# Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques

Senjuti Rahman, Mehedi Hasan, and Ajay Krishno Sarkar

***Abstract* — The brain is the human body's primary upper organ. Stroke is a medical disorder in which the blood arteries in the brain are ruptured, causing damage to the brain. When the supply of blood and other nutrients to the brain is interrupted, symptoms might develop. Stroke is considered as medical urgent situation and can cause long-term neurological damage, complications and often death. The World Health Organization (WHO) claims that stroke is the leading cause of death and disability worldwide. Early detection of the numerous stroke warning symptoms can lessen the stroke's severity. The main objective of this study is to forecast the possibility of a brain stroke occurring at an early stage using deep learning and machine learning techniques. To gauge the effectiveness of the algorithm, a reliable dataset for stroke prediction was taken from the Kaggle website. Several classification models, including Extreme Gradient Boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Neighbors, SVM - Linear Kernel, Naive Bayes, and deep neural networks (3-layer and 4-layer ANN) were successfully used in this study for classification tasks. The Random Forest classifier has 99% classification accuracy, which was the highest (among the machine learning classifiers). The three layer deep neural network (4-Layer ANN) has produced a higher accuracy of 92.39% than the three-layer ANN method utilizing the selected features as input. The research's findings showed that machine learning techniques outperformed deep neural networks.**

***Keywords* — Deep Neural Network, Extreme Gradient Boosting, Machine Learning, Stroke Prediction.**

## I. Introduction

The different body parts and how they function are the foundation of human life. A hazardous condition that ends human lives is stroke. After the age of 65, this condition is frequently discovered. Heart attacks influence the working of the heart, and strokes affect the brain similarly. One of these two conditions—a blood supply restriction to the brain or the rupture and bleeding of brain blood vessels—is what causes strokes. If there is a rupture or a blockage, blood and oxygen cannot reach the brain's tissues. In both industrialized and developing nations, it is currently the fifth greatest cause of death [1]. A stroke victim's chances of making a full recovery are improved the earlier they receive medical care. Any stroke victim needs to see a doctor right away. Otherwise, it will result in death, permanent disability, and brain damage. Patients can develop stoke for a variety of reasons. Diet, inactivity, alcohol, tobacco, personal history, medical history, and complications are the main causes of stroke, according to National Heart, Lung, and Blood Institute [2].

With the help of artificial intelligence, machine learning, and data science techniques, significant advancements have been made in the realm of clinical and medical services. The foundation of the present period is machine learning, which is utilized to anticipate numerous issues at an earlier stage. As a serious disease that may be treated if anticipated in the early stages, stroke is one of many that can be prevented if predicted early. In the health care sector, machine learning is crucial for the diagnosis and prognosis of diseases. Currently, stroke incidence is predicted using machine learning algorithms [3], [4]. Strong data analysis tools are needed for big amounts of medical data. A substantial area of research in this field is on the use of artificial intelligence (AI) in medicine. The system can recognize which patients are most likely to develop the illness based on a patient's medical history. Through analysis of a patient's medical history, including age, blood pressure, sugar levels, and other factors, the technology can predict the risk that they will develop a disease. When there are a lot of factors, classification algorithms are employed to predict disease. A feed-forward multi-layer artificial neural network-based deep learning model for predicting strokes was investigated in [5]. Similar research for developing an intelligent system to predict stroke from patient information was investigated in [6], [7].

Previous studies on strokes have centered on, among other things, heart attack forecasting. There haven't been many studies on brain stroke. This paper's major goal is to show how machine learning algorithms, boosting techniques, and artificial neural networks (ANN) can be used to predict when a brain stroke will occur. This research's main contribution is the application of various algorithms on a dataset that is freely available (from Kaggle website) and compare and identify the best approach for prediction of the onset of Stroke. In this study, the Neural networks and machine Learning Algorithms are employed as classification algorithms to predict the existence of stroke disease with a variety of associated characteristics. For lowering the dimensions, the principal component analysis (PCA) approach is utilized. After reducing the dimension, we have kept the most important features for the prediction of stroke.

Different performance metrics such as accuracy, precision, recall, f-1 score, auc curve (are under roc curve) of the classification models are determined and compared with each other to decide which one predicts more accurately on the dataset. The experimental results of the proposed methods were also compared with the existing works henceforth to show the novelty of the work.

Submitted on November 15, 2022.
Published on January 20, 2023.
S. Rahman, Electrical & Electronic Engineering, Ahsanullah University of Science & Engineering, Dhaka, Bangladesh.
(corresponding e-mail: senjuti.eee@aust.edu)
M. Hasan, Electronics & Telecommunication Engineering, RUET, Rajshahi-6204, Bangladesh.
(e-mail: mehedi.hasan28.bd@gmail.com)
A. K. Sarkar, Electrical & Electronic Engineering, RUET, Rajshahi-6204, Bangladesh.
(e-mail: sarkarajay139@gmail.com)

