Brain Stroke Prediction Using Learning Machines & Deep Learning

Samman Ashraf, Zunaira Akram, Umair Muneer Butt and Asia Sharif

August 15, 2023

Brain Stroke Using Learning Machines and Deep Learning

Samman Ashraf [1], Zunaira Akram [1], Umair Muneer Butt [1] and Asia Sharif [1]

[1] University of Chenab, Gujrat, Pakistan
Samanashraf1996@gmail.com, haremfatima49@gmail.com,
umair@cs.uchenab.edu.pk, asia.sharif1988@gmail.com

Abstract: A stroke can happen if blood flow suddenly to a region of the brain stops. Depending on damaged part of the brain, disability is caused by a lack of blood flow because progressively losing brain cells perish. Predicting strokes & promoting healthy living can both benefit substantially from early detection of symptoms. In this research, there are several models developed & evaluated using machine learning (ML) in sequence to provide a strong pattern for the stroke incidence risk prediction over the long run. The main contribution of this study is a stacking technique that performs well and is supported by numerous measures, consisting of. K nearest neighbor, logistic regression, XG boost, random forest classifier, decision tree classifier, ada-boost, catboost, etc. An gradually popular computer system that copycats human intellect is being used in numerous sectors, including medicine. One such field of AI practice is stroke medicine, which objects to increase the precision of diagnosis and the standard of patient treatment. An accurate analysis of stroke imaging is important for stroke therapy. We offer a fleeting summary of the use of AI in stroke imaging in this review, highlighting the technological basics, clinical solicitations, and upcoming perceptions.

Key words: data analysis, machine learning, risk prediction, and stroke

1 Introduction

Brain stroke means a brain attack occurs when something blocks providing blood to a portion of the brain. Around 5.5 million individuals will pass away due to brain strokes each year. It has a substantial influence on every aspect of life because it is the top an origin of mortality and disability in the globe. Stroke affects the sufferer as well as their loved ones, friends, and social network. In spite of widespread assumption, anybody may experience, it regardless of their gender or physical condition, at any age. A stroke is referred to being a severe neurological condition of the blood arteries when the blood supply to a specific region of the brain is cut off, the brain cells are deprived of the necessary oxygen, which causes brain damage. The angle of the mouth is reduced, ranging from modest to severely severe (crooked mouth). In situations of severe strokes, the patient ultimately goes unconscious and goes into a coma. Schemes and hemorrhagic strokes can be broadly classified into two categories. According to the

American Heart Association (AHA), ischemic strokes, which account for 87% of all strokes, occur when a blood clot blocks or stops a blood artery feeding the brain.

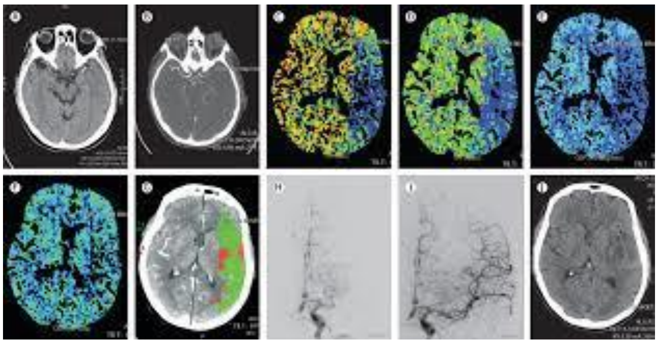

Fig.1. Brain Stroke Prediction in Real World.

When a patient experiences a stroke, a computed tomography (CT) scan is performed without delay. For ischemic stroke, magnetic resonance imaging (MRI) is useful. There are two more auxiliary diagnostic procedures: carotid triplex and cardiac triplex. Strokes can range in severity from slight to substantial. In the vast majority of situations, the first 24 hours are crucial. The diagnosis will be used to highlight the treatment, which is primarily medical—pharmaceutical—and, in certain cases, surgical. When a patient enters a coma, intubation and mechanical breathing in the intensive care unit are required [3]. A stroke must be anticipated in order to be treated in time to prevent deaths or long-term damage. As indications of the risk of stroke, we considered hypertension, obesity, heart disease, and the average blood glucose levels. Additionally, the decision-making processes of this prediction system can benefit greatly from machine learning.

