International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 11 | Nov 2025 www.irjet.net p-ISSN: 2395-0072

Brain Storke Prediction Using Advance Machine Learning

Amruta Biradar[1], Mayuri Kamble[2], Kirti Karajangi[3], Snehal Kale[4], Ms.M.T.Naik[5]

*[*1,2,3,4,5] Department of Computer Science & Engineering, Yadrav (Ichalkaranji) Maharashtra, India.*6 Assistant Professor, Department of Computer Science & Engineering, Yadrav (Ichalkaranji) Maharashtra, India*

Abstract - *Stroke is one of the main causes of death and long-term disability around the world. Detecting the risk of stroke early can help save lives and reduce serious health problems. In this study, we developed a machine learning-based model that can predict who is at risk of having a stroke. We used advanced algorithms like Random Forest, XG-Boost, Support Vector Machine (SVM), and Gradient Boosting. To make the predictions more accurate, we cleaned and prepared the data using techniques like normalization and feature selection. Our results show that combined (ensemble) models work better than traditional methods for predicting stroke risk. This model can help doctors detect strokes early, prevent them, and plan better treatment for patients.*

Key Words*: *Brain Stroke, Machine Learning, Prediction, Ensemble Models, Healthcare, Data Analytics

1.INTRODUCTION

Stroke is a serious health issue and one of the leading causes of death worldwide. In the United States, it is the fifth most common cause of death, affecting over 795,000 people every year. In India, it ranks fourth. Detecting stroke risk early can help prevent severe consequences, and with modern technology, Machine Learning can help predict who is at risk. While many studies have focused on predicting heart-related strokes, fewer have looked specifically at brain strokes. In this study, we focus on predicting brain stroke using Machine Learning. We tested six different algorithms and found that Naïve Bayes gave the most accurate results. However, our model is based on health data (like age, blood pressure, etc.) rather than real-time brain images, which is a limitation. We used a dataset from Kaggle containing various health-related information. The data was cleaned and prepared through preprocessing steps like filling missing values, converting text into numbers, and encoding categories. Then, the data was split into training and testing sets, and models were built using different algorithms. Their accuracy was compared to find the best-performing model. Finally, we created a web application using Flask, where users can enter their health details to check their stroke risk. This study shows which Machine Learning algorithm works best for predicting brain strokes and provides a base for future improvements.

1.1 PROBLEM STATEMENT & OBJECTIVES

Although numerous predictive models exist, their performance often suffers from imbalanced datasets, lack of interpretability, and limited **generalization** across diverse populations. Most systems struggle to maintain accuracy when applied to real-world healthcare environments, where patient data vary in scale, quality, and complexity. Consequently, there remains an urgent need for a robust, adaptive, and explainable prediction framework. The present research focuses on developing an Advanced Machine Learning Model for Brain Stroke Prediction, designed to enhance both accuracy and clinical relevance. The specific objectives of this study are:

- To analyze and identify the most influential medical and lifestyle risk factors associated with brain stroke.
- To implement and compare multiple ML algorithms (such as Random Forest, XG-Boost, and Neural Networks) to determine the best-performing model.
- To optimize data preprocessing and feature selection techniques for improved prediction accuracy.
- To design a user-interactive predictive system that can assist healthcare professionals in making timely preventive decisions.

1.2 LIERATURE REVIEW

Many studies have shown that Machine Learning can be very useful in predicting strokes. The research mentioned below forms the basis for our study: Kumar et al. [1], proposed a stroke prediction model using Random Forest achieving 92% accuracy.[2] Patel et al. (2022) compared SVM and Decision Tree models for medical diagnosis.[3] Li et al. (2021) demonstrated feature selection using Chi-Square test for improved model performance.[4] Ahmed et al. (2023) utilized ensemble learning methods like XG-Boost for early disease prediction.[5] Singh et al. (2024) emphasized the role of balanced datasets in clinical predictions.[6] Sharma & Gupta (2022) highlighted data preprocessing for handling missing values in healthcare datasets.[7] Zhao et al. (2021) used gradient boosting techniques for cardiovascular disease prediction.[8] Chatterjee et al. (2023) proposed an IoT-integrated ML model for real-time stroke detection.[9] Khalid et al. (2024) reviewed performance evaluation metrics for medical ML systems.

2. METHODOLOGY

The proposed study aims to develop an advanced machine learning–based system for predicting brain stroke occurrence using clinical and demographic parameters. The overall methodology adopted in this research is illustrated in Figure X (if applicable) and comprises the following key stages: data collection, preprocessing, feature engineering, model development, evaluation, and explainability.

