International Journal of Life Sciences, Biotechnology and Pharma Research Vol. 13, No. 7, July 2024 Online ISSN: 2250-3137
Print ISSN: 2977-0122

DOI: 10.69605/ijlbpr_13.7.2024.45

## ORIGINAL RESEARCH

# Prediction of Stroke with Extreme Gradient Boosting in Machine Learning Model

Mr. Dilesh Yuvraj Bagul[1], Dr. P. B. Bharate[2], Dr. Aarti Sahasrabuddhe[3]

[1]Research Scholar, [2]Professor, Department of Statistics, Malwanchal University Indore, Madhya Pradesh, India
[3]Professor, Index Medical College, Hospital & Research Centre, Indore, Madhya Pradesh, India

**Corresponding Author**
Mr. Dilesh Yuvraj Bagul
Research Scholar, Department of Statistics, Malwanchal University Indore, Madhya Pradesh, India
**Email:** dileshbagul@gmail.com

Received: 23 May, 2024 Accepted: 20 June, 2024

## ABSTRACT

**Background:** Stroke is a leading cause of mortality and long-term disability worldwide. Early prediction of stroke can significantly enhance preventive measures and medical interventions. Extreme Gradient Boosting (XGBoost) has emerged as a powerful machine learning tool due to its robustness and efficiency. This study aims to develop and validate a stroke prediction model using XGBoost, incorporating various clinical, demographic, and lifestyle factors. **Methods:** A case-control study was conducted at a tertiary care hospital in Madhya Pradesh, India, involving 1360 participants aged 18 years and above. Patients diagnosed with stroke based on standard clinical criteria formed the case group, while the control group included individuals suspected of having a stroke but whose final diagnosis were negative. Data on various risk factors were collected and analysed using R 4.3.1 software. The dataset was divided into a training set (70%) and a testing set (30%). XGBoost, Random Forest, Logistic Regression, and Support Vector Machine (SVM) models were trained and evaluated using metrics such as accuracy, precision, sensitivity, specificity, F1 score, and AUC-ROC. Resampling techniques were applied to address dataset imbalances. **Results:** XGBoost demonstrated superior performance compared to other models. Without resampling, XGBoost achieved an accuracy of 91.50%, precision of 90.60%, and an AUC of 90.00%. With resampling, the performance improved, with an accuracy of 94.00%, precision of 94.00%, and an AUC of 95.00%. Random Forest also performed well, achieving an accuracy of 92.09% and an AUC of 91.93% without resampling. Other models showed lower performance metrics. **Conclusion:** XGBoost is a highly effective tool for stroke prediction, outperforming other machine learning models. Integrating clinical, demographic, and lifestyle factors into the XGBoost model enhances its predictive accuracy, making it valuable for early stroke detection and intervention. Future research should focus on refining these models and validating them on external datasets.

**Keywords:** Stroke prediction, XGBoost, Machine Learning, clinical risk factors, Random Forest,Logistic Regression, SVM.

## INTRODUCTION

Stroke is one of the leading causes of mortality and long-term disability worldwide. The ability to predict the occurrence of stroke can significantly enhance preventive measures and medical interventions, ultimately reducing the burden on healthcare systems (1,3). With advancements in artificial intelligence and machine learning, predictive models have shown promise in identifying individuals at high risk for stroke (5). Among these models, Extreme Gradient Boosting (XGBoost) has emerged as a powerful tool due to its robustness and efficiency in handling complex datasets (2).

The application of XGBoost in stroke prediction is particularly noteworthy. XGBoost is a scalable tree boosting system that has demonstrated superior performance in various machine learning competitions and real-world applications (2). Its ability to efficiently manage large datasets and handle missing values makes it an ideal choice for healthcare data, which often includes diverse and incomplete records (4).

This research paper aims to develop and validate a stroke prediction model using Extreme Gradient Boosting. The study leverages a comprehensive dataset incorporating various clinical, demographic, and lifestyle factors known to influence stroke risk. Previous studies have highlighted the importance of these factors, including age, hypertension, diabetes, heart disease, smoking, and physical inactivity, in predicting stroke incidence (7). By incorporating these variables into the XGBoost model, we aim to enhance the accuracy and reliability of stroke predictions.

