Investigating Machine Learning Algorithms for Stroke Occurrence Prediction

1Kazeem B. Adedeji, 1Titilayo A. Ogunjobi, 2Thabane H. Shabangu, and 1Joshua A. Omowaye

1Department of Electrical and Electronics Engineering, Federal University of Technology Akure, Ondo State, Nigeria

2Department of Electrical Engineering, Tshwane University of Technology, Pretoria, South Africa

Corresponding Email: kbadedeji@futa.edu.ng

*Abstract*—Stroke is the leading cause of death and the principal cause of long term disability. Accurate prediction of stroke is highly valuable for early intervention of treatment. In this study, six (6) machine learning (ML) algorithms namely: Random Rorest (RF) classifier, Decision Tree Classifier (DTC), KNeighbour Classifier (KN), Support Vector Classifier (SVC), Logistic regression (LR) and Stacking Classifier (SC) were trained on 10 stroke risk factors to determine the most precise model for predicting the risk of stroke occurrence. The primary contribution of this work is the development of a stacking method that achieves high performance, as measured by various metrics such as Area under Curve (AUC), precision, recall, F1 score, and accuracy. The experimental results indicate that the stacking classification outperforms other methods, with an AUC of 98.80%, F1 score of 95.18%, precision of 95.08%, recall of 95.28%, and accuracy of 95.25%. The result revealed that the stacking classifier is an ideal classifier that can be used by physician for early detection of probability of stroke occurrence.

Keywords—accuracy; data processing; machine learning; stroke prediction.

# Introduction (*Heading 1*)

Stroke occurs due to the interruption of the flow of blood to a part of the brain as a result of blood cloth. Globally stroke is one of the most severe diseases and it is directly responsible for a considerable number of death. According to the World Stroke Organization, 15 million people suffer a stroke each year out of these approximately 5 million people die as a result and another 5 million are left permanently disabled [1-5]. It is therefore considered as the leading causes of death and disability worldwide. It not only affects patients but also impacts their social environment, family, and workplace. Contrary to popular belief, stroke can happen to anyone, at any age, regardless of gender or physical condition. Each year, millions of stroke survivors have to adapt to a life with restrictions in daily activities. Many face problems such as memory, concentration, attention issues, speech difficulties, emotional problems, loss of balance, and difficulty swallowing [6]. Depending on the cause of stroke, strokes can be categorized into three; ischemic stroke, hemorrhagic stroke, and transient ischemic attack (TIA) as shown in Fig. 1. In ischemic stroke, the arteries supplying blood to the brain completely become blocked. The hemorrhagic stroke occurs happens when an artery in the brain breaks leaks blood. As a result, the blood from that artery creates excess pressure in the skull and swells the brain, damaging brain cells and tissues. The TIA on the other hand is sometimes referred to as a mini stroke. It occurs when blood flow to the brain is blocked temporarily. While its symptoms are similar to those of hemorrhagic stroke, they typically disappear after a few minutes or hours when the blockage moves and blood flow is restored.



1. *Classification of stroke [7].*

The rising cost of hospitalization for stroke patients necessitates the development of advanced technologies to aid in clinical diagnosis, treatment, and prediction of clinical events. Early detection is critical for effective stroke treatment, making machine learning (ML) a vital tool. ML algorithms can learn complex patterns by integrating numerous variables from high-dimensional data. This capability enables health professionals to make informed clinical decisions and deliver accurate, rapid predictions [8-11]. To advance research in this area, this study employs six distinct machine learning algorithms to predict the likelihood of stroke occurrence in individuals based on certain parameters such as age, gender, body mass index (BMI), smoking status, heart disease, marital status, hypertension, and average glucose level. The machine learning models are trained from the data generated from these parameters. After training the models, their performance was evaluated based on the prediction accuracy, recall, F1 score, precision.