Without exact info around the context and scope of the paper on Brain Stroke Using Learning Machines and Deep Learning, it's hard to fix whether this kind of study has been tried before. The significance of the current work eventually resides in its addition to the larger field of research on deep learning and machine learning applications in stroke diagnosis and therapy.

The framework uses supervised and unsupervised learning, convolutional neural networks, and repeated neural networks, as well as other deep learning methods, in the context of artificial intelligence. We also yield into account the moral and legal complications of employing these technologies in healthcare, including issues with prejudgment and data privacy.

Therefore, utilizing machine and deep learning approaches, our goal is to forecast brain stroke. For our purposes, decision trees, random forests, stochastic gradient descent (SGD), naive Bayes, logistic regression, K-NN, stochastic gradient descent, and multilayer perception were all assessed.

Overall, the imaginary framework for this examination offers a thorough and multidisciplinary approach to knowing the potential of deep learning and machine learning in attractive stroke outcomes.

2 Related Work

Several research have been undertaken to use machine learning approaches to predict stroke. To determine the relationship between risk variables and their influence on stroke, Jeena et al. employed a regression-based technique. Other research has concentrated on finding key elements for stroke prediction and building algorithms for predicting stroke from possibly modifiable risk variables. However, factors such as data collection, feature selection, and data cleaning can all have an impact on the accuracy of these models. As a result, it is critical to investigate the interdependencies between risk variables collected in electronic health records and their influence on stroke prediction accuracy. Before applying a classification method, data mining practitioners should also eliminate redundant and useless characteristics.

To guess strokes, several investigators have previously organized machine learning-based methods. Data on 500+ patients were assembled for Govindarajan et al. [11]'s work to groups stroke illnesses using a text mining mixture and a machine learning classifier.

Amini et al. [4], [12] recruited 807 well and sick people for their study and personal 55 risk factors for stroke, counting diabetes, circulatory disease, and smoking, hyperlipidemia, and alcohol eating.

A study on the calculating of the ischemic stroke prognosis was published by Cheng et al. [13]. Two ANN models and data from 80 patients who had ischemic strokes were consumed in the study.

A research was conducted by Cheon et al. [14]–[16] to forecast stroke patient death. They employed 15098 participants in their study to determine the regularity of strokes. To identify strokes, they practical a deep neural network technique.

A study on artificial intelligence-based stroke prediction was conducted by Singh et al. [17]. In their work, scientists applied a novel method to the cardiovascular health study (CHS) dataset to predict stroke. They built the model with a neural network classification technique and attained 98% accuracy.

In a research, Monteiro et al. [19] used machine learning to predict the functional projection of an ischemic stroke.

3 METHODOLOGY

The accompanying author will disclose the data backing up the study's findings upon reasonable request. A prospective cohort that records patients with ischemic stroke who are hospitalized within 7 days after the beginning of symptoms was used in this retrospective analysis. All patients hospitalized between January 2010 and December 2014 were included in this research. Patients who had had recanalization therapy or had a pre-stroke modified Rankin Scale (mRS) score of >2 were not included in the study. A favorable result was defined as an mRS score of 0, 1, or 2 and functional outcome was assessed at 3 months (online-only Data Supplement).

The three sections that make up this section are:

- Description of data.
- Classifiers and evaluation matrices for machine learning.
- The procedures for implementation.

Below is an explanation of these three procedures: both its variations.

