Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Accurate Prediction Of Stroke Through Concatenated Gated Recurrent Unit And Adaboost Convolutional Neural Network

R.Punitha Lakshmi [1*], V.Vijayalakshmi [2], Praveen Zayaraz[3], Sunil Babu Melingi[4]
[1*]Department of ECE Puducherry Technological University Puducherry, Indiapunitha278@ptuniv.edu.in
[2]Department of ECE Puducherry Technological University Puducherry, Indiavvijizai@ptuniv.edu.in
[3]Department of CSE Puducherry Technological University Puducherry, Indiapraveenzayaraz@ptuniv.edu.in
[4]Department of ECE K L University Guntur, Indiasunil.babu.m.@gmail.com

ARTICLE INFO

Received: 20 Dec 2024

Revised: 12 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Stroke, the greatest threat causes an enormous number of death. The medical field employs a range of data mining tools to help with early illness detection and diagnosis. Machine Learning (ML) methods have gained popularity in the prediction, diagnosis, evaluation of this illness; however, because the data are gathered from multiple institutions, there are problems with data quality. The research objective is to enhance the accuracy of stroke data by applying a better pre-processing technique and enhance the prediction of stroke using hybrid adaboost convolutional neural network. An improved method for determining possible risk factors and forecasting the probability of stroke is done using an open access clinical dataset. The dataset has less precise categorization and an excess fitting issue. To minimize the expense of the traditional AdaBoost when working with large sets of training data, an AdaBoost-Convolutional Neural Network (AB-CNN) was developed. The AB-CNN was implemented with minimum number of learning epochs for categorization of different classes of stroke. The proposed work analyses both modifiable and non-modifiable risk factors. Modifiable factors include lifestyle habits like smoking and medical conditions such as high blood pressure, while non-modifiable factors include age, gender, and family history. A combined deep learning model using Gated Recurrent Units (GRUs) and Convolutional Neural Networks (CNNs) is employed to capture patterns from both time-based and spatial data. This approach supports both MRI and clinical data. Thereby the proposed technique reduces the overall processing time and identifies individuals at high risk, enabling healthcare providers to offer timely preventive care.

Keywords--- Stroke Prediction, , Gated Recurrent Units (GRU), Adaboost Convolutional Neural Networks (AB-CNN), Deep Learning, Modifiable Risk Factors, Non-Modifiable Risk Factors, Multimodal Healthcare Dataset

I.INTRODUCTION

The human brain is an important organ that regulates many bodily processes and needs oxygen to convey nerve impulses to every part of the body. Cerebrovascular Disease (CVD) is an illness which impacts the circulation and blood arteries in the brain. The frequent potentially fatal neurological event in the US is CVD, reported to the American Association of Neurological Surgeons [1]. Strokes are the main cause of death in the US amongst CVDs, including aneurysms, vascular , and intracranial stenosis [2]. Stroke mostly happens when a clot restricts blood vessels that nourish the brain, causing harm to the arteries that link to the brain. Still, 80% of strokes can be avoided with early intervention [3].

Depression, emotional dysregulation, and cognitive impairments are warning indicators that emerge days before a stroke [4]. Doctors are able to identify strokes based on these signs and treat patients quickly away to avoid permanent brain damage. Yet, by applying past understanding of risk variables, doctors can forecast the development of a stroke and lessen its repercussions. Among these changeable risk variables are diabetes, smoking, and alcohol intake; nonmodifiable risk variables are age and race [5]. According to the WHO's official study, stroke is the leading causes of disability worldwide, accounting for an estimated 17 million deaths from heart disease and strokes [6].

To eliminate blood flow blockages and restore blood flow to the ischemic tissues, intravenous medication and vascular procedures are the mainstays of current treatments for ischemic stroke research [7]. Magnetic resonance imaging and neuroimaging are the main methods employed for detection of stroke [8].

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In Acute ischemic strokes [9, 10] timely decision-making and early detection is the first step of treatment.
Stroke risk is strongly influenced by a combination of modifiable and non-modifiable risk factors. Modifiable factors include behaviors and conditions that can be altered through lifestyle changes, medication, or medical interventions. Examples include high blood pressure, diabetes, smoking, excessive alcohol use, physical inactivity, and obesity. Studies have shown that addressing these factors through interventions can significantly reduce stroke risk.
Non-modifiable risk factors, which cannot be altered, are also critical in predicting stroke. Age is a primary non-modifiable risk factor, with stroke risk increasing significantly with age. Gender and genetic predispositions also play substantial roles. For example, men are generally at higher risk for stroke than women, although women have a higher lifetime risk due to longevity. Family history of stroke and certain genetic mutations have been linked to increased stroke risk, indicating that non-modifiable factors provide valuable insights into baseline risk levels.
Machine Learning (ML) helps in early detection and remote patient monitoring [11]. It is also important to note that the majority of stroke victims experience motor impairments following their events. ML techniques were employed to predict the likelihood of those results according to evaluation of structural magnetic resonance imaging (MRI) of the heart and brain. Also, AI-driven medical solutions are developing quickly to enable the early detection of stroke. In the treatment of stroke, artificial intelligence is used in several ways, such as decision support, precise diagnosis, and quick diagnosis. By utilizing data from prior medical records, ML can recognize people who are at elevated risk of stroke [7].
The organisation of this paper is as follows: Section 1 gives an introduction on stroke and its types along with the general prediction techniques. An in-depth literature survey for different techniques in the prediction of stroke is discussed in Section 2. Section 3 proposes the integration of GRU and Adaboost Convolutional Neural Network. The performance analysis of the proposed GRU & AB-CNN is executed and compared with the existing results in Section 4. Section 5 concludes the paper. Finally, Section 6 gives the statements and information about the dataset used in the research work.

