Research Article

Journal of Scientific Enquiry
Vol: 4 (1), Pages: 07 - 14

Early Determination of Brain Stroke Leveraging Multiple Machine Learning Approaches

Pintu Pal[1*], Deblina Banerjee[1], Subhodeep Moitra[1], Bimal Dutta[1], Sanat Kumar Mahato[2]

[1] *Department of Computer Application, Techno College Hooghly, Hooghly-712101, West Bengal, India.*

[2] *Department of Mathematics, Sidho-Kanho-Birsha University, Purulia, 723104, India.*

*Corresponding author

Received: 07.08.2024; accepted: 06.09.2024; published online: May, 2025

Abstract

Brain stroke, also known as a cerebrovascular accident (CVA), happens as the blood supply to the brain's regions is disrupted, depriving the neural tissue of vital oxygen and nutrients. This can result in brain damage, incapacity, or even death. As a result, early detection of brain stroke is critical for saving lives, reducing long-term impairments, and lowering healthcare costs through on-time therapies. In our work, we studied the capability of different machine learning methods for predicting brain stroke at an early stage. Here we applied different machine learning methods such as logistic regression, gaussian naive byes, and Decision tree classifiers to produce significant results with 95% accuracy to identify it. Decision trees and naive byes give a significant probabilistic result to predict the chances of brain stroke in a person in the early stage.

Keywords: Machine learning, Brain Stroke, Logistic Regression, Decision Tree Classifier, Gaussian Nive Byes.

*Email: dr.ppal.aec@gmail.com (corresponding author)

1 Introduction

Brain stroke has emerged as a major cause of fatality and impairment globally. The World Stroke Organization's 2022 report states that over 12.2 million individuals worldwide are affected by stroke every year, which has fatal repercussions [1]. According to the World Health Organization reports stroke is the second leading cause of death and a major source of long-term impairment [2].

In a study that included several regions of India, an estimated population of 22,479,509 was studied, indicating 11,654 stroke incident cases. The annual incidence varied from 108 to 172 per 100,000 people, with a frequency of 26 to 757 per 100,000 and a one-month case fatality rate of 18% to 42% [3]. Strokes are classified into ischemic and hemorrhagic types. The most common type of stroke is ischemic, which occurs when an artery becomes blocked or restricted. The most prevalent type of stroke is ischemic, which takes place when an artery is blocked or restricted, while hemorrhagic strokes occur when a blood vessel in the brain ruptures. This applies to individuals with previous records of age, heart disease, hypertension, smoking, TIAs, average glucose level,

 DOI: https://doi.org/10.54280/jse.254102

BMI, and sedentary behavior. The symptoms could range from erratic paralysis and trouble communicating to disorientation and serious headaches. Prompt care is critical for preventing brain injury and improving patient outcomes. Modern developments in ML, as well as AI, promise possible solutions for estimating stroke risk and diagnosis. ML may investigate enormous amounts of personal information, which involves medical data and behaviors. This tool assists in early detection, facilitating healthcare practitioners to implement preventive measures and make timely interventions. *In 2021*, Sailasya and Kumari [4] used machine learning methods such as Logistic Regression (LR) and Naive Bayes Classification to predict brain stroke. They discovered that the Naive Bayes approach was the most successful, achieving an accuracy of around 82%. *In 2023*, Ojha and Jha [5] used a few ML algorithms—Naive Bayes, AdaBoost as well as Decision Table, k-NN, alongside Random Forest to identify strokes based on individuals' medical records and physical activity. The Decision Table method fared the best, obtaining an accuracy of 82.1%. *In 2023* Mridha et al. [6] created a computerized stroke prediction system by combining multiple machine learning algorithms with explainable methodologies such as SHAP and LIME. Their proposed framework obtained up to 91% accuracy, demonstrating the value of integrating intricate models with clarity for stroke estimation.