## II. Related Works

Earlier studies in the literature looked into a number of stroke prediction-related topics. Jeena *et al.* [3] provided a study of several risk factors to comprehend the likelihood of stroke. To determine the relationship between a factor and its related effect on stroke, a regression-based methodology was applied. In order to predict stroke, Adam *et al.* [8] conducted research using the decision tree method and the k-nearest neighbor algorithm. When predicting the occurrence of strokes in their study, medical professionals discovered that the decision tree method was more useful. The Cardiovascular Health Study (CHS) dataset was utilized by Singh and Choudhary [9] to predict stroke in individuals. Emon *et al.* [10] implemented the learning-based classification algorithms namely XGboost, Random Forest, Navies Bayes, Logistic Regression and Decision Tree on the dataset which is retrieved from Kaggle. The likelihood of stroke was investigated by Kansadub *et al.* [11] using decision trees, neural networks, and Naive Bayes analysis, and the study's authors attempted to predict strokes from the data. They evaluated the precision and AUC of their pointer during their study. Tazin *et al.* [12] proposed to predict stroke at an early stage by implementing Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier. Random Forest was the best performing algorithm for this task. Chetan Sharma *et al.* [13] proposed one supervised algorithm, random forest on the dataset obtained from a freely available source to predict the occurrence of a stroke shortly. A feed-forward multi-layer artificial neural network-based deep learning model for predicting strokes was also investigated in [5]. Similar research for developing an intelligent system to predict stroke using patient records was investigated in [14]. Hung *et al.* [15] compared machine learning and deep learning models constructing stroke prediction models from the electronic medical claims database. Fang *et al.* [16] applied three current Deep Learning (DL) approaches and compared these DL (CNN, LSTM, Resnet) approaches with machine learning algorithms (Deep Forest, Random Forest, Support Vector Machine, etc.) for performing in clinical prediction. Mahesh *et al.* [12] used various deep learning (DL) algorithms such as CNN, Densenet and VGG16 to evaluate the performance metrics used to predict the brain stroke automatically.

## III. Description of the Data Set

### A. Introduction to the Dataset

The research was carried out using the stroke prediction dataset available on the Kaggle website. In this dataset, there were 12 columns and 5110 rows. The description of the dataset is given in Table I. The output column's stroke has a value of either 1 or 0. A stroke risk was found when the value 1 was detected, but a stroke risk was not found when the value 0 was displayed. In this dataset, there is a greater chance of a 0 in the output column (stroke) than a 1 there. The stroke column alone has 249 rows with a value of 1, whereas 4861 rows have a value of 0. In the output column before preprocessing, Fig. 1 shows the Visualizing Count of classes along with the number. Preprocessing (oversampling) of the data is used to balance the data and increase accuracy. After Oversampling, the value of stroke with 1 has increased to the level of 4861 while the value of non-stroke data with 0 remains the same as before (i.e., 4861).

### B. Visualization of Feature Selection

A statistical indicator of the strength of the association between the relative movements of two variables is the correlation coefficient. The values are in the -1.0 to 1.0 range.

There was a measurement error in the correlation if the estimated value was larger than 1.0 or lower than -1.0.

TABLE I: Description of the Dataset used for the Proposed Approaches

| Attribute Name | Description |
|---|---|
| ID of Individuals | Unique identification number of 5110 patient |
| Gender | Male = '0', Female= '1' |
| Age (in years) | Age of the patient (1-82) |
| Hypertension | Indicating whether the patient has hypertension (1) problem or not (0) |
| Heart Disease | Demonstrating whether the unique id patient has heart disease problem (1) or not (0) |
| Ever married | Represents the marital status by yes (1) or no (0). |
| Work type | It indicates five category of work status of the patient.<br>Children = '0'<br>Government Job = '1'<br>Never worked = '2'<br>Private = '3'<br>Self Employed ='4' |
| Residence type | It denotes the residential area type, whether Rural ='0' or Urban = '1' |
| Average Glucose Level | It gives the average glucose level which is represented in numeric form (55.12-271.74) |
| BMI | Body Mass Index is represented in numeric form (10.3-97.6). |
| Smoking Status | It indicates four categories of smokers.<br>Formerly Smoked = '0'<br>Never Smoked = '1'<br>Smokes = '2'<br>Unknown = '3'<br>(There is no information/could not found about the unknown type) |
| Stroke | It indicates the target, whether have stroke (1) or non-stroke (0). |

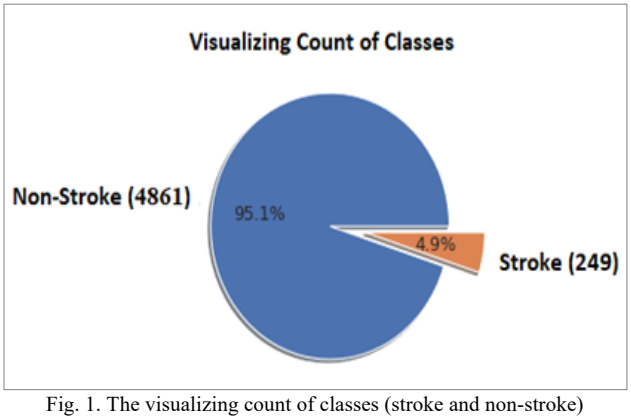


Fig. 1. The visualizing count of classes (stroke and non-stroke) along with the number.

Perfect negative correlation is shown by a correlation of -1.0, and perfect positive correlation is shown by a correlation of 1.0. A correlation of 0.0 indicates that there is no linear link between the two variables' movements. Fig. 3 depicts the feature selection procedure.

