fully correlated data. In practice, the only important PCA parameter that needs to be adjusted is the dimension of the projection space. This can be conveniently determined by examining the variance ratios of the principal components. Several types of improvements have been introduced for PCA. For example, possible principal component analysis (PPCA) was introduced to address the problem of missing values of features [19] or an extended PCA [20] was presented for applying on big data.

II) Information Bottleneck (IB)

IB introduced as an information-theoretic principle for extracting a compressed representation of the input data that maximizes a target prediction. It can be considered as an optimization problem that minimizes the mutual information I(Z; X) between the input variable X and its latent representation Z and it maximizes the mutual information I(Z;Y) between the output variable Y and the latent representation Z. In other words, it intends to maximize the following objective function:

$$\phi_{IB}^{\Theta} = I(Z;Y|\vartheta) - \beta I(Z;X|\vartheta) \quad (2)$$

where β ∈ [0, 1] manages the size of IB and θ is a Lagrange multiplier.

III) t-SNE

t-SNE represents a non-linear, unsupervised, and manifold-based feature extraction technique. It can map the high-dimensional data into a lower-dimensional space, typically comprising two or three dimensions, while maintaining the significant structure of the original data. Its primary application lies in the realms of data exploration and visualization. Although various feature extraction algorithms exhibit robust performance, they often struggle with visualizing high-dimensional data effectively and frequently fail to maintain both local and global data structures. In this context, t-SNE proves to be an advantageous tool for visualizing high-dimensional data by preserving the important structural attributes. The process begins with the application of Stochastic Neighbor Embedding (SNE), which transforms high-dimensional Euclidean distances into conditional probabilities that denote similarities between each pair of data points. Subsequently, a student t-distribution with one degree of freedom, similar to Cauchy distribution, is utilized to derive the second set of probabilities in the lower-dimensional space. Consequently, t-SNE aims to minimize the divergence between these two sets of probabilities across the high-dimensional and low-dimensional spaces [21].

IV) UMAP

The UMAP algorithm stands out as a strong competitor to t-SNE in terms of visualization quality, often demonstrating a greater ability to maintain global structure while offering enhanced computational efficiency. Additionally, UMAP imposes no limitations on the embedding dimension, rendering it a versatile option for dimension reduction in machine learning applications. While UMAP is similar to t-SNE, it also possesses significant differences that have led many practitioners to favor it for dimension reduction tasks. UMAP optimizes performance utilizing cross-entropy as the loss function in contrast to t-SNE's use of KL divergence, and employing stochastic gradient descent to optimize the cost function rather than the more time-consuming gradient descent method.

V) Autoencoder (AE)

The methods such as PCA may not fully succeed in

extracting the complex features of nonlinear datasets. In order to address this issue, AE as a deep learning model can be used. AE is trained to learn how to generate the original input with a minimum reconstruction error. It comprises two steps: the encoder, which transforms the d-dimensional input data into a latent representation, and the decoder, which reconstructs the representation to a vector resembling the original input. This process is known as reconstruction, with the difference between the decoder's output and the original input termed as reconstruction error.

Node layers identify input data patterns and use them to generate encrypted data representations. The network training algorithm adjusts the behavior of each node to be close to the configuration of the input data. If a linear activation function is applied, the AE becomes similar to a simple linear regression or PCA. But a nonlinear activation function, such as a rectified linear unit (ReLU) or a sigmoid function, makes the AE different from the PCA. The multiple types of AEs can be combined or modified to obtain new models for various applications. Among their widely used types, we can mention the types of variational AE (VAE), denoising AE (DAE) and sparse AE (SAE). VAE is enhanced with variational inference and parameterization to increase the model's ability in feature extraction and retain the diversity of the generated data. DAE takes a noisy input while training to recover the original undistorted input. By this means, the encoder can extract the most essential features and learn a robust representation of the input data. In SAE a sparsity constraint is imposed on the hidden nodes to mine essential information and avoid redundancy in large-scale datasets.

In this paper, we use a hybrid model based on VAE to enhance the model's ability in feature extraction while preserving the diversity of the generated data. The experimental results indicate that even with a 3 layered VAE, the model outperforms both IB and PCA.

C. Classifier Techniques

In the following, we review some common classifier methods that have high performance results, including MLP, SVM and Logistic Regression (LR) algorithms.

I) Multi-Layer Perceptron (MLP)

MLP is an artificial neural network that consists of an input layer, an output layer and multiple hidden layers instead of a single hidden layer. It is a feedforward network, meaning that each layer feeds the subsequent layer through a series of weights. MLP uses the backpropagation technique which is a supervised learning method. It has the capability to learn nonlinear models. Its multiple hidden layers and nonlinear activation function differentiate it from a linear perceptron. For applying MLP, several hyperparameters such as the number of hidden neurons, layers, and iterations must be adjusted.

II) Support Vector Machine (SVM)

SVM is a supervised machine learning technique used for classification regression and outliers detection. In SVM, a hyper-plane is created for separating different types of data. One of the advantages of SVM is that its training is computationally simple and unlike neural networks, it does not suffer from the problem of a local minimum. To accurately control the error rate, the kernel function and C parameter should be chosen correctly.