# Review of Related Studies

Several studies have explored the use of machine learning algorithms for stroke prediction. Elias *et* al [12] trained eight machine learning algorithms namely Naive Bayes (NB), RF, (LR), Stochastic Gradient Descent (SGB), Multilayer Perceptron (MLP), Majority Voting (MV) and Stacking Algorithm (SA) on the dataset from Kaggle to classify whether an individual will have a stroke or not. Out of all the algorithm used, Stacking classifier was the most efficient with a precision score of 97%, recall score of 978%, F1 score of 97.4%, accuracy score of 98%, and AUC score 98.9% respectively followed by Random forest classifier with a precision score of 95.5%, recall score of 97.6%, F1 score of 96.5%, accuracy score of 97% and AUC score of 98.6% then Majority voting classifier with a precision score of 92.3%, recall score of 93.8%, f1 score of 93.1%, accuracy score of 93% and AUC score of 93%. Gangavarapu *et* al, [13] used LR, DT, RF, KN, SVC and NB to predict the occurrence of stroke. They collected their dataset from Kaggle which contains records of 5110 patient out of which 249 patient had stroke and 4861 does not. Due to the high level of imbalance, an under sampling technique was used to handle the imbalance. Among all the algorithms evaluated, NB gave the best performance with an accuracy of 82%, precision score of 79.2%, recall score of 85.7% and f1 score of 82.3%. Then followed by support vector classifier with an accuracy score of 80%, precision score of 78.6%, recall score of 83.8% and f1 score of 81.1%, RF performed the least with accuracy score of 73%, precision score of 72%, recall score of 73.5% and f1 score of 72.7%.

Tazin *et* al, [14] used a DT, RF, LR, voting classifier for stroke prediction. The dataset used contains 5110 records and 249 records has stroke while 4861 record does not. Due to the imbalance in the dataset, synthetic minority oversampling technique was used to balance the dataset. Among all the algorithms used for the prediction, RF performed best with an accuracy score of 96%, precision score of 95%, recall score of 97% and f1 score of 96% respectively. In Emon et al. [15] stroke prediction was performed. In this study, 5110 patient record obtained from medical clinic of Bangladesh was used to predict the occurrence of stroke. For analysis, 10 different algorithms namely LR, SGD, DT, adaboost, gaussian, quadratic discriminant analysis, multilayer perceptron, knearest neighbour, gradient boosting, extreme gradient boosting were used for training purpose, also a weighted voting classifier was used to improve the accuracy of all the classifiers. The result presented showed that an accuracy of 97%, FPR of 0.9% and FNR of 3% was achieved. Wiryaseputra [16] used DT, RF, XGboost, LR for stroke prediction. In this study, the dataset used contains 5110 records and 249 out of this record has stroke while 4861 record does not. Similarly to the study presented in [14], SMOTE was used to balance the dataset. Among all the algorithms used for prediction, RF performed best with an accuracy score, precision score, recall score, F1 score of 99.27%, 98.56%, 100% and 99.27% respectively. In Islam *et* al. [17] DT, RF, KNN, LR were trained for stroke prediction. The dataset used by them contains 5110 records of stroke patient. The result presented show that RF performed best with precision score, recall score, f1 score of 96%, 96% and 96% respectively**.**

# Research Method

## Data Source and Dataset Description

Fig. 2 shows the graphical representation of the steps employed in this study for predicting stroke occurrence. The dataset used in this study was obtained from Kaggle (an open source data repository), and it contains the records of 5110 patients, their age is between 25 years to 82 years and majority of the patient are females with a count of 2994 as compared to the male with a count of 2116. All the other attributes (10 as input to the machine learning model is described in Fig. 3). Most of the features are categorical except for age, average glucose level and BMI which are numerical.

## Data Processing

This is a data mining technique which is used to clean, prepare the raw data order to make it more suitable for machine learning analysis. Since the dataset used in this study contains some categorical variables, it is then necessary to clean it in order for the models to produce a more accurate output. The data preprocessing steps include:

1. Data cleaning: Data cleaning was not necessary for the dataset set used in this study as there were no missing values, no null values and no Nan values as seen in Fig. 4. Fig. 4 shows that there is no missing or null value present in the dataset.
2. Data transformation: Categorical data were encoded using label encoding. Features were scaled using standardization.
3. Data balancing: SMOTE was used to address class imbalance, ensuring an equal distribution of stroke and no-stroke instances. Fig. 5 and Fig. 6 show the effect before and after applying SMOTE.

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) was employed to examine the datasets to summarize their key characteristics using statistical graphics and other visualization techniques. It also enhances understanding of the data and detects patterns that might not be immediately apparent from merely examination. EDA included visualizing the distribution of various attributes and their relationship with stroke. Fig. 7 and Fig. 8 show the distribution of the attributes in the dataset and correlation among the features.

