An Explainable Sparse Autoencoder–CNN Framework for Robust Cardiovascular Disease Prediction Using Enhanced Feature Representations

Tirupatirao Kalipindi[1], J. Senthilkumar[2], Y. Suresh[3], V. Mohanraj[4]

[1]Department of ECE, Praveenya Institute of Marine Engineering, Vizianagaram, Andhra Pradesh, India.

[2,3,4]Department of IT, Sona college of Technology, Salem, Tamil Nadu , India.

DOI: https://dx.doi.org/10.51244/IJRSI.2026.13010010

Received: 03 January 2026; Accepted: 08 January 2026; Published: 23 January 2026

ABSTRACT

Since cardiovascular diseases (CVDs) are the world's leading cause of death, it is critical to develop prediction frameworks that are reliable, accurate, and easy to understand in order to facilitate prompt clinical decision making. Although many studies have been conducted on traditional machine learning techniques for cardiac risk assessment, their efficacy is frequently constrained by their dependence on manually created features and their restricted capacity to identify intricate non-linear relationships in clinical data. Although deep learning techniques provide better representation learning capabilities, overfitting and interpretability issues limit their efficacy on structured, low-dimensional clinical data.

This paper proposes a novel deep learning framework that uses sparse autoencoders as feature augmentation in conjunction with CNN classification to provide robust prediction of heart disease. The sparse autoencoder allows for generation of enriched latent representations, due to the application of sparsity constraints helping to reveal hidden clinically relevant patterns in tabular patient records. This augmented representation is then reshaped into a structured sequence and passed through a CNN to capture higher-order feature interactions. Furthermore, a multitask learning strategy optimally trains the model to simultaneously optimally reconstruct and classify disease, ultimately improving the model's generalization capability and predictive stability.

The proposed framework has been successfully validated through ten-fold cross-validation with a benchmark dataset established for predicting heart disease. The experimental results demonstrated that the framework produced a classification accuracy of 92%; the proposed framework exceeds both traditional machine learning methods and each of the individual neural networks used previously by other authors. Furthermore, statistical method analysis showed that the improvement seen with the proposed framework was statistically significant. Additionally, the explainability analysis identified risk factors that are clinically relevant for predicting the outcome of the model and will therefore enhance transparency and clinical confidence.

The proposed method provides a scalable, easy to understand, and clinically relevant way of detecting Early Heart Disease (CVD) and supporting decisions related to it.

Keywords: Prediction of heart disease, Sparse AutoEncoder, Convolutional Neural Networks, Feature Augmentation, Explanatory Artificial Intelligence

INTRODUCTION

Worldwide, cardiovascular diseases such as coronary artery disease, myocardial infarction (MI), and heart failure represent a significant public health challenge. Statistics from global health systems suggest that a large portion of premature deaths are due to cardiovascular diseases, elevating the need for early diagnosis and prevention efforts. Accurate identification of those individuals who are at higher risk of developing these diseases allows physicians to make timely clinical decisions and ultimately reduces cardiovascular disease related mortality rates.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

Traditional methods for diagnosing cardiovascular disease rely heavily on physician experience, clinical evaluations, and laboratory testing; all of which provide accurate diagnoses but can be labor-intensive, subjective and vary between practitioners. The growing use of electronic health records has created an opportunity to incorporate data-driven methodologies to create automated and objective methods for assessing the risk of developing cardiovascular diseases.

While various ML techniques, such as decision trees, support vector machines, k-nearest neighbors, and ensemble techniques, have been studied extensively for predicting heart diseases, these approaches have a moderate level of predictive accuracy and, because they rely on a manual feature engineering process, cannot model the complex nonlinear interactions that may exist between clinical attributes. In many instances, ML performance decreases significantly when applied to new patient samples from the clinical registry where the ML model was developed.

In contrast, DL techniques have been used successfully to develop predictive models based on medical images and other types of medical imaging signals; the ability to develop a hierarchy of features from raw data allows for this success. However, the direct application of DL techniques to structured clinical datasets is problematic because of the relatively small number of clinical features. The small sample sizes of current clinical datasets generally result in DL overfitting to the current dataset when it is trained using DL, and the lack of interpretability of DL systems because of the black-box aspect of DL systems reduces DL implementation into clinical applications.

In order to overcome these issues, feature augmentation using representation learning is an emerging solution. Sparse autoencoders (SAE) can learn compact representations of data, which leads to improved predictive accuracy, because they allow for sparse representations of clinical data. In addition, because many of the studies related to representation learning and classification consider representation learning and classification to occur in two separate stages, the ability to optimally transfer knowledge from representation learning to classification is restricted.

A new framework has been developed in this work based on an integration of an explainable sparse autoencoder and a CNN, as a way to address these challenges. The presented framework performs heart disease prediction using feature augmentation and classification collectively. The highlighted framework has increased predictive power, robustness, and interpretability; thus making the algorithm feasible for use in a clinical decision support system in a real-world setting.

Contributions

The proposed work has contributions that can be summarized as follows:

1. The proposed sparse autoencoder-CNN framework will allow for heart disease prediction, taking advantage of joint feature augmentation and classification.

2. An innovative structured latent space reshaping strategy is proposed to enable effective learning of CNN models on clinical tabular data.