III) Logistic Regression

Logistic regression is a statistical method to classify an observation into one of two classes, or into one of many classes. It models the relationship between the independent features and the binary dependent variable (target) using the logistic function.

Architecture of the Model

Our proposed model combines feature reduction and data balancing. First, the initial data is preprocess to normalize. Then, the prepared data is employed in training the VAE, (t-SNE, UMAP, IB or PCA). After training, VAE (t-SNE, UMAP, IB or PCA) can differentiate between classes in the latent space, utilize BSM for interpolation of latent variables and synthesis new data. Finally, combination of the original and synthesized data is used for classification. The general architecture of the model is demonstrated in Fig. 1.

Since, VAE is a generative model, it can provide better performance and contribute for generating new samples using the latent variables. Therefore, we input the latent variables to the decoder to synthesis data. Subsequently, the decoder is eliminated and the encoder output is connected to the classifier and a combined network is created. The original data along with synthesized data are used to train the network. Algorithm 1 shows the pseudo code of the model based on VAE and BSM.

Algorithm 1. The proposed model based on VAE and BSM.

Input: Training data $T = \{t_1, t_2, \dots, t_n\}$;
Test data $S = \{s_1, s_2, \dots, s_m\}$

Output: predicated labels: $P = \{p_1, p_2, \dots, p_m\}$

1. Initialize VAE network;
2. Data preprocessing: remove missing data and normalize features to $[0,1]$. Output T' and S';
3. Feed T' to VAE's encoder and output $Z = \{z_1, z_2, \dots, z_n\}$.
4. Run BSM using input Z and output Z'.
5. Feed Z' to VAE's decoder and output new samples T_{new};
6. $T' = T' \; U \; T_{new}$;
7. Train the combined classification network by T'
8. For each $s_i, i = 1,2, \dots, m$:
   Return the prediction p_i;

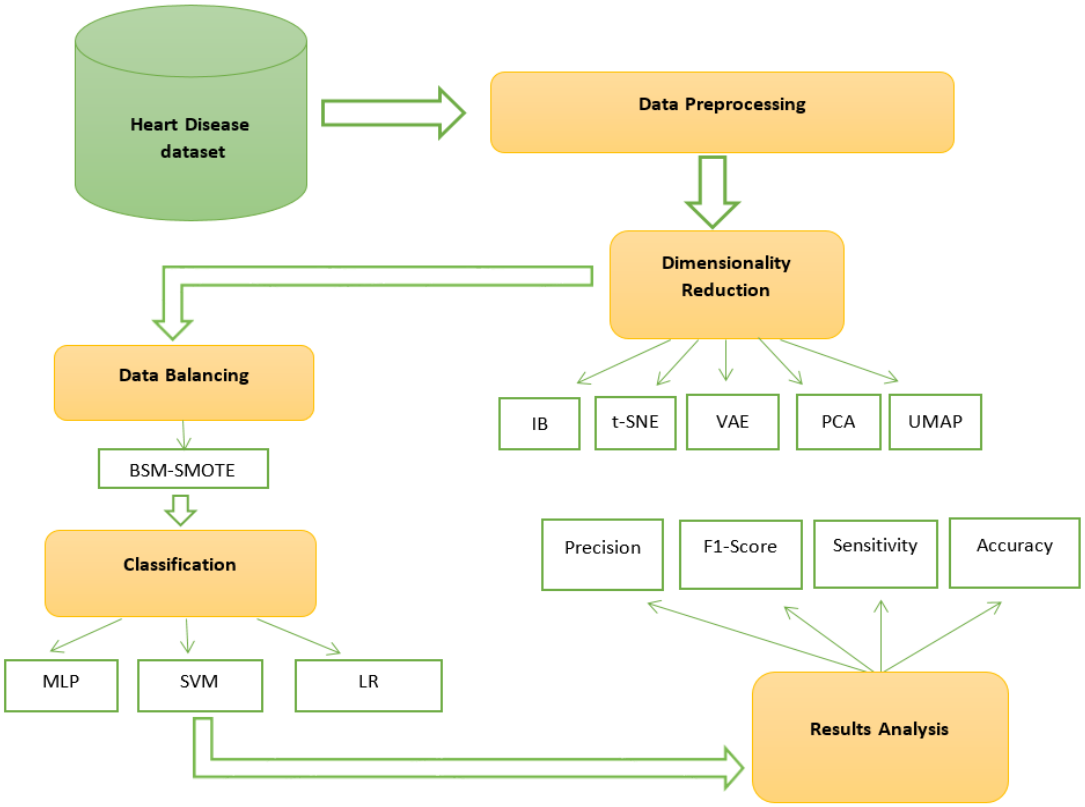

Fig. 1: Diagram of the proposed model.