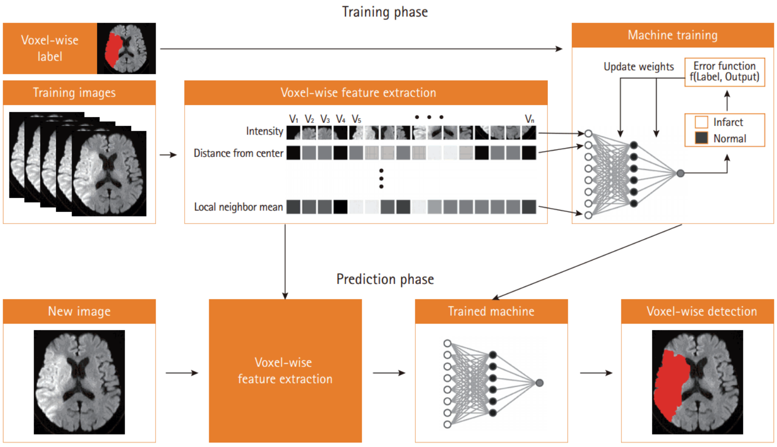

Fig.2. A proposed approach for our Brain Stroke Prediction

3.1 Description of data.

The Kaggle website provided the dataset for our study. We selected participants who were over 18 years old, resulting in a total of 3254 participants. The dataset consisted of 10 attributes, of which 9 were nominal and 3 were numerical. A description of each attribute is provided below:

- **Age (years)**: The age of participants who are over 18 is represented by this feature.
- **Gender:** With 1260 males and 1994 women in the sample, this feature denotes the individuals' gender.
- **Hypertension:** This attribute represents whether the participant has hypertension or not, with 12.54% of participants having hypertension.
- **Heart disease:** This attribute represents whether if the person has heart disease or not, with 6.33% of participants having heart disease.
- **Ever married**: With 79.84% of participants married, this characteristic reveals the individuals' marital status.

• **Work:** Four categories make up this feature, which describes the individuals' employment status: private sector (65.02%), self-employment (19.21%), government employment (15.68%), and never worked (0.1%).

• **Type of residence:** This characteristic, which divides people into two categories based on where they live, is urban (51.14%) and rural (48.86%).

• **Avg glucose level (mg/dL):** This attribute represents the average glucose level of the participants.

• **BMI (Kg/m2):** This attribute represents the body mass index of the participants.

• **Smoking Status:** Three categories are available for this feature, which indicates the individuals' smoking status: current smokers (22.37%), never smokers (52.63%), and past smokers (24.98%).

• **Stroke:** This characteristic indicates if the subject has ever had a stroke before; 5.53 percent of participants have.

3.2 Classifiers and evaluation matrices for machine learning.

In this section, we will discuss the machine learning classifiers that were utilised to develop stroke predictors. These classifiers were selected because of their well-established track record in developing vulnerability predictors and their widespread application in related research work.

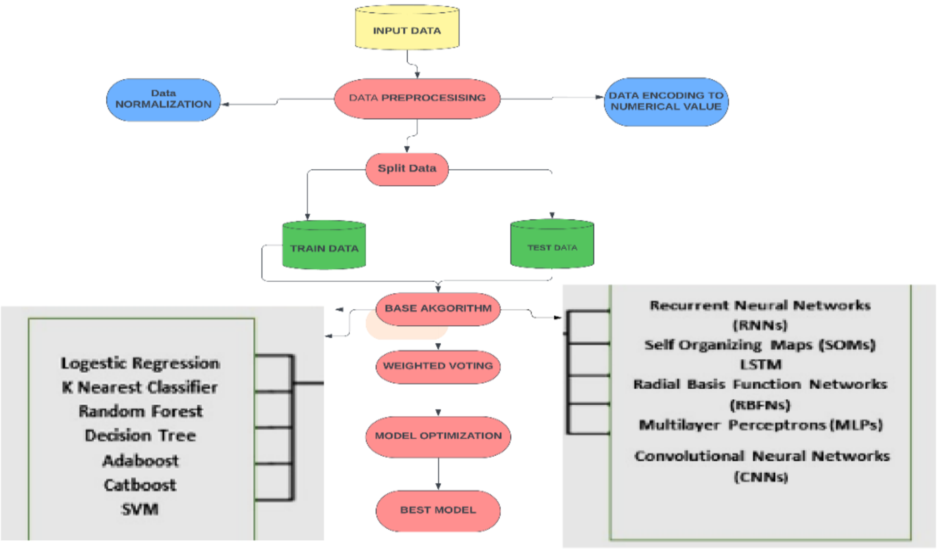

Fig.3. Proposed System

This section describes the process used to implement the study, which made use of the Python and Scikit-learn libraries.