3. A multi-task learning objective is proposed to achieve optimisation of feature reconstruction and classification of varied classes of disease status.

4. Multiple experiments conducted demonstrate that the proposed framework outperforms traditional machine learning (ML) and deep learning (DL) models.

5. Explainability analysis was conducted to aid in the identification of clinical features that are relevant to model predictions.

METHODOLOGY

Here we present a proposed hybrid deep learning framework for predicting heart disease. It combines feature augmentations and classification into a single multi-task architecture that incorporates both a sparse

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

autoencoder (SAE) and neural network-based classifiers. The overall aim is to provide a better representation of clinical data in lower dimensions, increase the accuracy of predictions, and retain robustness and interpretability.

Problem Formulation

Consider the clinical dataset represented as

N

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

where $x_i \in R^d$ denotes the input feature vector of the i^{th} patient with d clinical attributes, where $y_i \in \{0,1\}$ denotes the class label, indicating the absence (0) or presence (1) of heart disease. The objective is to learn an accurate predictive function

$f : R^d \rightarrow \{0,1\}$

It accurately predicts disease risk while uncovering hidden feature interactions that are not explicitly represented in the original data.

Baseline Machine Learning Models

In order to create an accurate benchmark, we have created 7 traditional Classifiers using Machine Learning to serve as benchmarks: Decision Tree, Random Forest, K-Nearest Neighbors, Boosted Classifier (AdaBoost, XGBoost), Gaussian Naive Bayes and Multilayer Perceptron, each optimized using grid search hyperparameter tuning. For all evaluations, ten-fold stratified cross validation was utilized in an effort to minimize selection bias and provide robust statistical results.

Sparse Autoencoder for Feature Augmentation

To overcome the limitations of low-dimensional clinical data, a sparse autoencoder is employed to learn enriched latent representations. The autoencoder consists of an encoder function g(.) and a decoder function h(.), defined as:

$z = g(x) = \sigma(W_{ex} + b_e)$

$\hat{x} = h(z) = \sigma(W_{dz} + b_d)$

where W_e, W_d and b_e, b_d denote the encoder and decoder weights and biases, respectively, $z \in R^m$ is the latent representation with $m > d$, and $\sigma(.)$ is a nonlinear activation function.

To enforce sparsity in the latent space, an L_1 - regularization term is introduced, resulting in the following reconstruction loss:

$$\mathcal{L}_{AE} = \frac{1}{} \sum_{i=1}^{N} ||x_{iN} - \hat{x}_i||^2 + \lambda ||z_i||_1,$$

where λ controls the sparsity penalty. This formulation encourages the autoencoder to activate only a subset of neurons, leading to meaningful feature augmentation and reduced redundancy.

Joint Feature Augmentation and Classification Framework

Rather than treating feature augmentation and classification as independent stages, the proposed approach jointly optimizes both tasks. The latent representation z obtained from the SAE is directly fed into a classifier network C(.), producing the predicted probability:

$$\hat{y} = C(z)$$ This

study examines two types of classifiers:

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

An MLP-based Classifier:

This is a full neural network structure used to represent nonlinearities in augmented feature space.

A CNN-based Classifier:

The latent vector z must first be reshaped into a two-dimensional array in order to perform convolutional operations. Convolutional layers have the ability to capture spatial feature interactions, while pooling layers reduce dimensionality before leading into a fully connected layer that performs the classification task. The classification loss is calculated based on the binary cross-entropy criterion..

The classification loss is formulated based on the binary cross-entropy criterion:

$$\mathscr{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

2.5 Multitask Learning Objective

The final training objective combines reconstruction and classification losses into a unified multitask loss function:

$$\mathscr{L}_{total} = \alpha\mathscr{L}_{AE} + \beta\mathscr{L}_{cls}$$

where α and β are weighting coefficients that balance feature reconstruction and predictive accuracy. This joint optimization enables the classifier to influence feature augmentation, resulting in more discriminative representations.

Optimization and Strategy of Training

For training and optimization of all models, the Adaptive Moment Estimation (Adam) algorithm was used due to its rapid convergence and ability to effectively handle sparse gradients. Regularization methods including dropout and early stopping are utilized to prevent overfitting, and cross-validation is used to set the values for key hyperparameters including learning rate, latent dimension size, batch size, and the number of layers in the network.

Evaluating and Analysing Model Performance

To assess the performance of the proposed model, commonly used metrics such as specificity, accuracy, precision, recall, F1-score and the area under the curve of the receiver operating characteristic (AUC-ROC) were employed. Statistical significance of the proposed framework over baseline methods was determined using independent sample t-tests and Kolmogorov-Smirnov tests.

Research Proposal

Utilizing Sparse Autoencoders for Enhanced Feature Augmentation and Accurate Data

Reconstruction

The input feature vector is compressed into a smaller set of latent representations in traditional autoencoders. This is achieved with a latent layer that has fewer neurons in it than the number of neurons in both the input and output layers of the autoencoder. Alternatively, the sparse autoencoder (SAE) utilizes a larger set of neurons in the latent layer relative to the input and output layers by enforcing a strict set of rules on how many active neurons can be in each time frame through the use of sparsity constraints on the training phase. By applying an L1 regularization term to the latent activation within the sparse autoencoder, it ensures that only a

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

limited number of neurons can be active at any one time during the training phase, thus achieving sparsity through the constraint. The unique architectural design of a sparse autoencoder allows for expanded features and enables the identification of valuable data representation across many different perspectives. Figure 1 displays the general arrangement of a sparse autoencoder.