In detail, the proposed model has mainly four phases:

1. Data preprocessing phase: this phase includes handling missing values, normalizing and shuffling data.

The dataset contains missing values that are handled by K-Nearest Neighbors imputation technique [22]. In the splitting data stage, we utilized 70% − 30% train-test data partitioning approach. The dimensionality reduction and data balancing steps are applied only to the training data.

2. Dimensionality reduction: In this phase, several dimensionality reduction methods, including PCA, IB, t-SNE, UMAP and VAE are applied to incorporate correlation between features. These dimensionality reduction methods are selected according to their performance and efficiency on heart disease prediction problem and the type of dataset.

3. Data balancing: Of all instances, 3596 are negative and 644 are positive. The lack of negative instances leads to low accuracy in predicting these cases (i.e., a high number of false negatives). Thus, we use an oversampling technique to generate samples of the minority class. However, most of these methods have limitations, as it can generate noise and redundant samples. Therefore, we use BSM which generates samples at the boundary of the sample distribution and avoids the synthesis of noise and redundant samples.

After balancing data, the performance of the models is improved. Since in most oversampling methods, the classification type data is discarded or examined separately, we balance the data after the dimensionality reduction.

4. Classification: At the end of the prediction process, the combined data is used as input to classification models. Among the classification methods, we selected three widely used categories that had higher performance than the others, MLP, SVM and LR.

5. Evaluation of methods: The performance of the models is evaluated and compared by the evaluation measures: accuracy, sensitivity, precision, F1-score. These measures are defined as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (5)$$

$$F1 = \frac{2*precision*sensitivity}{precision+sensitivity} \quad (6)$$

TN and TP denote true negative and true positive, i.e., they represent the number of patients and normal instances that are classified correctly. While FP and FN denote false positive and false negative, i.e., they represent the number of patients and normal instances that are incorrectly predicted.

Experimental Results

In this section, we use two publicly available datasets,

Kaggle heart dataset [23] and UCI heart disease repository [24]. First dataset consists of 4238 samples and 16 features. The second dataset consists of 597 samples and 13 features. Every feature is a risk factor that may be behavioral, demographic or medical. The features include age, knee joint pain status, waist circumference, neutral fat, BMI, SBP, TC, obesity status, frequency of eating out, HDL, weight change in one-year status, and marital status.

The heart disease database includes 15 features as input and its output is classified into two groups patient and normal. Since there is no general rule to adjust the parameters such as the number of hidden layers and the number of neurons in various layers, it is vital to obtain a good network structure with optimal performance.

D. Experiment Setup

We implemented the proposed model and determined the values of the parameters which provide optimal performance as follows:

VAE: The number of layers and neurons in the VAE is chosen based on a grid search over batch size (20, 30, 40), epochs (25, 50, 100, 200), neural network depth (2, 3, 4) and the dimensionality of the first hidden layer (40, 30, 20,12). we use Tanh and ReLU as activation functions and consider reduction size 4, 6, 8, 10 and 8. Also, we assume that the learning rate is 0.01, and ‘‘Adam’’ is used as the gradient descent algorithm. For example, we find the best architectures with two hidden layers listed in Table 1.

Table 1: The network structure parameters for VAEs with reduction sizes 4, 6, 8, 10

Reduction size	Architecture
4	15-30-10-4-10-40-15
6	15-20-8-6-8-20-15
8	15-20-10-8-10-20-15
10	15-20-12-10-12-20-15

IB: we consider the size of IB and Lagrange multiplier (θ) are 8 (size of reduction) and 0.95, respectively. Our implementation of IB is based on the Neural Network model for nonlinear information bottleneck [25].

PCA: it is enough to determine the optimal size to reduce the dimension.

t-SNE: We consider number of iterations and the value of α are 1000, respectively. Also, the perplexity is set to 30 to determines the number of nearest neighbors considered.

UMAP: We consider number of neighbors is 6 to balances local versus global structure in the data. Also, minimum distance is set to 0.3 to controls the minimum distance between points in the low-dimensional embedding.

MLP network: we apply five hidden layers with sizes 24, 30, 20, 15, 10, respectively. Also, the activation function and solver are selected Tanh and Adam, respectively.

However, in this method, adjusting the parameters is difficult and requires trial and error.

SVM: we use RBF as kernel function and consider parameters C = 1, gamma = 100000. With these selections, good results have been achieved.

LR algorithm: we apply it with the training parameter ridge estimator.

E. Performance Evaluation

In the following, we show performance measures with respect to the possible dimensionality reduction methods, data balancing, classification algorithms and the size of reduction. Observing the results of the methods with various reduction sizes, it was found that reduction size 8 has best performance measures values. Therefore, we have shown the performance parameters results of algorithms for reduction size 8.

Table 2-4 show the results of the performance measures obtained from models based on various dimensionality reduction methods, PCA, IB, t-SNE, UMAP and VAE, before and after data balancing for each UCI dataset and Kaggle dataset. We have shown the names of different methods as a combination of the dimension reduction method and classifier, so the possible methods will be the combination of PCA, IB, t-SNE, UMAP and VAE with MLP, PCA and LR.