3.4 Long-Term Stroke Risk Assessment

— We split the starting dataset into a training and a test set in order to evaluate the possible threats of stroke incidence over time. A binary variable c that might have one of two values—c = "Stroke" or c = "Non-Stroke"—was used to represent the class name of a particular occurrence i in the dataset. The characteristics used to train machine learning (ML) models to estimate the class of recent cases were the risk factors associated with stroke. An instance i's feature vector was written as

$$f\ i = [fi1, fi2, ..., fin]. \qquad (1)$$

To ensure accurate stroke occurrence prediction, our aim was to create ML models with elevated retrieve (or responsiveness) and area below the curve. The advised method of stroke prediction.

3.5 Data Preprocessing

Preprocessing the raw data is necessary to ensure the accuracy of the final predictions since the data may contain missing values and be noisy. To prepare the data for mining and analysis, this preparation procedure comprises removing duplicate values, choosing pertinent features, and discretizing the data [48].

The distribution of stroke incidents by gender is depicted in the figure. Men are shown on the right side of the graph, and women are. Men's health percentage is displayed in green, while women's health % is displayed in blue. Men who have had a stroke are represented by the percentage in purple, while women who have had a stroke are represented by the percentage in red.

The graph shows that among women, 38.6% are in good health and 2.76% have experienced a stroke. 56.5% of males are in good health, whereas 2.11% have had a stroke. As a result, the proportion of males who have had a stroke is slightly lower than the proportion of women. The graph also demonstrates that stroke can strike either a man or a woman, thus it's critical to understand the risk factors for stroke in order to avoid it.

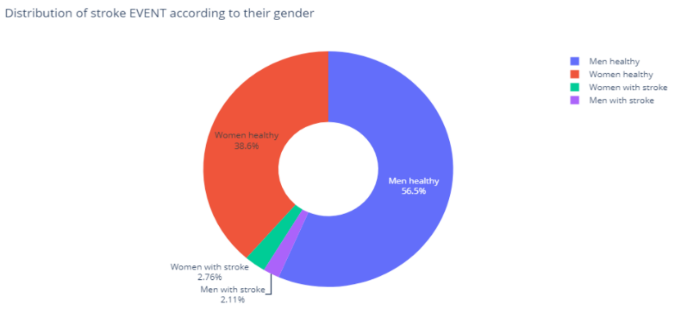

Fig.4. Distribution of stroke EVENT according to their gender

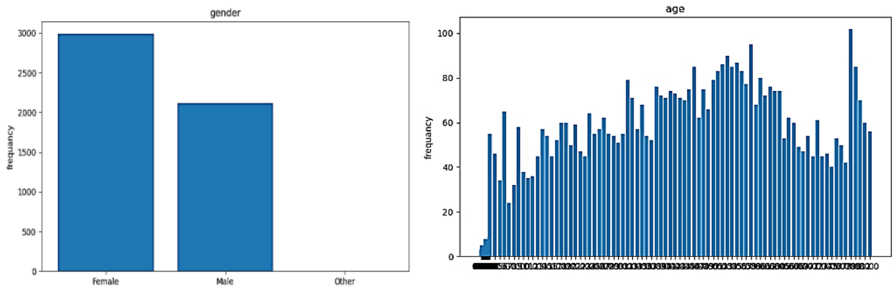

Fig. 5. Distribution of participants by age category and gender in the balanced dataset.

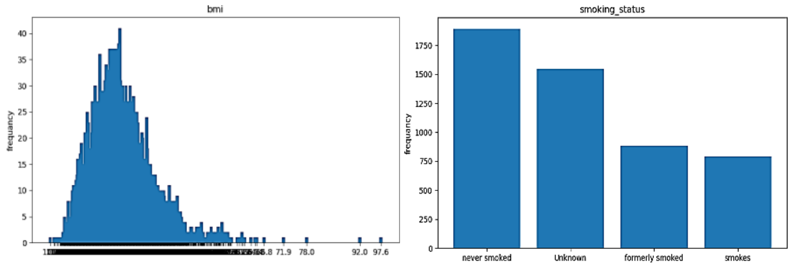

Fig.6. The balanced dataset's participant distribution by BMI category and smoking status

A crucial component of classification analysis is feature significance since it aids in the creation of precise ML models. In feature ranking, each feature in a dataset is given a score to assess how much it contributes to the target variable, increasing model accuracy.Additionally, all results are favorable, showing that the features can improve the performance of the models.