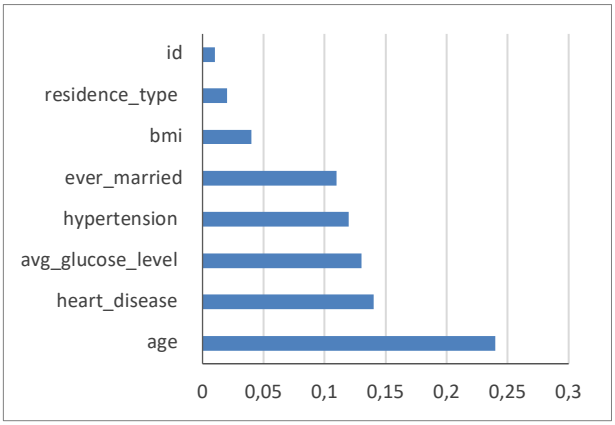


Fig. 2. The correlation between the attributes/features of the utilized stroke prediction dataset.

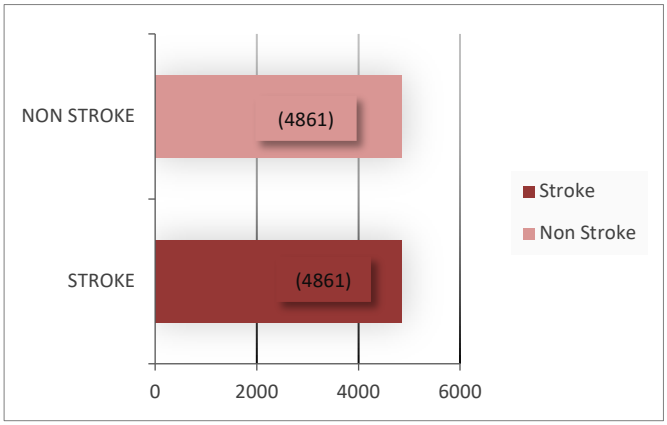


Fig. 3. Total count of stroke and non-stroke data after pre-processing.

Understanding how features are connected to one another is made easier by feature selection. Age, hypertension, average glucose level, heart disease, ever married, and BMI are all positively correlated with the target feature, as shown in Fig. 2. However, Fig. 2 does not depict how gender is adversely correlated (-0.0069) with stroke.

## IV. Proposed Methodology

Data pre-processing is necessary prior to model construction in order to eliminate a dataset's undesirable noise and outliers, which could cause the model to deviate from its intended training. This phase deals with all the issues that keep the model from operating more effectively. Data must be cleansed and processed for model development after the pertinent dataset has been collected. Twelve attributes make up the dataset, as was previously said. The column id is firstly ignored because its inclusion has no impact on model creation. After that, the dataset is checked for null values and filled if any are found. In this instance, the data column's "most frequent" value is used to fill in the null values in the BMI column. The string literals in the dataset are changed by label encoding into integer values that the computer can understand. It is necessary to transform the strings to integers because the computer is typically educated on numerical data. There are five columns (gender, ever married, work type, Residence type, smoking status with data of the type string in the obtained dataset. During label encoding, all strings are encoded, and the entire dataset is converted into a set of numbers. The stroke prediction dataset is severely unbalanced. There are 5110 rows in the -e dataset, 249 of which hint at the likelihood of a stroke and 4861 of which demonstrate its absence. While using such data to train a machine-level model may increase accuracy, other accuracy metrics like recall and precision are insufficient. The findings will be incorrect, and the forecast would be worthless if such uneven data is not handled properly. Therefore, this uneven data must be addressed first in order to produce an efficient model. This was accomplished using the Random Oversample (Ros approach). After implementing the Ros approach, the imbalanced data becomes balanced (both types have same dimension) which is shown in Fig. 3.

A MinMaxScaler was used to scale the features to between -1 and 1 to normalize them. After that, Principal component analysis (PCA) was utilized which chooses the minimum number of principal components such that 95% of the variance is retained. Following completion of data preparation and management of the unbalanced dataset, the model construction phase begins. The data is split into training and testing data with an 80/20 split, in order to increase the accuracy and efficiency of this job. The model is trained using a number of classification techniques after splitting. In this study, classification tasks were effectively completed using deep neural networks (3-layer and 4-layer ANN), Extreme gradient boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Nearest Neighbors, SVM - Linear Kernel and Naive Bayes. The workflow is shown in Fig. 4.

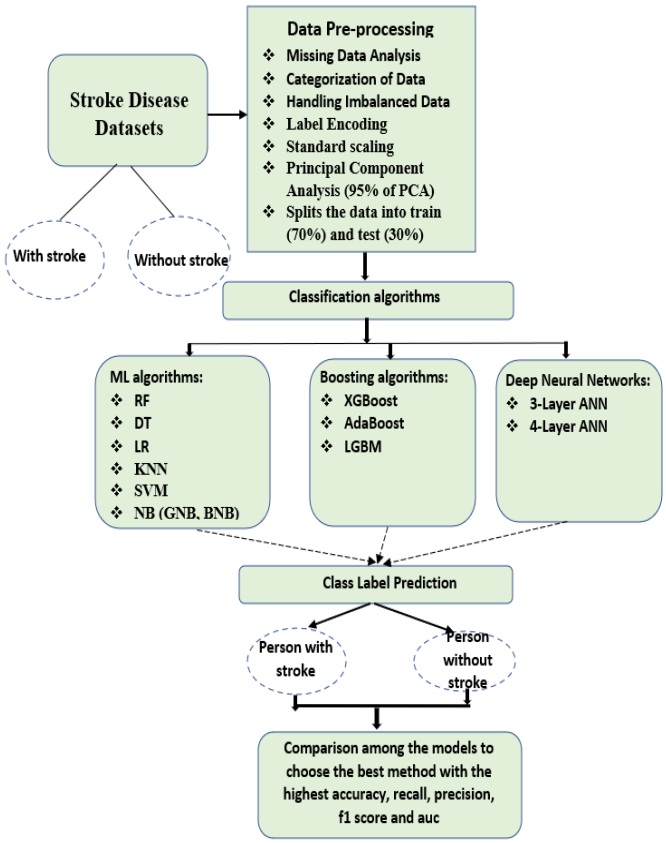


Fig. 4. The workflow of the proposed methodology.