II.RELATED WORKS

Previous researches have examined several aspects relating to stroke prognosis. Research by Jeena et al. examines a number of risk factors of stroke [12]. The study employed a regression-based methodology to describe the correlation among a given component and its effect on stroke. In Hanifa and Raja [13], the utilization of polynomial functions and the radial basis function in a non-linear support vector categorization system led to an increase in the precision of stroke risk prediction. Four categories were created from the risk factors found in this function: functional, lifestyle, medical, and demographic. In the same way, Luk et al. investigated 878 Chinese participants to see whether age affected the results of stroke recovery [14]. Min et al. [15] created a system to forecast stroke based on risk factors that may be changed. Using the CHS dataset, Singh and Choudhary [16] employed a decision tree method for estimating stroke risk in individuals. To forecast stroke, a DL algorithm built on a feed-forward multi-layer ANN was examined [17]. A comparable approach to develop a smart system to forecast stroke from patient information was discussed in [18–20].
Hung et al. investigated ML and DL systems for predicting strokes from an electronic health claims data [21]. Li et al. [22] employed ML techniques to forecast ischemia and thromboembolism. The outcomes derived from the diverse methodologies demonstrate that a multitude of circumstances can impact the findings of every investigation that is carried out.
The inference points collected from the literature survey are data's collection system, characteristics of the data employed, cleaning technique, restoration of missing values, randomization, and uniformity plays a vital role in the classification accuracy. Consequently, it is critical that scientists determine the relationships and effects of the various input elements in a digital medical record on the accuracy of stroke prediction. Studies in associated fields [3,19] show that determining the critical attributes affects the ML system's overall efficacy. It is important that, initially determine the ideal feature set rather than utilizing every feature in the feature space. Before using an algorithm for categorization, redundant and fully unrelated characteristics to a class to be discover and removed. Thus, from the literature survey it is clear that healthcare data specialists must relate the risk factors documented in electronic health records as it can influence the accuracy of stroke prediction.

III.PROPOSED METHODOLOGY

Fig. 1 shows the proposed flow diagram of Stroke prediction using Gated Recurrent Unit and Adaboost Convolutional Neural Network (GRU & AB-CNN) method. The input dataset is pre-processed using standard scalar approach. The AB-CNN structure relies on selecting the best hyper-parameters, which include

- The quantity of epochs,

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- The ideal quantity of neurons for every layer in the CNN,
- A suitable activation function for layer optimization,
- The suitable optimization method for the network,
- The quantity of precise layers in the CNN,
- The ideal learning rate.

From the literature review it is clear that providing raw datasets for training machine learning models would be inappropriate. In order to address imbalances, heterogeneity, and absent values, preprocessing is necessary in the current raw dataset. Initially, the mean of the element is used to fill in the absent values in the dataset. The subsequent phase involves utilising the Python Label Encoder method to convert the data values for each characteristic to numeric form. A resampling technique is implemented to correct any imbalances in the dataset following the modification of the data. The data will be normalised to a range of 0 to 1 using Standard Scalar at the end of the preprocessing phase.

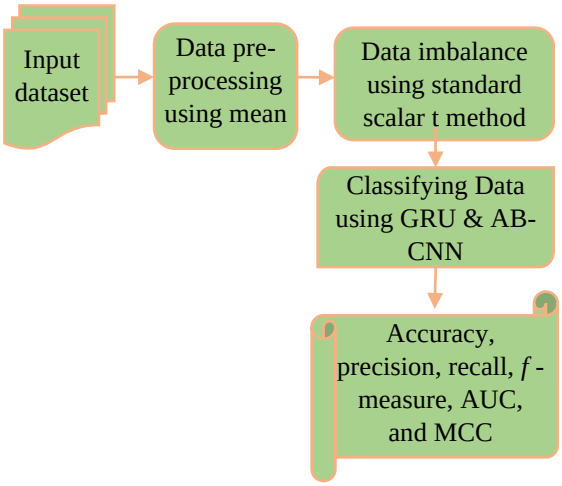

Fig.1: Proposed Methodology of Stroke prediction using GRU & AB-CNN

The novelty of the proposed GRU & AB-CNN shown in fig. 1 is explained as follows:

1. **Addressing Data Imbalances with Resampling**: The proposed work utilizes resampling technique in which the imbalances and diversity in the stroke dataset are eliminated. This ensures that the model is trained on a balanced dataset, which improves its generalization and prediction accuracy, especially when dealing with diverse and complex healthcare data.

2. **Advanced Data Preprocessing Techniques**: The combination of filling missing values with the attribute mean, using Python's Label Encoder, and applying the Standard Scalar for normalization creates an efficient preprocessing technique.
This systematic approach enhances data homogeneity and prepares the dataset for effective model training.