In 2021, Vempati Krishna et al. [7] used a variety of machine learning techniques to anticipate brain stroke, emphasizing factors such as high blood sugar, heart disease, diabetes, and high blood pressure. Tahia Tazin et al. (2021) [8] constructed stroke prediction models utilizing Logistic Regression, Decision Tree Classification, Random Forest Classification, and a Voting Classifier, highlighting the importance of robust model construction. In 2024, Ahmad A. Abujaber et al. [9] used the XGBoost model for early detection of posterior circulation stroke (PCS), with an AUC of 0.81 and accuracy of 0.79 to identify important factors such as the Body Mass Index, Random Blood Sugar, ataxia, dysarthria, systolic blood pressure, and thermal regulation through SHAP analysis to improve diagnostic value. In 2024, Ching-Heng Lin et al. [10] explored machine learning algorithms for stroke outcome prediction using a statewide disease registry, emphasizing the efficacy of leveraging extensive clinical data to improve prediction capacities regarding both acute and chronic strokes. In 2022, Nitish Biswas et al. [11] investigated ML classifiers for stroke prediction using imbalanced data. The authors utilized Random Over Sampling (ROS) and hyperparameter tuning to improve the performance of models throughout eleven classifiers, including Support Vector Machine, Random Forest, and Naïve Bayes. This proposed work focuses on three machine learning algorithms for predicting stroke occurrence: logistic regression (LR), Gaussian Naive Bayes algorithm, and decision tree classifier The performance of such models is evaluated using metrics that include accuracy, recall, F1 score, precision, ROC-AUC, and accompanying heat maps. The primary goal of this research is to analyze and compare various machine learning algorithms for predicting the chance of a stroke. This study intends to improve early detection capabilities, facilitate prompt medical interventions, and help contribute to a better knowledge of stroke prediction approaches. The ultimate objective is to deliver insights that will assist in enhancing stroke risk

assessment and contribute to the development of efficient preventative interventions.

2 Procedure and Methodology

2.1 Data Description

The data used in this investigation was obtained from Kaggle. The original dataset consists of 5110 rows and 12 columns. It comprises 5110 data points and 12 features, which were used to train and validate machine learning models. Table 1 describes several parameters incorporated into the dataset. The term "subject" implies the individual whose data was used.

Table 1: Description of Features Used in the Study

Feature	Description
Id	Specifies the ID of the subject
Age	Indicates whether the subject is male or female
Gender	Contains the age of the subject
Hypertension	Indicates whether the subject has hypertension
Heart Disease	Indicates whether the subject has any heart ailments
Ever married	Indicates whether the subject is married
Work type	Contains the job type of the subject
Residence type	Indicates where the subject resides
Avg_glucose level	Indicates the average glucose level of the subject
Bmi	Indicates the BMI of the subject
Smoking status	Indicates whether the subject smokes or not
Stroke	Tells whether the subject had a stroke or not

2.2 Data Preprocessing

Throughout the process of machine learning model training, substantial dataset preprocessing was needed to assure data accuracy and algorithm compatibility. The preprocessing processes included addressing missing values, converting categorical data to numerical values, and standardizing the attributes.

Handling **Missing Values**: Initially, the dataset contained missing values, particularly in the "BMI" variable. 201 entries were missing in this attribute, accounting for nearly 3.93% of the dataset. Rows having missing values for BMI were eliminated to ensure the dataset's integrity. Figure 1 and Figure 2 illustrate the distribution of missing values before and after this cleaning process.