**

1. *Activity diagram of the research method.*

**

1. *Feature attributes from the dataset.*

**

1. *No missing or null value present in the dataset.*

**

1. *The count of stroke occurrence in the dataset before applying SMOTE.*

**

1. *The count of stroke occurrence in the dataset after it was balanced with SMOTE.*

**

1. *The distribution of all variables in the dataset.*

**

1. *A heatmap showing the correlation among the stroke attributes.*

## Machine Learning Models

1. Logistic Regression: This predicts the probability of the target variable belonging to a certain class based on the values of the input feature (such as age, BMI, avg\_glucose\_level) according to (1).

 (1)

where *p*=*P*(*Y*=1) is the probability of an instance belonging to the stroke class, *Xi* are the input features, and *βi* are the coefficients.

1. Support Vector Machines (SVM): SVMs work by finding a N-dimensional hyper plane which can distinguish N-dimensional data points. The hyper plane is the line that best separates the data points with the largest margin which is the distance between the hyper plane and the closest data points (support vectors).When the data is not linearly separable, a kernel function was used. The kernel function transforms the data into a higher-dimensional space, where the data may be linearly separable.
2. Decision Tree: It works by recursively splitting the training dataset into smaller subsets based on the most significant feature that provides the most information gain or decrease in impurity.
3. Random Forest: Random forests are an ensemble of decision trees. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all trees.
4. K-Nearest Neighbors (KNN): The idea behind KNN is to identify the K nearest data points (i.e., neighbors) to the data point to be classified, and use the class labels of these neighbors to predict the class label of the data point.
5. Stacking Classifier: This is an ensemble learning technique that involves combining multiple classification models via a meta-classifier. Thereafter the output of several individual classifiers are combined and then fed into a meta-classifier to make the final prediction.

## Performance Metrics

The performance of all the classifier was accessed based on the following metrics

1. Accuracy (*A*):This is the ratio of correctly predicted instances (both positive and negative) to the total instances. This was estimated using (2).

 (2)

where *TP* is the true positive, that its, the number of positive instances correctly identified by the model,

*FP* is the false positive (the number of negative instances incorrectly identified as positive by the model), *TN* is the True Negative (the number of negative instances correctly identified by the model), and *FN* is the false negative (the number of positive instances incorrectly identified as negative by the model).

1. Precision (*P*): This is the ratio of correctly predicted positive instances to the total predicted positive instances. This was computed using (3).

 (3)

1. Recall (*R*): This is the ratio of correctly predicted positive instances to the total actual positive instances. This was computed using (4)

 (4)

1. F-measure (F1 Score): This is the harmonic mean of *P* and *R*, providing a balance between the two. This was estimated using (5).

 (5)

1. ROC-AUC Curve: The Receiver Operating Characteristic (ROC) is a graphical representation of a classifier's performance across all classification thresholds. The Area Under the Curve (AUC) represents the degree or measure of separability between classes.   
   The ROC curve plots True Positive Rate (Recall) against False Positive Rate (FPR). The false positive rate is calculated using (6).

 (6)

# Results And Discussions

Fig. 9 shows the training score recorded by the six algorithms. This bar chart compares the training scores of various machine learning models applied to the stroke dataset. The Stacking classifier achieved the highest training score of 99.86%, followed closely by RF with 99.07%. LR had the lowest score of 81.07%, indicating its lower performance in capturing patterns during training compared to other models.



1. *Comparison of the training score for the trained classifiers.*

Fig. 10 illustrates the accuracy scores of the six classifiers when predicting stroke occurrences. Stacking had the highest accuracy at 95.23%, indicating robust performance. SVC and Random Forest also performed well, with scores of 94.03% and 93.42%, respectively. Logistic Regression had the lowest accuracy at 79.97%, showing that it may not be the best choice for this dataset compared to the other classifiers.



Fig. 10. *Comparison of the accuracy score for the trained classifiers.*

In Fig. 11, the precision score shows how well each classifier identified positive cases of stroke without including too many false positives. SVC and Stacking performed best, with precision scores of 95.64% and 95.08%, respectively, indicating strong precision in their predictions. Random Forest also had a high precision of 92.1%. Logistic Regression had the lowest precision score of 79.06%, suggesting that it may misclassify more non-stroke cases as strokes compared to other models.