3. **Modification of AdaBoost for CNNs**: Adaptive Boosting, is a ML technique that improves classification accuracy by combining multiple weak models into a single model through iterative training. AdaBoost-CNN enhances CNN by incorporating AdaBoost's iterative weight adjustment strategy, addressing data imbalance, and improving classification accuracy and efficiency.

4. **Adaptive Re-weighting:** AdaBoost's iterative process of adjusting weights based on misclassifications is applied to CNNs. CNNs are generally trained on datasets where each sample is treated equally. In this novel method, each CNN in the ensemble is trained on a re-weighted dataset.

5. **Combining Outputs with a Weighted Sum:** Instead of a simple majority vote, the method utilizes a weighted sum of the CNN predictions, where the weights reflect each CNN's performance. This weighted aggregation allows the model to increase the strength of each CNN in its prediction.

A. Data pre-processing

The multimodal stroke dataset for the proposed technique is obtained from Kaggle's which is a publicly accessible data repository [23]. Data quality is important in predictive analytics because poor data quality

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

produces inaccurate predictions. Real-world health data sets are full of outliers, missing values, and superfluous features. Before utilizing data for modelling, pre-processing is necessary to increase the efficiency. Key features of Dataset include:
1.Demographic information: age, gender, family history
2.Lifestyle factors: smoking status, physical activity level
3.Medical history: hypertension, diabetes, history of cardiovascular diseases
4.Lab results and vitals: blood pressure readings, cholesterol levels, BMI
The dataset includes both modifiable risk factors (e.g., smoking, hypertension) and non-modifiable risk factors (e.g., age, gender), which are essential in developing a comprehensive predictive model.
Data extraction, anonymization, integration, cleansing, outlier identification, and duplication removal are the initial steps in data pre-processing. The minimum values of zero, are replaced with average values. To achieve an accurate detection, the substitute missing values operator is utilized to eliminate any blank entries present.

B. Model Training and Hyperparameter Tuning
The hybrid model is trained using backpropagation and optimized using the Adam optimizer, known for its efficient convergence. Key hyperparameters, such as learning rate, batch size, and dropout rate, are tuned using grid search and cross-validation to maximize model performance. Dropout layers are added to prevent overfitting, ensuring that the model generalizes well on unseen data.

C. Data imbalance using standardscaler T method
The data is normalized to an interval of 0 to 1 utilizing standard scaler. Standard Scaler removes the mean and scales to unit variance to equalize the initial features and provide standardized data.

$$\text{standardscaler} = X' = \frac{X - X_{mean}}{X_{stdev}} \quad (1)$$

where,
X' - Transformed data point
X – Original data point
X_{mean} - Mean of the feature
X_{stdev} – Standard deviation of the feature

D. Proposed GRU & AB-CNN Model Architecture
The proposed stroke prediction model combines Gated Recurrent Units (GRUs) and Convolutional Neural Networks (CNNs) to increase the strength of each approach in analyzing both sequential and structural data. This hybrid architecture is designed to capture temporal changes and spatial relationships among features, which is especially useful in healthcare prediction tasks involving complex, multimodal data. GRUs are a type of Recurrent Neural Network (RNN) that are particularly suited for handling time-series data. They are designed to manage sequential data by maintaining a memory of past information, which is crucial for tasks where prior data can impact future predictions. Unlike traditional RNNs, GRUs have a gating mechanism that helps control the flow of information, making them less prone to issues of vanishing gradients.
CNNs are widely used in image processing and feature extraction due to their ability to capture spatial and structural patterns. CNN layers, through convolution and pooling operations, help reduce data dimensionality while preserving important features, contributing to more efficient learning and higher predictive accuracy. In stroke prediction, traditional AdaBoost algorithms might not be sufficient due to the complex nature of the data. The proposed AdaBoost-CNN modifies the stagewise additive modelling method to be compatible with CNNs, enabling the iterative training of CNNs. This method resolves the distinctive problems associated with CNN architecture, including the computational expense of training and its tendency to overfit on small data subsets. AdaBoost-CNN improves the model's capacity by predicting strokes in complex and imbalanced datasets by concentrating on misclassified cases.
AdaBoost's iterative learning approach is helpful in stroke prediction, as it can be used for improving the accuracy of the prediction model. This approach helps to identify patients with borderline symptoms or unusual data patterns.
The block diagram for Stroke Prediction using AdaBoost's Iterative Learning Approach is explained in the below fig. 2. A detailed explanation of the step-by-step AdaBoost implementation with CNN for stroke prediction is given below:

Step 1: Initial Dataset Collection
Dataset: Multimodal Healthcare Dataset Stroke Data from Kaggle with 43,400 records and 12 features.

