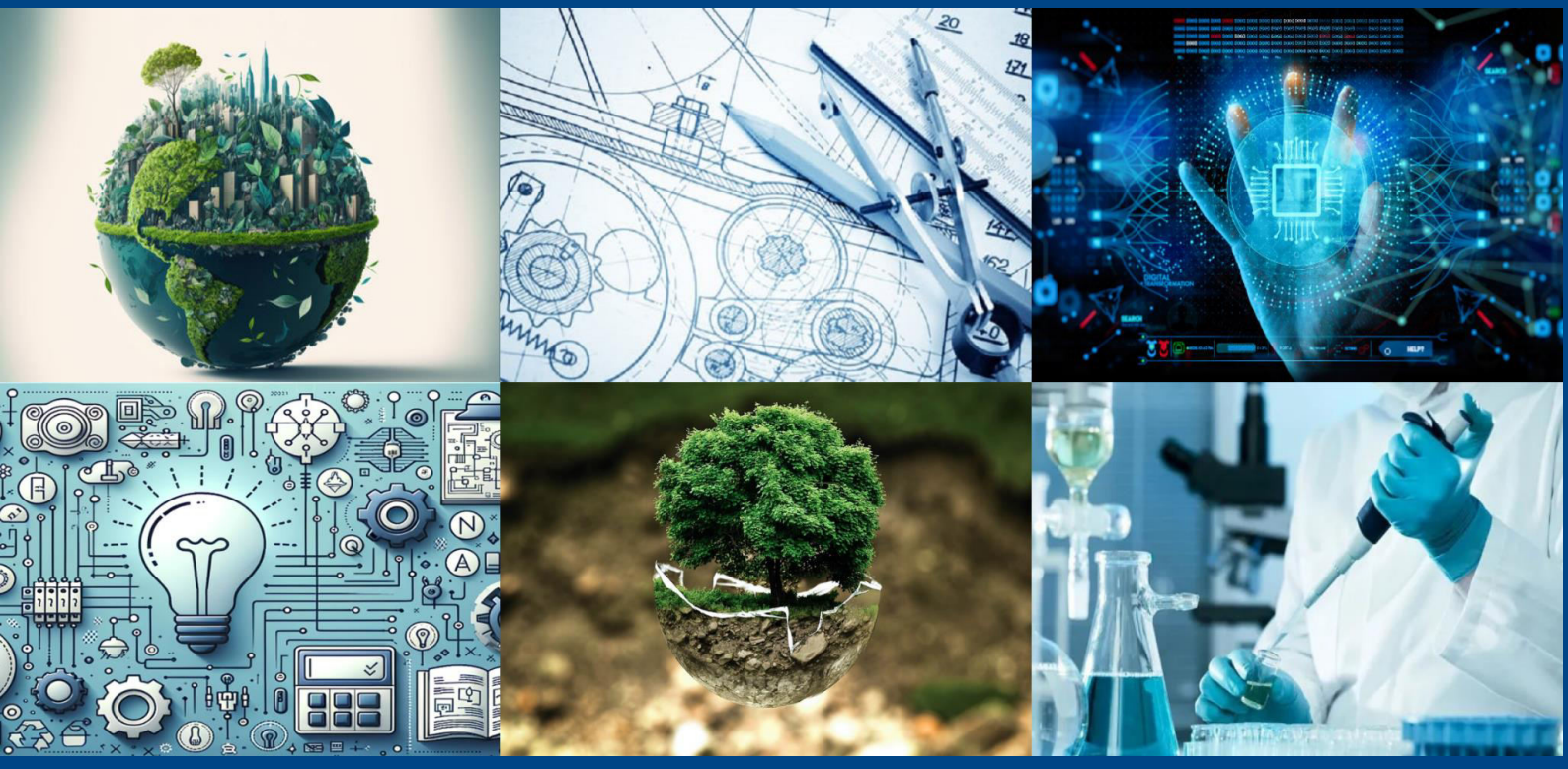
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805124

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Brain Stroke Prediction using Machine Learning

Vedashri Girish Bandewar [1]