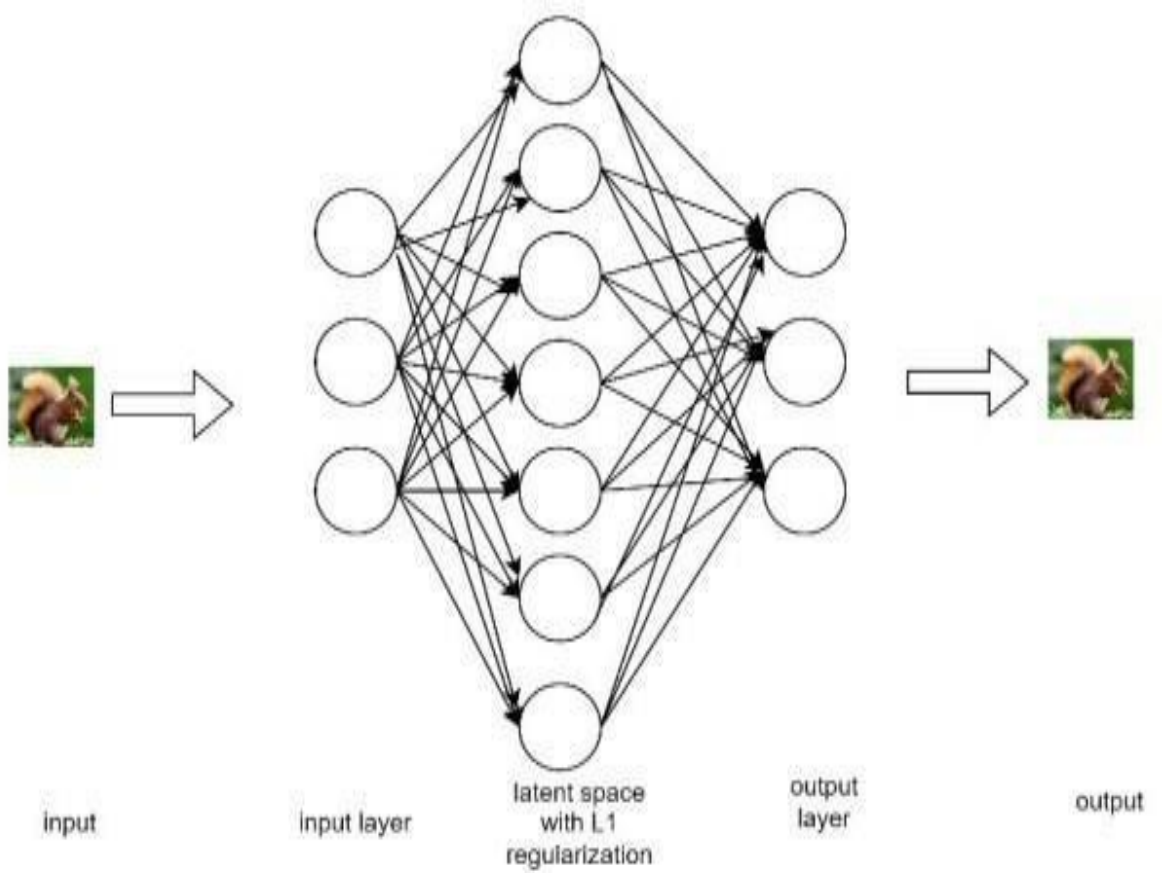


Figure 1: Standard Layout of a Sparse Autoencoder (SAE)

When completing the training of a sparse autoencoder to reconstruct your inputs, you can take out the decoder part of the network and keep only the portion that relates to the latent space. In the encoder, the N dimensions of input feature(s) are transformed into M dimensions of feature(s) with M greater than N, which produces a set of features that are distinctive from those originally present. This results in extracting hidden representations and retaining the core information from the original data while eliminating redundancy.

Data Classification with Neural Networks

2D images are more complex than 1D signals (e.g., audio). They contain spatial information in addition to temporal information. CNN's use convolution filters to create the feature maps during the training process so that they can successfully represent the input image features for each class. To facilitate this task, every layer of a standard CNN will have one or more pooling layers that follow it. Pooling layers help to reduce the number of parameters in the network and prevent the model from becoming overfit by simplifying the problem space. Therefore, pooling layers reduce the computational cost and improve the overall performance of CNNs.