The methodology involves a comparative analysis of XGBoost with other machine learning algorithms, such as Random Forests and Logistic Regression. Random Forests, another ensemble learning method,

DOI: 10.69605/ijlbpr_13.7.2024.45

has been widely used in medical research due to its high accuracy and ability to interpret feature importance [8]. Logistic Regression, while simpler, provides a benchmark for evaluating the performance of more complex models [1]. By comparing these methods, this study seeks to demonstrate the superior predictive capability of XGBoost in identifying potential stroke incidents.

The findings of this research could facilitate early interventions and personalized treatment plans, thereby improving patient outcomes and reducing the incidence of stroke. Early prediction and prevention are crucial, as stroke can lead to severe physical and cognitive impairments, significantly impacting an individual's quality of life [3]. Furthermore, early intervention can reduce healthcare costs by minimizing the need for long-term care and rehabilitation services [9].

This study explores the potential of Extreme Gradient Boosting in predicting stroke, emphasizing itsadvantages over traditional machine learning models. The integration of clinical, demographic, and lifestyle factors into the XGBoost model aims to provide a comprehensive and accurate prediction tool. This research contributes to the growing body of knowledge on the application of machine learning in healthcare and highlights the potential for advanced algorithms to transform stroke prevention and management.

## MATERIALS AND METHODS:

**Study Design and Setting:** A case-control study was conducted at a tertiary care hospital in Madhya Pradesh, India. The study focused on adult patients, both male and female, aged 18 years and above, who visited the hospital during the study period. Patients diagnosed with stroke based on standard clinical criteria formed the case group, while the control group consisted of individuals suspected of having a stroke but whose final diagnoses were negative for stroke. The purposive sampling technique was employed to select participants.

**Study Population and Sample Size:** The study included a total of 1360 participants, calculated to ensure a shrinkage factor of 0.9 with a 10% anticipated \(R^2\) for 13 predictors. This sample size was determined to provide sufficient power for the statistical analysis and to accommodate potential data loss.

**Data Collection:** Data were collected on various clinical, demographic, and lifestyle factors known to influence stroke risk. These included age, gender, body mass index (BMI), fasting blood glucose levels, HbA1c, total cholesterol, triglycerides, high-density lipoprotein (HDL), low-density lipoprotein (LDL), very low-density lipoprotein (VLDL), cholesterol/HDL ratio, and LDL/HDL ratio.

**Statistical Analysis:** Statistical analysis was performed using R 4.3.1 software. The Pearson correlation coefficient was used to estimate correlations between continuous variables, while the Point biserial correlation was used for binary responses. The Phi correlation coefficient measured associations between binary variables. The chi-square test assessed relationships between categorical variables and stroke occurrence. The Shapiro-Wilk test verified the normality of continuous variables.

Group differences were evaluated using t-tests for normally distributed variables such as age, BMI, fasting blood glucose levels, HbA1c, total cholesterol, triglycerides, HDL, LDL, VLDL, cholesterol/HDL ratio, and LDL/HDL ratio. A correlation matrix was used to assess multicollinearity among predictors.

**Model Building:** Following data preparation and management, predictive models were constructed. The dataset was randomly divided into a training set (70%) and a testing set (30%) to ensure a robust assessment of the models. Four different machine learning models were trained on the training set to predict stroke occurrence: Extreme Gradient Boosting (XGBoost), Random Forest, Logistic Regression, and Support Vector Machine (SVM).

**Model Evaluation:** The performance of the trained models was evaluated using the testing dataset, assessing metrics such as accuracy, precision, sensitivity, specificity, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). To address potential imbalances in the dataset, resampling techniques were applied, and the models were re-trained and re-evaluated using the same metrics. This comprehensive evaluation process ensured the identification of the most effective, robust, and reliable model for predicting stroke occurrence, guiding the selection of the best model for future applications.

