Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 8s, https://carcinogenesis.com/

Optimal Prediction of Cardiovascular Disease Using Stochastic Layer-wise Autoencoder

Irfan Khan[1], Pinaki Ghosh [2]

[1]School of Engineering and Technology, Sanjeev Agrawal Global Educational University, Bhopal, India

Email ID : irfank92@gmail.com , ORCHID : 0009-0004-2704-8853

[2] School of Advanced Computing, Sanjeev Agrawal Global Educational University, Bhopal, India

Email ID : pinaki.g@sageuniversity.edu.in , ORCHID: 0000-0002-4598-9508

ABSTRACT

Today, the world's leading cause of death is cardiovascular disease. Based on medical records, the machine learning approach has been used in some studies to predict cardiovascular diseases. However, there is still a need for an optimal approach because medical records have a high correlation with data. Sometimes, data with high dimensions leads to inefficiency. An effective dimensionality reduction technique is necessary to reduce noise and high-dimensional data. This paper proposed a novel stochastic layer-wise sparse autoencoder model with a deep feed-forward neural network (SLSAE-DFFN). Unlike traditional sparse autoencoders, which use a fixed sparsity across the entire network, our method introduces probabilistic layer-wise sparsity. This approach reduces overfitting and improves generalization with diverse latent feature representations. After training, lower-dimensional features from the bottleneck layer of the autoencoder are used to train a deep feedforward neural network for classification. This proposed model shows great promise in predicting cardiovascular disease, achieving an impressive 97% accuracy. The SLSAE-DFFN model is carefully evaluated using various performance metrics, including accuracy, precision, recall, and the F1-score.

Keywords: *Machine Learning, Artificial Neural Networks, Cardiovascular Disease, Autoencoder*

How to Cite: Irfan Khan, Pinaki Ghosh, (2025) Optimal Prediction of Cardiovascular Disease Using Stochastic Layer-wise Autoencoder, *Journal of Carcinogenesis*, *Vol.24, No.8s*, 317-326

1. INTRODUCTION

Cardiovascular diseases are the main reason for numerous deaths that are taking place globally. Developing nations experience a high mortality level because of the delay in disease detection. Early detection can reduce the severity of a heart attack. Therefore, further studies on cardiovascular disease are needed [1]. The machine learning approach involves creating prediction models based on historical data to forecast outcomes for new data. The accuracy of these models improves with more data, as larger datasets lead to more precise predictions. The process hinges on two essential steps: training and testing. During training, the model learns from data by applying various ML algorithms, such as logistic regression, decision trees, naive Bayes, random forests, and support vector machines. However, traditional machine learning methods are not capable of dealing with data complexity and nonlinearity. However, this requires a significant amount of human effort, which is costly and relies on expertise [2]. That means deep learning can address these issues better than machine learning.

Deep learning is primarily used in supervised learning for various types of data. It is also used in unsupervised learning, such as feature learning. In feature learning, the deep learning methods are designed to extract the underlying latent variables. Feature learning reduces the need for manual feature engineering, enabling machines to learn features by examining the features themselves [3].

Machine learning models face two significant challenges: first, dealing with higher dimensions can lead to inefficiency and negatively impact the model's performance, and second, limited generalisation when trained on a small dataset. An effective technique is needed to reduce high dimensions and improve the generalization. Feature extraction using autoencoders is a method for addressing this issue. Feature extraction reduces data complexity and identifies relevant features [4]. Autoencoders learn compact, latent representations that retain essential information while discarding noise. Traditional sparse autoencoders encourage sparsity by penalizing activations using fixed L1 regularization or KL divergence. While effective, these methods apply uniform sparsity constraints across layers, which may not optimally capture hierarchical feature representations [5].