## V. Classification Algorithms

### A. Machine Learning Approaches

#### 1) Decision Tree (DT)

Classification with DT addresses both regression and classification issues. This approach uses a supervised learning model and an output variable. It has tree-like characteristics. The decision tree's two components are the decision node and the leaf node. The data is split at the first node, then it is combined at the second node that generates the output. Since DT mimics the stages, a human goes through when creating a real-world object, it is simple to understand. The algorithm assumes that the existence or absence of a given feature in the dataset depends on others and helps to categorize the target towards a particular class [17], [18].

#### 2) Random Forest (RF)

Classification and regression problems can be resolved using ensemble learning techniques like Random Forests (sometimes referred to as random choice forests). They work through distributed training of a large number of decision trees. When attempting to solve classification problems, a random forest's output is the class that majority of the trees select. In order to tackle complex issues and improve the performance of a model, ensemble learning, which is a technique, employs several classifiers. The Random Forest classifier takes into account predictions from all of the trees rather than just one, as stated in its name, "combining a large number of decision trees on different subsets of a given dataset and taking an average to boost the projected accuracy". To determine the outcome, the most popular forecasts are used. The Random Forest demonstrated the highest accuracy (criterion=entropy) with the utilized Stroke prediction dataset.

#### 3) Naive Bayes (NB)

The supervised learning technique known as naive Bayes, which is based on the naïve theorem, is used in machine learning (ML). The Naive Bayes algorithm is founded on the notion that the presence of one feature or parameter does not preclude the presence of another, i.e., that one feature's existence is unrelated to that of other features. The Bayes theorem is a conditional probability theorem in mathematics that estimates the chance that a certain event will take place assuming that a particular condition has already been satisfied. With the use of the Bayes theorem, conditional probability is contrasted; it is the likelihood that a particular event has occurred, given the premise that some event has already occurred. There are several variations of NB found in the literature [19], with the key distinction being how the likelihood of the intended class is calculated. These variations include simple Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, and Multi-variant Poisson Naive Bayes. Gaussian Naive Bayes and Bernoulli Naive Bayes were used for this study.

#### 4) K-Nearest Neighbors (KNN)

K-NN is a kind of slow learning in which there is no specific pre-processing stage and all computations are saved for classification. The nearest training data points on the feature map are used to make decisions using this data categorization technique. The Euclidean distance metric is used by the K-NN classifier to predict the target class. The dataset defines the best value for the parameter k, which controls how well the classifier performs. The ideal value is then determined once the impacts have been studied. K = 3 was applied in our investigation.

#### 5) Support Vector Machine (SVM)

The Support Vector Machine is a type of supervised learning system that uses labeled data to categorize unknown data [3]. To express decision boundaries, it uses the concept of decision planes or hyperplanes [19]. A hyperplane is used to separate the collection of data objects into the various classifications. The Radial Basis Function (RBF) kernel, with a setting of 1, was applied. SVM attempts to classify the data by creating a function that assigns each data point to its appropriate label with the minimum amount of error and the largest (possible) margin.

#### 6) Logistic Regression (LR)

LR is one of the most widely used ML algorithms in the supervised learning approach [12]. It is a forecasting technique that forecasts a categorical dependent variable using a number of independent factors. Logistic regression and linear regression are fairly similar, with the exception of how they are used. While logistic regression is used to address classification issues, regression issues are addressed by linear regression. A model tuning method called ridge regression can be used to analyze any multicollinear data. This method performs L2 regularization. For this task, we utilized solver='liblinear' and max iter=100.

#### 7) XGBoost

A good application of the gradient augmentation technique is XGboost. Although there may not be any groundbreaking mathematical discoveries in this case, the gradient gain alternative can be carefully planned for accuracy and optimization. It consists of a linear version, and the newborn tree may be a method that makes use of different AI algorithms to check whether a fragile newbie would produce a trustworthy newbie in order to increase the version's accuracy. For instance, random forest can be learned from (impulsive) and parallel learning (bagging). Data gathering is a technique that can be utilized to regulate the presentation of an advanced AI version whose precision processing is quicker than improving gradients. These techniques for filling the data gap are built in. Few parameters (base score=0.5, max bin=256, learning rate=0.9, max depth=25, min child weight=1) were chosen in this investigation which gives the second highest accuracy from this classifier. The parameters for this classifier are shown in Table II.

#### 8) AdaBoost

Machine learning ensemble methods use the boosting technique known as the AdaBoost algorithm, sometimes referred to as Adaptive Boosting. The weights are redistributed to each instance, with examples that were incorrectly classified receiving higher weights, hence the term "adaptive boosting." The 100 DT algorithm was combined with the CART (Classification and Regression Tree) algorithm. Each time the boosting procedure was repeated, each classifier was given a weight of 1. For the investigation, we utilized learning rate = 0.9, n estimators = 20, and random state = 42.

