

# Brain Stroke Prediction Using ML

**Bhavyashree H D<sup>1</sup>, Amarnath R<sup>2</sup>, Ankith T P<sup>3</sup>, Jagath Ponnanna P M<sup>4</sup>, Sandesh K M<sup>5</sup>**

Assistant Professor, Department of ISE, MITM, Mysore, VTU Belagavi, India<sup>1</sup>

UG Students, Department of ISE, MITM, Mysore, VTU Belagavi, India<sup>2-5</sup>

**Abstract:** This study investigates the complex interplay between general health parameters—most notably blood pressure and the risk of brain stroke through a data-driven approach using machine learning algorithms. By leveraging a comprehensive dataset and applying various classification models including Random Forest (RF), Decision Tree (DT) and Artificial Neural Networks (ANN), the research evaluates each model's effectiveness in predicting stroke occurrences. Feature engineering, data pre-processing, and statistical validation techniques are employed to enhance model performance and accuracy. The study not only identifies key predictors of stroke but also offers a comparative analysis of algorithmic performance, paving the way for intelligent diagnostic systems that support early detection and preventive healthcare strategies.

**Keywords:** Brain stroke prediction, Machine learning, including Random Forest (RF), Decision Tree (DT) and Artificial Neural Networks (ANN), Predictive analytics, Health informatics, Feature engineering, Data pre-processing, Stroke risk assessment.

## I. INTRODUCTION

Brain stroke remains one of the leading causes of death and longterm disability worldwide, presenting a significant challenge to public health systems. Early prediction and diagnosis of stroke are critical to mitigating its devastating impacts, enabling timely intervention and improving patient outcomes. Among the many physiological indicators associated with stroke, blood pressure and general health metrics such as cholesterol levels, diabetes status, and cardiovascular history have emerged as pivotal risk factors. However, the interplay among these variables is often complex and nonlinear, necessitating advanced analytical techniques for effective interpretation and decision-making. In recent years, machine learning (ML) has gained prominence as a transformative tool in medical diagnostics, offering the ability to analyse large-scale health datasets and uncover patterns that traditional statistical methods may overlook. This research builds upon the growing body of work that applies machine learning to stroke prediction, utilizing models such as including Random Forest (RF), Decision Tree (DT) and Artificial Neural Networks (ANN) and Artificial Neural Networks (ANN). These models are selected for their robust classification capabilities and their adaptability to diverse data types and structures.

The objective of this study is to perform a comparative evaluation of these algorithms to determine their effectiveness in predicting brain stroke based on key health indicators. Through rigorous feature engineering, preprocessing, and validation, we aim to develop a predictive framework that not only demonstrates high accuracy but also contributes to clinical decision support systems. By doing so, this research supports the broader goal of personalized medicine and proactive healthcare, wherein data-driven insights can guide early diagnosis and preventive strategies for at-risk populations.

## II. MOTIVATION

The global burden of brain stroke continues to rise, with increasing incidence rates linked to aging populations, sedentary lifestyles, and undiagnosed comorbid conditions such as hypertension and diabetes. Despite advancements in medical imaging and treatment protocols, stroke remains a condition where prevention is far more effective than cure. Early identification of individuals at risk can significantly reduce mortality and improve quality of life, yet conventional diagnostic systems often fail to leverage the full spectrum of patient data available in electronic health records. A Machine learning offers a powerful solution to this challenge by enabling the automated analysis of complex, high-dimensional datasets.

It facilitates the discovery of non-obvious correlations between health parameters and stroke risk, which can be invaluable in developing predictive models for early intervention. However, with numerous ML algorithms available, there is a pressing need for comparative evaluation to determine which models provide the best balance of accuracy, interpretability, and computational efficiency in the context of stroke prediction.