To address this limitation, we propose a novel stochastic layer-wise sparse autoencoder (SLSAE). This method applies probabilistic sparsity constraints on each layer of the autoencoder. Each layer has its own Bernoulli-distributed sparsity probability, which stochastically decides which activations are penalized during training. This prevents the network from converging to a single, overly sparse representation and encourages robust,

generalizable features. The latent features are then passed to a deep feedforward neural network (DFFN) for classification. We have also used the cardiovascular disease dataset, an open-source dataset found on Kaggle. This dataset comprises the Cleveland, Hungary, Switzerland, and Long Beach datasets [6].

2. RELATED WORK:

Many authors have published their work in this domain. In recent decades, various machine learning and deep learning techniques have been used for predicting cardiovascular disease. Machine learning techniques utilize historical data or medical records. In one study, Shah et al. [7] showed that KNN outperforms other methods, including naive bayes, random forests, and decision trees. Ali et al. [8] proposed that decision trees, random forests, and KNN algorithms were more effective than logistic regression, AdaboostM1, and MLP algorithms. It is observed that these studies report high accuracy on the essential features of the UCI dataset, which has a relatively small number of instances. Many researchers have also proposed many effective hybrid models of machine learning techniques using different datasets. In their model, El-Shafiey, Mohamed G., et al. [9] used the University of California, Cleveland, and Statlog datasets. They suggested random forest, genetic algorithm, and particle swarm optimization (GAPSO). R. Rajendran et al. [10] used ensemble algorithms, naive bayes, and logistic regression for classification and entropy-based FE for preprocessing to remove the outliers. In comparison to several machine learning algorithms, this model performed better. Chi-square statistical tests were suggested by K. Karthick et al. [11] to select specific characteristics from the heart disease dataset using traditional methods.

Additionally, many authors contributed novel techniques for feature extraction. Motiur Rahman et al. [13] proposed a convolutional LSTM model for predicting diabetic disease. Using an LSTM-based neural network, they have successfully identified diabetic patients by utilizing features extracted from a time series dataset. Anna Karen Garate et al. [12] employed random forest for classification and CHI-PCA (chi-square and principal component analysis) for feature extraction. Awais Mehmood et al. [14] introduced a traditional neural network for predicting cardiovascular disease. This model extracts hidden features from ECG signals and classifies them as normal or abnormal with higher accuracy.

Some researchers included autoencoders in their research work to perform both linear and non-linear projections on datasets over principal component analysis (PCA). Byeon, Haewon, et al. [15] developed an ABL model that combines broad learning with noise reduction of denoising autoencoders. They developed feature extraction techniques for medical environments and achieved high accuracy. Abdellatif, Abdallah, et al. [16] introduced a heart disease model called SPFHD, which uses tree-based ensemble learning algorithms and a support vector machine algorithm to enhance detection accuracy. A new conditional variational autoencoder method and Bayesian optimization are used. García-Ordás, María Teresa, et al. [17] proposed an architecture called a sparse autoencoder. In their work, the latent space is fed to classifiers. They evaluated two classifiers, a traditional multilayer perceptron and a convolutional neural network. According to the findings, the autoencoder was employed to identify latent space features in the neural network and classify them, resulting in low variance and improved accuracy.

3. RESEARCH METHODOLOGY

In this study, we have followed some steps to forecast the performance of cardiovascular disease. Initially, datasets were collected from Kaggle and then preprocessed using feature scaling. Afterwards, the data was split into training and testing datasets. An autoencoder will train datasets to learn essential features. An autoencoder contains two encoder and decoder components. The encoder uses latent space to represent the input data after compressing it to a reduced dimension. The decoder uses the encoder's output to try to reconstruct the input. After training, we extracted the encoded output from the bottleneck layer of the autoencoder, which was then used in supervised learning techniques to create a predictive model. This phase produces a model that will be used for testing or evaluation. Figure 1 illustrates the steps involved in predicting cardiovascular disease.