TABLE II: DESCRIPTION OF THE PARAMETERS USED FOR THE XGBOOST

| Name of the parameter | Default Value | Description of the parameters |
|---|---|---|
| learning_rate | 0.9 | Reduce the weights with each step. |
| n_estimators | 100 | Number of trees to fit |
| objective | binary | logistic regression for binary classification |
| booster | gbtree | Select the model for each iteration |
| nthread | max | Input the system core number |
| min_child_weight | 1 | Minimum sum of weights |
| max_depth | 25 | Maximum depth of a tree |
| gamma | 0 | The minimum loss reduction needed for splitting |
| reg_lambda | 1 | L2 regularization term on weights |
| reg_alpha | 0 | L1 regularization term on weights |

#### 9) *Light gradient-boosting machine (LightGBM)*

A gradient boosting system called LightGBM makes use of tree-based learning techniques. With a higher training efficiency, lower memory utilization, and better accuracy, it is intended to be distributed and efficient.

### B. *Deep Learning Approaches*

Deep learning is a term used to describe artificial neural networks used in machine learning. Examples of deep learning architectures include convolutional neural networks, deep belief networks, recurrent neural networks, and deep neural networks (DNN). A group of algorithms known as neural networks are created to recognize patterns and are roughly modeled after the human brain [21]. These are widely used in many different research areas, including as speech recognition, natural language processing, audio recognition, computer vision, gaming, and many more. A DNN is made up of an input layer, many hidden layers, and an output layer [22]. Backpropagation is used to train the network and reduce the difference between the desired and actual output. Fig. 5 depicts the flowchart of the two proposed ANN techniques.

Two ANN models have been put into practice for the analysis. The output of the two models, three- and four-layer ANNs has a sigmoid function. The data collection and preprocessing are the same as for machine learning approaches.

## VI. PERFORMANCE METRICS AND RESULTS

### A. *Performance Metrics*

Five statistical variables were utilized in this study to evaluate the performance and usefulness of the classifiers: accuracy, precision, recall/sensitivity (recall and sensitivity are the same in binary classification), F1 score, and AUC curve [12]. The following are the definitions of the statistical parameters.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2*(Recall*Precision)}{(Recall+Precision)} \quad (4)$$

where,
TP = true positive
FN = false negative,
FP = false positive
TN =true negative.

In this work, AUC curves were used to determine how well the probabilities from the positive classes and the negative classes could be separated. The degree of True Positive and False Positive rate is represented by the AUC curve, which also shows the model's overall performance.

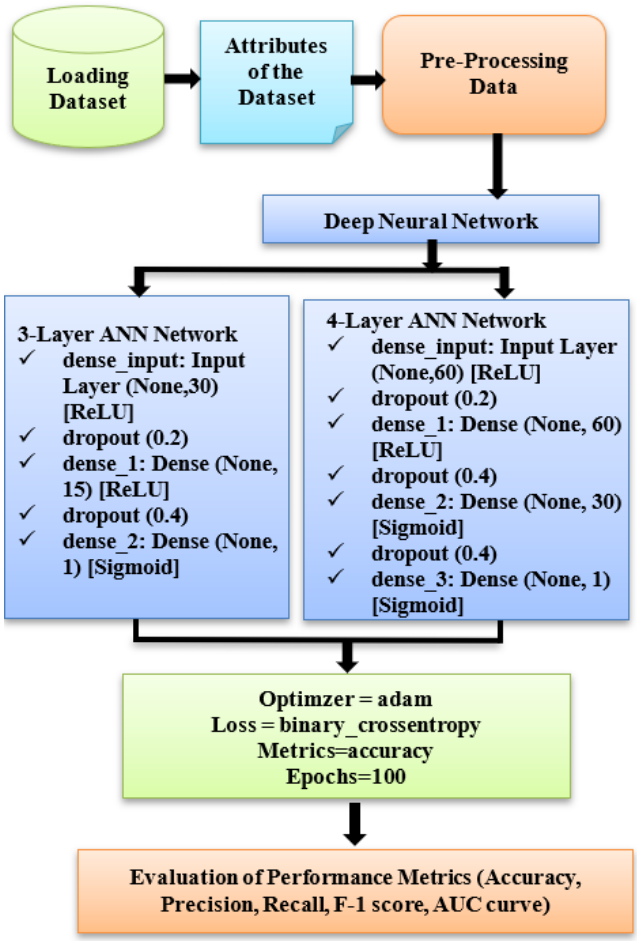


Fig. 5. The flowchart of the two proposed ANN techniques.

### B. *Results*

Random Forest outperforms other classifiers in terms of accuracy (0.99) which is calculated using equation (1). It shows the highest accuracy among the machine learning algorithms, whereas 3-layer ANN demonstrated promising results among deep learning techniques. The comparison among the machine learning approaches is shown in Table III and a chart is shown (in Fig. 6) to represent the superiority of RF method over the other ML algorithms using the performance metrics which were calculated using equation (1-4). The area under roc curve for Random Forest method is given in Fig. 7. The performance metrics for the ANN approaches are described in Table IV. The Area under roc curve for the 4-layer ANN is shown in Figure 8. The comparison between the Random Forest and 3-layer ANN method is depicted in the bar chart of Fig. 9. From the comparison, it is clear that RF algorithm outperforms all the Boosting algorithms and deep neural approaches in all aspects.