Table 2 demonstrates the results using the MLP network. It can be seen, applying dimensionality reduction methods on MLP not only does not increase the performance, but also has a negative effect on it. However, data balancing improved performance significantly and the t-SNE-based method is better than other dimensional reduction algorithms. Table 3 indicates that the VAE-based method improves performance metrics of SVM algorithm, with an accuracy of 81.3, while the other dimensionality reduction algorithms, PCA and IB, have a negative effect on the performance. The BSM data balancing method, due to the synthesis of minority class samples in the boundary, reduces the value of FN and significantly strengthens the performance parameters.

As can be seen in Table 4, the effect of dimensionality reduction on LR method is similar to SVM, and VAE-based method improves the performance metrics. Also, applying data balancing on VAE-LR increases performance measures such as accuracy to 91.7 and sensitivity to 97.3. Comparing all the results in Table 2-4, shows that the best performance is achieved with the VAE-SVM method after data balancing. It improves accuracy to 97.1% and sensitivity to 99.2%.

Table 2: Performance comparison between methods based on different dimension reduction techniques before and afterdata balancing while we use MLP classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-MLP	77.1	49.06	46.2	58.9	52.3
		IB-MLP	76.0	48.33	45.3	57.6	51.8
		tSNE-MLP	84.1	24.5	33.1	50.4	32.4
		UMAP-MLP	83.1	21.8	28.1	48.2	30.7
		VAE-MLP	79.0	50.18	47.1	60.4	53.7
		MLP	86.0	70.26	80.91	68.38	62.08
	Balanced data	PCA-MLP	88.5	89.29	86.3	83.4	92.5
		IB-MLP	86.6	82.07	77.2	74.1	87.6
		tSNE-MLP	95.5	95.1	95.1	95.5	95.1
		UMAP-MLP	89.1	88.3	95.4	89.5	79.1
		VAE-MLP	89.2	88.7	94.6	89.0	92.8
UCI	Original data	PCA-MLP	73.2	47.9	45.8	58.2	51.8
		IB-MLP	72.1	46.5	44.6	54.6	50.7
		tSNE-MLP	83.2	23.5	32.7	49.6	31.2
		UMAP-MLP	82.6	21.1	27.6	48.1	30.2
		VAE-MLP	77.3	48.2	47.3	58.3	52.2
		MLP	82.7	69.3	78.5	67.9	61.0
	Balanced data	PCA-MLP	87.8	89.1	84.6	85.3	90.5
		IB-MLP	86.2	81.5	76.7	75.3	86.5
		tSNE-MLP	944.4	93.9	93.8	93.9	94.5
		UMAP-MLP	88.6	87.5	93.1	88.5	78.0
		VAE-MLP	87.9	86.5	86.8	86.9	91.1

Table 3: Performance comparison between methods based on different dimension reduction techniques before and after data balancing while we use SVM classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-MLP	77.1	49.06	46.2	58.9	52.3
		IB-MLP	76.0	48.33	45.3	57.6	51.8
		tSNE-MLP	84.1	24.5	33.1	50.4	32.4
		UMAP-MLP	83.1	21.8	28.1	48.2	30.7
		VAE-MLP	79.0	50.18	47.1	60.4	53.7
		MLP	86.0	70.26	80.91	68.38	62.08
	Balanced data	PCA-MLP	88.5	89.29	86.3	83.4	92.5
		IB-MLP	86.6	82.07	77.2	74.1	87.6
		tSNE-MLP	95.5	95.1	95.1	95.5	95.1
		UMAP-MLP	89.1	88.3	95.4	89.5	79.1
		VAE-MLP	89.2	88.7	94.6	89.0	92.8
UCI	Original data	PCA-MLP	73.2	47.9	45.8	58.2	51.8
		IB-MLP	72.1	46.5	44.6	54.6	50.7
		tSNE-MLP	83.2	23.5	32.7	49.6	31.2
		UMAP-MLP	82.6	21.1	27.6	48.1	30.2
		VAE-MLP	77.3	48.2	47.3	58.3	52.2
		MLP	82.7	69.3	78.5	67.9	61.0
	Balanced data	PCA-MLP	87.8	89.1	84.6	85.3	90.5
		IB-MLP	86.2	81.5	76.7	75.3	86.5
		tSNE-MLP	944.4	93.9	93.8	93.9	94.5
		UMAP-MLP	88.6	87.5	93.1	88.5	78.0
		VAE-MLP	87.9	86.5	86.8	86.9	91.1

Table 4: Performance comparison between methods based on different dimension reduction techniques before and after data balancing while we use LR classification and reduction sizes 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively

