



BRAIN STROKE PREDICTION

A Phase 1 Report on Developing a Predictive Model for Brain Stroke Prediction

¹Ms.AMRUTHA, ²Dr. Shashidhar Kini K

¹Student, ²Professor & Head

¹Department of Master of Computer Applications,

¹Srinivas Institute of Technology, Valchil Mangaluru, Karnataka, India

Abstract : This study presents the development of a Brain Stroke Prediction System using supervised machine learning algorithms to aid in the early diagnosis and prevention of brain strokes. The system uses patient health records with features such as age, gender, hypertension, heart disease, average glucose level, BMI, smoking status, and work type to predict stroke occurrence. The dataset was obtained from open-source repositories like Kaggle and underwent rigorous preprocessing to handle missing values and encode categorical variables. The modeling framework compares the performance of Logistic Regression, K- Nearest Neighbors (KNN), Decision Tree, and Random Forest classifiers using metrics such as accuracy, precision, recall, F1- score, and AUC-ROC.

IndexTerms - Brain Stroke, Machine Learning, Classification, Logistic Regression, Random Forest, Health Prediction, Medical Analytics, Feature Engineering

I. INTRODUCTION

Brain stroke is a leading cause of death and long-term disability worldwide. Early detection of individuals at risk is vital for timely intervention and treatment. Strokes occur due to disrupted blood flow to the brain, often resulting from blocked or burst blood vessels. They can lead to severe neurological damage, loss of motor functions, and, in many cases, death.

Healthcare professionals often rely on diagnostic tools and patient history, which may not always lead to accurate early detection. Therefore, there is a pressing need for data-driven models that can assist in stroke risk prediction using historical patient data.

This project aims to build a machine learning-based stroke prediction system that can evaluate a patient's likelihood of having a stroke based on clinical and demographic data. The benefits of such a predictive system include:

- **Clinicians:** Early intervention and tailored treatment planning.
- **Patients:** Awareness and lifestyle changes to reduce stroke risk.
- **Healthcare Institutions:** Efficient resource allocation and risk-based triaging.
- **Government & NGOs:** Data-backed stroke awareness and prevention campaigns.

II. EASE OF USE

The final stroke prediction model is intended for easy integration into hospital management systems and mobile applications. The interface allows healthcare professionals or patients to enter details such as age, hypertension status, average glucose level, and smoking habits. Upon submission, the model returns a binary prediction (stroke/no stroke) along with a confidence score. Through API deployment, the model can be integrated into healthcare portals, enabling seamless usage without deep technical knowledge. The system is designed for high interpretability and performance, ensuring its usability across diverse demographic and clinical environments. Users interact with the system through a simple, intuitive interface—either as a web form or a mobile application—where they can input basic patient details such as age, gender, hypertension status, heart disease history, average glucose level, BMI, smoking status, work type, and residence type. Upon submitting these inputs, the model instantly returns a binary

classification indicating whether the patient is at risk of a stroke or not, along with a confidence score that quantifies the certainty of the prediction. This provides clinicians with an immediate, data-driven second opinion to assist in early diagnosis and preventive care. To further enhance usability, the model is scalable and adaptable. It can be retrained periodically with hospital- specific data to account for demographic variations or changing health trends. Moreover, a lightweight version of the model is optimized for mobile and offline use, making it suitable for deployment in remote or underserved areas where internet connectivity may be limited. The interface can also support local languages to improve accessibility in multilingual regions.

Prepare Your Paper Before Styling

Before finalizing the structure and format of the paper, we focused on collecting and preparing accurate and complete content for the brain stroke prediction project. The first step was gathering data from reliable sources like Kaggle. The dataset included important details such as age, gender, blood pressure, heart disease, glucose level, BMI, smoking status, and more—factors that influence the risk of stroke. Next, the data was cleaned and preprocessed. Missing values, especially in BMI and smoking status, were filled in using suitable methods. Categorical data (like gender or work type) was converted into numbers so that machine learning models could understand it. We also scaled the numerical values and removed outliers to improve model accuracy.

We then trained several machine learning models. Logistic Regression was used as a basic model because it's easy to interpret. More advanced models like Decision Tree, K-Nearest Neighbors (KNN), and Random Forest were also tested to see which gave the best results. We used accuracy, precision, recall, and F1-score to measure how well each model performed.