### III. LITERATURE SURVEY

This study presents a comprehensive review of various machine learning techniques used for brain stroke prediction. It highlights the critical role of features like age, BMI, hypertension, and smoking status. Multiple algorithms—such as SVM, Decision Trees, Random Forest, Naive Bayes, and Neural Networks—were compared, showing that ensemble and hybrid models often yield higher predictive accuracy. The study also emphasizes preprocessing and data balancing methods like SMOTE and under-sampling, revealing Random Forest and stacking models as top performers in accuracy and recall metrics [1].

This paper focuses on predicting stroke using machine learning models, especially KNN. It uses a wide set of patient data— demographics, lifestyle, and clinical attributes—to train and evaluate various classifiers. The study finds age, hypertension, cholesterol levels, and blood pressure as key predictors. The authors cite a literature gap in using image data and propose integrating both structured and unstructured data in future models. The work supports real-time deployment in healthcare systems while stressing ethical and regulatory considerations [2].

This research develops a robust ANN model, leveraging SMOTE for class balance and hyperparameter tuning for optimization. It addresses challenges from past studies like small sample sizes, missing data, and model interpretability. The ANN model showed superior performance in precision and AUC. It also includes detailed exploratory data analysis and visual correlation to support feature selection. A notable conclusion is the need for real-time clinical integration and the importance of interpretability in healthcare settings [3].

Focusing on image-based diagnosis, this study uses Convolutional Neural Networks (CNN) to detect brain stroke from CT scan images. The model achieved 90% accuracy and showed improved precision through image preprocessing and architectural tuning. Comparative analysis with other models (e.g., LeNet-5, Seg Net) suggests CNNs are powerful for image classification tasks in stroke detection. The study highlights the importance of using visual data alongside structured health records for comprehensive stroke prediction [4].

This work evaluates and compares multiple machine learning models—AdaBoost, RF, DT—alongside an ensemble voting classifier. Using a healthcare dataset with demographic and clinical attributes, the study found the ensemble model outperformed individual classifiers with 90% accuracy. It also provides insights into model interpretability and parameter tuning. The literature review component analyses related works, including deep learning, NLP based MRI analysis, and hybrid models, suggesting future directions in using ensemble techniques for clinical decision support [5].