Dataset		Methods	Accuracy (%)	F1-score (%)	Precision (%)	ROC-AUC score (%)	Sensitivity (%)
Kaggle	Original data	PCA-LR	77.9	47.69	47.1	48.0-	48.3
		IB-LR	76.3	45.75	45.7	46.8	45.8
		tSNE-LR	68.8	78.3	73.2	50.0	69.9
		UMAP-LR	71.8	80.1	74.8	50.5	71.8
		VAE-LR	79.6	54.6	55.0	54.1	54.2
		LR	78.5	55.98	61.6	55.3	53.0
	Balanced data	PCA-LR	83.2	84.65	99.8	78.5	73.5
		IB-LR	81.6	87.65	82.8	77.6	93.1
		tSNE-LR	85.4	84.7	83.2	73.9	85.1
		UMAP-LR	86.4	85.1	84.6	75.0	86.4
		VAE-LR	**91.7**	**93.07**	**89.2**	**89.4**	**97.3**
UCI	Original data	PCA-LR	75.8	46.2	45.	46.2	46.2
		IB-LR	74.5	44.2	43.9	44.8	44.1
		tSNE-LR	71.3	80.0	73.9	50.0	70.9
		UMAP-LR	79.2	54.1	54.0	53.9	54.1
		VAE-LR	79.2	53.8	54.2	53.6	53.0
		LR	77.1	53.5	60.2	53.5	51.7
	Balanced data	PCA-LR	81.2	83.1	99.1	77.3	71.8
		IB-LR	80.8	86.5	81.3	76.2	91.8
		tSNE-LR	85.1	83.6	82.9	73.3	84.5
		UMAP-LR	85.8	84.7	84.5	74.2	86.1
		VAE-LR	90.5	92.1	88.4	88.3	96.0

Also, it can be seen, in all experiments, performance parameters were enhanced after data balancing. The reason is that, besides increasing the samples of the minority class, the applied data balancing algorithm (BSM) does not consider the noise data. Furthermore, applying the model to both datasets has the similar effect on performance. But since the number of samples of Kegel dataset is more than other dataset, the results are more reliable.

Fig. 2 shows comparison of accuracy values between various sizes of dimension reduction while VAE algorithm is used. It can be found that the accuracy of all methods is greatly enhanced after increasing the reduction size to 4 and it is maximized in 8.

F. Time Complexity

In this section, we provide experiments for computational efficiency. For each combination of the dimensionality reduction methods with the mentioned ML algorithms, the computational time of the network training is calculated. We use a computer with this specification: Intel Core i7 7700HQ, 2.60GHz, and 8GB RAM and also, we utilize Python 3.9 as the programming language. We also emphasized that the execution conditions are the same for all methods.

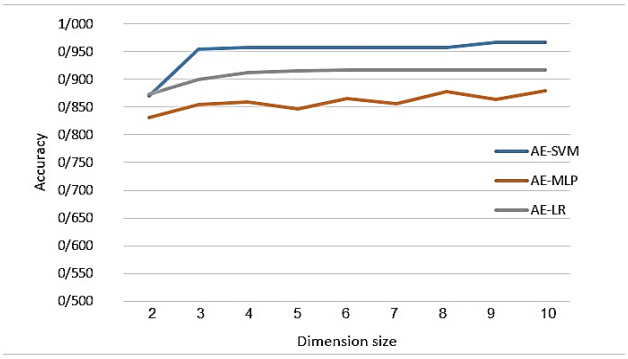

Fig. 2: Comparison of accuracy measure between various sizes of dimension reduction using VAE-based method for Kaggle dataset.

As can be seen in Fig. 4, the computational time for the LR in all cases is lower than the other classification methods. In addition, reducing dimension by using VAE usually improves computational time of the training while also increasing the performance.

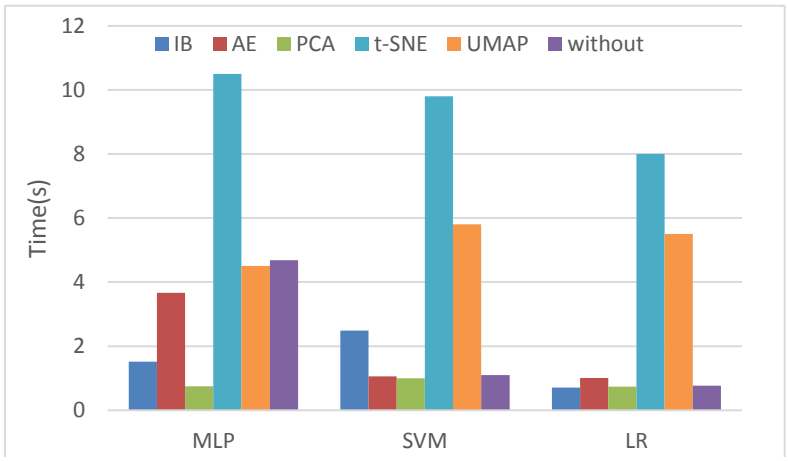

Fig. 4: Evaluation of time efficiency in the proposed model using various classifiers for Kaggle dataset while reduction sizes are 8, 8, 8, 6, 3 for PCA, IB, VAE, UMAP and t-SNE, respectively.