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

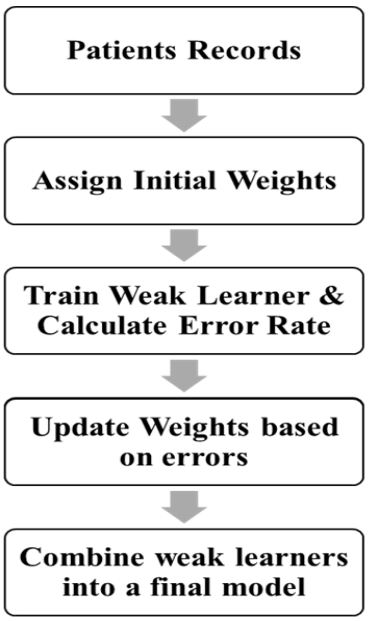

Fig.2 AdaBoost's Iterative Learning Approach for Stroke Prediction

Features: ID, gender, hypertension, medical history, marriage status, occupation, residence category, smoking status, average glucose level, BMI, stroke

Step 2: Preprocess the Data
Preprocessing the data involves three steps like feature scaling, handling missing values, Encoding categorical variables. Feature scaling is essential to ensure that all features, such as glucose levels and BMI, contribute equally to the model by normalizing or standardizing them. Handling missing values is done to prevent issues during training, which can be done by assigning missing values with mean.
Encoding categorical variables is necessary to convert non-numeric features, such as gender and smoking status, into numerical values using one-hot encoding. The encoding results are shown in fig.3. Finally, splitting the dataset into training and validation sets.

	ID	Gender_Female	Gender_Male	Smoking Status_Former Smoker	S
0	1	0	1	0	
1	2	1	0	0	
2	3	1	0	0	
3	4	0	1	1	

Fig. 3. Encoding results of Gender, smoking status using one-hot technique

Step 3: Assign Initial Weights
Initial Weight Assignment: Each patient record is assigned an equal weight initially. There are 43,400 records, each record's initial weight [24], given by eqn. (2)

$WT_i = \frac{1}{N}$ ----- (2)

where,
WT_i – Initial weight assigned
N – Total number of samples

$WT_i = \frac{1}{43,400} \approx 0.000023$

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 4 : Train the First Weak Learner (CNN)

Training: The Convolutional Neural Network for stroke prediction is designed with an input layer comprising 12 neurons, each corresponding to one of the 12 features. This is followed by two dense hidden layers with 64 and 32 neurons. The model utilizes ReLU activation functions to capture complex patterns in the data.
The output layer consists of a single neuron with a sigmoid activation function to provide a probability estimate of stroke occurrence. The model is compiled with binary cross-entropy as the loss function and the Adam optimizer to adjust the network's weights. During training, the CNN is trained on the dataset with initial equal weights for 10 epochs and a batch size of 64.

Step 5: Calculate Error Rate

Compute the error rate of the CNN by summing the weights of the incorrectly classified samples. If the CNN misclassifies 8,680 out of 43,400 records, the error rate given by eqn. (3) $Error_1$ is,

$$Error_1 = \frac{\sum i \text{ misclassified } WTi}{\sum iWTi(0)} \text{ ----------(3)}$$

where,
WT_i – Initial weight assigned

$$Error_1 = \frac{43{,}400 \times 0.000023}{8{,}680 \times 0.000023} \approx 0.2$$

Step 6: Update Weights

Increase the weights of the misclassified samples and decrease the weights of the correctly classified samples using the Weight Update Factor given by eqn. (4)

$$WF_1 = 0.5 \ln \left(\frac{1-Error1}{Error1} \right) \text{ ---------- (4)}$$

where,
WF_i – Weight updated in second iteration

Update Weights for Next Iteration: Continue this process for a total of 10 iterations where each new CNN is trained to correct the errors of its predecessors by focusing on samples that were misclassified in previous rounds.
For misclassified samples weights are updated based on eqn. (4) and for correct samples weights are updated based on eqn. (5)
For misclassified samples:
$WT\ (t+1) = WT(t) \times \exp(WF_1)$ ----------(4)
For correctly classified samples:
$WT\ (t+1) = WT(t) \times \exp(-WF_1)$ ----------(5)

Step 7: Iterative Training

Train a new CNN on the dataset with the updated weights. Repeat the training process for a total of T=10 iterations, where each subsequent CNN is trained with weights adjusted to focus more on previously misclassified samples. Fig. 4 shows the flowchart for Adaboost (Adaptive Boosting) Convolutional Neural Network.

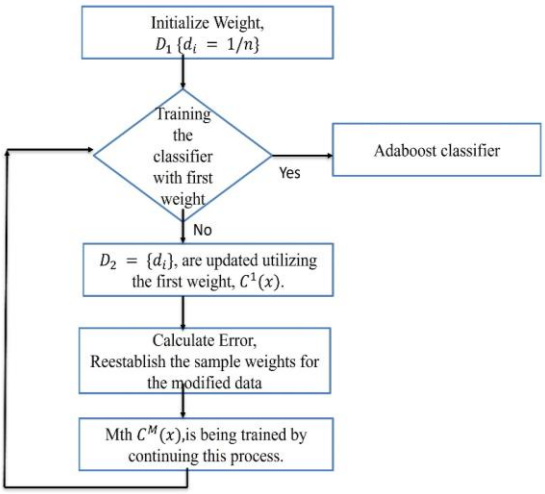

Fig. 4 Flowchart for Adaptive Boosting Convolutional Neural Network

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 8: Combining Weak Learners
Train a new CNN on the dataset with the updated weights. Repeat the training process for a total of T=10 iterations, where each subsequent CNN is trained with weights adjusted to focus more on previously misclassified samples. After training all 10 CNNs, combine their predictions using a weighted sum:
Final Score = $\sum_{i=1}^{T}$ WF ∗ CNN x(i) ---------- (6)
where,
WF – Weight updated
CNN x (i=1 to T) – Total number of iterations (T=10)