### IV. BLOCK DIAGRAM AND SYSTEM ARCHITECTURE

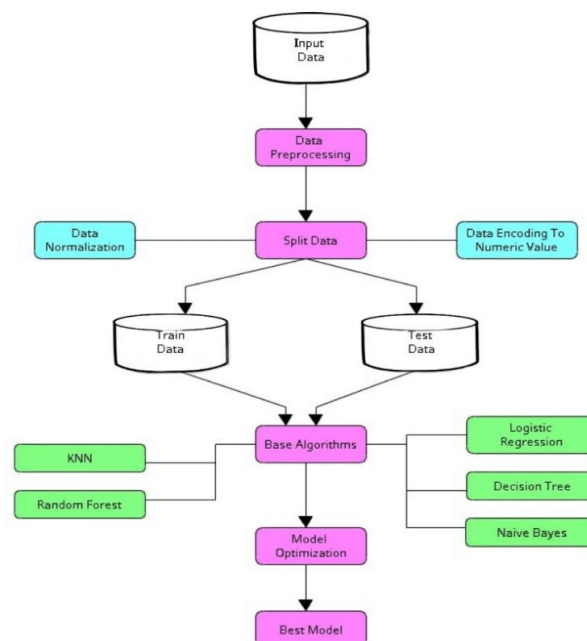


Fig 1. BLOCK DIAGRAM

## **V. ALGORITHM USED**

- 1. Artificial Neural Network (ANN):** A multi-layer neural network using ReLU and Sigmoid activations to classify stroke risk based on health data.
- 2. Support Vector Machine (SVM):** A supervised learning model using various kernels (Linear, RBF, Polynomial, Sigmoid) to separate stroke and non-stroke cases.
- 3. K-Nearest Neighbours (KNN):** distance-based algorithm that classifies stroke risk by majority vote among the nearest neighbours in the dataset.
- 4. Logistic Regression:** Logistic Regression is a statistical model used for binary classification tasks such as predicting stroke or no stroke. It estimates probabilities using a logistic function.
- 5. Decision Trees (DT):** Decision Trees (D) classify data by splitting it into branches based on feature values until a decision is made at a leaf node. They are easy to interpret and visualize, which is helpful in healthcare decision making.
- 6. Random Forests (RF):** Random Forests (RF) are ensembles of decision trees where each tree is built on a random subset of data and features. The final prediction is made by averaging the results (for regression).
- 7. Gradient Boosting Machines (GBM):** Gradient Boosting Machines (GBM) build models in a sequential manner, where each new model attempts to correct errors made by the previous one.
- 8. AdaBoost (Adaptive Boosting):** AdaBoost (Adaptive Boosting) improves model performance by giving more weight to incorrectly classified samples in each iteration.
- 9. XG Boost (Extreme Gradient Boosting):** XG Boost (Extreme Gradient Boosting) is a highly optimized implementation of gradient boosting that includes regularization, missing value handling, and efficient computation.
- 10. Naive Bayes (NB):** Naive Bayes (NB) is a probabilistic classifier based on Bayes' Theorem, assuming independence among features. It is simple, fast, and surprisingly effective in many cases, especially with categorical data.

## **VI. METHODOLOGY**

The methodology adopted in this research involves the application of various machine learning algorithms to predict the occurrence of brain strokes using patient health records. The overall process comprises data acquisition, preprocessing, exploratory analysis, model selection, training, evaluation, and performance comparison.

**A. Data Collection** The dataset used in this study was sourced from publicly available electronic health records (EHRs). It includes both numerical and categorical features such as age, gender, hypertension, heart disease, average glucose level, BMI, smoking status, and work type. The target variable indicates whether a patient has experienced a stroke.

**B. Data Preprocessing** To ensure data quality, preprocessing steps such as handling missing values, encoding categorical features, and normalization were performed. Null values in columns like BMI and Smoking Status were imputed using median and mode values respectively. The Standard Scaler () function was used to normalize the numerical features, standardizing the mean to 0 and standard deviation to 1.

**C. Algorithm Implementation** In this research, three primary machine learning algorithms— Artificial Neural Network (ANN), Random Forest (RF), and Decision Tree (DT)—were implemented to predict brain stroke occurrences. The ANN was constructed using a sequential model with two hidden layers activated by ReLU and a sigmoid function in the output layer, optimized using Stochastic Gradient Descent (SGD) and backpropagation. The SVM model was tested with different kernels, including Linear, RBF, Polynomial, and Sigmoid, with the linear kernel yielding the best results in terms of accuracy and ROC-AUC. KNN was used as a distance-based classifier, where the optimal number of neighbours 'k' was selected using the formula  $k = \sqrt{N}$ , and Euclidean distance was used as the metric. All models were trained on normalized data using the Standard Scaler function to ensure uniformity across features. 5-fold cross-validation was employed to validate model generalization, and performance was measured using accuracy, precision, recall, F1-score, and ROC-AUC to identify the most effective model for stroke prediction.

D. Comparative Analysis: Each model's results were compared across key evaluation metrics. ANN achieved the highest accuracy and ROC-AUC score, while SVM with a linear kernel performed strongly in both classification and generalization. KNN also showed robust performance, particularly after scaling and cross-validation.

E. Metrics Evaluation Accuracy: Measures the proportion of correctly predicted instances among the total predictions. High accuracy values (e.g., 98.47% for Random Forest with linear kernel) reflect strong overall performance. Precision: Evaluates the proportion of true positives among all positive predictions. For example, the ANN model achieved a precision of 86%, indicating it correctly identifies most stroke cases it predicts. Recall (Sensitivity): Reflects the model's ability to identify actual positive cases. The ANN model attained a recall of 90%, showing high sensitivity in stroke prediction. F1-Score: Harmonic mean of precision and recall, balancing both metrics. ROC-AUC Score: Indicates the ability of the model to distinguish between classes. ANN and DT achieved scores of 0.845 and 0.821, respectively, while the best RF score was 0.681 with a linear kernel. Log-Loss: Captures the confidence of predictions. Lower values indicate better performance; e.g., the ANN model had a log-loss of 0.066 on validation.