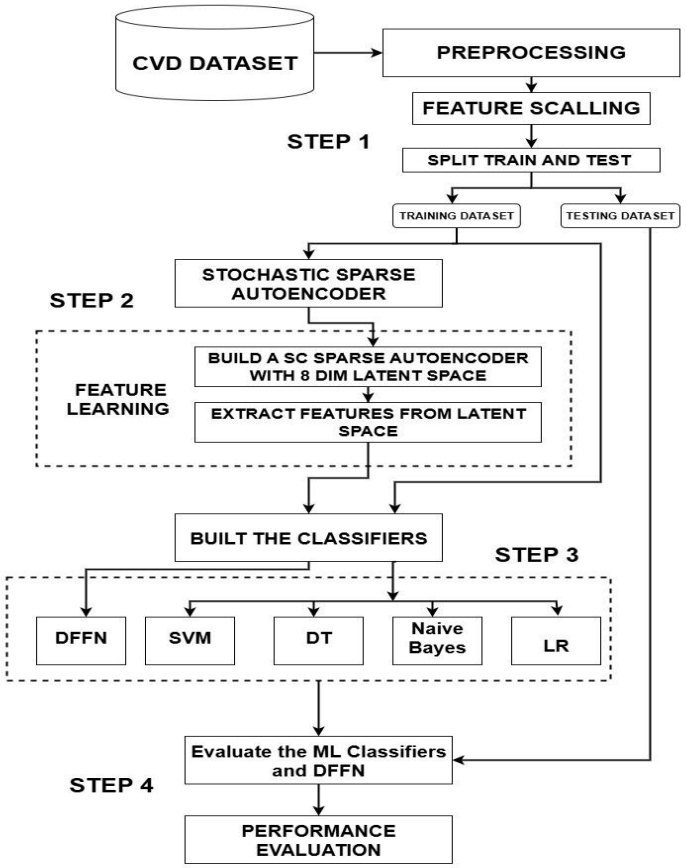

Figure 1: Steps of the proposed model methodology

3.1 Dataset

In this study, we utilized a dataset downloaded from Kaggle. The dataset comprises 1,025 records of cardiovascular patient data, featuring 14 attributes, including one target feature as the dependent variable and 13 independent variables. This dataset consists of 4 different datasets: Cleveland, Hungary, Switzerland, and Long Beach V. The dataset's characteristics include five numerical features—age, maximum heart rate, serum cholesterol, ST depression, and resting blood pressure—as well as four binary features—sex, fasting blood sugar, exercise-related angina, and the target variable of cardiovascular disease—as well as five categorical features—chest pain type, resting electrocardiography, number of major vessels, slope of the peak exercise, and thalassemia [6]. The features of the dataset used in this study are illustrated in Table I

TABLE I. Description of features of the cardiovascular dataset

Sr. No.	Attributes	Description	Input Type
1	Age	Age of Patients	Number
2	Sex	Sex (Male_1, Female_0)	Binary
3	Cp	Chest Pain: "1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic."	Categorical
4	Chol	Serum_Cholesterol	Number
5	Trestbps	Resting Blood Pressure (in mmHg)	Number
6	Restecg	Result Resting electrocardiographic	Categorical
7	Fbs	Fasting blood sugar	Binary
8	Exang	Exercise-related angina (1 for Yes, 0 for No)	Binary

9	Thalch	Maximum HeartRate	Number
10	Oldpeak	ST_depression	Number
11	Slope	ST_segment (peak exercise slope)	Categorical
12	Ca	Fluoroscopy colored (No of major vessels: "0–3")	Categorical
13	Thal	Thalassemia	Categorical
14	Targets	1 for heart disease and 0 for no heart disease	Binary

3.2 Preprocessing and Feature Scaling

Preprocessing is a method used to transform raw datasets into a well-organized collection of information. Proper data format is essential for improving machine learning outcomes. Some machine learning models need data in a specific format. To effectively represent datasets that require training and testing, data preprocessing is necessary. In this work, we have used the Cleveland, Hungary, Switzerland, and Long Beach V datasets. In our work, we also applied feature transformations, such as normalization and standardization, to improve accuracy. These data transformations are vital for bringing data onto the same scale, especially since data often differ in units, magnitude, and bias. Techniques such as normalization, scaling, and standardization are commonly employed in feature scaling to enhance the model's performance. [18].

