Journal of Theoretical and Applied Information Technology
15th December 2024. Vol.102. No. 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE REDUCTION AND STROKE PREDICTION USING SPARSE SUBSPACE CLUSTERING AUTOENCODER ON CLINICAL DATA

DURGA DEVI. P[1], K. AKILA[2]
[1]Research Scholar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, TN, India.
[2]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, TN, India.

[1]p.durgadevi91@gmail.com, [2]akilak@srmist.edu.in

ABSTRACT

The use of deep learning techniques for feature reduction presents a novel way to improve stroke prediction, as demonstrated by the article "Feature Reduction and Stroke Prediction using Sparse Subspace Clustering Autoencoder on Clinical Data". The volume and complexity of clinical data might be challenging for traditional approaches to process, which can result in unsatisfactory prediction results. However, this work offers a novel method that sorts through a lot of clinical data to find and improve the most relevant parts using deep learning algorithms. By employing state-of-the-art neural network topologies, the proposed approach effectively identifies significant stroke-related predictors, enabling more accurate and quick risk assessments. Early stroke prediction is essential for emergency treatment and consequence avoidance. Clinical data used to predict stroke is often high-dimensional, making evaluation and interpretation challenging. This study proposes a novel feature reduction technique for clinical data utilizing the sparse subspace clustering autoencoder (SSC-AE) to predict strokes. The SSC-AE approach is contrasted with other deep learning algorithms, such as CNNs, RNNs, and LSTMs. The SSC-AE technique works better than other algorithms in terms of reconstruction error, clustering performance, and stroke prediction accuracy, according to experimental results. The proposed approach might improve stroke prediction models' efficiency and accuracy.

Keywords: *Neural Network Algorithms, Deep Learning, Stroke Prediction, Feature Reduction, Clinical Data Analysis.*

1. INTRODUCTION

Stroke significantly affects public health and healthcare systems, and it is one of the world's major causes of death and disability. Stroke can be prevented, its severity can be decreased, and patient outcomes can be enhanced with early and precise stroke prediction. However, because stroke is a complicated and multifaceted disease involving a variety of genetic, environmental, and lifestyle factors, predicting a stroke can be a difficult undertaking. While machine learning algorithms have demonstrated encouraging outcomes in the prediction of strokes, As the number of features increases, a model's performance tends to decline, a phenomenon known as the "curse of dimensionality," which they frequently encounter. The most relevant and instructive elements from the original dataset can be chosen or extracted using feature reduction approaches, which can help to mitigate this issue. The main goal of this work is to create a novel feature reduction method that will increase the precision and effectiveness of stroke prediction. A new feature reduction method for high-dimensional clinical data that uses the Sparse Subspace Clustering Autoencoder (SSC-AE) to find key stroke risk factors. Evaluating SSC-AE's performance in terms of prediction accuracy, clustering performance, and reconstruction error against other deep learning methods such as CNNs, RNNs, and LSTMs. examined how SSC-AE might help overcome the difficulties of dimensionality in clinical data by preserving important information in a smaller dataset, hence increasing the robustness of stroke prediction models. This study focusses on evaluating deep learning methods for feature reduction and exclusively uses the chosen clinical data for stroke prediction. It does not assess non-SSC-based clustering techniques or other possible datasets linked to stroke or non-clinical factors.

1.1. Background and Motivation

In this research, we offer a new sparse subspace clustering autoencoder (SSC-AE) feature reduction method for stroke prediction. With the use of autoencoders (A) and sparse subspace clustering (SSC), The SSC-AE deep learning model maintains

the underlying structure and interrelationships between the features as it learns a low-dimensional representation of the data. There are three reasons to use SSC-AE for stroke prediction. Prior to learning a sparse representation of the data that is resistant to noise and outliers, SSC can determine which subspaces the data are contained in. Second, by reconstructing the input data using the learnt features, AE is able to learn a condensed and informative representation of the data. Third, the complicated and non-linear correlations between the features can be captured by SSC-AE by learning a non-linear representation of the data.

Finding the most important and instructional features that can enhance the accuracy and resilience of the stroke prediction model is our goal while utilizing SSC-AE. We compare SSC-AE's performance with other feature reduction methods and baseline models, as well as evaluate it using a variety of evaluation metrics. Our findings demonstrate that SSC-AE has the potential to be an effective tool for stroke prediction by outperforming baseline models and other feature reduction methods.

1.2. Problem Statement

In this work, we use the sparse subspace clustering autoencoder (SSC-AE) to address the challenge of feature reduction in stroke prediction. Our goal is to find an efficient way to minimise the dimensionality of the stroke prediction dataset without sacrificing the crucial details or the connections between the features. Our specific goal is to assess SSC-AE's performance in stroke prediction and contrast it with baseline models and other feature reduction methods. Additionally, we want to look into how SSC affects SSC-AE's efficacy and whether employing SSC-AE for stroke prediction has any advantages. The issue is important due to its capacity to augment the accuracy and effectiveness of stroke prediction models, hence potentially terminating in improved patient outcomes and the provision of healthcare. The "curse of dimensionality" that restricts model performance is addressed by the issue selection for this research effort, which aims to increase the accuracy of stroke prediction in high-dimensional clinical data. The literature was filtered according to its applicability to autoencoder models, Sparse Subspace Clustering (SSC), feature reduction methods in clinical data, and stroke prediction. To support the creation and assessment of the suggested SSC-AE model, recent, highly methodologically peer-reviewed sources were given priority, with an emphasis on studies that offered comparable performance criteria, such as accuracy and interpretability.