This weighted aggregation method leverages the strengths of each CNN, especially those trained on challenging examples, making the final model more accurate and less susceptible to overfitting. Final score is calculated by using eqn. (6)
Prediction Process
Initialize final_score = 0
For t = 1 to T do:
1. Make a prediction using CNN_t
2. Update final_score: final_score = final_score + alpha_t * prediction_t
Determine the final prediction
If final_score >= threshold : Predict "stroke"
Else:
Predict "no stroke"

Step 9: **GRU and AB-CNN layers**
GRU layers process sequential data, with each layer learning temporal dependencies from previous outputs. Dropout layers are incorporated to prevent overfitting and improve generalization.
CNN layers handle spatially-related data, such as relationships between various health indicators. Convolution and pooling layers are applied to extract high-level features, enabling the model to understand how risk factors interact.

Step 10: Fully Connected Layers
Outputs from the GRU and AB-CNN components are concatenated and passed through fully connected layers, enabling the model to combine temporal and spatial features for comprehensive stroke risk prediction.
The final output layer is a sigmoid activation function, which outputs a probability score indicating the likelihood of stroke.

Step 11: Integration of Modifiable and Non-Modifiable Risk Factors
In the hybrid model, modifiable and non-modifiable risk factors are incorporated as distinct inputs. The model processes them separately before combining them in the fully connected layers.
This approach enables the model to recognize the relative importance of different factors, such as modifiable lifestyle habits versus fixed demographic factors, in influencing stroke risk.
The hybrid GRU-AB CNN model architecture thus captures both time-dependent health changes and complex spatial relationships among risk factors. This design is well-suited to handle the diverse data types present in the dataset, offering an effective approach to stroke prediction based on modifiable and non-modifiable risk factors.

IV.SIMULATED RESULTS AND DISCUSSION

The following metrics are employed to assess the suggested framework:
1. **Accuracy:** The classifier's performance is measured by how well it determines the class label for newly entered data values. It was calculated by dividing the overall number of forecasts provided for the input values by the proportion of correct forecasts.
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions made}} \text{ --(7)}$$
2. **Precision:** False Positives are situations in ML and data mining wherein a system mistakenly classifies a negative scenario as a positive one. By calculating the proportion of True Positives to the total quantity of True and False Positives,
$$Precision = \frac{True\ positives}{True\ positives + False\ Negatives} \text{ --- -(8)}$$

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. **Recall:** The recall value is useful in determining how many real Positive situations the framework finds and classifies as True Positives. By calculating the percentage of true positives to the total quantity of false negatives and true positives.

$$Recall = \frac{True\ positives}{True\ positives + False\ Negatives} \quad \text{----(9)}$$

4. **F1 Score:** The F1 score attained by computing the harmonic mean of the precision and recall scores, is employed to assess the test's reliability.

$$F1 = 2 * \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \quad \text{-----(10)}$$

5. **Sensitivity:** The percentage of real positives that are anticipated to be True Positives is known as sensitivity. The percentage of False Positive to the total quantity of True Negative and False Positive is calculated.

$$Sensitivity = \frac{False\ positives}{False\ positives + True\ Negatives} \quad \text{------(11)}$$

6. **Specificity:** The percentage of true negatives that are anticipated from actual negatives is known as specificity. It is calculated by dividing the entire amount of False Negative and True Positives by the proportion of True Positives.

$$Specificity = \frac{True\ positives}{False\ Negatives + True\ Positives} \quad (12)$$

A. Pre-processing Results

The ID of an individual, which is the initial attribute in the stroke dataset, is deleted because it has no impact on whether a patient has had a stroke or not. Now that eleven variables have been added to the dataset, it can be processed further.

The dataset has some missing value cases, which have a important part on the categorization accuracy. The property mean has been utilized to substitute the absent values in the dataset. Substantial imbalances and categorical values for several of the variables are also present in the examined dataset.

ML methods are known to have difficulty with categorical data. As a result, the stroke dataset's must be converted to numerical values. Re-sampling is performed to address the imbalance after every characteristic values. They are converted to numerical values utilizing a Label Encoder. There are 42617 occurrences in which the class label (stroke) value is "0," while only 783 cases have the class label "1." This unbalanced dataset will almost certainly produce biased outcomes if ML methods are applied. By assigning each element an equal weight, normalization attempts at rendering every attribute significant. Simulated results conducted on this pre-processed data are covered in the following subsection.

Table 1 & Fig. 5 & 6, gives the comparison between the suggested design and other systems that are currently in practice. Comparing to existing meta-heuristic techniques for improving system efficiency, the suggested approach chooses the optimal hyper-parameters in an amazing quantity of time.

B. Stroke detection performance

In the stroke detection process each input attribute's provides valuable information about the patient. All the 10 variables are used for the prediction process. The multimodal dataset obtained is found to be imbalanced. Out of 29, 072 patients, only 548 had positive stroke; the remaining 28, 524 patient records have no stroke status.