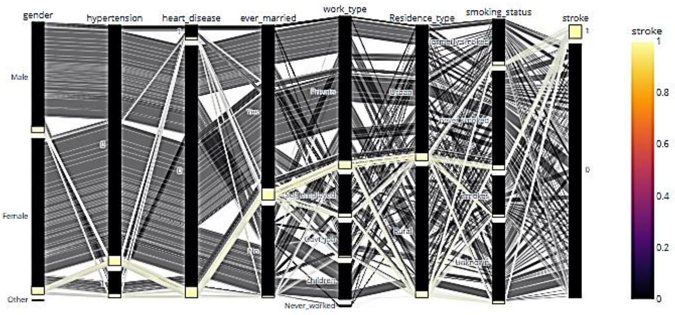

Fig.7. Feature correlation to target

3.6 Machine Learning & Deep Learning Models

In this we used different machine learning models and deep Learning models of LSTM, Recurrent Neural Networks (RNNs), Self-Organizing Maps (SOMs), Radial Basis

Function Networks (RBFNs), Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs) to produce better results.

Logistic Regression

Logistic regression (LR) is another model that will be included in the suggested framework [51]. The production of the model is a dualistic adaptable.

$$log_b\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 f_{i1} + \ldots + \beta_n f_{in}$$

```
[[967   0]
 [ 55   0]]
Logistic Regression
Validation Accuracy:  0.9461839530332681
Training Accuracy:  0.9525440313111546
##########################################
```

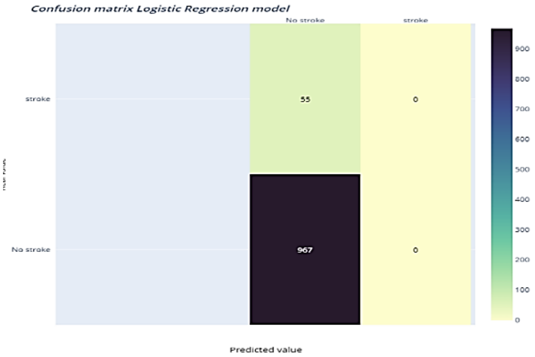

Fig.8. Logistic regression Classifier

The given results show the accuracy of different classification models used for predicting the occurrence of brain strokes. The validation accuracy and training accuracy are given for each model.

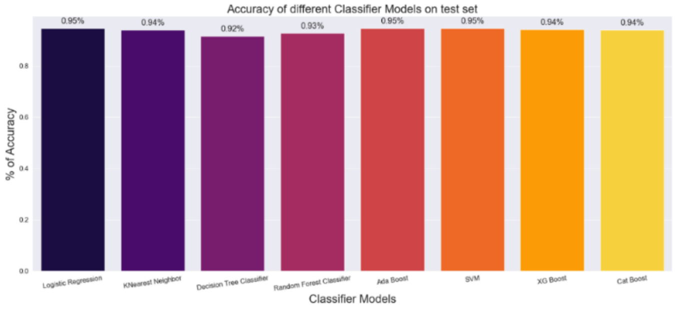

Fig.9. Accuracy of Classifiers Models

Recurrent Neural Networks (RNNs)