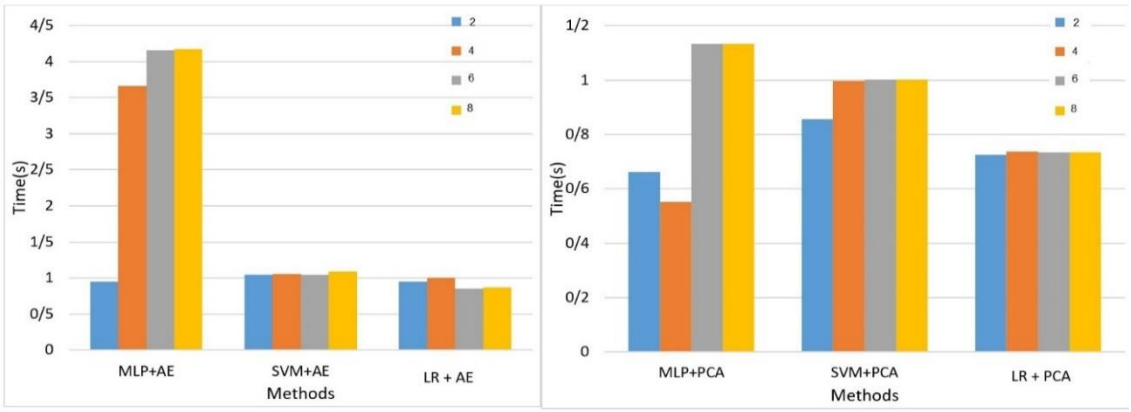

Fig. 3: The impact of reduction size on computational time using VAE (left) and PCA (right) for Kaggle dataset.

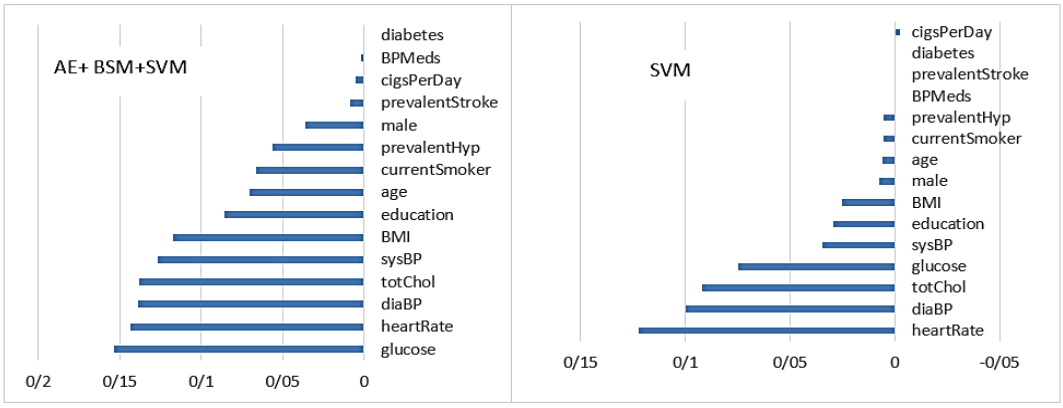

Fig. 5: Drop-column importance analysis on the proposed model for Kaggle dataset.

For large training sets, it was found that the difference in processing time is considerable. Also, t-SNE greatly increases the computational time.

Fig. 3 shows the impact of reduction size on computational time and indicates that in general, reducing the dimension decreases computational time, especially when the dataset is large.

I) Interpretability

Interpretability is a crucial aspect when it comes machine learning algorithms, especially deep learning algorithms, as they often produce models that are difficult to understand. These models, commonly known as black-box models, offer improved performance at the expense of complexity, making it challenging to comprehend the underlying mechanisms. Without it, even if accuracy is enhanced, the lack of transparency and accountability in the model may not be acceptable in medical settings. Research on interpretability has evolved

significantly due to the intricate nature of deep learning models, with various methods being employed to shed light on how these models operate. These methods include estimating feature importance, analyzing feature interactions, determining the contribution of specific layers or neurons, and interpreting models using high-level concepts that are more understandable to humans than low-level input features. We employed drop-column importance values to interpret the importance of the features in the proposed model, which has provided essential insights into the underlying mechanisms of disease prediction.

This information has the potential to assist clinicians in developing personalized treatment plans and risk management strategies for patients, ultimately leading to improved clinical outcomes. Fig. 5 shows the visualization results through drop-column importance on the VAE-based method and SVM method. The results indicate key features such as glucose, heartRate, diaBP, totChol and sysBP possess the highest importance value in VAE-based method.

It indicates that these variables play a crucial role in predicting heart disease. Similarly, variables heartRate, diaBP, totChol, glucose and sysBP have high importance value in SVM method. Therefore, the feature importance analysis discovered a consistent set of top 6 features, namely, glucose, heartrate, diaBP, totChol and sysBP which were very important in the prediction process. The results of the test indicate that there is no notable variance among the algorithms tested, as the dataset is limited and the supervised algorithms used are effective in yielding similar results.

Comparison

Now, we performed comparison study of our proposed model with the results from previous studies. The comparison results of the proposed model compared to the results given in other similar studies on Kaggle dataset is shown in Table 5. It can be seen that the proposed model demonstrated the high accuracy compared to previous studies results. In conclusion, our method outperformed most of these studies in accuracy, sensitivity, precision, F1-score, and AUC.