After completing all the technical work, we arranged the paper in IEEE format. Sections like Abstract, Introduction, Methodology, Results, and Conclusion were organized clearly. Charts, tables, and references were added, and the document was carefully proofread to avoid any errors before final submission.

2. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even if they have already been defined in the abstract. Common abbreviations such as IEEE, SI, and units of measurement (e.g., kg, km, and sqft) do not need to be defined. Do not use abbreviations in the paper title or section headings unless absolutely necessary.

In this paper, the following abbreviations and acronyms are used:

- **ML** – Machine Learning
- **KNN** – K-Nearest Neighbors
- **AUC** – Area Under the Curve
- **ROC** – Receiver Operating Characteristic
- **BMI** – Body Mass Index
- **CSV** – Comma-Separated Values
- **TP** – True Positive
- **FP** – False Positive
- **FN** – False Negative
- **TN** – True Negative

RESEARCH METHODOLOGY

This section outlines the methodology adopted to develop a machine learning model for predicting the likelihood of brain stroke in individuals. It includes the scope and target population of the study, sources of data, theoretical basis for selecting predictive features, and the statistical tools and machine learning techniques used to analyze and model the dataset.

3.1 Population and Sample

The population of this study consists of individuals who are at potential risk of experiencing a brain stroke. These include both healthy individuals and patients with various medical conditions commonly linked to strokes, such as hypertension, diabetes, heart disease, and high cholesterol levels. The study focuses on analyzing relevant health parameters and lifestyle factors that can help identify stroke risk. From this population, a sample dataset of approximately 5,000 patient records was collected from publicly available sources such as Kaggle and healthcare research repositories. The data includes patients from diverse age groups, genders, and geographic backgrounds to ensure a balanced and representative sample.

The dataset specifically includes medical attributes like age, gender, hypertension, heart disease, average glucose level, body mass index (BMI), smoking habits, and stroke status (target variable). The sampling strategy aimed to include entries with complete and consistent information. Records with missing, contradictory, or unrealistic values were excluded during preprocessing. The data used is cross-sectional, capturing a snapshot of patient conditions, and is suitable for training supervised machine learning models to predict stroke risk.

3.2 Data and Sources of Data

The data used in this study was sourced from publicly available health record databases and survey-based datasets intended for academic and research purposes. The primary dataset was obtained from Kaggle's *Stroke Prediction Dataset*, which contains anonymized health data of individuals, including their medical history and lifestyle factors. Key features extracted and used for modeling include:

- **Age** (in years)
- **Gender** (Male, Female, or Other)
- **Hypertension** (0 = No, 1 = Yes)
- **Heart Disease** (0 = No, 1 = Yes)
- **Ever Married** (Yes or No)
- **Work Type** (e.g., Private, Self-employed, Govt job)
- **Residence Type** (Urban or Rural)
- **Average Glucose Level** (in mg/dL)

- **Body Mass Index (BMI)**
- **Smoking Status** (Never smoked, Formerly smoked, Smokes)
- **Stroke** (Target variable: 0 = No stroke, 1 = Stroke)

The dataset underwent a series of preprocessing steps including handling missing BMI values, encoding categorical variables (e.g., gender, work type), and normalizing continuous features like glucose level and BMI. Outliers were filtered using statistical thresholds to ensure that the training data remained representative and unbiased.

3.3 Theoretical framework

The objective of this study is to build predictive models that estimate the likelihood of a brain stroke occurring in individuals based on various health and lifestyle factors. The dependent variable is whether the individual has had a stroke (binary classification: 0 = No stroke, 1 = Stroke), while the independent variables include:

- **Age** (numerical)
- **Gender** (categorical: Male, Female, Other)
- **Hypertension** (binary: 0 = No, 1 = Yes)
- **Heart Disease** (binary: 0 = No, 1 = Yes)
- **Ever Married** (binary: 0 = No, 1 = Yes)
- **Work Type** (categorical: Private, Self-employed, Government, etc.)
- **Residence Type** (categorical: Urban, Rural)
- **Average Glucose Level** (numerical, in mg/dL)
- **Body Mass Index (BMI)** (numerical)
- **Smoking Status** (categorical: Never smoked, Formerly smoked, Smokes)

The relationship between these variables and the likelihood of a stroke is expected to be complex and non-linear. This justifies the use of advanced machine learning algorithms, such as logistic regression, random forest, and XGBoost, which can model such non-linear relationships effectively.