F. Summary of Methodology The methodology of the study involved acquiring a publicly available electronic health record (EHR) dataset comprising 11 features and a target variable indicating stroke occurrence. The data underwent preprocessing, including handling missing values, encoding categorical variables, and standardizing features using scaling techniques. Data Analysis (DA) was performed to identify key predictive features such as age, blood pressure, and cardiovascular disease. Feature engineering involved selection and normalization to enhance model accuracy. Three machine learning algorithms— Artificial Neural Networks (ANN), RF, and DT—were implemented to classify stroke risk. The models were trained, validated, and tested through iterative processes, incorporating techniques like cross validation and hyperparameter tuning to ensure robustness and generalizability.

## VII. RESULT AND PERFORMANCE ANALYSIS

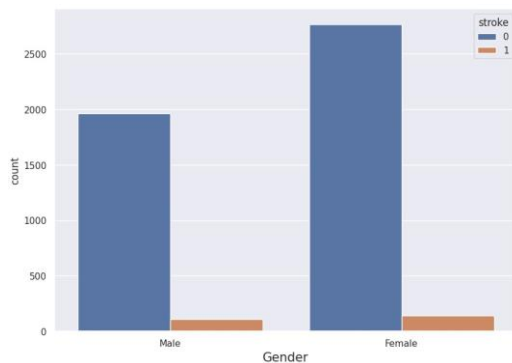


Fig 2. Stoke occurrence in Men vs Women

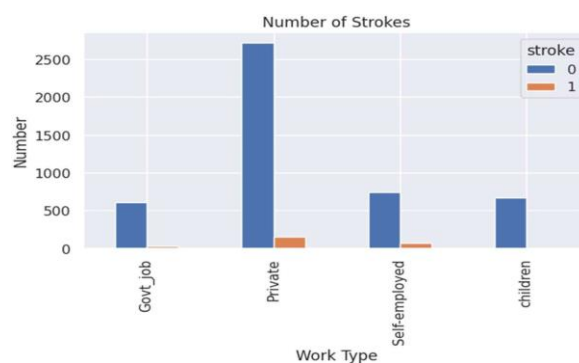


Fig 3. Stoke occurrence by Work Type

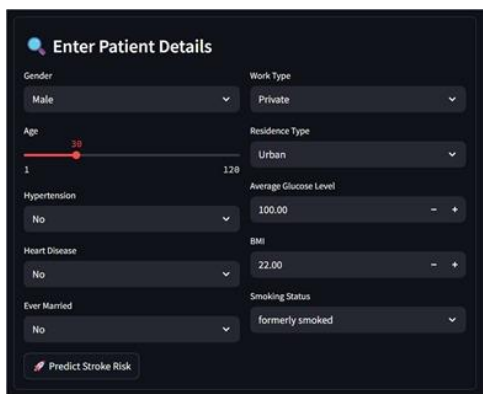


Fig 4. GUI snapshot

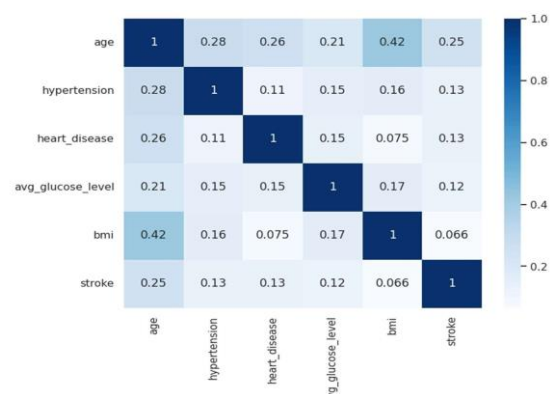


Fig 5. Correlation Analysis of Dataset Attributes

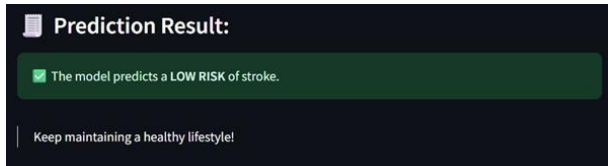


Fig 6. Instance of low probability of Stroke

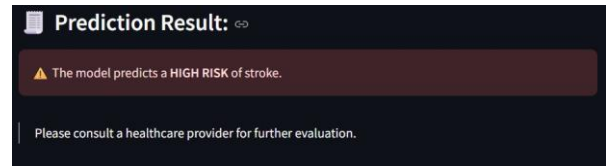


Fig 7. Instance of high probability of Stroke

## VIII. CONCLUSION

The study demonstrates the effectiveness of machine learning techniques in predicting brain stroke by analysing critical health indicators such as age, blood pressure, and cardiovascular conditions. Among the models evaluated, Artificial Neural Networks (ANN) showed the highest accuracy and ROC-AUC score, followed by K-Nearest Neighbours (KNN) and Support Vector Machines (SVM). The research highlights the importance of data preprocessing, feature selection, and proper model validation in achieving reliable predictions. These findings offer valuable insights for developing early diagnostic tools and personalized healthcare strategies, potentially aiding in timely stroke prevention and improved patient outcomes.