Table 5: Comparison the proposed model with other methods in recent studies

Authors	Approach	Accuracy	Precision	Sensitivity	F1-score	AUC
Saqlain et al. [26]	MFSFSA SVM	81.19	-	72.92	0.85	0.83
Mohan et al. [27]	HRFLM	88.4	90.1	92.8	90.0	-
Gupta et al. [28]	FAMD –RF	93.44	-	89.28	92.59	0.93
Fitriyani et al. [29]	DBSCAN-SMOTEE- XGBOOST	98.40	98.57	98.33	98.32	1.00
Bharti et al. [30]	DL-based Classifier	94.2	93.1	82.3	-	-
Hossain et al. [31]	Hybrid CNN-LSTM	74.15	81.82	72.04	76.62	73.95
Manikandan et al. [32]	Boruta feature selection	88.52	87.88	90.62	89.23	-
Proposed model	VAE- BSM - SVM	97.7	95.8	99.4	97.5	96.3

It is important to highlight that a direct comparison of the results may not be accurate due to application of different data pre-processing and training/testing methods. Moreover, the effectiveness of the prediction model is influenced by various factors including feature selection, data types and size, noise reduction, hyperparameters, data sampling, and model selection. Therefore, the overall comparison provided in Table 5 should not be solely relied upon to assess the performance of the prediction models. Instead, it can serve as a general comparison between the proposed model and previous research studies.

Conclusion and Future Work

Dimensionality reduction is a feature selection method that usually increases performance measures and computational speed of training. In this paper, we investigate the impact of some dimensionality reduction methods, including PCA, IB and VAE, on several machine learning algorithms in terms of performance measures and computational time. After implementing the model and reviewing the obtained results, we found that deep learning methods such as VAEs enhance the efficiency and the performance of the system. However, the effect of applying feature reduction on performance is negligible in some models. In addition, applying dimensionality reduction sometimes improves speed up to five times and sometimes does not affect. In an effort to better balance our training data, we use BSM data augmentation method. Finally, the hybrid model based on VAE and SVM achieves accuracy and sensitivity of 97.7% and 99.4% using Kaggle dataset.

In future works, the performance of the method can be enhanced to handle huge numbers of features and large volume of records. Additionally, the increasing emphasis on privacy, security, and time-sensitive applications shows the need to explore deeper into edge computing in order to enhance medical clinical decision support system.

Author Contributions

Vahidreza Afshin simulated the proposed method in Python. Seyed Hamid Zahiri and Saiedeh Kabirirad supervised and consulted in the design, implementation and results of this research. All authors discussed important sections and contributed to the final text.

Acknowledgment

We sincerely thank the respected referees for their accurate review of this paper.

Also, we sincerely thank the ICT Research Institute and Connectivity and Communication Technologies Development Headquarters and Dr. Kharrat.

Conflict of Interest

The authors announce no potential conflict of interest regarding the publication of this paper. Also, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission and redundancy have been completely witnessed by the authors.

Abbreviations

PCA	Principal Component Analysis
IB	Information Bottleneck
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder
SMOTE	The Synthetic Minority Oversampling Technique
BSM	the Borderline Synthetic Minority Oversampling Technique
LR	Logistic regression
DNN	Deep Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
RF	Random Forest
DT	Decision Tree
AUC	Area Under the ROC Curve
CM	Confusion Matrix
ROC	Receiver Operating Characteristic
DCNN	Deep CNN
MLP	Multi-Layer Perceptron
AE	Autoencoder
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative

References

[1] U. Kose et al., "A practical method for early diagnosis of heart diseases via deep neural network," in Deep Learning for Medical Decision Support Systems, pp. 95-106, 2021.

[2] A. A. Ali, H. S. Hassan, E. M. Anwar, A. Khanna, "Hybrid technique for heart diseases diagnosis based on convolution neural network and long short-term memory," in Applications of Big Data in Healthcare: Elsevier, pp. 261-280, 2021.

[3] D. Pratella, S. Ait-El-Mkadem Saadi, S. Bannwarth, V. Paquis-Fluckinger, S. Bottini, "A survey of autoencoder algorithms to pave the diagnosis of rare diseases," Int. J. Mol. Sci., 22(19): 10891, 2021.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., 16: 321-357, 2002.

[5] H. Han, W. Y. Wang, B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Proc. International Conference on Intelligent Computing: 878-887, 2005.

[6] C. M. Bhatt, P. Patel, T. Ghetia, P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," Algorithms, 16(2): 88, 2023.

[7] A. Khan, M. Qureshi, M. Daniyal, K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," Health Social Care Community, 2023(1): 1406060, 2023.

[8] D. Hassan, H. I. Hussein, M. M. Hassan, "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis," Biomed. Signal Process. Control, 79: 104019, 2023.

[9] S. Kabirirada, H. Kardanmoghaddamb, V. Afshin, "Heart disease prediction by using artificial neural networks," Int. J. Comput. Sci. Inf. Secur., 14(1), 2016.