Postgraduate Student, Dept. of Master of Computer Application, Anantrao Pawar College of Engineering and Research, Pune, India[1]*

Dr. Atul D. Newase [2]

Asst. Professor and Head, Dept. of Master of computer Application, Anantrao Pawar College of Engineering and Research, Pune, India[2]*

ABSTRACT: Brain is an important part our body. Brain Stroke is a disease that occurs due to bleeding within or outside the vessels of the brain that destroy neural tissue. Additionally, it occurs when there is obstruction of blood and other nutrients necessary for the brain. It's an extremely prevalent issue today and the number of cases of stroke is increasing-I believe it will just keep on increasing. causes of stroke are many most important among them is blood supply to nearby structures of brain. For physicians, brain stroke is one among them, but its proper diagnosis is most challenging.

Cerebral stroke is the most pathological condition sentencing death and irreversible disability everywhere all over the world in the assessment of WHO. The effects of stroke can be reduced in most cases if the warning signs are noticed and appropriately responded to in early stages. In predicting stroke, various machine learning (ML) algorithms that use various differing neural decision tree and regression methods can be utilized to estimate the probability of stroke in an individual. In this project, the physical attributes of an individual are examined with the assistance of ML classifiers Logistic Regression, Random Forest Classifier, XGBoost to establish four extremely accurate prediction models. Random Forest performed the task best with an accuracy rate of around 96 percent. The data utilized in developing the method was the open-source Stroke Prediction dataset. Models utilized in this study have a higher percentage of accuracy compared to that described in previous studies, which means models utilized in this study are more reliable. Certain model comparisons have been found to be robust, and the framework can be derived from the findings of the study.

KEYWORDS: brain, stroke, logistic regression, random forest, XGBoost, prediction, machine learning algorithms, diagnosis, blood supply

I. INTRODUCTION

Stroke is a significant cause of death and disability on an international level, and in most cases, it is caused by ischemic or haemorrhagic stroke. Prediction based on machine learning can be applied to improve at-risk population prediction through early intervention and treatment. The current stroke predictive tools are statistical models, and they do not hold true when they are applied to complex and nonlinear health data in the majority of cases. Machine learning methods, though, can detect underlying patterns in patient history, and predictive capability is boosted by this. This research makes an effort to compare and test some machine learning models for predicting stroke and give recommendations on the most suitable models to be applied clinically.

Stroke brain is a deadly disease due to the disruption of cerebral blood supply. The World Health Organization (WHO) characterizes stroke as one of the leading causes of death worldwide. The existing methods of diagnosis do not allow early intervention. The aim of the present research is to employ machine learning in an attempt to forecast stroke risk via health factors such as age, blood pressure, body mass index (BMI), and blood glucose level.

The research is based on the implementation of the latest ML techniques like XGBboosting, random forest classifier, and logistic regression .

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE SURVEY

- **Breakthroughs in Predictive Healthcare Through AI**
  Machine learning (ML) has transformed healthcare, with the ability to analyse complex medical data and determine patterns that are hard to spot through conventional methods. Stroke prediction research has mostly been based on using supervised learning algorithms, including logistic regression and random forest, to evaluate risk factors such as age, hypertension, blood glucose, and smoking history. Though these models show reasonable accuracy, they tend to perform poorly in managing non-linear relationships and imbalanced data. Current research indicates that ensemble learning algorithms like Random Forest and XGBoost provide better performance, with higher prediction accuracy.

- **Feature Engineering And Model Selection Challenges**
  Optimization of predictive models for stroke prediction is dependent on feature engineering. Methods such as one-hot encoding and normalization improve dataset quality and model accuracy. But difficulties in choosing the most effective features remain owing to the multivariate nature of stroke ethology. Experiments have demonstrated that adding medical imaging information, like MRI scans, to clinical parameters greatly enhances predictive accuracy. Hyperparameter optimization of state-of-the-art models, namely XGBoost and has also emerged as a critical component in enhancing their scalability and versatility across different data sets.