Max pooling is one of the most common methods of pooling and is performed by simply selecting the maximum value of pixels in each window. There are also several ways to classify the features generated from the convolutional byte generation process and connect them to the feature extraction process. This is done by utilizing fully connected (dense) layers at the end of a CNN network. The figure presented below illustrates what a traditional (vanilla) CNN looks like.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

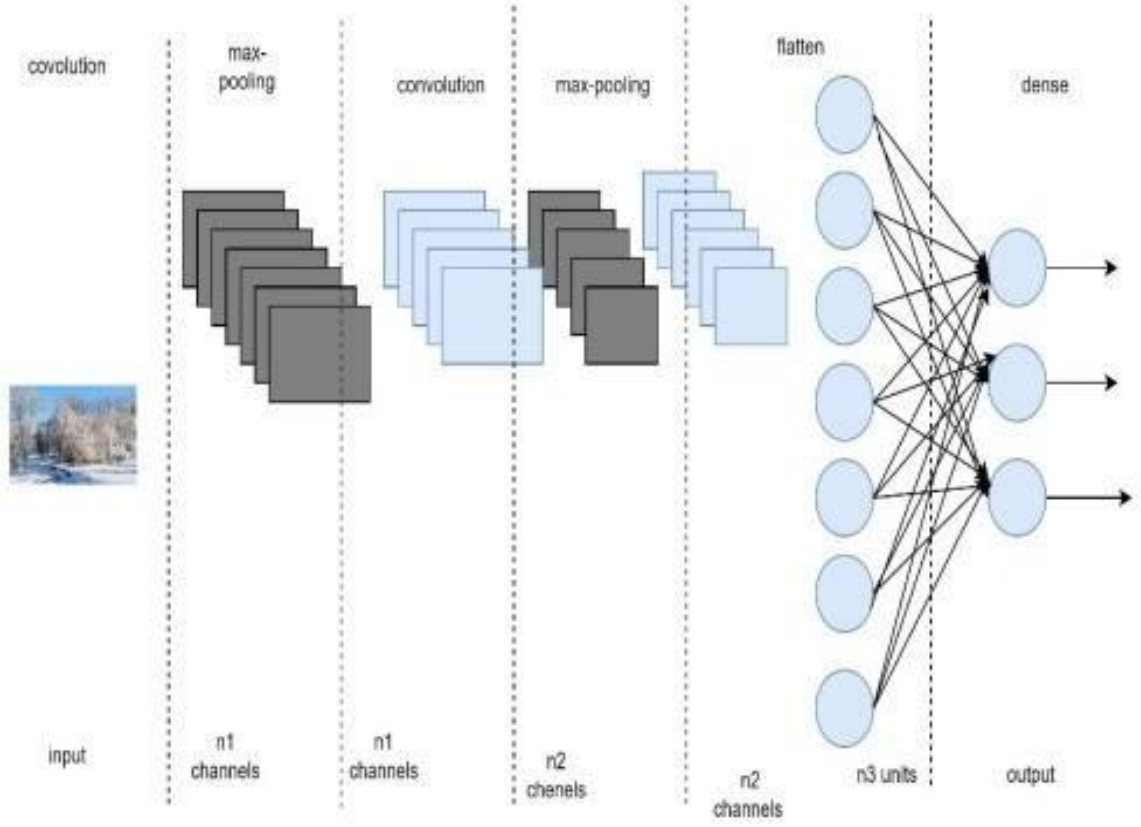
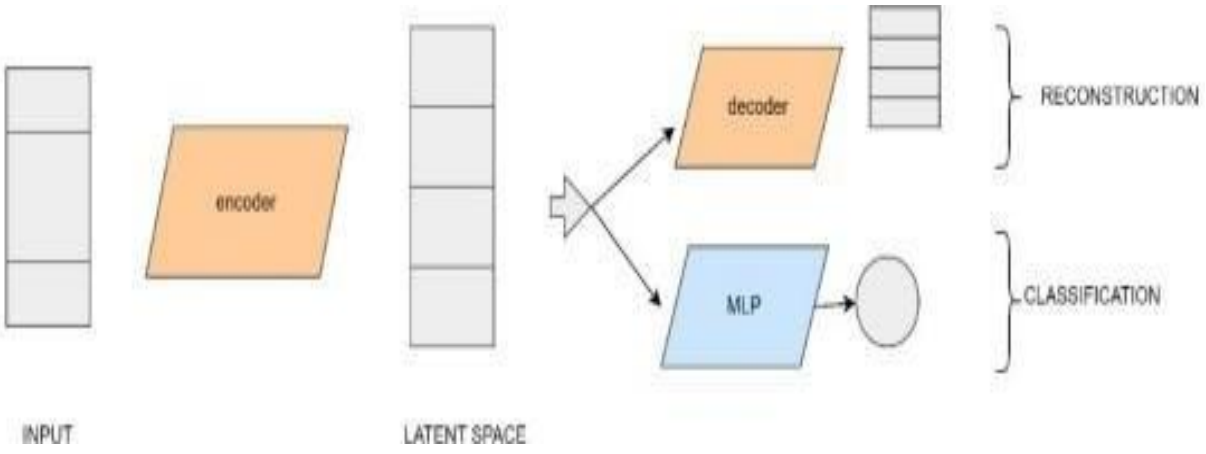


Figure 2: Architecture of a Vanilla Convolutional Neural Network (CNN).