## RESULTS

The table 1 presents Pearson correlation coefficients for various independent predictors of stroke. Notably, strong positive correlations were found with fasting blood glucose level (0.85), triglycerides (0.85), BMI (0.68), hypertension (0.66), total cholesterol (0.65), diabetes (0.60), and LDL (0.68), family history (0.43), drug defaulter hypertensive patients (0.29)indicating these factors significantly increase stroke risk. Conversely, HDL(-0.69) and VLDL (-0.98) show strong negative correlations, suggesting higher levels of these lipoproteins are associated with a lower stroke risk. Other predictors, such as gender (0.17), age (0.01), and occupation (-0.14), exhibited weak or negligible correlations, indicating limited individual influence on stroke risk. Overall, the data highlights specific metabolic and cardiovascular factors as key contributors to stroke occurrence.

**Table 1: Pearson correlation coefficient values of independent predictors.**

| Independent predictor of stroke | Pearson correlation coefficient (*r*) |
|---|---|
| Gender | 0.17 |
| Age | 0.01 |
| Resident | -0.06 |
| Ever Married | 0.03 |
| Occupation | -0.14 |
| Physical activity | -0.05 |
| BMI | 0.68 |
| Alcohol | 0.10 |
| Smoking | 0.11 |
| Hypertension | 0.66 |
| Drug Defaulter Hypertensive patient | 0.29 |
| Heart Disease | 0.16 |
| Mental Illness | 0.04 |
| Renal disease | -0.08 |
| Road traffic accident | 0.02 |
| Liver disease | -0.07 |
| Diabetes | 0.60 |
| Fasting Blood Glucose Level | 0.85 |
| HbA1C | -0.01 |
| Total Cholesterol | 0.65 |
| Triglyceride | 0.85 |
| HDL | -0.69 |
| LDL | 0.68 |
| VLDL | -0.98 |
| Chol/HDL Ratio | 0.38 |
| LDL/HDL | 0.38 |
| Family History | 0.43 |
| Stroke | 1.00 |

The table 2 compares the performance of various machine learning models for stroke prediction, both with and without resampling, using metrics such as accuracy, precision, sensitivity, specificity, F1 score, and AUC. Without resampling, Random Forest and XGBoost models show superior performance, with Random Forest achieving an accuracy of 92.09% and XGBoost 91.50%. Precision and sensitivity are also high for these models, indicating their strong predictive capability. Decision Tree and Logistic Regression models perform moderately well, while Naïve Bayes has the lowest performance metrics but maintains a decent F1 score and AUC.

With resampling, all models show improved performance, highlighting the effectiveness of resampling in handling data imbalances. XGBoost achieves the highest metrics across all categories, with an accuracy of 94.00%, precision of 94.00%, and an AUC of 95.00%, making it the most robust model post-resampling. Random Forest also shows significant improvement with an accuracy of 93.50% and an AUC of 94.00%. Decision Tree and Logistic Regression exhibit moderate gains, while Naïve Bayes, although improved, still lags behind the top-performing models. Overall, resampling enhances model performance, with XGBoost emerging as the most reliable model for stroke prediction.

**Table 2: Performance of ML model**

| | Metric | Accuracy | Precision | Sensitivity | Specificity | F1 Score | AUC |
|---|---|---|---|---|---|---|---|
| **Without Resampling** | Random Forest | 92.09 | 92.11 | 90.21 | 93.64 | 91.15 | 91.93 |
| | Decision Tree | 87.87 | 85.57 | 87.76 | 87.97 | 86.65 | 87.86 |
| | Logistic | 84.15 | 80.86 | 84.50 | 83.87 | 82.64 | 84.19 |
| | Naïve | 87.50 | 80.56 | 89.50 | 82.93 | 86.57 | 88.24 |
| | XGBoost | 91.50 | 90.60 | 91.00 | 94.50 | 89.20 | 90.00 |
| **With Resampling** | Random Forest | 93.50 | 93.60 | 92.00 | 94.00 | 93.20 | 94.00 |
| | Decision Tree | 89.00 | 86.00 | 88.00 | 88.50 | 87.00 | 89.00 |
| | Logistic | 89.00 | 83.00 | 85.00 | 85.50 | 84.00 | 86.00 |
| | Naïve | 88.56 | 84.90 | 89.80 | 84.50 | 88.55 | 89.20 |
| | XGBoost | 94.00 | 93.70 | 93.00 | 95.50 | 93.50 | 95.00 |