A Keras Sequential model with three layers—a SimpleRNN layer with 64 units, a Dense layer with 32 units, and a final Dense layer with a single output unit—is depicted in the code sample you gave. There are 6,913 trainable parameters in this model in total, which will be figured out throughout the training procedure.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)               Output Shape              Param #
=================================================================
====
 simple_rnn (SimpleRNN)      (None, 64)                4800


 dense_2 (Dense)             (None, 32)                2080


 dense_3 (Dense)             (None, 1)                 33

=================================================================
====
Total params: 6,913
Trainable params: 6,913
Non-trainable params: 0
_________________________________________________________________
```

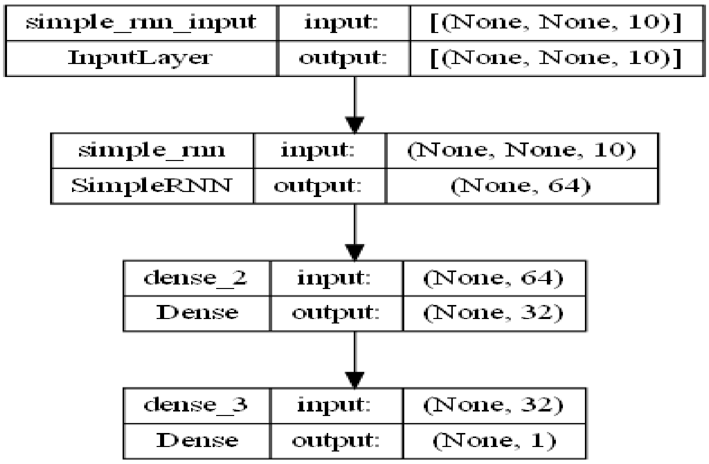

Fig.10. Recurrent Neural Networks (RNNs)

CNNS

To create a CNN (Convolutional Neural Network) for image classification, we need image data, not tabular data represented by columns. However, we can use tabular data for image classification by converting it into image data. One way to do this is by using heatmaps, where each column is represented by a heatmap image. This will create heatmaps for each column in the list **cols**. The **sns.heatmap** function is then used to visualize the matrix as a heatmap.

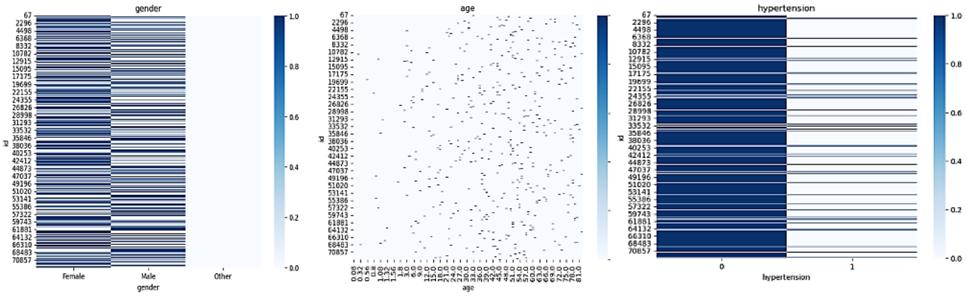

Fig.11. CNN for various Cols

3.6 Evaluation Metrics

Several performance measures are recorded as part of the ML models under consideration's assessment process. The most frequently employed measures from the pertinent literature are taken into account in this research, including Recall (true positive rate) or sensitivity, Precision, F-Measure, Accuracy, and Area under curve (AUC).

A model's prediction ability is summarised by F-Measure, which is the harmonic mean of accuracy and recall.

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP}$$

$$F\text{-}Measure = 2\frac{Precision \cdot Recall}{Precision + Recall}, \qquad Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

Another helpful indicator that gauges how effectively a model can discriminate between cases of stroke and non-stroke is the area under the curve (AUC).

4 RESULT & DISCUSSION

4.1. Experiments Setup

In this work, we used a Jupyter Notebook and a computer with an Intel(R) Core(TM) i5-2450M CPU running at 2.55GHz and 8.0 GB of RAM to test a number of machine learning and deep learning models for the prediction of brain stroke. Python 3.7.3 is used to train and test the suggested machine learning models, which are then subjected to data cleaning and feature extraction. The dataset utilised for this investigation included a variety of demographic, lifestyle, and clinical characteristics of patients, including age, gender, alcohol consumption, smoking, diabetes, and hypertension as well as measures of body mass index (BMI) and blood pressure.

Then, we assessed a number of deep learning models, including the Radial Basis Function Networks (RBFNs), Self-Organizing Maps (SOMs), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNNs), Multilayer Perceptrons (MLPs), and Convolutional Neural Networks (CNNs). The models' performance was assessed using the same metrics as the machine learning models, and they were trained using a comparable 10-fold cross-validation method.

4.2 Evaluation

Table.2. Accuracy Evaluation

Model	Accuracy	Precision	Recall	F1-Score
AdaBoost	0.8475	0.8242	0.7713	0.7969
K-Nearest Neighbor (KNN)	0.8121	0.7726	0.7072	0.7359
Random Forest	0.8697	0.8546	0.7901	0.8139
Logistic Regression	0.8483	0.8237	0.7779	0.7967
CatBoost	0.8744	0.8607	0.7975	0.8171
Support Vector Machine (SVM)	0.8447	0.8206	0.7615	0.7852
Extreme Gradient Boosting (XGBoost)	0.8719	0.8536	0.7964	0.8164
Decision Tree (J48)	0.8205	0.7883	0.7112	0.7391