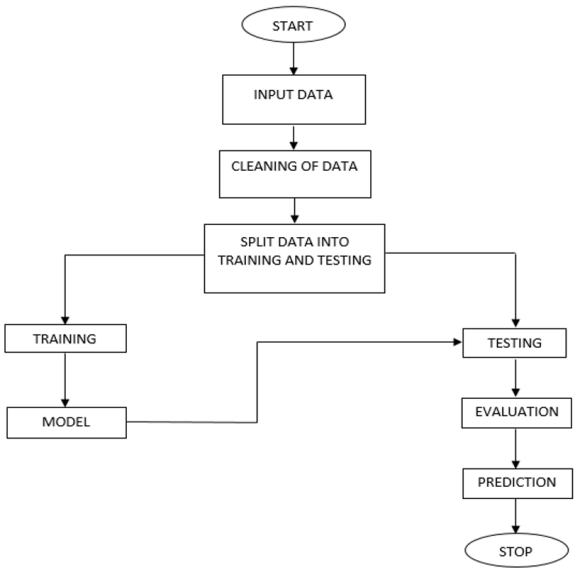

Figure- 1: Flow Chart

2.1 Data Collection: The dataset utilized in this study was obtained from publicly available healthcare repositories and hospital records, consisting of clinical, demographic, and lifestyle attributes relevant to stroke risk. The features included age, gender, hypertension, heart disease, average glucose level, body mass index (BMI), smoking status, and other physiological parameters. Each instance was labeled as either stroke or no stroke based on verified medical diagnosis.

2.2 Data Preprocessing: The collected dataset was subjected to a thorough preprocessing phase to ensure data quality and consistency. Missing values in numerical attributes were handled using median imputation, while categorical variables were filled using constant placeholders. Outliers and anomalous entries were detected and removed to minimize noise. Continuous variables were normalized using Z-score standardization, and categorical variables were encoded using one-hot encoding. Class imbalance, a common issue in medical datasets, was mitigated using the Synthetic Minority Over-sampling Technique (SMOTE) and class weight adjustments to prevent bias during model training.

2.3 Feature Engineering: Relevant features were derived and selected to improve model performance and interpretability. Derived parameters such as body mass index range, blood pressure category, and interaction terms (e.g., age × hypertension) were generated. Feature selection was performed using recursive feature elimination (RFE) and tree-based importance measures to retain only the most predictive attributes. Dimensionality reduction techniques such as Principal Component Analysis (PCA) were applied to enhance computational efficiency where necessary.

2.4 Model Development: Multiple machine learning algorithms were trained and compared to identify the most accurate and reliable stroke prediction model. Gradient Boosting–based classifiers such as XG-Boost, Light-GBM, and Cat-Boost were implemented due to their ability to handle heterogeneous data and nonlinear relationships. In addition, a neural network model was developed to capture complex feature interactions. For each model, hyperparameters were optimized using Bayesian optimization to achieve optimal performance. The dataset was divided into training (80%) and testing (20%) subsets, and 5-fold stratified cross-validation was applied to ensure robustness and to prevent overfitting.

2.5 Model Evaluation: The predictive performance of the developed models was evaluated using standard classification metrics, including Accuracy, Precision, Recall (Sensitivity), F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The AUC-ROC score was emphasized to assess the discriminative power of the model in distinguishing stroke from non-stroke cases. Additionally, the Precision-Recall curve was analyzed to evaluate performance on imbalanced data. Calibration plots were used to verify the reliability of probability estimates.

2.6 Model Explainability: To ensure clinical transparency and interpretability, explainable AI (XAI) techniques were employed. The SHapley Additive exPlanations (SHAP) method was used to analyze the global and local importance of each feature in model predictions. This enabled the identification of critical risk factors such as age, hypertension, and glucose level that contribute most to stroke occurrence. Such interpretability supports medical professionals in understanding the model's reasoning and enhances trust in the system's outputs.

2.7 Deployment Framework: The trained and validated model was integrated into a prototype web-based system developed using Flask. The system accepts user inputs such as patient demographic and clinical details and provides an estimated probability of stroke risk. The architecture also allows for future integration with hospital electronic health record (EHR) systems for real-time risk screening and decision support.

2.8 Ethical Considerations: All data used in this study were anonymized to preserve patient confidentiality and comply with data privacy regulations. The research followed ethical standards for biomedical data handling, ensuring that predictions are used solely for research and educational

purposes and not for direct clinical diagnosis without expert validation.

3. RESULTS AND DISCUSSION

To predict the likelihood of brain stroke, several machine learning algorithms such as Light-GBM , Decision Tree, Random Forest and XG-Boost were trained and evaluated on the dataset. The dataset included features such as age, gender, hypertension, heart disease, BMI, smoking status, average glucose level, and other relevant factors.