3.2.1 Normalization. Normalization is a machine learning data preprocessing technique for scaling numerical features to a standard range. This approach guarantees that each feature has an equal impact on the model, thereby enhancing the performance of most algorithms. Normalization defines the range of data between 0 and 1.

$$X = X - (\frac{Xmin}{Xmax} - Xmin) \qquad (1)$$

3.2.2 Standardization. Standardization is a machine learning preprocessing technique that normalizes numerical data by setting the mean to zero and the standard deviation to 1. Standardization, also known as Z-score normalization, transforms data using the following formula:

$$X' = X - \frac{\mu}{\sigma} \qquad (2)$$

Here, the mean and standard deviation are denoted by μ *and* σ.

3.3 Stochastic Layer-Wise Sparse Autoencoder

The autoencoder is a technique used to extract efficient features of datasets. With this unsupervised technique, data is compressed into lower dimensions, and input data is reconstructed closely to the original data. The autoencoder process is carried out using the encoder-decoder paradigm. The encoder compresses the input data into a lower-dimensional representation [5, 17].

$$h = f(Wx + b) \qquad (3)$$

The input data is denoted by *x*, *W* is a weight, *b is* a bias, *f* is an activation function, and the latent representation by h. Using the latent representation, the decoder reconstructs the input data.

$$x\hat{} = g(W'h + b') \qquad (4)$$

Here, *x*^ is reconstructed data, *W* and *b* are decoder weights and biases, and *g* is an activation function.

In contrast to autoencoders, stochastic layer-wise sparse autoencoders include a probabilistic penalty that encourages the hidden layers to activate only a few neurons. This means the network will be forced to learn representations where most neurons will be inactive. This is usually achieved by adding a penalty term to the loss function during training, along with a probability that applies a penalty. This penalty discourages the hidden units from being active simultaneously. In this study, we applied this principle during the training of autoencoders, as shown in Figure 2. This verifies that the autoencoder learned meaningful latent representations from the input data [17].

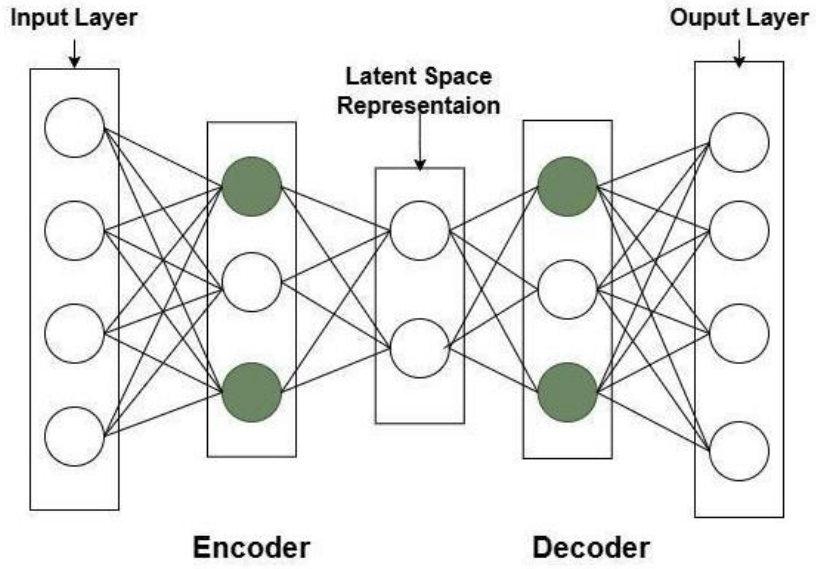

Figure 2: Architecture of Sparse Autoencoder

Sparsity can be constructed using either the KL divergence or the L1 regularization. In this work, we used L1 regularization. L1 regularization applies a penalty that is linear in the magnitude of the coefficient values. This technique is typically used to precisely set certain coefficients to zero, thereby facilitating easier feature selection [5, 16].

$$L1 = Error(x, x\hat{}\,) + \lambda \sum_i a_i \qquad (5)$$

Stochastic L1 regularization is a novel variation of traditional L1 regularization where the penalty on each weight is applied probabilistically, rather than deterministically, at each layer using a Bernoulli distribution. The modified equation is given below.

$$L1 = Error(x, x\hat{}\,) + \lambda \sum_i a_i \; p_i \qquad (6)$$

Where wi is a weight, λ controls the regularization strength, and pi is a stochastic binary mask per weight using a Bernoulli distribution.