TABLE III: COMPARISON AMONG THE MACHINE LEARNING APPROACHES

| Algorithm | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| LR | 0.71 | 0.69 | 0.73 | 0.71 | 0.79 |
| DT | 0.98 | 1.00 | 0.95 | 0.97 | 0.98 |
| RF | 0.99 | 1.00 | 0.98 | 0.99 | 1.00 |
| KNN | 0.96 | 1.00 | 0.92 | 0.96 | 0.98 |
| SVM | 0.82 | 0.84 | 0.77 | 0.81 | - |
| GaussianNB | 0.70 | 0.68 | 0.73 | 0.70 | 0.78 |
| BernoulliNB | 0.67 | 0.69 | 0.59 | 0.63 | 0.71 |
| XGBoost | 0.97 | 1.00 | 0.93 | 0.97 | 0.98 |
| AdaBoost | 0.78 | 0.75 | 0.82 | 0.78 | 0.76 |
| LGBM | 0.95 | 1.00 | 0.90 | 0.95 | 0.96 |

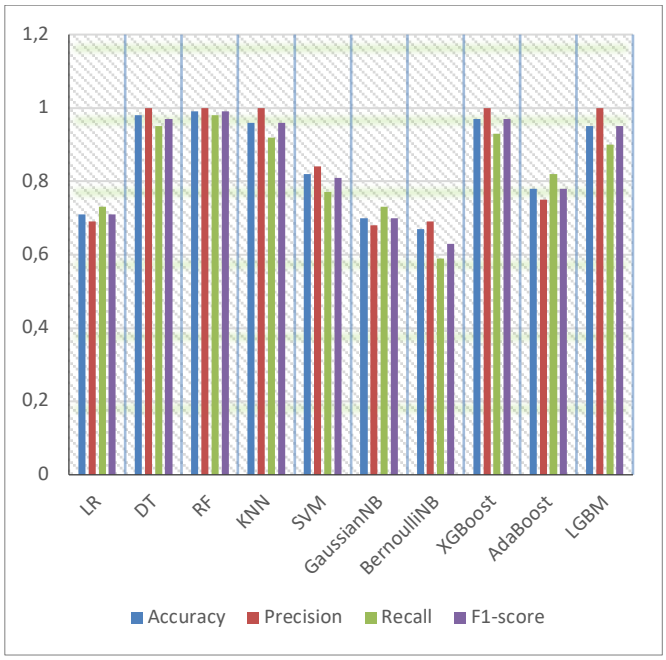


Fig. 6. A comparison chart of evaluation metrics of machine learning algorithms.

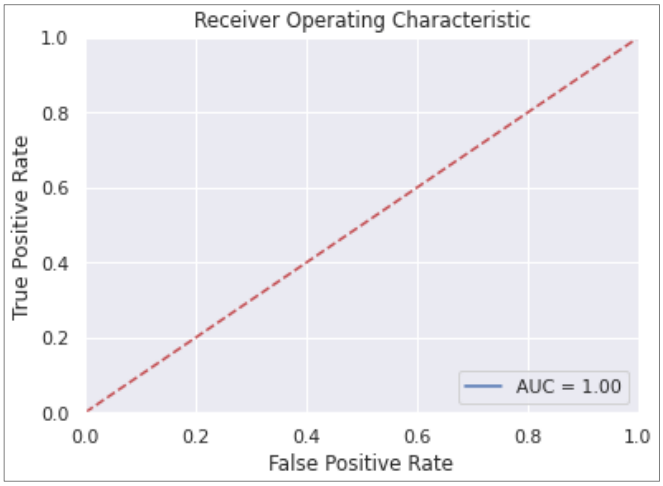


Fig. 7. The area under roc curve for Random Forest method.

TABLE IV: COMPARISON BETWEEN THE DEEP LEARNING APPROACHES

| Algorithm | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| 4 -layer ANN | 0.9239 | 0.8867 | 0.992 | 0.9364 | 0.97 |
| 3-layer ANN | 0.8401 | 0.7709 | 0.974 | 0.8606 | 0.91 |

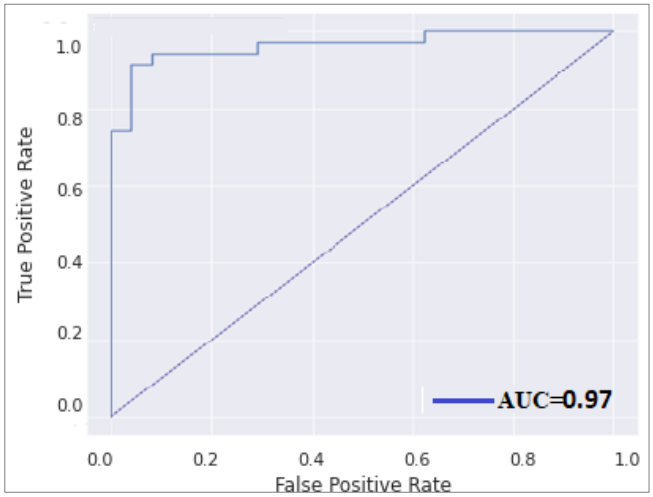


Fig. 8. The area under roc curve for 4-layer ANN method.

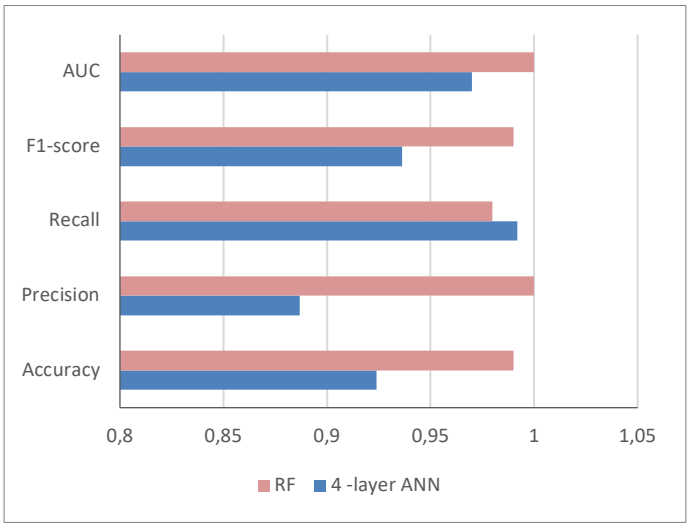


Fig. 9. A comparison chart of evaluation metrics of machine RF and 4-layer ANN algorithms.