- **Tackling Imbalanced Data and Ethical Issues**
  The skewness in stroke datasets, with most of the samples belonging to non-stroke cases, is the biggest challenge to predictive modelling. Methods such as Synthetic Minority Over-sampling Technique (SMOTE) have proved effective in dealing with skewness so that models can more easily pick out stroke-positive cases. But ethical issues arise over patient privacy and algorithmic bias in prediction, which could unfairly target particular demographic groups. These challenges highlight the need for ethical reasoning and open validation procedures for implementing ML-based stroke prediction systems in practice.

  This review of the literature emphasizes the potential and challenges of applying machine learning to brain stroke prediction, opening the door to future research in creating accurate, fair, and clinically relevant solutions.

III. METHODOLOGY

Brain Stroke Prediction Machine Learning Method Being a step-by-step procedure, data preprocessing and collection are done initially; it gives quality inputs to the training model; one-hot encoding of categories, one-hot encoding of continuous features, and normalization for continuous variables are given to give improved performance to the model Brain Stroke Predicting Machine Learning Method is a proven step-by-step procedure beginning with data collection, preprocessing, and quality inputs in data to training the model. The Kaggle stroke prediction dataset is used which holds 5110 instances of critical health indicators like age, hypertension, glucose level, bmi, and smoking status are utilized in this research work, missing values of bmi that are utilized for data augmentation and replaced by median value; category features are one-hot encoded and continuous features are supplied with normalization to give good performance to the model as there is high class imbalance due to which synthetic minority over-sampling technique (smote) is utilized with stroke-positive samples getting over-sampled in a way that is adaptable for recall values; According to surveys and studies the best predictors are age, hypertension, blood glucose level, and smoking; the research work is conducted using three machine learning algorithms namely logistic regression , xgboost ,random forest and the model trained on an 80: 20 train-test split utilizing 10-fold cross validation to validate the model stabilizement is attained by checking critical parameters like accuracy, precision, recall, f1-score and auc-roc of which it was observed that the best predictor is xgboost hyperparameter tuning is executed by grid search and randomized search which tuned the learning rate, number of estimators, and maximum tree dept for boosting the model for efficient execution; the proposed model is implemented in a flask-based web application for real-time prediction using user-supplied health data. The subsequent work will include integration of deep learning (integration of medical imaging data) and clinical validation to achieve higher accuracy and usability in clinic which is beneficial for goal such system-based solution provides reliability, scalability, and effectiveness of predictive stroke evaluation model.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS AND ANALYSIS

We have performed our experiments on the Kaggle Stroke Prediction Dataset of 5110 records and 11 important features (age, hypertension, glucose levels, BMI, and smoking). Following careful preprocessing with median filling in absent data points, Binary vector representation for categorical variables, normalization, SMOTE-based class balancing, we have trained three classifiers: Logistic Regression, Random Forest, and XGBoost. For robust evaluation, we employed 80-20 train-test splitting along with 10-fold Cross validation.

Following table gives an overview of the three models' performance:

Models	Accuracy	Precision	Recall	F1-score	AUC-Score
Logistic regression	78.9%	76.7%	74.5%	75%	0.80
Random forest	82.3%	81.8%	79.4%	80.3%	0.85
XGBoost	88.9%	87.3%	86.8%	87.0%	0.91

The results indicate that XGBoost significantly outperforms the other classifiers. With an accuracy of 88.9% and an AUC-ROC of 0.91, XGBoost effectively distinguished between stroke and non-stroke cases. Its superior performance can be attributed to its inherent ability to model complex, non-linear relationships and enhanced feature weighting mechanisms. In contrast, Logistic Regression, while simpler and faster, struggled to capture the non-linear dependencies between predictors.

Feature importance analysis using SHAP (Shapley Additive explanations) further validated the findings, identifying age, hypertension, glucose levels, and smoking status as the most critical predictors of stroke risk. For instance, older age and the presence of hypertension were consistently linked to higher stroke risk, which is fully aligned with established medical knowledge. This insight not only adds interpretability to our model but also reinforces the clinical relevance of the selected features.