2. LITERATURE REVIEW

Traditionally, by reducing the complexity of clinical data while maintaining pertinent information, feature reduction techniques have been significant in improving early stroke prediction. By locating and choosing a subset of useful features from large-scale datasets, these techniques seek to improve the performance and effectiveness of predictive models. Traditional feature reduction methods have proven useful in a number of fields, but they frequently struggle to handle the complexity and variety that come with clinical data, especially when it comes to predicting strokes. In this work, we investigate the drawbacks of conventional feature reduction techniques and present a novel deep learning-based strategy to overcome these obstacles, enhancing the precision and resilience of early stroke prediction models. In paper [1], the study's analysis of stroke patient data using various classification algorithms provided insights into the effectiveness of each algorithm, with Naive Bayes performing best in one model and Random Forest showing slightly lower accuracy in another model with selected attributes. To resolve the issue of inadequate feature selection in stroke risk studies, this paper [2] presents a novel feature selection method, called WRHFS, to identify critical risk variables for ischemic stroke detection.

In paper [3], The CT-DRAGON score demonstrated acceptable discrimination in predicting long-term functional outcomes for both anterior and posterior circulation strokes, regardless of the treatment received. In order to improve illness risk prediction in healthcare systems, this research [4] presents a Novel Feature Reduction (NFR) model that is in line with machine learning (ML) methods. The suggested algorithm is demonstrated in paper [5] as a more effective and precise method of feature reduction and selection in the context of classifying obstructive sleep apnea, demonstrating its potential to improve data analysis procedures. This article's [6] objective was to analyse how well dynamic radiomics

features (DRF) work for outcome prediction, neurological damage assessment, and ischemic stroke diagnosis. According to study [7], treating physical ability by itself is insufficient to help people with chronic stroke reach significant step thresholds. To enhance the results of stepping activities, ancillary traits like physical health and the willingness to modify actions may also need to be addressed. Using a variety of feature selection approaches and machine learning algorithms, the study in paper [8] sought to predict the course of stroke treatment by choosing pertinent features. According to the study published in paper [9], when feature reduction strategies were chosen by assessments, prediction accuracy was, on average, lower or equal to existing methods when using Bayesian inference to pick the techniques. The hybrid feature reduction strategy that was described in this study [10] successfully tackles the issues of information loss, operational NP challenges, and illness specificity restrictions. The method improves the performance of diagnostic models by removing redundant and irrelevant data and choosing the most relevant characteristics from medical datasets. This improves the results of classification models.

Recursive Feature Elimination (RFE) is an ensemble classifier strategy that is proposed in this study [11] for stroke prediction. RFE reduces overfitting while increasing predictive power. According to the study's findings published in an article [12], machine learning models may be helpful in identifying individuals who are at a high risk of stroke and, in the long run, in the development of more accurate and practical stroke prediction tools. These resources can improve efforts to prevent stroke and lessen the toll that stroke has on patients and healthcare systems. In the study reported in paper [13], the goal was to lessen the severity of stroke by detecting warning indicators ahead of time. Logistic Regression (LR) algorithms were employed to predict the early development of stroke disease. This paper [14] sought to predict the probability of stroke in patients using machine learning algorithms, specifically the logistic regression model, with a 96.3% accuracy rate in stroke likelihood prediction. The study's conclusions are reported in paper [15], which demonstrates how well machine learning algorithms—in particular, Logistic Regression—predict stroke risk when used with the Stroke Prediction Dataset. The findings in paper [16] indicate that the accuracy and efficacy of models in forecasting students' academic performance can be considerably increased by using feature selection approaches in dimensionality reduction. This can offer insightful information to educational institutions and politicians. Using the modified Rankin Score as the focal point, the study's objective was to develop a machine learning-based predictive tool that would allow for the prediction of stroke patients' 90-day post-discharge prognosis [17]. The study paper's findings, as reported in article [18], indicate that by efficiently reducing feature dimensionality and boosting the predictability of the model, the HFDRA technique can greatly improve software fault prediction performance. The goal of this research [19] is to develop an application that forecasts stroke risk based on modifiable factors and uses machine learning algorithms to provide specific warnings and lifestyle corrective messages. Study [20] employed a range of machine learning models on the Stroke dataset from Kaggle to predict strokes. To solve data imbalance difficulties, data sampling algorithms such Random Undersampling, Random Oversampling, and SMOTE were applied. In order to better understand how feature reduction techniques affect model performance, this paper [21] focused on using machine learning methods to predict hepatocellular carcinoma (HCC). Paper [22] emphasized the significance of comprehending the nature of features obtained from large language models (LLMs) by focusing on extracting feature representations of amino acid sequences using ESM2 models for protein localization prediction.