The dataset's imbalance made it difficult to train any machine learning models. To reduce the impact of the uneven data, a random down sampling technique was utilized. It created two classes, with the remaining 28, 524 observations are taken as the majority class and the remaining 548 observations as the minority class. A balanced dataset of 548 majority and 548 minority data is produced.

There are now 1096 observations total in the balanced dataset. Thirty percent of the balanced dataset was utilized for performance testing, while seventy percent of the dataset was employed for training. The findings are summarized in Table 2 and Fig 7. . Every value is the mean calculated from one hundred trials utilizing random sampling. The majority of the features performed similarly over all metrics.

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

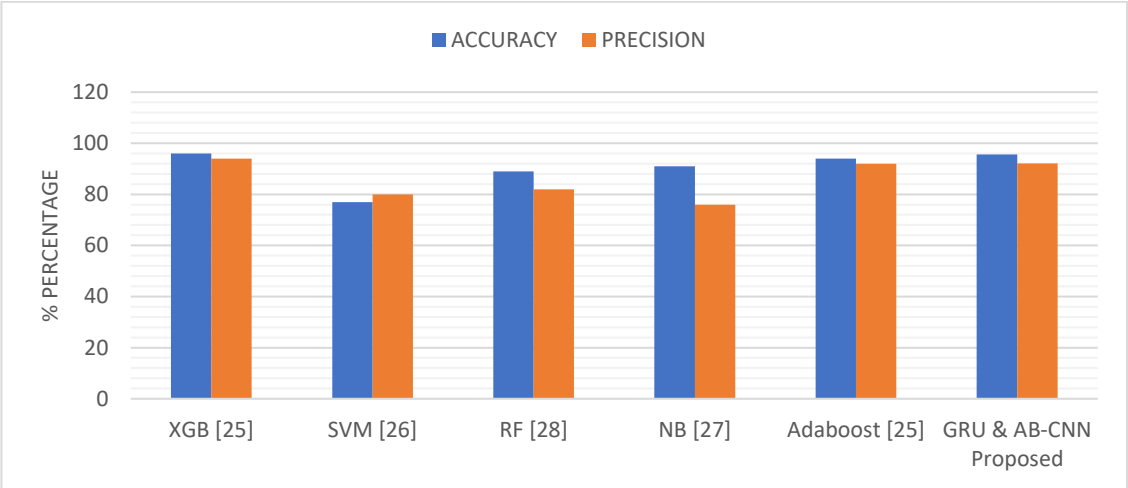

Fig. 5: Comparative Analysis of the various approaches

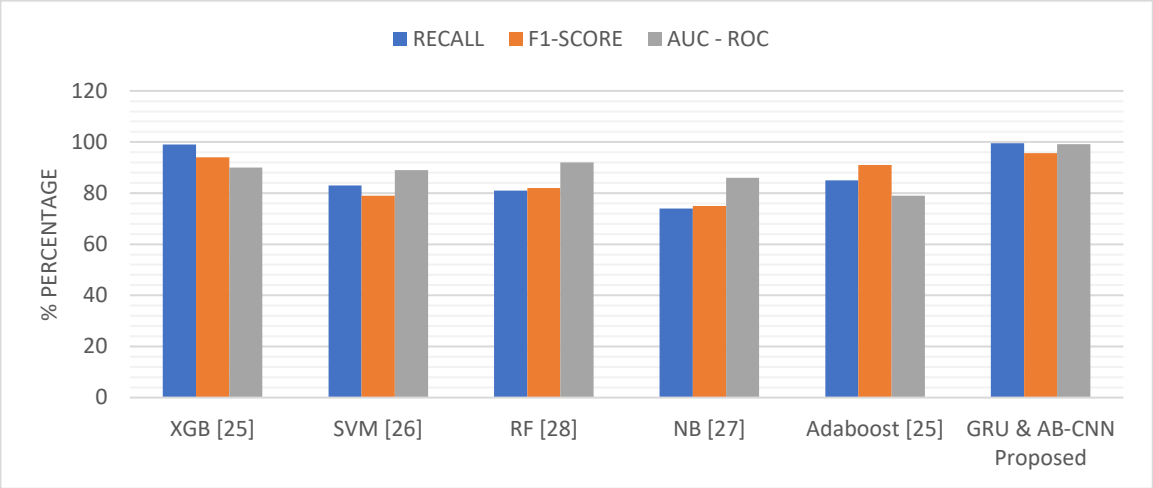

Fig. 6: Comparative Error performance Analysis of the various approaches

Table 1: Comparative Analysis of the proposed with other methods

ALGORITHMS	ACCURACY (in %)	PRECISION (in %)	RECALL (in %)	F1-SCORE (in %)	AUC - ROC (in %)
XGB [25]	96	94	99	94	90
SVM [26]	77	80	83	79	89
RF [28]	89	82	81	82	92
NB [27]	91	76	74	75	86
Adaboost [25]	94	92	85	91	79
GRU & AB-CNN Proposed	95.56	92.12	99.58	95.71	99.2

Table 2: An assessment of the various qualities' efficacy in identifying strokes from medical records. Every value is calculated utilizing the average value derived from 100 tests with random sampling on a balanced dataset.