[10] A. Ahmed, S. A. Hannan, "Data mining techniques to find out heart diseases: an overview," Int. J. Innovative Technol. Exploring Eng. (IJITEE), 1(4): 18-23, 2012.

[11] S. I. Ayon, M. M. Islam, M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," IETE J. Res., 68(4): 2488-2507, 2022.

[12] K. Li, A. Zhu, W. Zhou, P. Zhao, J. Song, J. Liu, "Utilizing deep learning to optimize software development processes," arXiv preprint arXiv:2404.13630, 2024.

[13] S. Shilaskar , A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert Syst. Appl., 40(10): 4146-4153, 2013.

[14] Y. E. Shao, C. D. Hou, C. C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," Appl. Soft Comput., 14: 47-52, 2014.

[15] R. R. Ema, P. C. Shill, "Integration of fuzzy C-Means and artificial neural network with principle component analysis for heart disease prediction," in Proc. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT): 1-6, 2020.

[16] I. D. Mienye, Y. Sun, Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," Inf. Med. Unlocked, 18: 100307, 2020.

[17] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, Y. Piao, K. H. Ryu, "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets," IEEE Access, 9: 135210-135223, 2021.

[18] S. A. Ebiaredoh-Mienye, E. Esenogho, T. G. Swart, "Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis," Electronics, 9(11): 1963, 2020.

[19] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, S. A. Hussain, "Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis," Physica A, 482: 796-807, 2017.

[20] T. Zhang, B. Yang, "Big data dimension reduction using PCA," in Proc. 2016 IEEE International Conference on Smart Cloud (SmartCloud): 152-157, 2016.

[21] F. Anowar, S. Sadaoui, B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," Comput. Sci. Rev., 40: 100378, 2021.

[22] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," J. Syst. Software, 85(11): 2541-2552, 2012.

[23] "Kaggle Cardiovascular Disease Dataset," Accessed 1 November 2022.

[24] "UCI Machine Learning Repository. Uci.edu.," Accessed 14 June 2022.

[25] A. Kolchinsky, B. D. Tracey, D. H. Wolpert, "Nonlinear information bottleneck," Entropy, 21(12): 1181, 2019.

[26] S. M. Saqlain et al., "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," Knowl. Inf. Syst., 58: 139-167, 2019.

[27] S. Mohan, C. Thirumalai, G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, 7: 81542-81554, 2019.

[28] A. CS, S. Lal, V. PRABHU GURUPUR, P. P. Saxena, "Multi-modal medical image fusion with adaptive weighted combination of NSST bands using chaotic grey wolf optimization," IEEE Access, 7: 40782-40796, 2019.

[29] N. L. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee, "HDPM: an effective heart disease prediction model for a clinical decision support system," IEEE Access, 8: 133034-133050, 2020.

[30] 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," Comput. Intell. Neurosci., 2021(1): 8387680, 2021.

[31] M. M. Hossain et al., "Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI," Inf. Med. Unlocked, 42: 101370, 2023.

[32] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. Karthikeyan, R. Manikandan, A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," Inf. Med. Unlocked, 44: 101442, 2024.

Biographies

Saiedeh Kabirirad is an Assistant Professor in the Department of Computer Science, at Birjand University of Technology. She received her M.Sc. and Ph.D. in Computer Science from Shahid Beheshti University in 2010 and 2019, respectively. Her research area includes cryptography, information security, and data mining.

- Email: kabiri@birjandut.ac.ir
- ORCID: 0000-0003-1503-138X
- Web of Science Researcher ID: NA
- Scopus Author ID: 57204363739
- Homepage: NA

Vahidreza Afshin received the B.S. and M.S. degrees in Electrical Engineering in 2005 and 2012 from Islamic Azad University respectively. He is currently pursuing the Ph.D. degree in the Department of Electronics Engineering, University of Birjand, Iran. His research interest includes soft computing and optimization of control systems by machine learning algorithms.

- Email: vahidreza.afshin@birjand.ac.ir
- ORCID: 0000-0003-2621-515X
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

Seyed Hamid Zahiri received the B.Sc., M.Sc. and Ph.D. degrees in Electronics Engineering from Sharif University of Technology, Tehran, Tarbiat Modarres University, Tehran, and Mashhad Ferdowsi University, Mashhad, Iran, in 1993, 1995, and 2005, respectively. Currently, he is a Professor with the Department of Electronics Engineering, University of Birjand, Birjand, Iran. His research interests include pattern recognition, evolutionary algorithms, swarm intelligence algorithms, and soft computing.

- Email: hzahiri@birjand.ac.ir
- ORCID: 0000-0002-1280-8133
- Web of Science Researcher ID: NA
- Scopus Author ID:NA
- Homepage: NA

How to cite this paper:
S. Kabirirad, V. Afshin, S. H. Zahiri, “An effective heart disease prediction model using deep learning-based dimensionality reduction on imbalanced data,” J. Electr. Comput. Eng. Innovations, 13(2): 317-330, 2025.

DOI: 10.22061/jecei.2024.10847.742

URL: https://jecei.sru.ac.ir/article_2208.html