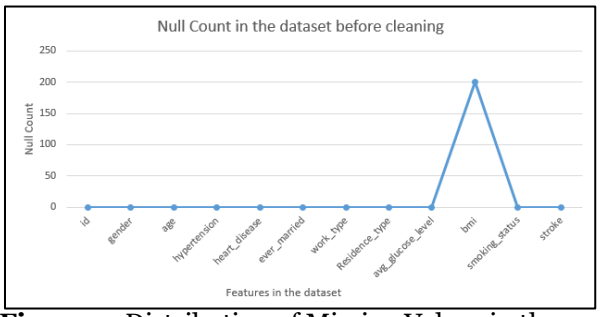

Figure 1: Distribution of Missing Values in the Dataset

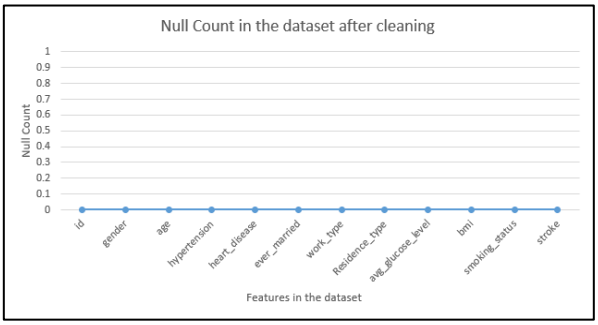

Figure 2: Dataset after Removing the Null values

Categorical Data Transformation: Several non-numeric columns were encoded into numerical values using various techniques.

Binary Variables: Gender and Ever Married have been converted to binary values. For example, 'Female' was encoded as 0 'Male' as 1 for Gender, 'No' as 0, and 'Yes' as 1 for Ever Married.

Categorical Variables with Multiple Categories: Work Type, Residence Type, and Smoking Status were encoded with one-hot encoding, resulting in separate binary columns for each category. For example, Work Type was encoded into columns such as 'Private', 'Self-employed', 'Govt_job', 'Children', and 'Never_worked', each one indicating a distinct category with binary values. A graphical representation showing the average stroke percentage concerning gender, residence type, smoking status, and work type is presented in Figure 3.

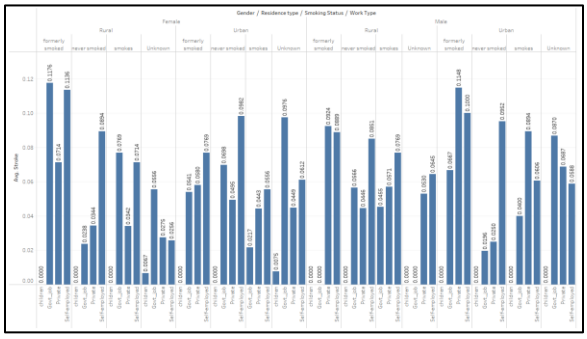

Figure 3: Illustration of the average stroke percentage concerning gender, residence type, smoking status, and work type

Feature Standardization: Standard scaling was used on the numerical features to achieve consistent scalability across all input variables. This procedure norm features a mean of zero as well as a standard deviation of one, hence improving the efficiency and convergence of machine learning models.

Data Splitting: The set of data was split into subsets for training and testing using an 80/20 split ratio. This strategy helps ensure models are trained on a comprehensive dataset while being evaluated on previously unseen data, allowing for a more effective evaluation of model performance.

2.3 Proposed Models

Stroke is the most prevalent disease diagnosed in the medical field, and its prevalence is increasing year after year. The proposed study compares three popular machine learning algorithms for predicting the chance of a stroke. The approaches are:

- Logistic Regression,
- Gaussian Naive Bayes
- Decision Tree.

2.3.1 Logistic Regression:

Logistic Regression (LR) is a simple yet powerful statistical technique for binary classification applications [12]. Using a logistic function, it determines if a sample belongs to the stroke or non-stroke class:

$$log_b\ (\frac{p}{1-p}) = \beta_0 + \beta_1 f_{i1} + ... \beta_n f_{in} \qquad (2.3.1.1)$$

where p is the probability of an occurrence being categorized as a stroke, and βi is the model parameters. The output in (2.3.1.1) is a binary variable with p denoting the likelihood of an instance belonging to the "Stroke" class and 1 - p representing the probabilities of an instance belonging to the "non-stroke" class. This model is

renowned for its simplicity and interpretability and acts as a foundation for more sophisticated models.