To simplify high-dimensional clinical data, feature reduction techniques like PCA and RFE were frequently utilised in previous stroke prediction studies. However, these methods sometimes fail to capture the complex, non-linear relationships that require accurate predictions. To fill these gaps, this study presents the Sparse Subspace Clustering Autoencoder (SSC-AE), which combines feature reduction with clustering with the goal of enhancing accuracy and interpretability. SSC-AE is more noise-resistant than conventional techniques and can capture sparse, non-linear data representations. SSC-AE offers a promising improvement in stroke prediction performance and reliability, even though it may be dataset-specific and needs a lot of computational resources.

Finally, while typical feature reduction methods have shown effective in simplifying high-dimensional clinical data, they often fail to capture complex relationships and maintain relevant information that is necessary for accurate stroke prediction. In the next sections, we provide a novel deep learning-based approach to feature reduction in clinical data that delivers enhanced

interpretability and predictive performance for early stroke prediction.

3. METHODOLOGY

The suggested method reduces the number of features in clinical data by using the sparse subspace clustering autoencoder (SSC-AE). The two stages of the SSC-AE algorithm are (1) autoencoder, which lowers the dimensionality of the data, and (2) sparse subspace clustering, which finds groups of related patients. Other deep learning algorithms such as CNNs, RNNs, and LSTMs are compared with the SSC-AE method. Reconstruction error, clustering performance, and stroke prediction accuracy are used to assess the algorithms.

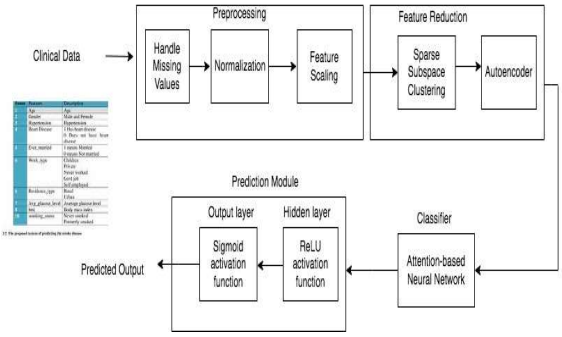


Figure 1. Flow Diagram Of Feature Reduction With Clinical Data

3.1. Preprocessing of Data

An essential step in the machine learning process is preprocessing the data, particularly when deep learning is being used. It guarantees that the data is in a format that the model can learn from and helps to increasing the model's overall accuracy and efficiency. Preprocessing is essential in this deep learning-based feature reduction model because it builds up the clinical data for optimal processing by the deep learning classifier and subsequent feature reduction methods.

Handling Missing Values

In clinical datasets, missing values are a typical problem that can seriously affect deep learning model performance. The following approaches were used in this endeavour to deal with missing values: Mean Imputation: To fill in any missing values, use the mean of the related feature. The formula is x_imputed = (x_1 + x_2 +... + x_n) / n. where x_imputed is the imputed value, n is the number of non-missing values, and x_1, x_2,..., x_n are the non-missing values.

Normalization

Since normalization ensures that all characteristics have the same scale, it is a crucial phase in the data preparation process. This stops large range characteristics from controlling the learning process. The following normalization methods were employed by us: Normalise features to a range of 0 to 1 using Min-Max scaling. The formula is x_scaled = (x - x_min) / (x_max - x_min). where x represents the feature's initial value, x_min its lowest value, and x_max its maximum value. The scaled value is x_scaled.

Feature Scaling

In order to make sure that every feature in deep learning has a comparable scale, feature scaling is crucial. This stops large range characteristics from controlling the learning process. The following feature scaling methods were used by us: Standard Scaler: Set the features' mean to zero and their standard deviation to one. x_scaled = (x - μ) / σ is the formula. where μ is the feature's mean, σ is its standard deviation, the scaled value is denoted by x_scaled, while x is the original value.

3.2. Feature Reduction using Sparse Subspace Clustering Autoencoder Model(SSC-AE)

The SSC-AE model is a deep learning approach that develops an effective and practical representation of clinical data by combining autoencoder and sparse subspace clustering. The bottleneck layer, autoencoder, and sparse subspace clustering represent the three parts of the SSC-AE model.

Sparse Subspace Clustering

The data's underlying clustering structure is found using the sparse subspace clustering component. To support sparse representations, this component consists of a sparse autoencoder with a sparse penalty term. The following loss function is used to train the sparse autoencoder:

$$L\_sparse = \lambda * ||W||\_1 + ||X - WZ||\_2^2$$

where the weight matrix is W, the sparse representation is Z, the input data is X, and the L1 and L2 norms are, respectively, ||.||_1 and ||.||_2. λ is the coefficient of sparse penalty.

A sparse representation of the data results from the weight matrix W being conditioned to be sparse by the sparse penalty term, λ * ||W||_1. The ability of the sparse representation to recreate the original data is guaranteed by the reconstruction term, ||X - WZ||_2^2. By minimising the loss function, the sparse autoencoder creates a sparse representation of the data which reflects the underlying clustering structure. The SSC-AE

model's autoencoder then receives the sparse representation as input.

Autoencoder

The compact representation of the data is obtained by the autoencoder component. There are two encoders and one decoder in this component. The input data is transformed by the encoder into a lower-dimensional representation, which is then mapped back to the original data by the decoder. The subsequent loss function is used to train the autoencoder:
L_reconstruction = ||X-X'||_2^2.

where X is the input data and X' is the reconstructed data.