3.4 Statistical tools and econometric models

This section explains the methods and models used to predict the likelihood of a stroke.

3.4.1 Descriptive Statistics

First, we analyzed the data to understand its distribution. For numeric variables like age and blood pressure, we calculated basic stats like mean, median, and standard deviation. We also looked at the frequency of categories (e.g., hypertension or smoking status) to see how common certain conditions were. This helped us identify any patterns or imbalances in the data.

3.4.2 Machine Learning Models Used

We used and compared several machine learning models to predict strokes:

a) Logistic Regression (Baseline Model)

This simple model helps us understand the relationship between health factors and stroke risk. It predicts whether a person is likely to have a stroke based on the input features.

b) Random Forest Classifier

This model builds many decision trees and combines their results. It's good at handling complex relationships and avoids overfitting, meaning it's less likely to make mistakes on new data.

c) XGBoost Classifier

XGBoost is an advanced model that works well with complex data. It's fast, accurate, and can handle missing values, making it ideal for stroke prediction.

d) K-Nearest Neighbour

Predicting brain strokes often involves analyzing medical data using algorithms like K-Nearest Neighbour (KNN) or other machine learning models. These predictions take factors such as age, blood pressure, cholesterol levels, heart rate, and lifestyle habits into account. Models can classify a person into risk categories by comparing their data to existing cases.

3.4.3 Evaluation Metrics

We used the following to measure how well each model predicted strokes:

- **Accuracy:** The percentage of correct predictions.
- **Precision:** How many of the predicted strokes were actually strokes.
- **Recall:** How many of the actual strokes were correctly predicted.
- **F1-Score:** A balance between precision and recall.
- **ROC-AUC:** Measures how well the model distinguishes between stroke and no stroke.

3.4.3 Comparison of the Models

We compared the models based on their performance and chose the one with the best F1-Score and ROC-AUC. We also looked at which features (like age or blood pressure) were most important in making predictions.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
Age	0.08	82	43.1	22.3
Glucose Level	55	270	106	45.7
BMI	10	60	28.7	7.4

Table 4.1 Logistic Regression performed well on balanced data but struggled with recall. Random Forest achieved the highest AUC (0.89) and demonstrated robustness to outliers and feature interactions. The model was able to generalize well and showed consistent results across different folds.

III. ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Project Guide and Head of the Department of MCA, Dr. Shashidhar Kini K, for his invaluable guidance, constant encouragement, and kind support throughout this research work. Appreciation is also extended to the Principal, Dr. Shrinivasa Mayya D, for fostering an environment conducive to completing this project within the institution. The author thanks the management of Srinivas Institute of Technology for their direct and indirect support. Gratitude is also due to all the faculty members and non-teaching staff of the MCA department for their constant help and support. Finally, the author is indebted to parents and friends for their unwavering support and belief throughout this endeavor.

REFERENCES

1. Kaggle. (2023). Stroke Prediction Dataset. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- [2] Aslam, A., et al. (2021). Machine Learning Based Stroke Prediction: A Comparative Study. *International Journal of Biomedical Engineering and Technology*, 37(1), 25-34.
- [3] Sharma, S., & Gupta, R. (2022). Predicting Stroke Using Data Mining Techniques. *International Journal of Computer Applications*, 183(34).
- [4] Kaur, G., & Singh, D. (2021). Comparative Analysis of Machine Learning Algorithms for Stroke Prediction. *Journal of Medical Systems*, 45(3), 1-9.
- [5] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," *International Journal of Information Engineering and Electronic Business (IJIEEB)*, vol. 12, no. 2, pp. 15–20, 2020, doi: 10.5815/ijieeb.2020.02.03.
- [6] S. Dabreo, S. Rodrigues, V. Rodrigues, and P. Shah, "Real Estate Price Prediction," *Fr. Conceicao Rodrigues College of Engineering, Mumbai*, 2021.