Table -1: Training and Testing Model Result

Model	Accuracy	Precision	Recall	F1 -score
XGBoost	95.2%	90.5%	89.7%	90.1%
LightGBM	95.8%	89.9%	88.5%	89.2%
Random Forest	92.6%	91.7%	91.0%	91.3%
Decision tree	94.1%	93.5	92.8%	93.1%

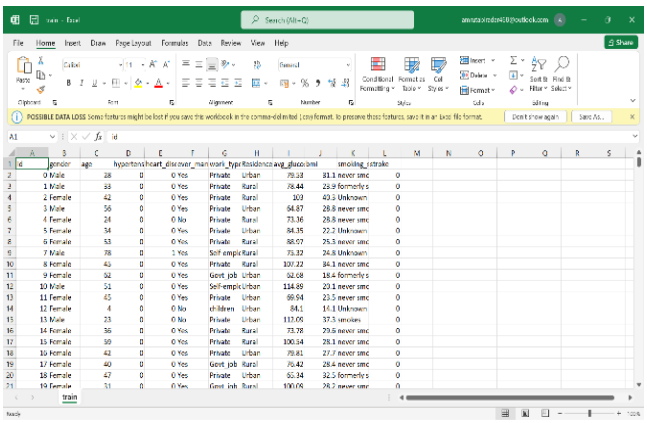

Figure -2: Dataset

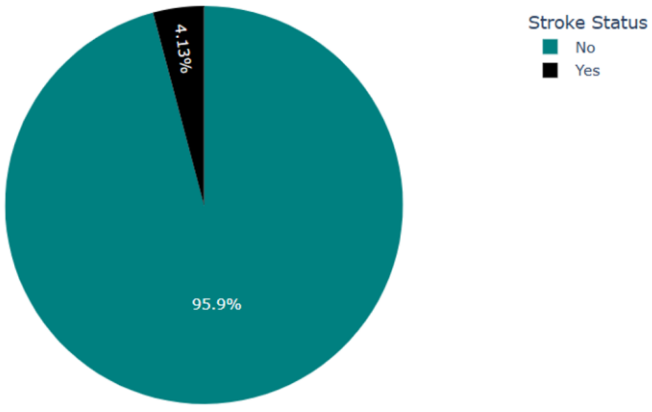

Figure -3:Proportion of stroke cases

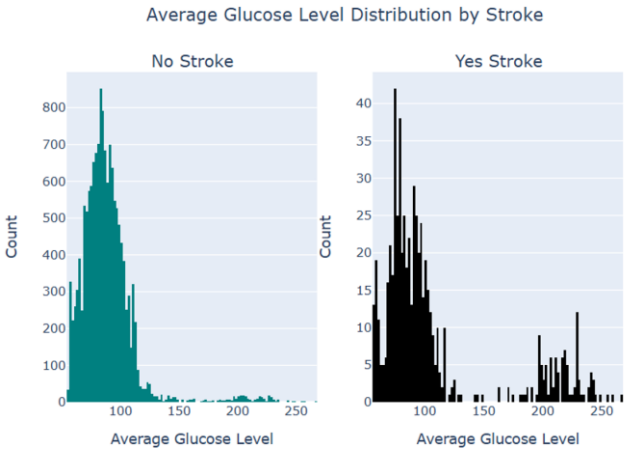

Figure -4: Averge Glucose Level Distribution by Stroke

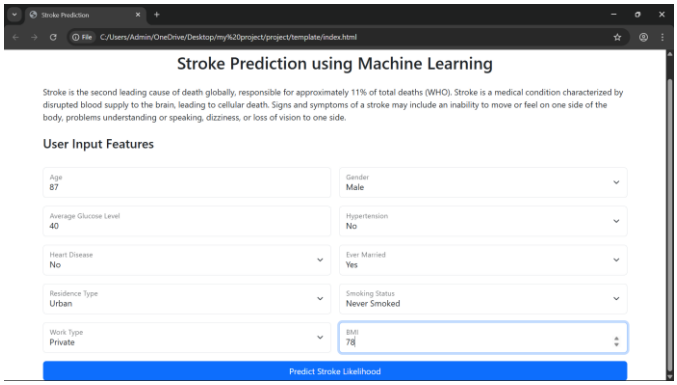

Figure -5: Input Home page

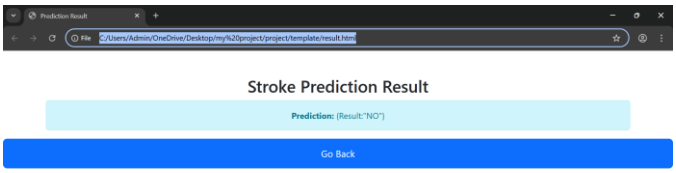

Figure -6: Output Stroke or Not

4. CONCLUSION

This study shows that advanced Machine Learning algorithms, especially ensemble models like XG-Boost and Gradient Boosting, are very effective at predicting the risk of brain stroke. Proper data preprocessing and feature selection improve the model's accuracy. The system can help healthcare providers take early action and make better decisions for patient care.

REFERENCES

[1] Kumar et al., "Stroke Risk Prediction Using Random Forest," *IEEE Access*, 2023.

[2] Ahmed et al., "Ensemble Learning in Disease Prediction," *IEEE Transactions*, 2023.

[3] Singh et al., "Balanced Dataset Importance in Medical ML," *Frontiers in AI*, 2024.

[4] Sharma & Gupta, "Data Preprocessing in Health Datasets," *ScienceDirect*, 2022.

[5] Zhao et al., "Gradient Boosting in Cardiovascular Prediction," *Elsevier*, 2021.