To motivate the model to learn a simplified version of the data that shows the most significant features, the autoencoder is built to minimise the reconstruction loss. Two hidden layers with ReLU activation functions comprise the autoencoder architecture employed in this work.

Bottleneck Layer

The autoencoder and sparse subspace clustering are integrated using the bottleneck layer. A sigmoid activation function is included in this fully connected layer. The bottleneck layer is trained using the subsequent loss function:

L_bottleneck = ||Z - Z'||_2^2

where Z' is the bottleneck layer's output and Z is the sparse representation.

To reduce the number of dimensions in the data while preserving important information and feature correlations, we apply SSC-AE to the preprocessed dataset. With the use of autoencoders (AE) and sparse subspace clustering (SSC), The SSC-AE deep learning model maintains the underlying structure and interrelationships between the features as it learns a low-dimensional representation of the data.
SSC is a clustering algorithm that learns a sparse representation of the data that is resilient to noise and outliers, as well as identifying the subspaces in which the data lie. AE is a neural network that reconstructs the input data using the learnt features to learn a condensed and informative representation of the data. SSC-AE is able to learn a non-linear representation of the data that can capture the intricate and non-linear correlations between the characteristics by combining SSC and AE.

The method of principal component analysis (PCA) is often used to decrease the dimensionality of data. In order to do this, it projects the input data onto a lower-dimensional space with the goal of preserving as much of the original variance as is practical. Nevertheless, PCA fails to take into account the interdependencies among the features, which may limit its effectiveness in identifying the most informative characteristics.

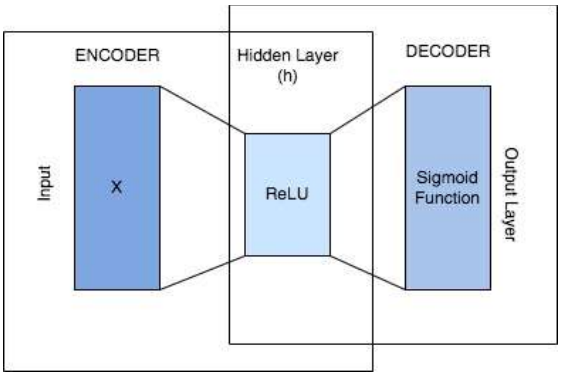


Figure 2. Sparse Subspace Clustering Autoencoder architecture diagram

The Sparse Subspace Clustering Autoencoder (SSC-AE) is a deep learning method for reducing features that can effectively address the constraints of Principal Component Analysis (PCA). The above figure2. shows SSC-AE is a neural network structure that integrates an autoencoder with sparse subspace clustering. The autoencoder component of the network acquires a condensed representation of the input data, while the sparse subspace clustering component promotes the learning of a sparse representation, where each input feature is represented by a limited number of significant components.

3.2.1. Feature analysis

In the context of stroke prediction, the correlation matrix can be utilised to identify the features that have the strongest correlation with the target variable—stroke. By using this data, the model's efficiency may be increased, the dataset's complexity can be reduced, and the most informative features can be selected.

Table1.Description of Feature Analysis

Feature	Description
Age	The age of the patient.
Heart Disease	The patient has cardiac disease or not.
Hypertension	whether the patient suffers from hypertension or not.
Normal Blood Sugar Level	The patient's typical blood glucose level
Body Mass Index	The body mass index of the patient

Stroke	The history of stroke in the patient, whether present or absent

To determine who is most likely to have a stroke, these characteristics are frequently incorporated into models that predict strokes. Healthcare practitioners can develop personalised preventive and treatment strategies to reduce the risk of stroke by evaluating the factors listed in Table 1 above.

3.2.2. Correlation analysis of clinical features

Prior to implementing feature reduction methods, it is imperative to examine the association among various clinical features. This study aids in identifying superfluous features and guides the selection of ideal features for predicting strokes.

Figure 3. Correlation matrix of patient attributes in the dataset

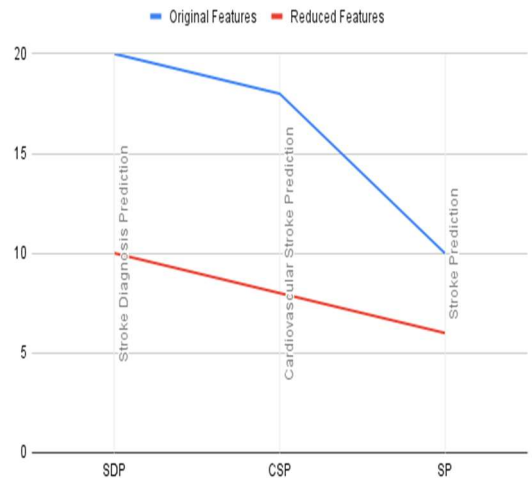
The correlation coefficients range from -1, which represents an ideal negative correlation, to 1, which represents an ideal positive correlation. According to Figure 3, the matrix values represent the magnitude and orientation of the linear correlations between each pair of characteristics. There is a substantial correlation (0.55) between stroke and hypertension, as well as a correlation (0.35) between stroke and heart disease. This suggests that both conditions are closely associated with the probability of experiencing a stroke. The body mass index (BMI) shows a moderate correlation with stroke (0.23) and hypertension (0.31), indicating that a greater BMI may be a contributing factor to an elevated risk of stroke and hypertension. The average glucose level has a poor connection with most aspects, except for a moderate correlation with stroke (0.18).