3.4 Feedforward Neural Network

A feedforward neural network is a deep learning architecture that operates in a forward direction. In a feed-forward model, information flows strictly in one direction, and the input data is processed only once, without any backward flow or loops [19]. This design makes feed-forward networks the fundamental type of architecture illustrated in Figure 3.

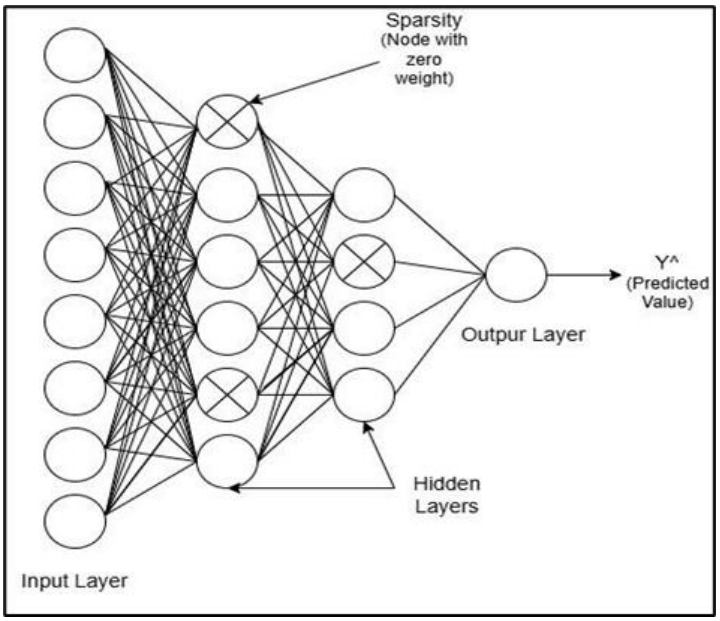

Figure 3: Deep Feed-Forward Network

This study utilized 64 and 32 nodes in the hidden layers, 13 nodes in the input layer, and one node in the output layer. No limit exists on hidden layers, with at least one. The hidden layer used ReLU activation, and the output layer used sigmoid activation.

$$fx(\sum_{i=1}^{n} xi\omega i + b) \qquad (7)$$

Here x is an input: w is the weight, b is the bias parameter, and fx is the activation function.

4. RESULT AND DISCUSSION

A stochastic layer-wise sparse autoencoder model with a deep feed-forward neural network (SLSAE-DFFN) is utilized for feature extraction and prediction of cardiovascular disease. The performance of the model was evaluated using a cardiovascular dataset. This dataset is available on Kaggle and consists of four types of databases: Cleveland, Hungary, Switzerland, and Long Beach V. The dataset has 76 features, including one target feature. However, many articles included only 14 attributes in their work. The "target" refers to the presence or absence of disease. The standard classification model, DFFN, was used to test the proposed methodology. The 14 features are fed to all classifiers to evaluate their performance.

4.1 Analysis of Cardiovascular Disease Dataset

This study utilizes the SAE-DFFN model to analyze the heart disease dataset, which comprises 14 features [6]. The dataset is first preprocessed to find the linearity among these features. The correlation of the features is presented in the heat map.