## VII. COMPARISON WITH THE EXISTING WORK

When the proposed method for recognizing Stroke patients is compared with the previous studies available in this field, its novelty becomes apparent. Table V presents the comparison. It can be clearly seen from Table V, the values of the proposed work are higher for the performance metrics for the Random Forest, XGBoost and 4-layer ANN model. This shows the novelty of the work.

## VIII. CONCLUSION

Stroke is a potentially fatal medical condition that needs to be treated right away to prevent future consequences. The creation of a machine learning (ML) and Deep Learning model could help with stroke early diagnosis and subsequent reduction of its severe consequences. This study examines how well different machine learning (ML) as well as Boosting algorithms predict stroke based on various biological factors. With a classification accuracy of 99%, and AUC of 1, random forest classification exceeds the other investigated techniques. According to the study, the random forest method performs better than other methods when forecasting brain strokes using cross-validation measures.

TABLE V: COMPARISON OF THE PROPOSED METHODS WITH THE RELATED WORK

| Existing Work | Objectives | Source of data | Machine learning/Deep learning approaches | Outcomes |
|---|---|---|---|---|
| [15] | Prediction of future stroke occurrence | National Health Insurance Research Database (NHIRD) | DNN, gradient boosting decision tree (GBDT), LR and SVM | DNN and GBDT algorithm's Value of AUC is 0.915 and 0.918. |
| [20] | Prediction of stroke type | Retrieved dataset from Kaggle | XGboost, RF, NB, LR and DT | An accuracy =97.56% is obtained using XGBoost. |
| [23] | Stroke Risk Prediction | Retrieved dataset from Kaggle | LR, DT, RF, KNN, SVM, and NB Classification | Naïve Bayes performed best in the task that gave an accuracy of 82%. |
| [24] | Stroke Risk Prediction | Collected dataset from Kaggle | LR, DT, KNN, RF, NB | Highest accuracy =95.4% utilizing Random Forest |
| [25] | Prediction of stroke type | Dataset retrieved from Kaggle was used | DT, LR, NB, KNN, RF, ANN, SVM, XGBoost (Second highest accuracy achieved). | Highest accuracy= 92.32% achieved using Random Forest (AUC= 0.975) |
| [26] | Prediction of stroke type | - | DT, KNN, LR, NB, RF, SVM and neural network | Highest accuracy =93% using Random Forest |
| Proposed Work (With Random Forest, 4-layer ANN and XGBoost) | Classification of with and without stroke patient | Collected dataset from Kaggle | XGBoost, Ada Boost, LightGBM, RF, DT, LR, KNN, NB, SVM, 3 and 4-layer ANN. | Random Forest, XGBoost and 4-layer ANN Accuracy= 99%, 97% and 92.39% |

## IX. FUTURE SCOPE

The study's future objectives could involve employing a bigger dataset or applying the same model to several distinct datasets. Convnet and ChronoNet can be used to improve Deep Learning framework models. The artificial intelligence architecture may aid the general public in assessing the possibility of a stroke occurring in an adult patient, the risk level connected with it, as well as the determination of the chance of the condition recurring, in exchange for merely supplying some simple information. In a perfect world, it would assist patients in receiving early stroke treatment and recovering from the incident.

## REFERENCES

[1] Pikula A, Howard BV, Seshadri S. Stroke and Diabetes. In: Cowie CC, Casagrande SS, Menke A, et al., editors. *Diabetes in America.* (3rd ed.). Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases (US), 2018, ch.19.

[2] Gary H, Gibbons L. *National Heart, Lung and Blood Institute.* 2022 [updated 2022 March 24]. Available from: https://www.nhlbi.nih.gov/health/stroke.

[3] Jeena RS, Kumar S. Stroke prediction using SVM, International Conference on Control. *Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2016: 600–602.

[4] Hanifa SM, Raja SK. Stroke risk prediction through non-linear support vector classification models. *Int. J. Adv. Res. Comput. Sci.,* 2010; 1(3).

[5] Chantamit-o P, Madhu G. Prediction of Stroke Using Deep Learning Model. *International Conference on Neural Information Processing*, 2017: 774-781.

[6] Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction, *in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010: 183–192.

[7] Hung CY, Lin CH, Lan TH, Peng GS, Lee CC. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. *PLOS ONE*, 2019;14(3):e0213007. https://doi.org/10.1371/journal.pone.0213007.

[8] Adam SY, Yousif A, Bashir MB. Classification of ischemic stroke using machine learning algorithms. *International Journal of Computer Application*, 2016;149(10):26–31.

[9] Singh MS, Choudhary P. Stroke prediction using artificial intelligence. *8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON),* 2017:158–161.

[10] Emon MU, Keya MS, Meghla TI, Rahman MA, Mamun SA, Kaiser MS. Performance Analysis of Machine Learning Approaches in Stroke Prediction, *International Conference on Enumerative Combinatorics and Applications*, Nov. 2021.