2.3.2 Gaussian Naive Bayes:

Gaussian Naive Bayes (GNB) has its foundation in Bayes' theorem and assumes feature independence. It uses this premise to determine the conditional probability of every category based on a set of attributes [13]. The model optimizes the stipulated likelihood of a class given the property vector:

$$P(c|f_{i1}, \dots, f_{in}) = \frac{P(f_{i1,\dots,}f_{in}|c)\,P\,(c)}{p(f_{i1,\dots,}f_{in})} \qquad (2.3.2.1)$$

where $P(f_{i1,\dots,}f_{in}|c) = \prod_{j=1}^{n} P\,(f_{ij}|c)$ is the probability of the characteristics given the class, $P(f_{i1,\dots,}f_{in})$ is the prior probability of the features, and $P(c)$ is the prior probability of the class. The subset with an elevated posterior probability is selected as the forecast shown in the equation (2.3.2.1). GNB is especially effective with small datasets and offers a probabilistic approach to categorization.

2.3.3 Decision Tree Classifier:

The Decision Tree (DT) is a supervised learning model that may be applied to both regression and classification applications. It separates the data into subgroups depending on the value of each characteristic and makes decisions by asking a series of queries about the features. Each internal node reflects a feature, each branch resembles a decision rule, and each leaf node provides the result. A decision tree consists of two parts: decision nodes and leaf nodes [14].

Decision Node: Where data is divided depending on a specific attribute.

Leaf Node: The leaf node contains the class label following the data that has been separated using the decision node.

The model determines the appropriate split at each node based on Gini impurity or information gain. This method is repeated until all the data has been classified or a stopping requirement is satisfied. Decision trees are well-known for their ease of interpretation and ability to explain complicated relationships in data.

2.3.4 Evaluation Metrics:

The performance of the algorithms was evaluated using the subsequent metrics:

- Accuracy is the ratio of accurately anticipated instances to total instances shown in (2.3.4.1).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.3.4.1)$$

- Precision (2.3.4.2) is the ratio of accurately predicted positive results to all expected positives.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2.3.4.2)$$

- Recall (sensitivity) is the ratio of accurately predicted positive outcomes compared to every observation in the actual class. The equation is shown in (2.3.4.3).

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2.3.4.3)$$

- F1 Score is the calculated mean of Precision and Recall. The F1 score is illustrated in (2.3.4.4).

$$\text{F1 Score} = 2.\ \frac{Precision\ .\ Recall}{Precision+Recall} \qquad (2.3.4.4)$$

- ROC-AUC: AUC assesses the model's ability to distinguish between the stroke and non-stroke categories via TPR and FPR. The letters TP, FP, TN, and FN stand for True Positive, False Positive, True Negative, and False Negative, respectively.

3 Results:

The subsections follow a detailed analysis concerning Logistic Regression, Gaussian Naive Bayes, and Decision Tree classifiers. The results entail confusion matrices, ROC curves, and a comparison of each model's accuracy and AUC metrics.

3.1 ROC – AUC of Curve Logistic Regression, Gaussian Naïve Byes and Decision Tree:

Logistic Regression achieved the result with a ROC curve in the following ***Figure 4,*** Gaussian Naïve Byes achieved in ***Figure 5***, and Decision Tree Classifier obtained the following ROC - AUC curve in ***Figure 6.***