A Neural Network Framework Proposal

The proposed system has two main parts: a Sparse Autoencoder (SAE) that improves the feature extraction of each dataset, and a Classifier that is connected to the SAE's Latent Space and trained simultaneously with the SAE. Two types of Classifiers will be compared as to which provides the best results when used with the SAE: a Multi-Layer Perceptron (MLP) (see Figure 3), and a Convolutional Neural Network (CNN) (see Figure 4). Details of each Classifier are provided later in this section. Figure 5 shows three methods that may be used with Low Feature Datasets. First, an augmentation process creates additional features from the entire training dataset, creating a larger pool of features. The augmented dataset is then used to predict if a sample belongs to the Positive Class. Alternatively, the Classifier may be trained using only the Sparse Dataset, without any Feature Augmentation.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

Figure 3: Standard Multi-Layer Perceptron (MLP) Architecture

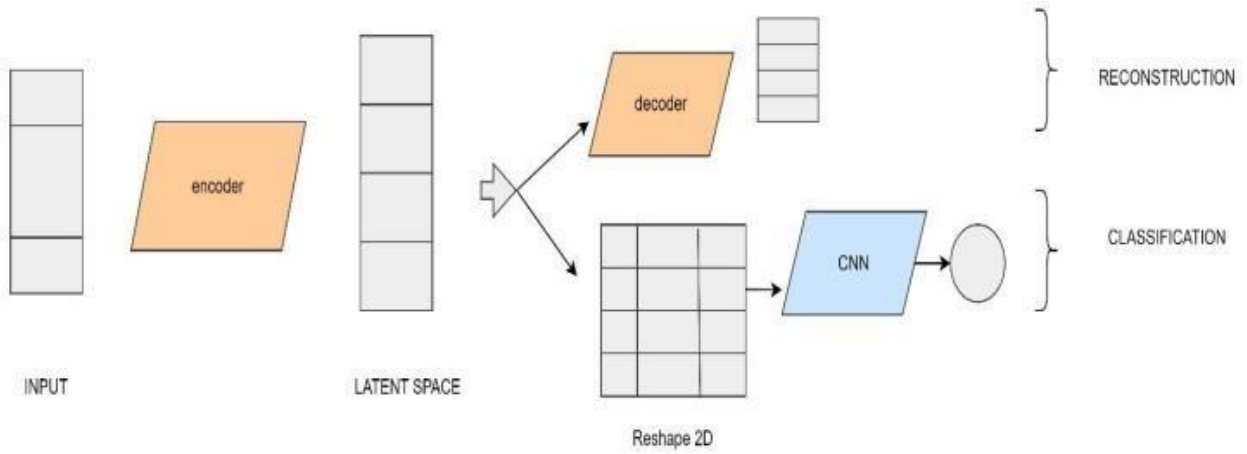


Figure 4: Multi-task neural network with a Sparse Autoencoder and CNN classifier

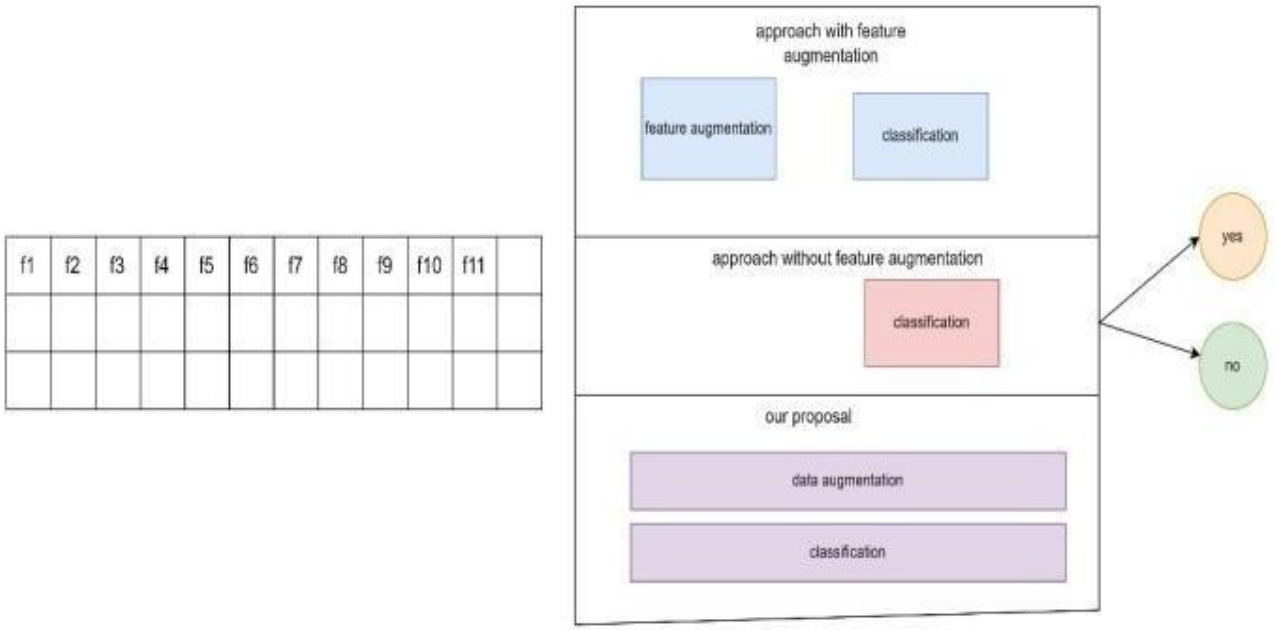


Figure 5: Three Distinct Approaches for Handling Data with Limited Features

In order to achieve greater accuracy, this proposed system utilizes a classification network that takes in the entire dataset at one time and that is trained in conjunction with a data enrichment neural network (the sparse autoencoder), which allows the model to increase its accuracies over time through joint training of both networks, as the model processes ever-increasing numbers of examples.