V. DISCUSSION

Machine learning-based brain stroke prediction offers a revolutionary method for stroke early detection and prevention based on patient health information. In our research, we trained models on the Kaggle Stroke Prediction Dataset, which embrace 5110 patient records with 11 critical aspect such as age, hypertension, glucose, BMI, smoking, and heart disease. The XGBoost algorithm proved to be the most efficient with 88.9% accuracy, surpassing the conventional classifiers like Logistic Regression (78.9%) and Random Forest (82.3%) because XGBoost can efficiently tackle the non-linear relationships present in medical datasets.

Feature importance analysis with SHAP (Shapley Additive explanations) identified age, hypertension, and glucose levels as the Fundamental contributing factors of stroke risk. These results are consistent with medical literature, where older patients with high blood pressure and abnormal glucose levels are much more likely to experience stroke events. Although encouraging, issues like data imbalance—where stroke-positive samples account for only 5% of the dataset—were addressed using SMOTE (Synthetic Minority Over-sampling Technique) to improve model sensitivity.

A second limitation of this research is that there is no real-time monitoring of health status, e.g., continuous blood pressure or genetic predisposition indicators, which might further improve the accuracy of prediction. Overcoming these issues in future work may include deep learning methods, with the inclusion of medical imaging information (CT/MRI scans), and model validation in clinical environments to guarantee realistic applicability. In addition, moral issues like model bias need to be rigorously explored in order to avoid unequal healthcare AI choices. Machine learning is here recognized as a valuable tool for predicting strokes, though, by necessitating comprehensive multi-source health information for ideal application in real life.

VI. CONCLUSION

Brain stroke prediction using machine learning is revolutionizing healthcare by offering early risk detection through data-driven insights. This study utilized the Kaggle Stroke Prediction Dataset with 5110 patient records, applying

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

preprocessing techniques such as median addressing missing data. BMI values, converting categorical data into numerical data, Standardizing the data ensures each variable contributes equally to the outcome, and SMOTE-based class balancing to enhance model effectiveness. The analysis identified age, hypertension, glucose levels, and smoking status as the most critical predictors, aligning with established medical findings. Three model like Logistic Regression, Random Forest, and XGBoost were trained and evaluated using an 80-20 train-test split also 10-fold cross-validation, with XGBoost achieving the highest accuracy of 88.9% due to its ability to handle non-linear relationships efficiently. Feature importance analysis using SHAP (Shapley Additive explanations) further reinforced the relevance of selected attributes, providing interpretable predictions. Despite these promising results, challenges such as class imbalance, limited real-time health monitoring, and potential biases remain areas for future research. Incorporating deep learning techniques, genetic predisposition factors, and medical imaging data could refine model predictions and elevate stroke risk assessments to clinical applicability. Ethical concerns surrounding fairness, transparency, and data security must also be addressed before real-world implementation. This study Showcases the pivotal role of machine learning in anticipating healthcare trends ,setting the stage for AI-driven, personalized, proactive stroke intervention strategies that can achieve better patient health and reduce mortality rates.

REFERENCES

1. Indian Council of Medical Research (ICMR). (2022). Artificial Intelligence in Stroke Prediction and Prevention: A Review. Retrieved from https://www.icmr.nic.in
2. National Institute of Mental Health and Neurosciences (NIMHANS). (2021). Machine Learning-Based Stroke Risk Assessment: An Indian Perspective. Journal of Neurological Research, 58(2), 134–147.
3. Machine Learning for Healthcare" by John D. Kelleher, Brendan Tierney, and Elaine Toland
4. Artificial Intelligence in Healthcare" by Rajendra Akerkar
5. Kaggle Stroke Prediction Dataset. (2021). Public Stroke Dataset for Predictive Modeling. Retrieved from https://www.kaggle.com
6. World Health Organization (WHO). (2022). Global Stroke Burden and AI-Based Intervention Strategies. Retrieved from https://www.who.int
7. National Health Mission (NHM), India. (2023). Advancements in AI for Stroke Diagnosis and Prevention in Indian Healthcare. Retrieved from https://nhm.gov.in

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com