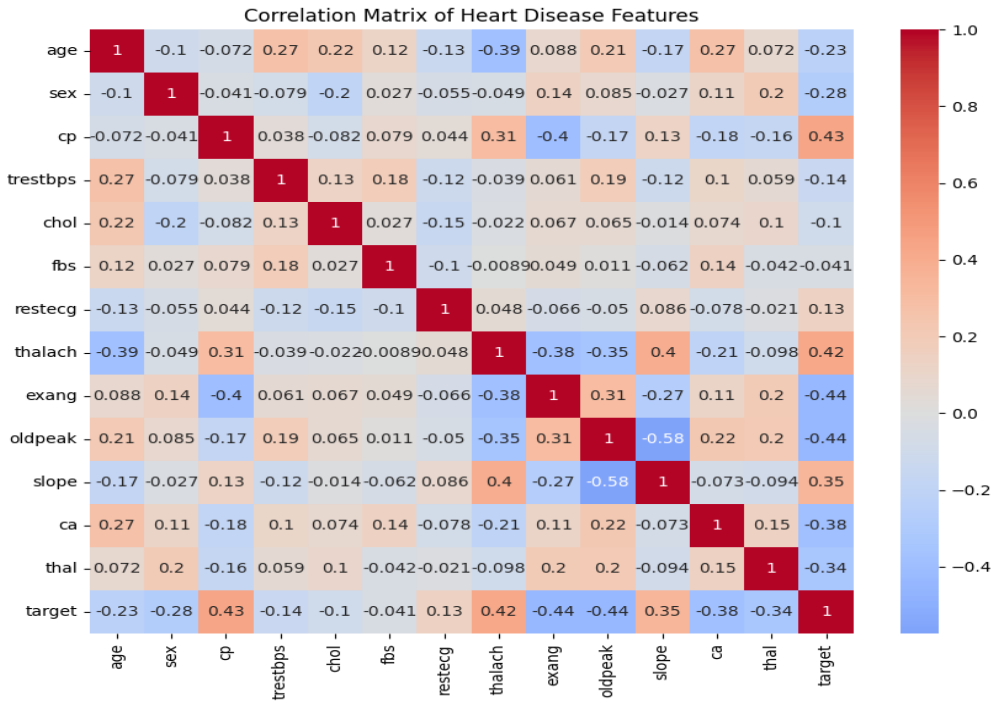

Figure 4: Correlation Matrix

The features and their connections are displayed in the correlation matrix, which ranges from -1 to 1. Values close to 1 indicate that the features are highly dependent on one another, whereas values close to -1 indicate that the correlation between the features is weak. The correlation matrix displays highly independent features, with a value of 1 indicating self-correlation [20].

The categorical features are illustrated in Figure 4, and the graphs highlight the notable variations among the categories of each feature.

4.2 Results and Analysis

The performance of the SLSAE-DFFN model is assessed and compared with various machine learning algorithms, such as Naive Bayes, Support Vector Machine, logistic regression, Random Forest, and ANN. The performance evaluation is calculated in terms of precision, recall, F1 score, and accuracy based on the confusion matrix depicted in Table III.

TABLE III. Confusion Matrix

		Predicted	
		Positive(1)	**Negative(0)**
Actual	**Positive(1)**	**101**	**4**
	Negative(0)	**1**	**98**

Accuracy is the measure of the model that can be defined as

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (8)$$

Precision is defined as the percentage of correctly identified positive instances (including false positives) out of all instances predicted to be positive. Conversely, recall indicates the percentage of actual positive cases identified in the database compared to all those that were positive. Recall and precision are frequently inversely correlated. Increasing precision might lead to a decrease in recall [21].

$$Precision = TP/(TP+FP) \quad (9)$$

$$Recall = TP/(TP+FN) \quad (10)$$

Table IV and Figure 5 present a comparison of the suggested model with SVM, Naive Bayes, LR, RF, and ANN.