Fig. 12 shows the recall score recorded by the six algorithms. The K-Nearest Neighbors (KNN) model achieved the highest recall score of 99.19%, followed by Random Forest and Stacking with scores of 95.31% and 95.28%, respectively.



Fig. 11. *Comparison of the precision score for the trained classifiers.*

LR had the lowest recall score of 81.72%, suggesting it misses a higher proportion of true stroke cases compared to other models. These results highlight the superior recall performance of KNN, RF, and Stacking classifiers in detecting strokes.



Fig. 12. *Comparison of the recall score for the trained classifiers.*

Fig. 13 shows a comparison of the F1 scores for six different machine learning models.



Fig. 13. *Comparison of the F1 score for the trained classifiers.*

As shown in Fig. 13, the Stacking classifier achieved the highest F1 score of 95.18%, indicating it had the best balance between precision and recall for this stroke dataset. Random Forest also performed well with an F1 score of 93.68%, followed by SVC at 94.07%. LR had the lowest score at 80.37%, showing relatively poorer performance compared to the others.

Fig 14 shows the AUC score recorded by the six algorithms. Stacking classifier performs the best with an AUC score of 98.8%, followed closely by SVC (Support Vector Classifier) at 98.58%. K-Nearest Neighbors (KNN), Random Forest, and Decision Tree also perform well, with scores in the 93%-96% range. Logistic Regression has the lowest performance, with an AUC score of 88.35%.



Fig. 14. *Comparison of the AUC score for the trained classifiers.*

TABLE I presents the summary of the performance of the six classifiers. Based on the results from the table, it is evident that the stacking classifier outperforms other models in predicting stroke occurrence, achieving the highest testing accuracy score of 95.2348%. The support vector machine also performed well with a high accuracy score of 94.0349%. Notably, the KNN classifier had the lowest false negative rate, indicating its effectiveness in correctly identifying individuals at high risk of stroke. The random forest classifier also demonstrated strong performance with an accuracy score of 93.4178% and a high AUC score of 98.2212. Conversely, the logistic regression classifier had the lowest accuracy score of 79.967%, showing it was less effective compared to the other models. In summary, the stacking classifier stands out as the best algorithm for predicting stroke.

1. Summary of the overall performance of the classifiers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classifiers | Accuracy | AUC | TP | FP | TN | FN |
| KNN | 89.88 | 95.95 | 1468 | 283 | 1154 | 12 |
| SVM | 94.03 | 98.59 | 1381 | 63 | 1362 | 111 |
| DT | 89.4 | 93.89 | 1346 | 163 | 1262 | 146 |
| RF | 93.41 | 98.22 | 1422 | 122 | 1308 | 70 |
| LR | 79.96 | 88.35 | 997 | 264 | 947 | 223 |
| Stacking | 95.2 | 98.8 | 1373 | 71 | 1405 | 66 |

# V. Conclusion and Future Studies

This study has conclusively demonstrated the prowess of machine learning algorithms, particularly the stacking classifier, in predicting stroke risk with remarkable accuracy. The achieved high performance on metrics such as AUC, precision, recall, and F-measure underscores the transformative potential of these models in facilitating early stroke detection and intervention. Nevertheless, it is crucial to acknowledge the limitations of the current approach. Firstly, the models are not yet capable of predicting the specific type of stroke an individual might experience. Secondly, the research primarily relies on text-based data, which constrains the model's ability to fully capture the multifaceted nature of stroke risk factors.

Moving forward, further study will be to transcend these limitations by integrating brain imaging data, such as MRI or CT scans to enhance the model's predictive power and enable the identification of specific stroke types and implementing advanced techniques to elucidate the model's predictions, providing deeper insights into the contributing factors and mechanisms underlying stroke risk.