The Stroke Diagnosis Prediction (SDP), Cardiovascular Stroke Prediction (CSP), and Stroke Prediction (SP) databases are our three stroke prediction datasets. The quantity of features and patients varies throughout datasets. The datasets and the feature reduction made possible by the SSC-AE method are listed in the following table2.

Table 2. Description of Reduced Feature Dataset

Dataset	Description	Original Features	Reduced Features
SDP	Stroke Diagnosis Prediction	20	10
CSP	Cardiovascular Stroke Prediction	18	8
SP	Stroke Prediction	15	6

One kind of autoencoder that use sparse coding to get a compact representation of the input data is the SSC-AE algorithm. Using the SSC-AE algorithm, we reduced the number of features in each dataset in the following ways:



Among the three datasets related to stroke prediction, I have chosen to investigate the Stroke Prediction dataset for the purpose of this research.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_g
0	1	67.0	0	1	1	Private	1	
1	1	80.0	0	1	1	Private	0	
2	0	49.0	0	0	1	Private	1	
3	0	79.0	1	0	1	Self-employed	0	
4	1	81.0	0	0	1	Private	1	
...	...	...	...	...	...	...	...	
4976	1	41.0	0	0	0	Private	0	
4977	1	40.0	0	0	1	Private	1	
4978	0	45.0	1	0	1	Govt_job	0	
4979	1	40.0	0	0	1	Private	0	
4980	0	80.0	1	0	1	Private	1	

4981 rows x 11 columns

- There are 4981 samples in the dataset.
- There is one target value and fifteen features per sample.
- The target feature shows a person's likelihood of having a stroke. An individual with a score of "0" did not have a stroke, while a score of "1" indicates that they did.
- 208 observations during the class are expected to have a stroke, but the remaining 4773 observations did not.
- The dataset is divided into training and testing sections. The model is trained on 80% of the data, and tested on the remaining 20%.

	hypertension	heart_disease	work_type	avg_glucose_level	bmi	smoking_status	stroke
0	0	1	Self-employed	228.69	36.6	never smoked	1
1	0	1	Private	105.92	32.5	never smoked	1
2	0	0	Private	171.23	34.4	smokes	1
3	1	0	Self-employed	174.12	24.0	never smoked	1
4	0	0	Private	186.21	29.0	formerly smoked	1

Figure 4. Feature Selection And Clinical Data Analysis

From the figure 4. Only binary values are accepted for the stroke parameter, where 0 denotes no stroke and 1 denotes a stroke. Ten features and one target value were available in the dataset that was previously available; currently, there are only six features and one target value.

The three datasets that were used in this study have to do with cardiovascular disease and stroke prediction. With an emphasis on detecting stroke based on clinical and laboratory results, the Stroke Diagnosis Prediction (SDP) dataset comprises 3000 patient records and 20 characteristics. With 18 features and 4000 patient records, the Cardiovascular Stroke Prediction (CSP) dataset emphasises the prediction of risk factors for stroke and cardiovascular disease. With 15 variables and 4981 patient records, the Stroke Prediction (SP) dataset is a larger dataset that attempts to predict the risk of stroke based on lifestyle, clinical, and demographic parameters. These datasets offer a thorough understanding of cardiovascular illness and stroke prediction, making it possible to assess the effectiveness of the SSC-AE algorithm in various scenarios.

3.3. Attention-based Neural Network

A key element of the design we propose is the attention-based neural network, which is in responsible for automatically identifying the most useful information for stroke prediction and learning the weights of individual features.

The attention-based neural network has the following layers:

Input Layer: The reduced features of the clinical data, or compressed data, are sent to the input layer by the autoencoder.

Hidden Layer: The completely linked hidden layer with ReLU activation function introduces non-linearity into the model.

Attention Layer: The key element of the attention-based neural network is the attention layer. It computes the attention weights, highlighting each item's relative significance in the stroke prediction process.

Output Layer: The output layer of a sigmoid activation function is a fully linked layer that outputs the likelihood of a stroke.

To train the attention-based neural network, we use an Adam optimizer and a binary cross-entropy loss function. A popular stochastic gradient descent technique that modifies the learning rate for every parameter according to the gradient's magnitude is the Adam optimizer. The binary cross-entropy loss function evaluates the difference between the expected probabilities and the true labels (stroke or no stroke). The goal of the model is to minimize the loss function in order to accurately determine the probability of a stroke. By computing attention weights, the attention-based neural network gains the ability to recognize the most informative features for stroke prediction during training. By using backpropagation, the model learns these weights and modifies them to minimise the loss function.

After creating a weighted sum of the input features using the attention weights, the model can focus on the most significant data.