Table IV: Comparison of the proposed SAE-DFFN with Logistic Regression, SVM, Naive Bayes, RF, and ANN classifiers

Algorithms	Accurac (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	81	76	92	83
NB	80	76	90	82
LR	80	76	87	81
RF	92	91	95	93
ANN	95	96	93	95
SSLAE-DFFN	97	98	96	97

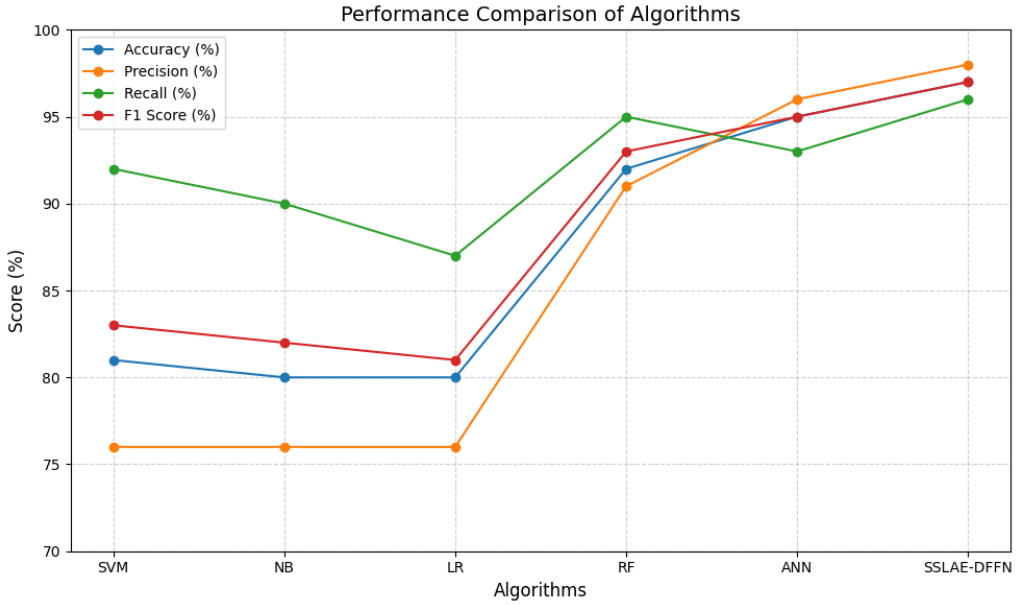

Figure 5. Comparison of the performance of Classifiers

5. CONCLUSION AND FUTURE WORK

Predicting cardiovascular disease is a critical challenge in this era. This research proposes an optimized prediction of cardiovascular disease using the SLSAE-DFFN model. In this proposed work, the hidden layer of the autoencoder incorporates probabilistic sparse constraints to ensure that only a small number of neurons are active during training. This study's significant contribution is that we extracted dimensionally reduced features from the autoencoder's latent space representation and fed them to a feed-forward neural network for classification, which improved the model's performance. The impact of this research is enhanced because the proposed system outperforms the majority of existing techniques in terms of accuracy, precision, sensitivity, and F1 score. Comparative studies are also covered in the paper. The F1 score, accuracy, precision, and sensitivity were, in that order, 98.58%, 1.0%, 97.08%, and 98.52%.

This study employs a small sample size of 1025 instances. In the future, advanced machine learning techniques could enhance the diagnosis process by utilizing fusion datasets, which combine two or more distinct datasets, and large-scale datasets with numerous attributes. We can also explore optimization methods, such as particle swarm optimization, ant colony optimization, and artificial bee colony, in conjunction with various machine learning classifiers.

Funding: All authors declare no funding sources to disclose.

Data and Code Availability: Available upon request from the corresponding author

Declarations

Competing Interests The authors declare no competing interests.