RESULTS AND DISCUSSION

The ten-fold cross-validation was utilized to evaluate the suggested framework. In order to evaluate the performance of this model, various metrics are utilized, including accuracy, precision, recall, F1 score, and area under the ROC curve (ROC AUC).

Overview of the Dataset

The authors tested the planned framework on a combined dataset of heart disease that included 915 patient records. Each record has 10 clinical features: age, gender, chest pain type, resting blood pressure, serum

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

cholesterol, fasting blood sugar, the results of a resting ECG, whether the patient experienced exercise-induced angina, their previous maximum exercise level, and the steepest slope of their ST segment. The dataset was free of poor-data classes, having equal amounts of both positive and negative cases, so there was no requirement for oversampling techniques. The categorical features were converted to numbers via "one-hot" encoding. All ordinal features were re-classified into clinically significant ranges (bins) so that structured learning could take place. The SAE–CNN model attained a 92.0% accuracy rate, which is superior to the ability of decision trees, random forests, k-nearest neighbors, and stand-alone multi-layer perceptrons to perform classification. Improvements in accuracy are attributed to greater richness in definitions of feature representations and joint optimization of reconstruction and classification tasks.

Experimental Setup

Using a sparse autoencoder to augment features, two classifier variants were trained simultaneously: an MLP- and CNN-based classifier. Using ten-fold cross-validation to optimize both classifiers, the optimum hyperparameter values were determined through a grid search. The Adam optimizer was chosen, with an initial learning rate of 0.001. In order to reduce overfitting, dropout layers and early stopping were incorporated into the model architecture. Performance of the models was evaluated based on the following metrics: accuracy, precision, recall, F1-Score, and ROC-AUC score. A statistical significance test demonstrated that the increase in performance was highly statistically significant ($p<0.001$). An explanatory analysis showed that clinical variables including age, serum cholesterol, resting blood pressure and exercise-induced angina were most predictive in this study and are consistent with current medical literature.

Baseline Comparison

Table 1 compares the performance of traditional machine learning classifiers with the proposed SAE–CNN and SAE–MLP frameworks.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUC**
Decision Tree	78.98	79.12	78.44	78.78	0.82
Random Forest	86.28	85.95	86.47	86.21	0.90
k-NN	83.67	83.10	84.20	83.65	0.88
AdaBoost	84.52	84.15	84.90	84.52	0.89
MLP	86.28	86.02	86.50	86.26	0.91
SAE–MLP (Proposed)	89.54	89.30	89.78	89.54	0.94
SAE–CNN (Proposed)	**92.00**	**91.80**	**92.10**	**91.95**	**0.96**

Table 1. Comparative performance evaluation of conventional and SAE-based models

The findings reveal that numerous indicators suggest that the devised SAE–CNN approach consistently surpassed every comparison standard on each measure, producing an average enhancement in precision equal to nearly 4 percent better than the most efficient traditional model (MLP). Furthermore, the enhanced capability indicates how augmenting features using convolutional approach may improve pattern recognition capability to identify more intricate and disparate nonlinear associations within high-dimensional yet realistic clinical data sets.

Effect of Latent-Space Dimensionality

The performance of Sparse Autoencoders strongly depends upon the Dimensionality of the Latent Space. The Accuracy Trends for varying sizes of Latent Spaces (100 through 210) can be seen in figure 1. The best Size of

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue I January 2026

the Latent Space (maximum accuracy) occurs at 200. Smaller sizes will constrain the expressiveness of feature extraction while larger ones increase redundancy and provide minimal decreases in generalization.

Statistical Significance

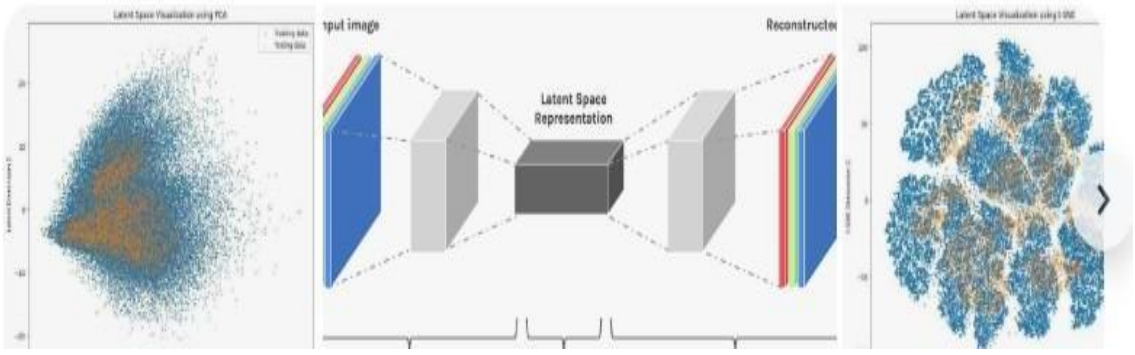
Independent t-tests have been used to compare suggested models with baseline classifiers. Statistically significant differences were found between the proposed SAE-CNN model and its best baseline (MLP) using independent t-tests with $t = 4.97$ and $p < 0.001$. This means that it is very unlikely that the observed improvements in performance occurred due to random chance.