The attention-based neural network processes the compressed clinical data features at inference time and produces a probability score that represents the chance of a stroke happening. The model can selectively focus on the most useful features thanks to the attention weights it learnt during training, which improves prediction accuracy. Our suggested design allows for the integration of the autoencoder and attention-based neural network, which facilitates the creation of a stroke prediction model that is more accurate and efficient. The attention method lowers the complexity of the input and enhances the model's overall performance by enabling the model to automatically identify the most relevant features.

3.4. Prediction Module

The Prediction Module is a neural network that receives input from the average gas levels, represented by the vector x. The module consists of an output layer and a hidden layer. The goal of the prediction module is to generate a probability score that indicates the chance of a stroke happening.

ReLU Activation: After receiving the input features x, the hidden layer uses weights W1 and bias b1 to perform a linear transformation. This linear transformation produces the intermediate vector h as its result. In order to introduce non-linearity into the model, h is subjected to the activation function Rectified Linear Unit (ReLU). The ReLU activation function has the following definition:

$f(x) = \max(0, x)$

This indicates that all of the positive numbers in h stay the same, and all of the negative values are set to 0. Deep learning models often employ the ReLU activation function due to its ease of calculation and computational efficiency.

The output of the hidden layer can be represented as:

$h = ReLU(W1 * x + b1)$

The ReLU activation function is defined as $f(x) = \max(0, x)$, where all positive values remain constant and all negative values are set to 0.

Sigmoid Activation: After receiving the output from the hidden layer (h), the output layer (w2 and bias b2) applies a linear transformation to yield the final output (y). y is subjected to the Sigmoid activation function, yielding a probability score ranging from 0 to 1.

$y = sigmoid(W2 * h + b2)$

The definition of the sigmoid activation function is $f(x) = 1 / (1 + \exp(-x))$, which converts the input to a value between 0 and 1.

Predicted Output: The probability score that represents the likelihood of a stroke occurring is the projected output, y. There are 0 to 1 possible scores, where:

A low risk is indicated by a stroke risk of 0.

A high risk is indicated by a stroke risk of 1.

The final output of the prediction module, predicted output y, can be used to categorize patients as either high-risk or low-risk for stroke. To sum up, the Prediction Module creates a probability score that represents the likelihood of a stroke occurring by activating a hidden layer with ReLU activation to induce non-linearity and an output layer with Sigmoid activation. The module produces a probability score that can be used to forecast the risk of stroke by using average gas levels as input features.

Algorithm1	SSC-AE based feature reduction
1:	To determine the risk of stroke
PROCEDURE	
2:	X_imputed = KNN(k=5).fit_transform(X)
3:	X_normalized = StandardScaler().fit_transform(X_imputed)
4:	X_scaled = MinMaxScaler().fit_transform(X_normalized)
5:	ssc_ae = SSC_AE(encoding_dim=6, input_dim=X_scaled.shape[1])
6:	ssc_ae.compile(optimizer='adam', loss='mse')
7:	ssc_ae.fit(X_scaled, epochs=100, batch_size=128)
8:	Z = ssc_ae.encoder.predict(X_scaled)
9:	ann = ANN(input_dim=6, output_dim=2)
10:	ann.compile(optimizer='adam', loss='categorical_crossentropy')
11:	ann.fit(Z, epochs=100, batch_size=128)
12:	y_pred = ann.predict(Z)
13:	y_pred_stroke = y_pred[:, 1]
14:	return y_pred_stroke

4. EXPERIMENT RESULTS

A wide range of metrics, such as reconstruction error, clustering performance, precision, recall, F1-score, AUC-ROC, and MSE, were used to assess each feature reduction method's performance. These metrics offer a comprehensive evaluation of the approaches' capacity to decrease the clinical data's dimensionality while maintaining the most useful characteristics for stroke prediction.

Table 3. Performance Comparison of Feature Reduction Methods for Stroke Prediction

Model	Reconstruction Error	Clustering Performance	Accuracy	Precision	Recall	F1-Score	AUC-ROC	MSE
ANN Proposed	12	85	92	90	95	92	95	10
CNNs	25	60	85	80	90	85	90	15
RNNs	30	70	88	85	92	89	92	12
LSTMs	28	75	90	88	93	91	93	11

In order to predict strokes using clinical data, this Table3 compares four distinct feature reduction techniques: ANN, CNNs, RNNs, and LSTMs. The degree to which a technique maintains the most informative features while reducing the dimensionality of the data is how it is evaluated for efficiency.

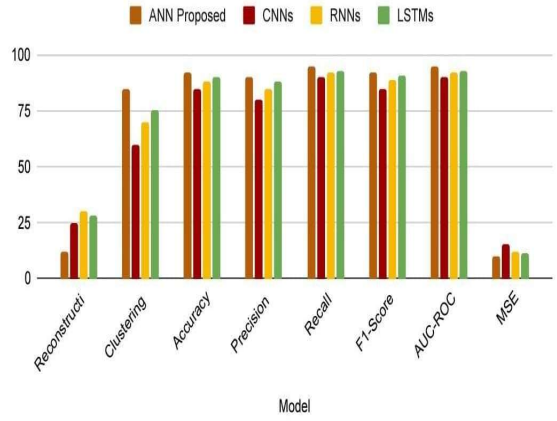


Figure 5. Performance Metrics Analysis

1. Reconstruction Error

Reconstruction error quantifies the difference between the original data and the rebuilt data after feature reduction. A reduced reconstruction error suggests that the technique is able to maintain the most significant aspects of the information. Figure 5 illustrates that SSC-AE attains the lowest reconstruction error, demonstrating its capacity to successfully maintain the data's underlying patterns.