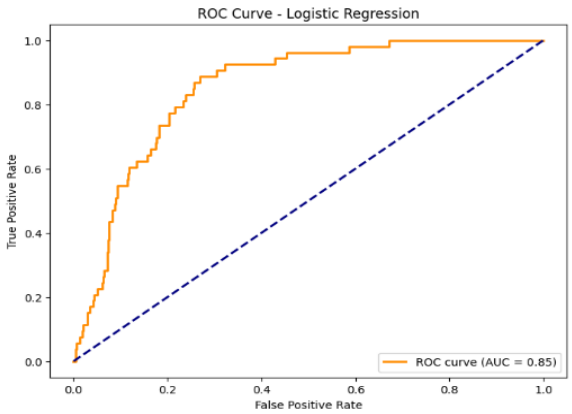

Figure 4. ROC-AUC Curve of Logistic Regression

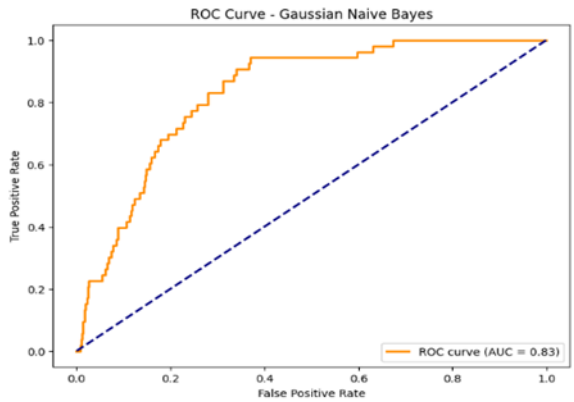

Figure. 5 Gaussian Naïve Byes ROC-AUC Curve

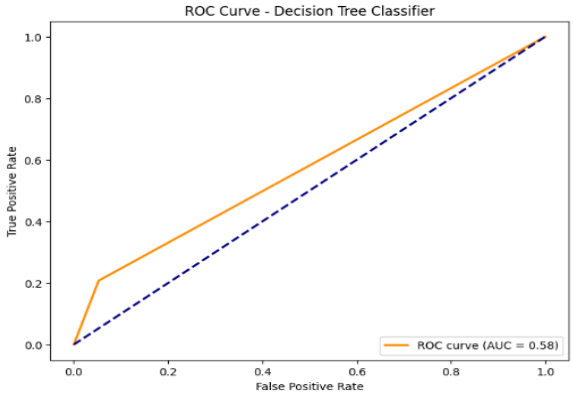

Figure. 6 Decision Tree Classifier ROC-AUC Curve

3.2 Comparative Analysis

A comparison between performance Metrics is shown in Table 2.

Table 2. Comparative analysis of performance

	Precision		**Recall**		**F1 Score**		**Accuracy**	**AUC**
	Stroke	**No Stroke**	**Stroke**	**No Stroke**	**Stroke**	**No Stroke**		
Logistic Regression	0	0.95	0	1	0	0.97	0.95	0.85
Gaussian Naïve Byes	0.18	0.96	0.43	0.88	0.25	0.92	0.86	0.83
Decision Tree Classifier	0.18	0.95	0.21	0.95	0.19	0.95	0.91	0.58

The accuracy and AUC comparisons, as well as the Heatmap of confusion matrices for the three models, are shown below in **Figure 7** and **Figure 8** for a better understanding.