Feature	Precision	Recall	F-Score	Accuracy
Gender	0.60	0.33	0.31	0.61
Age	0.56	0.43	0.49	0.55
Hypertension	0.54	0.55	0.50	0.54
Heart disease	0.52	0.51	0.39	0.56
Ever married	0.65	0.07	0.13	0.57
Work type	0.55	0.78	0.65	0.55
Residence type	0.55	0.86	0.66	0.56

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Glucose level	0.53	0.53	0.50	0.58
BMI	0.56	0.40	0.35	0.58
Smoking status	0.51	0.41	0.34	0.59

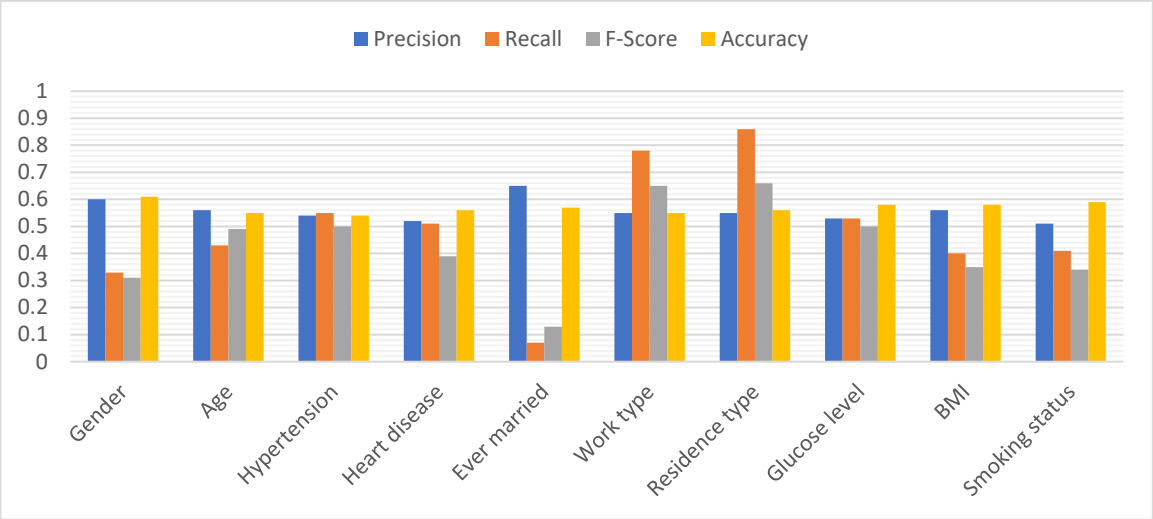

Fig. 7: Risk factor Analysis of the proposed technique.

V.CONCLUSION & FUTURE WORK

The proposed work demonstrates the effectiveness of a hybrid deep learning approach using Gated Recurrent Units (GRU) and Convolutional Neural Networks (CNN) for stroke prediction. By integrating temporal patterns from sequential health data with spatial patterns from structural health features, the hybrid model achieved high accuracy. The proposed work has concentrated on creating a better data pre-processing method that removes the majority of issues related to the integrity of data in the dataset. Improved learning efficiency and higher prediction accuracy are two benefits of the multimodal data in the dataset. In summary, attribute values are used to substitute any missing values in the dataset to start the pre-processing procedure. After that, LabelEncoder is employed to modify the data. The ABCNN model has better performance across multiple metrics, with an accuracy of 95.56%, precision of 92.12%, recall/sensitivity of 99.58%, F1-score of 95.71%, and specificity of 99%. These results indicate that GRU - ABCNN is highly effective in accurately predicting outcomes, correctly identifying true positives, and minimizing both false positives and false negatives. Designed to reduce the processing costs associated with traditional AdaBoost, ABCNN achieves efficiency by lowering the number of learning epochs required, making it particularly suitable for large datasets.
Future work could enhance its performance further by developing a robust, cross-institutional dataset to improve classification and accuracy. Stroke risk prediction could be improved by incorporating more longitudinal data that tracks patients' health indicators over extended periods. Real-time data, such as continuous monitoring of blood pressure, heart rate, and activity levels through wearable devices, could be integrated to provide an up-to-date assessment of stroke risk. Longitudinal and real-time data would allow the model to better capture fluctuations in health status, potentially increasing the accuracy and timeliness of predictions.

VI.STATEMENTS AND DECLARATIONS

The study is conducted employing a 500 GB hard drive, 8GB RAM, and Python 3.7 to execute on a personal laptop running Windows 10. The "Multimodal Healthcare Dataset Stroke Data [32]" was gathered from Kaggle and utilized in this study. With 43400 records, the dataset contains 12 multimodal patient features: ID, gender, hypertension, medical history, marriage status, occupation, residence category, smoking status, average glucose level, BMI, and stroke. Conflict of Interest: The authors declare that they have no funding and conflict of interest.