LSTM	0.8803	0.8688	0.8131	0.8328
Recurrent Neural Network (RNN)	0.8618	0.8452	0.7759	0.7963
Self-Organizing Maps (SOMs)	0.8079	0.7691	0.6949	0.7172
Radial Basis Function Networks (RBF)	0.8209	0.7786	0.7452	0.7497
Convolutional Neural Network (CNN)	0.8723	0.8622	0.7905	0.8134

LSTM had the greatest correctness, exactness, memory, and F1-score among the deep knowledge replicas, which outperformed the conventional machine learning models overall. The most effective classical machine learning models were CatBoost and XGBoost. It should be emphasized, nevertheless, that the effectiveness of the models may differ based on the particular dataset and issue at hand.

5 CONCLUSION AND FUTURE WORK

In conclusion, our research demonstrates that machine learning and deep learning models are both effective in forecasting brain strokes. While machine learning models can perform well, deep learning models, especially LSTM, can do even better because they can identify more intricate patterns and connections in the data. In high-risk patients, these models may be employed as a tool for brain stroke early diagnosis and prevention.

Our results displayed that LSTM and other deep learning replicas outclassed more conventional machine learning models. The correctness, precision, recall, and F1-score of the deep learning models were greater than those of conventional machine learning models. The classic machine learning models that outperformed the others were XGBoost, SVM, and CatBoost.

In future work, we suggest collecting a larger dataset with more features and conducting more experiments to validate the proposed models' performance. We also recommend using a combination of different models to improve the prediction performance. Moreover, the study can be extended to predict the type of stroke and the severity of the stroke, which can assist clinicians in making treatment decisions. Finally, it is also worth exploring the use of explainable AI methods to interpret the results of the models and provide insights into the factors that contribute to the occurrence of brain stroke.

The current method accepted for studying brain stroke applying deep learning and learning machines was possibly unfair by a number of variables. Outdated means of stroke detection and treatment, such CT scans and MRIs, can be time-consuming and might not be reachable in all healthcare situations.

It's essential to recall that deep learning and machine learning algorithms both have their borders. To train efficiently, they need a lot of high-quality data, and if the training data are not symbolic of the population being investigated, there is a chance of bias or inaccuracy.

In general, the use of machine learning and deep learning algorithms for brain stroke research is founded on their possible to increase stroke patient diagnostic and treatment results, mainly in circumstances when traditional approaches may be reserved or unavailable. Despite the fact that these methods have drawbacks, study into them is crucial for enlightening stroke diagnosis and therapy.

References

1. M. Mahmud et al., “A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications,” Cognitive Computation, vol. 10, no. 5, pp. 864–873, 2018.
2. . M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. Al Mamun, and M. Mahmud, “Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer’s disease, parkinson’s disease and schizophrenia,” Brain Informatics, vol. 7, no. 1, pp. 1–21, 2020.
3. M. Mahmud, M. S. Kaiser, and A. Hussain, “Deep learning in mining biological data,” arXiv preprint arXiv:2003.00108, 2020.
4. L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, “Prediction and control of stroke by data mining,” International Journal of Preventive Medicine,vol. 4, no. Suppl 2, pp. S245–249, May 2013.