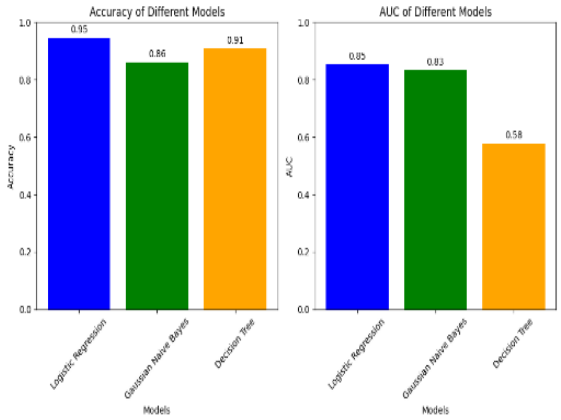

Figure. 7 Accuracy and AUC comparison

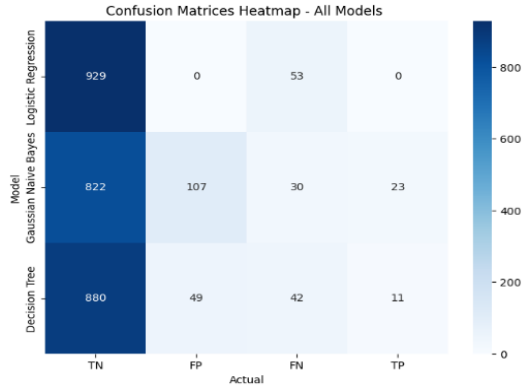

Figure. 8 Heatmap of all models

4 Discussion and Conclusion

Using a stroke diagnosis dataset, three machine learning models were compared: logistic regression, Gaussian Naive Bayes, and decision trees. Logistic Regression obtained 95% accuracy and 85% AUC, indicating high reliability in discriminating stroke from non-stroke cases. The Decision Tree likewise performed admirably, with 91% accuracy and 58% AUC. Confusion matrices had high true positive rates, whereas Logistic Regression and Decision Tree generated fewer false positives and negatives than Gaussian Naive Bayes. While Logistic Regression and Decision Trees are popular due to their accuracy, Gaussian Naive Bayes are prized for their ease of use and effectiveness. Future studies could look into ensemble approaches for improving performance.

Acknowledgements:

We gratefully acknowledge the constructive comments from the editor and the anonymous referees.

5 References

[1] WSO (n.d.). Global Stroke Fact Sheet 2022. World Stroke Organization (WSO). https://www.worldstroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf Cortes, C., & V. Vapnik, (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

[2] WHO (2020, December 9). The top 10 causes of death. World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[3] S. P. Jones, K. Baqai, A. R. Clegg, Georgiou, C. Harris, E. J. Holland, Y. Kalkonde, C. E. Lightbody, P. K. Maulik, P. M. Srivastava, J. D. Pandian, P. Kulsum, P. Sylaja, C. L. Watkins, & M. L. Hackett, (2021). Stroke in India: A systematic review of the incidence, prevalence, and case fatality. Sage Journals, 17(2). https://doi.org/10.1177/17474930211027834

[4] G. Sailasya, & G. L. A. Kumari, (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. The Science and Information Organization, 12(6). https://doi.org/10.14569/IJACSA.2021.0120662

[5] M. F. Ojha, R. T. & A. K. Jha, (2023). Analyzing the Performance of the Machine Learning Algorithms for Stroke Detection. The Science and

Information Organization, 13(2). https://doi.org/10.5815/ijeme.2023.02.04

[6] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. Uddin, & M. F. Mridha, (2023). Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study with a Web Application for Early Intervention. IEEE, 11(2). https://doi.org/10.1109/ACCESS.2023.3278273

[7] V. Krishna, J. S. Kiran, P. P. Rao, G. C. Babu, & G. J. Babu, (2021). Early Detection of Brain Stroke Using Machine Learning Techniques. IEEE Explore. https://doi.org/10.1109/ICOSEC51865.2021.9591840

[8] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, & M. M. Khan, (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering. https://doi.org/10.1155/2021/7633381

[9] A.A. Abujaber, Y. Imam, I. Albalkhi, S. Yaseen, A. J. Nashwan, &, N. Akhtar (2024). Utilizing machine learning to facilitate the early diagnosis of posterior circulation stroke. Springer,24. https://doi.org/10.1186/s12883-024-03638-8

[10] C. H. Lin, K. C. Hsu, K. R. Johnson, Y. C Fann, C. H. Tsai, Lien E, Y. S., L. M., Chang, Chen W. L., P. L., C. L Lin, & C. Y. Hsu, (2020). Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Computer Methods and Programs in Biomedicine, 190. https://doi.org/10.1016/j.cmpb.2020.105381

[11] N. Biswas, K. M. Mohi Uddin, S. T. Rikta, & S. K. Dey, (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2. https://doi.org/10.1016/j.health.2022.100116

[12] A. Das, (2023). Logistic Regression. In: Maggino, F. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Cham. https://doi.org/10.1007/978-3-031-17299-1_1689

[13] T. M. Mitchell, (2020). Generative And Discriminative Classifiers: Naive Bayes And Logistic Regression (2017th ed.). McGraw Hill.

[14] B. M. Greenwell, (2022). Tree-Based Methods for Statistical Learning in R: A Practical Introduction with Applications in R (1st ed.). CRC Press.