[11] Kansadub T, Ammaboosadee S, Kiattisin S, Jalayondeja C. Stroke risk prediction model based on de-mographic data, *in Proceedings of the 2015 8th Biomedical Engineering International Conference (BMEiCON)*, Pattaya, -Thailand, November 2015: 1-3.

[12] Tazin T, Alam MN, Dola NN, Bari MS, Bourouis S, Khan M. Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*, 2021:1-12. doi: 10.1155/2021/7633381.

[13] Sharma C, Sharma S, Kumar M, Sodhi A. Early Stroke Prediction Using Machine Learning. *International Conference on Decision Aid Sciences and Applications;* Mar. 2022.

[14] Teoh D. Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making,* 2018; Dec.(1): 1–11. doi: 10.1186/s12911-018-0702-y.

[15] Hung CY, Lin CH, Lan TH, Peng GS, Lee CC. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE*, 2017: 3110–3113.

[16] Fang G, Huang Z, Wang Z. Predicting Ischemic Stroke Outcome Using Deep Learning Approaches. *Front Genet.* 2022 Jan 24;12:827522. doi: 10.3389/fgene.2021.827522.

[17] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 1991 May-June; 21(3): 660-674. doi: 10.1109/21.97458.

[18] Navada A, Ansari AN, Patil S, Sonkamble BA. Overview of use of decision tree algorithms in machine learning. *IEEE Control and System Graduate Research Colloquium, ICSGRC*, 2011: 37–42.

[19] Rahman MM, Rana MR, Alam NAA, Khan MSI. *A web-based heart disease prediction system using machine learning algorithms*; 2022 June; 12. 64-80.

[20] Dhillon S, Bansal C, Sidhu B. Machine Learning Based Approach Using XGboost for Heart Stroke Prediction. in *International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications*, September 06–07, 2021.

[21] Akash K, Shashank HN, Srikanth S, Thejas AM. Prediction of Stroke Using Machine Learning. June 2020.

[22] Aiello S, Cliff C, Roark H, Rehak L, Stetsenko P, and Bartz A. *Machine Learning with Python and H2O.* (5th Ed.). H2O. ai Inc. Nov. 2017.

[23] Sailasya G and Kumari G. L. A. Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science And Applications*. 2021; 12(6): 539–545.

[24] Gurjar R, Sahana K, Sathish BS. Stroke Risk Prediction Using Machine Learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* 2022: 20-25. doi: 10.32628/CSEIT2283121.

[25] Tavares J-A. Stroke prediction through Data Science and Machine Learning Algorithms. 2021; doi: 10.13140/RG.2.2.33027.43040.

[26] Martín J.R, Ayala J.L, Roselló G.R and Camarasaltas J. M. Comparison of Different Machine Learning Approaches to Model Stroke Subtype Classification and Risk Prediction. *Spring Simulation Conference (SpringSim),* pp. 1-10, 2019.

**Senjuti Rahman** received a B.Sc. degree from Rajshahi University of Engineering and Technology, Bangladesh, in 2016 in electronics and telecommunication engineering. Currently, she is pursuing an M.Sc. degree in electrical and electronic engineering from the same university. From 2018 to 2022, she was working as a lecturer in the department of EEE in Eastern University (EU), Dhaka, Bangladesh. From 25th July 2022 she joined Ahsanullah University of Engineering and Technology as a Lecturer in the EEE department. She has a total of 5 conference papers in ICECTE 2016, ICEEE 2017, ICRPSET 2022, 4IREF 2022 etc. Her research interests include Biomedical Engineering, Machine Learning, and Deep Learning.

**Md Mehedi Hasan** received the B.Sc. degree from Rajshahi University of Engineering and Technology, Bangladesh, in 2016 in electronics and telecommunication engineering. From 2017 to 2019, he was working as a Junior Software Engineer at Smart Aspects Ltd. In Dhaka, Bangladesh. Recently, from December 2019 he joined as a Software Engineer at Zantrik, Dhaka, Bangladesh. His research interests include Biomedical Engineering, Machine Learning, and Deep Learning Biomedical Engineering, Machine Learning, Deep Learning, Computer Vision, Data Science. His working area covers Android Mobile Application Development, Web Development, Database Management, Machine Learning, Deep Learning. He has a good grasp of some programming languages, such as Java, C#, Python, C++, SQL. He has completed two online courses, AI for Medical Diagnosis- an online non-credit course authorized by Deep Learning.AI and offered through Coursera and Machine Learning Projects for Healthcare on Udemy. He has two international conference papers in ICEEE 2017, 4IREF 2022 and four of his recent research works are under process.

**Dr. Ajay Krishno Sarkar** has received Ph. D in Electronic and Computer Engineering from Griffith University, Australia, M. Sc in EEE from Japan, and B. Sc. in EEE from RUET, Bangladesh. He is currently working as a professor in the department of Electrical and Electronic Engineering (EEE) at Rajshahi University of Engineering and Technology (RUET), Rajshahi-6204, Bangladesh. He is currently a member of IEEE; Institute of Engineers Bangladesh (IEB) and he had a position in different organizing and technical committees at different international conferences in Bangladesh and abroad. He is a reviewer of several journals such IEEE Access, IEEE Photonics Journal, Computers and Electronics in Agriculture etc., and technical papers submitted in different international conferences in Bangladesh and abroad. His research interests include Sports and Biomedical Engineering, Photonic Crystal Fiber and Biosensors, Microwave and RF circuits &amp, Devices, Microwave absorptions, and Thin Films.