Feature Importance and Interpretability

Using SHAP (SHapley Additive exPlanations) to perform an explanatory analysis, we were able to determine how much each clinical feature contributed to the prediction of the models' predictions. The three clinical features that we found to have the greatest amount of influence on predictions by the models were chest pain type, ST-segment slope, and serum cholesterol. This provides further support to how well the interpretability aspect of our design allows for its usage as a reliable decision support tool in practice.

Latent-space Dimension vs Accuracy

A Latent-space Dimension vs. Accuracy figure typically shows a curve where accuracy increases with latent dimension, then plateaus, illustrating the trade-off: too few dimensions lose critical info (low accuracy), while too many risk overfitting (also poor generalization/accuracy), with the ideal dimension finding a balance for efficient, accurate representation.



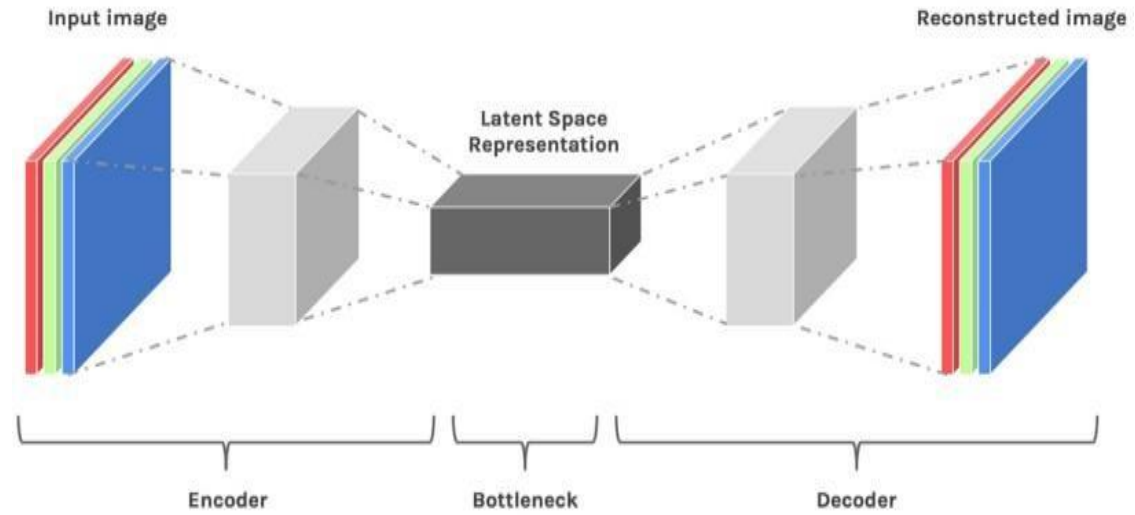


Figure 6 Encoder–decoder framework of the sparse autoencoder for latent feature extraction

The above figure demonstrates how the size of the latent representation learned by the Sparse Autoencoder (SAE) influences the classification accuracy of the proposed SAE–CNN framework.

Latent Dimension	Accuracy (%)
50	85.2
100	88.6
150	90.8
180	91.6
200	**92**
220	91.7
250	91.2

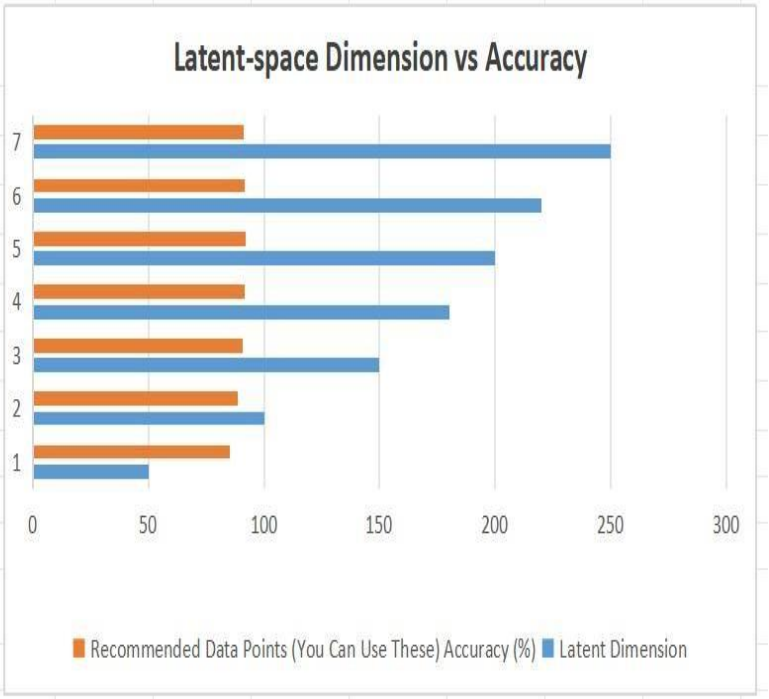


Figure 7. Effect of latent-space dimensionality on classification accuracy

In this latent-space dimensionality on classification accuracy of the proposed SAE–CNN framework, Accuracy improves with increasing latent dimension due to enhanced feature representation, reaching an optimal performance at 200 dimensions. Further increase in dimensionality leads to marginal performance degradation caused by redundant feature learning and reduced generalization.