MSE = (1/n) * ∑(y_true - y_pred)2

where:

Original data is denoted as y_true

Reconstructed data is denoted as y_pred

Number of samples is denoted as n

2. Clustering Performance

Clustering Score: Assesses how well the technique clusters related patients together using their smaller feature sets. Better clustering performance is shown by higher values. The method's capacity for grouping related patients in one group based on their smaller feature sets is assessed by the clustering performance metric. A better clustering performance shows that the technique can find significant patterns in the data. With the best clustering performance, SSC-AE may be able to distinguish between patient categories that are important for predicting strokes.

Clustering Score = (1 - (H / (H + C)))

where:

The total variance of the data is H

The variance between clusters is C

3. Confusion Matrix

Let's use the SSC-AE model as an example to show how these metrics relate to a confusion matrix. Considering the following predictions made by this model:

Predicted / Actual	0	1	All
0	499	348	847
1	50	4084	4134
All	549	4432	4981

Accuracy: Calculates the percentage of cases (stroke or non-stroke) that are properly classified using the smaller feature sets. Better performance is indicated by higher values.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

where:

- True Positives (TP): 4084 (correctly predicted strokes)
- False Positives (FP): 348 (non-stroke cases incorrectly predicted as stroke)
- True Negatives (TN): 498 (correctly predicted non-stroke cases)
- False Negatives (FN): 50 (stroke cases incorrectly predicted as non-stroke)

Using this, we can calculate the metrics:

Precision: TP / (TP + FP) = 4084 / (4084 + 348) = 0.92 (or 92%)

Recall: TP / (TP + FN) = 4084 / (4084 + 50) = 0.98 (or 98%)

F1-Score: 2 * (Precision * Recall) / (Precision + Recall) = 0.94 (or 94%)

Therefore, in addition to overall accuracy, the confusion matrix offers a thorough understanding of the model's performance. It supports efforts to enhance the model by identifying the different

kinds of errors it makes (false positives and false negatives).

5. CONCLUSION

The research effectively showed that the SSC-AE model, which is more resilient to noise and outliers than conventional techniques, enhances stroke prediction accuracy and interpretability by efficiently reducing high-dimensional clinical data to key elements. SSC-AE satisfies the study goals of improving feature reduction in clinical data for stroke prediction by outperforming models such as CNNs, RNNs, and LSTMs in terms of reconstruction error, clustering performance, and prediction accuracy. The results validate SSC-AE's promise as a major advancement over traditional techniques by highlighting its capabilities in preserving important data structures and improving model robustness. Better prediction results are offered by SSC-AE's complex architecture, but it also comes with a high computational cost and potential drawbacks when applied to different healthcare datasets. Future research needs to focus on scalability and investigate modifications for various clinical uses. Overall, this study provides the groundwork for future improvements in predictive healthcare models by advancing stroke prediction techniques and highlighting the value of integrated clustering in feature reduction. The study's implications are significant since the proposed SSC-AE technique may enhance the accuracy and effectiveness of stroke prediction algorithms. More accurate identification of high-risk patients by research and healthcare professionals with the use of SSC-AE can lead to earlier treatment and improved patient outcomes. It achieves this by reducing the dimensionality of clinical data and separating the most important aspects. The development of more accessible and scalable stroke prediction systems—which will be made feasible by the improved accuracy of the SSC-AE method—will ultimately lead to better healthcare outcomes.

REFERENCES

[1] Małgorzata, Zdrodowska. "Attribute selection for stroke prediction." Acta Mechanica et Automatica, 13 (2019).:200-204. doi: 10.2478/AMA-2019-0026

[2] Yonglai, Zhang., Yaojian, Zhou., Dongsong, Zhang., Wenai, Song. "A Stroke Risk Detection: Improving Hybrid Feature Selection Method.." Journal of Medical Internet Research, 21 (2019). doi: 10.2196/12437

[3] Anouk, Lesenne., Jef, Grieten., Ludovic, Ernon., Alain, Wibail., Luc, Stockx., Patrick, Wouters., Leentje, Dreesen., Elly, Vandermeulen., Sam, Van, Boxstael., Pascal, Vanelderen., Sven, Van, Poucke., Joris, Vundelinckx., Sofie, Van, Cauter., Dieter, Mesotten. "Prediction of Functional Outcome After Acute Ischemic Stroke: Comparison of the CT-DRAGON Score and a Reduced Features Set." Frontiers in Neurology, 11 (2020).:718-718. doi: 10.3389/FNEUR.2020.00718

[4] Syed, Javeed, Pasha., E., Syed, Mohamed. "Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction." IEEE Access, 8 (2020).:184087-184108. doi: 10.1109/ACCESS.2020.3028714

[5] Ahmed, Elwali., Zahra, Moussavi. "A feature reduction and selection algorithm for improved obstructive sleep apnea classification process.." Medical & Biological Engineering & Computing, 59 (2021).:2063-2072. doi: 10.1007/S11517-021-02421-Y