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

VII.REFERENCES

[1] American Association of Neurological Surgeons, "Cerebrovascular disease–classifications, symptoms, diagnosis and treatments." [Online]. Available: https://www.aans.org/.
[2] Centers for Disease Control and Prevention, USA, "Prevalence and most common causes of disability among adults--United States, 2005," *MMWR Morb. Mortal Wkly. Rep.*, vol. 58, no. 16, pp. 421–426, 2009.
[3] A. Z. Sherzai and M. S. V. Elkind, "Advances in stroke prevention," *Ann. N. Y. Acad. Sci.*, vol. 1338, no. 1, pp. 1–15, 2015. [Online]. Available: https://doi.org/10.1111/nyas.12723.
[4] S. Mayor, "Warning signs often occur hours or days before a stroke," *BMJ*, vol. 330, no. 7491, p. 556, 2005. [Online]. Available: https://doi.org/10.1136/bmj.330.7491.556-b.
[5] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, "Stroke risk factors, genetics, and prevention," *Circ. Res.*, vol. 120, no. 3, pp. 472–495, 2017. [Online]. Available: https://doi.org/10.1161/CIRCRESAHA.116.308398.
[6] Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, Chiuve SE, Cushman M, Delling FN, Deo R, "heart disease and stroke statistics-2018 update: a report from the American Heart Association," *Circulation*, vol. 137, no. 12, p. e67, 2018.
[7] K. Kamal, V. Lopez, and S. A. Sheth, "Machine learning in acute ischemic stroke neuroimaging," *Frontiers in Neurology*, vol. 9, p. 945, 2018.
[8] Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, Cheripelli B, Cho T-H, Fazekas F, Fiehler J, "MRI-guided thrombolysis for stroke with unknown time of onset," *N. Engl. J. Med.*, vol. 379, no. 7, pp. 611–622, 2018.
[9] L. Feng, A. Ali, M. Iqbal, A. K. Bashir, S. A. Hussain, and S. Pack, "Optimal haptic communications over nanonetworks for e-health systems," *IEEE Trans. Ind. Informatics*, vol. 15, no. 5, pp. 3016–3027, 2019.
[10] S. Kutia, S. H. Chauhdary, C. Iwendi, L. Liu, W. Yong, and A. K. Bashir, "Socio-technological factors affecting user's adoption of ehealth functionalities: A case study of China and Ukraine ehealth systems," *IEEE Access*, vol. 7, pp. 90777–90788, 2019.
[11] Z. Qin, H. Li, and Z. Liu, "Multi-objective comprehensive evaluation approach to a health system based on fuzzy entropy," *Math. Struct. Comput. Sci.*, vol. 24, no. 5, 2014.
[12] R. S. Jeena and S. Kumar, "Stroke prediction using SVM," in *Proc. Int. Conf. Control, Instrum., Commun. Comput. Technol. (ICCICCT)*, 2016, pp. 600–602. [Online]. Available: http://dx.doi.org/10.1109/ICCICCT.2016.7988020.
[13] S.-M. Hanifa and K. Raja-S, "Stroke risk prediction through non-linear support vector classification models," *Int. J. Adv. Res. Comput. Sci.*, vol. 1, no. 3, 2010.
[14] J. K. Luk, R. T. Cheung, S. Ho, and L. Li, "Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects," *Cerebrovasc. Dis.*, vol. 21, no. 4, pp. 229–234, 2006.
[15] S. N. Min, S. J. Park, D. J. Kim, M. Subramaniyam, and K.-S. Lee, "Development of an algorithm for stroke prediction: A national health insurance database study in Korea," *Eur. Neurol.*, vol. 79, no. 3–4, pp. 214–220, 2018.
[16] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *2017 8th Annu. Ind. Autom. Electromech. Eng. Conf. (IEMECON)*, IEEE, 2017, pp. 158–161.
[17] P. Chantamit-o, "Prediction of stroke disease using deep learning model."
[18] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2010, pp. 183–192.
[19] C.-Y. Hung, C.-H. Lin, T.-H. Lan, G.-S. Peng, and C.-C. Lee, "Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database," *PLoS One*, vol. 14, no. 3, p. e0213007, 2019.
[20] D. Teoh, "Towards stroke prediction using electronic health records," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, pp. 1–11, 2018.
[21] C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, and C.-C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, IEEE, 2017, pp. 3110–3113.
[22] X. Li, H. Liu, X. Du, P. Zhang, G. Hu, G. Xie, S. Guo, M. Xu, X. Xie, "Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation," in *AMIA Annu. Symp. Proc.*, 2016, p. 799.
[23] "Stroke prediction (2020 (Accessed on January 22, 2020))." [Online]. Available: https://www.kaggle.com/swatis1/strokeprediction.

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[24] Q. Zhang, Q. Liu, and H. Yang, "A Comprehensive Review of Boosting Methods for Deep Learning," *Pattern Recognition*, vol. 133, p. 108788, 2023. doi: 10.1016/j.patcog.2022.108788.
[25] Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Mamun, M. S. A., & Kaiser, M. S, " Performance Analysis of Machine Learning Approaches in Stroke Prediction*", 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1464-1469 2020*
[26] Feng, Yifan, "Support Vector Machine for Stroke Risk Prediction", *Journal of Highlights in Science, Engineering and Technology*, vol. 38. pp.917-923, 2023 10.54097/hset.v38i.5977.
[27] Sudha, Gayathri, Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", *International Journal of Computer Applications* Volume 43– No.14,pp. (0975 – 8887, 2012
[28] Asit Subudhi, Manasa Dash, Sukanta Sabut, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier", *Journal of Biocybernetics and Biomedical Engineering*, vol. 40, pp. 277-289, 2020