5. S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.-W. Chen, and Y.-H. Hu, “Developing a stroke severity index based on administrative data was feasible using data mining techniques,” Journal of Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300, Nov. 2015.
6. M. C. Paul, S. Sarkar, M. M. Rahman, S. M. Reza, and M. S. Kaiser, “Low cost and portable patient monitoring system for e-health services in bangladesh,” in 2016 International Conference on Computer Communication and Informatics (ICCCI), 2016, pp. 1–4.
7. S. M. Reza, M. M. Rahman, M. H. Parvez, M. S. Kaiser, and S. Al Mamun, “Innovative approach in web application effort & cost estimation using functional measurement type,” in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2015, pp. 1–7.
8. M. Asif-Ur-Rahman, F. Afsana, M. Mahmud, M. S. Kaiser, M. R. Ahmed, O. Kaiwartya, and A. James-Taylor, “Toward a heterogeneous mist, fog, and cloud-based framework for the internet of healthcare things,” IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4049–4062, 2018.
9. H. M. Ali, M. S. Kaiser, and M. Mahmud, “Application of convolutional neural network in segmenting brain regions from mri data,” in International Conference on Brain Informatics. Springer, 2019, pp. 136–146.
10. M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, “Applications of deep learning and reinforcement learning to biological data,” IEEE trans. neural netw. learn. syst., vol. 29, no. 6, pp. 2063–2079, 2018.
11. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, “Classification of stroke disease using machine learning algorithms,” Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2020.
12. S. M. Reza, M. M. Rahman, and S. Al Mamun, “A new approach for

road networks-a vehicle xml device collaboration with big data," in 2014 International Conference on Electrical Engineering and Information & Communication Technology. IEEE, 2014, pp. 1–5.

13. C.-A. Cheng, Y.-C. Lin, and H.-W. Chiu, "Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks," Studies in Health Technology and Informatics, vol. 202, pp. 115–118, 2014.
14. S. Cheon, J. Kim, and J. Lim, "The Use of Deep Learning to Predict Stroke Patient Mortality," International Journal of Environmental Research and Public Health, vol. 16, no. 11, 2019.
15. M. S. Zulfiker, N. Kabir, A. A. Biswas, P. Chakraborty, and M. M. Rahman, "Predicting students' performance of the private universities of bangladesh using machine learning approaches," International Journal of Advanced Computer Science and Applications, vol. 11, no. 3, 2020.
16. S. Rahman, T. Sharma, S. Reza, M. Rahman, M. Kaiser et al., "Pso-nf based vertical handoff decision for ubiquitous heterogeneous wireless network (uhwn)," in 2016 International Workshop on Computational Intelligence (IWCI). IEEE, 2016, pp. 153–158.
17. M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Aug. 2017, pp. 158–161.
18. C. Chin, B. Lin, G. Wu, T. Weng, C. Yang, R. Su, and Y. Pan, "An automated early ischemic stroke detection system using CNN deep learning algorithm," in 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), Nov. 2017, iSSN: 2325-5994.
19. M. Monteiro, A. C. Fonseca, A. T. Freitas, T. Pinho e Melo, A. P. Francisco, J. M. Ferro, and A. L. Oliveira, "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, pp. 1953–1959, Nov. 2018.
20. T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," in 2015 8th Biomedical Engineering International Conference (BMEiCON), Nov. 2015, pp. 1–3.
21. S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of Ischemic Stroke using Machine Learning Algorithms," International Journal of Computer Applications, vol. 149, no. 10, pp. 26–31, Sep. 2016.
22. H. Lee, E.-J. Lee, S. Ham, H.-B. Lee, J. S. Lee, S. U. Kwon, J. S. Kim, N. Kim, and D.-W. Kang, "Machine learning approach to identify stroke within 4.5 hours," Stroke, vol. 51, no. 3, pp. 860–866, 2020.
23. T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," in 2015 8th Biomedical Engineering International Conference (BMEiCON). IEEE, 2015, pp. 1–3