## REFERENCES

- [1]. "About stroke," Centers for Disease Control and Prevention, 02-Nov2022. [Online]. Available: <https://www.cdc.gov/stroke/about.htm> [Accessed: 16- Nov2022].
- [2]. "Stroke," Mayo Clinic, 20-Jan-2022. [Online]. Available: <https://www.mayoclinic.org/diseasesconditions/stroke/symptomscauses/syc-20350113>. [Accessed: 09-Jan-2023].
- [3]. World Health Organization. [Online]. Available: <http://www.emro.who.int/healthtopics/strokecerebrovascular-accident/index.html>. [Accessed: 16- Nov-2022].
- [4]. "Risk factors for stroke," Risk Factors for Stroke — Johns Hopkins Medicine, 13-Dec-2022. [Online]. Available: <https://www.hopkinsmedicine.org/health/conditionsanddiseases/stroke/riskfactors-for-stroke>. [Accessed: 08-Jan2023].
- [5]. "Ischemic stroke," Medline Plus. [Online]. Available: <https://www.medlineplus.gov/ischemicstroke.html>. [Accessed: 17Nov-2022].
- [6]. "The top 10 causes of death," World Health Organization, 09Dec2020. [Online]. Available: <https://www.who.int/newsroom/factsheets/detail/the-top-10-causes-of-death>. [Accessed: 08Feb-2023].
- [7]. "Hemorrhagic strokes (bleeds)," [Online]. Available: <https://www.stroke.org/en/about-stroke/typesofstroke/hemorrhagic-strokesbleeds>. [Accessed: 17-Nov-2022].
- [8]. Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial Intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, 2022.
- [9]. S. Das, A. Dey, A. Pal, and N. Roy, "Applications of artificial intelligence in machine learning: Review and Prospect," *International Journal of Computer Applications*, vol. 115, no. 9, pp. 31–41, Apr 2015.
- [10]. D. Petersson, "What is supervised learning?" *Enterprise AI*, 26- Mar-2021. [Online]. Available: <https://www.techtarget.com/searchenterpriseai/definition/supervised-learning>. [Accessed: 12Jan2023].
- [11]. R. Choudhary and H. K. Gianey, "Comprehensive Review on Supervised Machine Learning Algorithms," 2017 International Conference on Machine Learning and Data Science (MLDS), pp. 37– 43, Dec 2017.
- [12]. K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in Intelligent surveillance system," *International Journal of Information Technology*, vol. 14, no. 1, pp. 397–410, 2019.