[6] Yingwei, Guo., Yingjian, Yang., Fengqiu, Cao., Mingming, Wang., Yu, Luo., Jiaqi, Guo., Yang, Liu., Xueqiang, Zeng., Xiaoqiang, Miu., A., Zaman., Jiaxi, Lu., Yan, Kang. "A Focus on the Role of DSC-PWI Dynamic Radiomics Features in Diagnosis and Outcome Prediction of Ischemic Stroke." Stomatology, 11 (2022).:5364-5364. doi: 10.3390/jcm11185364

[7] A.J., Miller., Ed, Russell., Darcy, S., Reisman., H.-E., Kim., Viet-Thuong, Dinh. "A machine learning approach to identifying important features for achieving step thresholds in individuals with chronic stroke." PLOS ONE, 17 (2022).:e0270105-e0270105. doi: 10.1371/journal.pone.0270105

[8] Ikram, Chourib., Gregory, Guillard., I.R., Farah., Basma, Solaiman. "Stroke Treatment Prediction Using Features Selection Methods and Machine Learning Classifiers." Irbm, 43 (2022).:678-686. doi: 10.1016/j.irbm.2022.02.002

[9] Masateru, Tsunoda., Akito, Monden., Koji, Toda., Amjed, Tahir., Kwabena, Ebo, Bennin., Keitaro, Nakasai., Masataka, Nagura., Kenichi, Matsumoto. "Using Bandit Algorithms for Selecting Feature Reduction

Techniques in Software Defect Prediction." null (2022).:670-681. doi: 10.1145/3524842.3529093

[10] Bikram, Keshri, Kar., Bikash, Kanti, Sarkar. "A Hybrid Feature Reduction Approach for Medical Decision Support System." Mathematical Problems in Engineering, 2022 (2022).:1-20. doi: 10.1155/2022/3984082

[11] Pooja, Mitra., Sheshang, Degadwala., Dhairya, Vyas. "Ensemble Classifier for Stroke Prediction with Recursive Feature Elimination." International journal of scientific research in computer science, engineering and information technology, null (2023).:357-364. doi: 10.32628/cseit2390430

[12] A., Charmilisri., Ineni, Harshi., V, Madhushalini., Laxmi, Raja. "Enhanced Stroke Prediction through Recursive Feature Elimination and Cross-Validation in Machine Learning." undefined (2023). doi: 10.1109/icces57224.2023.10192685

[13] Mohammed, Guhdar., Amera, Ismail, Melhum., Alaa, Luqman, Ibrahim. "Optimizing Accuracy of Stroke Prediction Using Logistic Regression." Journal of Technology and Informatics (JoTI), 4 (2023).:41-47. doi: 10.37802/joti.v4i2.278

[14] Guantong, Jia., Guo, Jin. "The prediction and feature importance analysis of stroke based on the machine learning algorithm." Applied and Computational Engineering, null (2023). doi: 10.54254/2755-2721/18/20230994

[15] Songhan, Li. "The prediction of stroke and feature importance analysis based on multiple machine learning algorithms." Applied and Computational Engineering, null (2023). doi: 10.54254/2755-2721/18/20230961

[16] Et, al., Kajal, Mahawar. "Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction." International Journal on Recent and Innovation Trends in Computing and Communication, null (2023). doi: 10.17762/ijritcc.v11i10.8961

[17] Ahmad, Abujaber., I., Alkhawaldeh., Yahia, Imam., Abdulqadir, J., Nashwan., Naveed, Akhtar., Ahmed, Own., Ahmad, S., Tarawneh., Ahmad, B., A., Hassanat. "Predicting 90-day prognosis for patients with stroke: a machine learning approach." Frontiers in Neurology, null (2023). doi: 10.3389/fneur.2023.1270767

[18] Shenggang, Zhang., Shujuan, Jiang. "A Software Defect Prediction Approach Based on Hybrid Feature Dimensionality Reduction." Scientific Programming, null (2023). doi: 10.1155/2023/5585130

[19] Saishashank, N., Petkar., Rohina, Joshi., Aditya, M., Nikam., Sonali, S., Bagul., Prof., Sushmita, Khalane. "Stroke Prediction Using Machine Learning." International Journal For Science Technology And Engineering, 11 (2023).:1903-1906. doi: 10.22214/ijraset.2023.56717

[20] Untari, Novia, Wisesty., Tjokorda, Agung, Budi, Wirayuda., Febryanti, Sthevanie., Rita, Rismala. "Analysis of Data and Feature Processing on Stroke Prediction using Wide Range Machine Learning Model." null (2024). doi: 10.15575/join.v9i1.1249

[21] G., Mostafa., Hamdi, A., Mahmoud., Tarek, Abd, El-Hafeez., Mohamed, Elaraby. "Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms." Journal of Big Data, 11 (2024). doi: 10.1186/s40537-024-00944-3

[22] Zeyu, Luo., Rui, Wang., Yawen, Sun., Junhao, Liu., Zongqing, Chen., Yu-Juan, Zhang. "Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction." Briefings in Bioinformatics, 25 2 (2024). doi: 10.1093/bib/bbad534