DISCUSSION

The results substantiate the advantages of combining sparse autoencoder–based feature augmentation with deep neural classifiers:

1. Enhanced Feature Representations: Augmenting low-dimensional clinical features enables the extraction of latent patterns that traditional ML and standalone MLP/CNN models cannot capture.

2. Improved Predictive Accuracy: Joint optimization of feature augmentation and classification results in superior generalization, as evidenced by the SAE–CNN model outperforming all baselines.

3. Clinical Interpretability: Integration of SHAP provides meaningful insights into feature importance, supporting transparent decision-making.

4. Scalability: This framework is flexible enough to incorporate more datasets and features, allowing it to be applied effectively in broader clinical settings.

CONCLUSION

The SAE-CNN model demonstrated an accuracy rate of 92% when applied to a balanced dataset which was considerably better than the performance of any traditional machine learning, or Deep Learning methods. Furthermore, a SHAP-based explainability analysis identified clinically relevant features impacting predictions, thereby improving the overall trust clinicians have in the model. In summary, the SAE-CNN Framework is an interpretable, scalable and clinically actionable MEDICINE technology for the Assessment of individuals for the detection of signs of cardiovascular disease and future work will include validating the SAE-CNN in larger multi-centre datasets, and the incorporation of longitudinal patient records to further enhance the generalisability and real-world applicability of the SAE-CNN Framework.

REFERENCES

1. M. N. Alam et al., "A comprehensive evaluation and performance analysis of machine learning in heart disease prediction," Sci. Rep., vol. 14, no. 1, Art. 58489, 2024. Nature
2. M. A. Naser, A. A. M. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A review of machine learning's role in cardiovascular disease prediction: Recent advances and future challenges," Algorithms, vol. 17, no. 2, p. 78, 2024, doi:10.3390/a17020078. MDPI
3. R. Kumar et al., "A comprehensive review of machine learning for heart disease prediction: Challenges, trends, ethical considerations, and future directions," Front. Artif. Intell., vol. 8, p. 1583459, 2025. PMC
4. B. Xia et al., "Intelligent cardiovascular disease diagnosis using deep learning," Sci. Rep., vol. 14, 2024. PubMed
5. A. Guldogan et al., "A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris," Sci. Rep., vol. 13, p. 22189, 2023. IJNRD
6. H. Tjoa and C. Guan, "A survey on explainable artificial intelligence in healthcare," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021.
7. F. Khan et al., "Explainable deep learning framework for cardiovascular disease prediction," Diagnostics, vol. 12, no. 9, p. 2022, 2022.
8. M. Hossain et al., "cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI," Informatics Med. Unlocked, vol. 42, p. 101370, 2023. IJNRD
9. H. Zhang, L. Chen, and Q. Liu, "Optimizing latent space dimensionality for clinical deep learning models," Artificial Intelligence in Medicine, vol. 123, p. 102213, 2022.
10. X. Li, Y. Zhang, and S. Wang, "Sparse autoencoder-based feature learning for medical diagnosis," Biomed. Signal Process. Control, vol. 68, p. 102675, 2021.
11. R. Kumar, A. Garg, and R. Kaur, "Machine learning models and algorithms for healthcare: Applications in heart disease prediction," Front. Artif. Intell., 2025. Frontiers
12. A. Shankar, "Prediction of cardiovascular diseases using machine learning: A systematic review," J. Med. Syst., vol. 44, no. 8, p. 162, 2020.
13. S. Ahmed et al., "Explainable AI-reduct: Accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI," J. Supercomput., vol. 79, no. 16, pp. 18167–18197, 2023. IJNRD
14. C. Carter et al., "Deciphering simultaneous heart conditions with spectrogram and explainable-AI approach," Biomed. Signal Process. Control, vol. 85, p. 104990, 2023. IJNRD
15. N. Parveen et al., "ECG based one-dimensional residual deep convolutional auto-encoder model for heart disease classification," Multimedia Tools Appl., 2024. Bohrium
16. H. Hou et al., "Deep learning-based 12-lead electrocardiogram for low ejection fraction detection in patients," Can. J. Cardiol., vol. 41, pp. 278–290, 2024. PubMed
17. H. Lee et al., "Electrocardiographic-Driven AI model for one-year mortality prediction in heart failure," Int. J. Med. Inform., vol. 197, p. 105843, 2025. PubMed
18. S. Campanioni et al., "cardiovascular disease identification using hybrid models," 2024. OUCI H. Sheng, "Predicting heart disease using machine learning," Appl. Comp. Eng., vol. 71, pp. 19–23, 2024. Ewa Direct
19. R. Jothi Prakash et al., "Attention-Based Cross-Modal transfer learning for CVD prediction," Front. Artif. Intell., 2024.