Ethics and Consent to Participate declarations are not applicable

REFERENCES

[1] E. O. Lopez, B. D. Ballard, and A. Jan, "Cardiovascular disease," StatPearls Publishing, 2023 [Online]. Available: https://doi.org/10.1016/j.neuroimage.2010.11.004

[2] Z. Chen, N. Li, and X. Li, "Introduction to Machine Learning for Predictive Modelling I," Materials Informatics II: Software Tools and Databases, Springer Nature Switzerland, 2025, pp. 3–30. [Online]. Available: https://doi.org/10.1007/978-3-031-78728-7_1

[3] Y. Li et al., "Deep learning in bioinformatics: introduction, application, and perspective in the big data era," Methods, vol. 166, pp. 4–21, 2019. [Online]. Available: https://doi.org/10.1016/j.ymeth.2019.04.008

[4] W. H. L. Pinaya, S. Vieira, R. Garcia-Dias, and A. Mechelli, "Autoencoders," Machine Learning, Academic Press, 2020, pp. 193–208 [Online]. Available: https://doi.org/10.1016/B978-0-12-815739-8.00011-0

[5] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: a survey," Artif. Intell. Rev, vol. 57, no. 2, p. 28, 2024 [Online]. Available: https://doi.org/10.1007/s10462-023-10662-6

[6] Heart Disease Dataset [Online] - Kaggle. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset [Accessed: Mar. 2, 2025].

[7] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," SN Comput. Sci., vol. 1, no. 6, p. 345, 2020. [Online]. Available: https://doi.org/10.1007/s42979-020-00365-y

[8] M. M. Ali et al., "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput. Biol. Med., vol. 136, p. 104672, 2021 [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104672

[9] M. G. El-Shafiey, A. Hagag, E. S. A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," Multimed. Tools Appl., vol. 81, no. 13, pp. 18155–18179, 2022 [Online]. Available: https://doi.org/10.1007/s11042-022-12425-x

[10] R. Rajendran and A. Karthi, "Heart disease prediction using entropy-based feature engineering and ensembling of machine learning classifiers," Expert Syst. Appl., vol. 207, p. 117882, 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.117882

[11] K. Karthick et al., "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," Comput. Math. Methods Med., p. 6517716, 2022 [Online]. Available: https://doi.org/10.1155/2023/9764021

[12] A. K. Garate-Escamilla, A. H. E. Hassani, and E. Andres, "Classification models for heart disease prediction using feature selection and PCA," *Med. Unlocked*, 2020 [Online]. Available: https://doi.org/10.1016/j.imu.2020.100330.

[13] M. Rahman, D. Islam, R. J. Mukti, and I. Saha, "A deep learning approach based on convolutional LSTM for detecting diabetes," *Comput. Biol. Chem.*, vol. 88, p. 107329, oct. 2020. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2020.107329

[14] A. Mehmood et al., "Prediction of Heart Disease Using Deep Convolutional Neural Networks," Arab. J. Sci. Eng., 2021. [Online]. Available: https://doi.org/10.1007/s13369-020-05105-1

[15] H. Byeon et al., "Deep neural network model for enhancing disease prediction using autoencoder-based broad learning," SLAS Technol., vol. 29, no. 3, p. 100145, 2024. [Online]. Available: https://doi.org/10.1016/j.slast.2024.100145

[16] A. Abdellatif et al., "Computational detection and interpretation of heart disease based on conditional variational auto-encoder and stacked ensemble-learning framework," Biomed. Signal Process. Control, vol. 88, p. 105644, Feb. 2024. [Online]. Available: https://doi.org/10.1016/j.bspc.2023.105644

[17] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation," Multimed. Tools Appl., vol. 82, no. 20, pp. 31759–31773, 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-14817-z

[18] H. Alshaher, "Studying the effects of feature scaling in machine learning," Doctoral dissertation, North Carolina Agricultural and Technical State Univ., 2021

[19] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Netw., vol. 61, pp. 85–117, 2015. [Online]. Available: https://doi.org/10.1016/j.neunet.2014.09.003

[20] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature analysis of coronary artery heart disease data sets," Procedia Comput. Sci., vol. 65, pp. 459–468, 2015. [Online]. Available: https://doi.org/10.1016/j.procs.2015.09.132

[21] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure-driven view," Inf. Sci., vol. 507, pp. 772–794, 2020. [Online]. Available: https://doi.org/10.1016/j.ins.2019.06.064.