##### References

1. “About stroke by cdc,” [Online]. Available from : https:// [www.cdc.gov/stroke/index.html](http://www.cdc.gov/stroke/index.html). [Accessed: 24/10/2024].
2. E.S. Donkor, "Stroke in the 21st century: a snapshot of the burden, epidemiology, and quality of life," Stroke Research and Treatment, vol. 2018, no. 1, p. 3238165, 2018.
3. R.O. Akinyemi, B. Ovbiagele, O.A. Adeniji, F.S. Sarfo, F. Abd-Allah, T. Adoukonou, O.S. Ogah, Naidoo, A. Damasceno, R.W. Walker, and A. Ogunniyi, “Stroke in Africa: profile, progress, prospects and priorities,” Nature Reviews Neurology, vol. 17, no. 10, pp.634-656, 2021.
4. J.N. Fernandes, V.E. Cardoso, A. Comesaña-Campos, and A. Pinheira, “Comprehensive review: Machine and deep learning in brain stroke diagnosis, Sensors, vol. 24, no. 13, pp. 1-27, 2024.
5. S. Gupta, and S. Raheja (2022). “Stroke prediction using machine learning methods,” In: Proceedings of the 12th IEEE International Conference on Cloud Computing, Data Science and Engineering, 27-28 January, 2022, India.
6. B. Delpont, C. Blanc, G. Osseby, M. Hervieu-Bègue, M. Giroud, and Y. Béjot, “Pain after stroke: A review,” Rev. Neurol., vol. 174, pp. 671–674, 2018.
7. Healthline 2024. Everything you eed to know about stroke.https://www.healthline.com/health/stroke
8. M. Javaid, A. Haleem, R.P. Singh, R. Suman, and S. Rab, “Significance of machine learning in healthcare: Features, pillars and applications,” International Journal of Intelligent Networks, vol. 3, pp. 58-73, 2022.
9. A. Rahman, M. Karmakar, and P. Debnath, “Predictive analytics for healthcare: Improving patient outcomes in the US through machine learning,” Revista de Inteligencia Artificial en Medicina, vol. 14, no. 1, pp.595-624, 2023.
10. R. Avula, “Data-driven decision-making in healthcare through advanced data mining techniques: A survey on applications and limitations,” International Journal of Applied Machine Learning and Computational Intelligence, vol. 12, no. 4, pp.64-85, 2022.
11. M. Nasiruddin, S. Dutta, R. Sikder, R. Islam, A.A. Mukaddim, and M.A. Hider, “Predicting heart failure survival with machine learning: Assessing my risk, “ Journal of Computer Science and Technology Studies, vol. 6, no. 3, pp.42-55, 2024.
12. D. Elias,and M. Trigka, “Stroke risk prediction with machine learning techniques”, Sensors, vol. 22, no. 13, pp. 1-13, 2022.
13. S. Gangavarapu, and G.L.A. Kumari, “Analyzing the performance of stroke prediction using ML classification algorithms,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 145-149, 2021.
14. [T. Tazin](https://pubmed.ncbi.nlm.nih.gov/?term=Tazin+T&cauthor_id=34868531), [M. Alam](https://pubmed.ncbi.nlm.nih.gov/?term=Alam+MN&cauthor_id=34868531) , [N.N. Dola](https://pubmed.ncbi.nlm.nih.gov/?term=Dola+NN&cauthor_id=34868531), [M.S. Bari](https://pubmed.ncbi.nlm.nih.gov/?term=Bari+MS&cauthor_id=34868531), [S. Bourouis](https://pubmed.ncbi.nlm.nih.gov/?term=Bourouis+S&cauthor_id=34868531), and [N.M. Khan](https://pubmed.ncbi.nlm.nih.gov/?term=Monirujjaman+Khan+M&cauthor_id=34868531), “Stroke disease detection and prediction using robust learning approaches,” [J Healthc Eng.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8641997/), vol. 2021, p. 7633381, 2021.
15. M.U. Emon, M.S. Keya, T.I. Meghla, M. Rahman, S. Al Mamun, M.S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," In: Proceedings of the 4th IEEE International Conference on Electronics, Communication and Aerospace Engineering, 5-7 November, 2020, India, pp. 1464-1469.
16. M. Wiryaseputra, "Stroke prediction using machine learning classification algorithm," International Journal of Scientific and Engineering Research, vol. 8, no. 1, 2017.
17. M. Islam, S. Akter, M. Rokunojjaman, J.H. Rony, "Stroke prediction analysis using machine learning classifiers and feature technique," International Journal of Electronics and Communication System, vol. 1, no. 2, pp. 17-22, 2021.