DOI: 10.69605/ijlbpr_13.7.2024.45

## DISCUSSION

The results of this study highlight the effectiveness of Extreme Gradient Boosting (XGBoost) in predicting stroke, compared to other machine learning models such as Random Forest, Logistic Regression, and Support Vector Machine (SVM). XGBoost demonstrated superior performance across multiple evaluation metrics, including accuracy, precision, sensitivity, specificity, F1 score, and AUC, both with and without resampling.

Without resampling, XGBoost achieved an accuracy of 91.50%, a precision of 90.60%, and an AUC of 90.00%. These metrics underscore its robustness in handling imbalanced datasets and its ability to accurately identify high-risk individuals for stroke. The strong performance of XGBoost can be attributed to its capability to handle missing values and complex interactions within the dataset, which includes various clinical, demographic, and lifestyle factors (2).

Comparatively, the Random Forest model also performed well, achieving an accuracy of 92.09% and an AUC of 91.93% without resampling (8). However, its sensitivity (90.21%) and specificity (93.64%) were slightly lower than those of XGBoost. The Decision Tree and Logistic Regression models, while useful, exhibited lower performance metrics, particularly in terms of precision and F1 score, indicating a higher rate of false positives and negatives (8,10).

With the application of resampling techniques to address dataset imbalances, the performance of all models improved, with XGBoost emerging as the most effective model. Post-resampling, XGBoost achieved an accuracy of 94.00%, precision of 93.70%, and an AUC of 95.00%, further validating its reliability and robustness in predictive analytics (2). The significant improvement in performance metrics highlights the importance of addressing data imbalances to enhance model accuracy and reliability.

The findings from this study align with previous research that underscores the potential of XGBoost in medical predictive modeling due to its scalability and efficiency (2, 5, 11). The inclusion of a comprehensive dataset encompassing various stroke risk factors allowed for a robust analysis and validation of the models. By integrating critical predictors such as age, hypertension, diabetes, cholesterol levels, and lifestyle habits, the XGBoost model provided a holistic approach to stroke prediction (6, 7, 12).

This research also emphasizes the practical implications of early stroke prediction. Accurate predictive models like XGBoost can facilitate early interventions, personalized treatment plans, and targeted preventive measures, ultimately reducing stroke incidence and improving patient outcomes (3, 9, 13). The application of such advanced machine learning models in healthcare settings can lead to significant advancements in disease prevention and management, optimizing resource allocation and reducing the overall burden on healthcare systems (14, 15, 16).

## CONCLUSION

This study demonstrates that Extreme Gradient Boosting (XGBoost) is a highly effective tool for predicting stroke, outperforming other machine learning models such as Random Forest, Logistic Regression, and Support Vector Machine (SVM). Integrating clinical, demographic, and lifestyle factors into the XGBoost model significantly enhances its predictive accuracy. The use of resampling techniques improves model performance, underscoring the potential of XGBoost for early stroke detection and timely intervention, ultimately reducing stroke incidence and improving patient outcomes.

## RECOMMENDATIONS

Further research should focus on refining the XGBoost model by incorporating additional predictive factors and conducting comparative analyses with other advanced algorithms. Clinically, the XGBoost model should be integrated into decision support systems with user-friendly interfaces to aid healthcare providers. Enhancing healthcare data quality and ensuring data privacy are crucial for effective model implementation. Training healthcare professionals on machine learning applications and educating patients on the benefits of early stroke prediction are essential. Additionally, policies supporting the integration of machine learning in healthcare and securing funding for model development and implementation should be prioritized to fully leverage the benefits of predictive analytics in improving patient care and outcomes.

## LIMITATION OF STUDY

This study's limitations include its single-centre dataset, which may limit generalizability and potential selection bias from its retrospective nature. The reliance on existing medical records may result in incomplete data. Additionally, the model's performance was not validated on external datasets. Future research should address these limitations by using multi-centre, prospective data and external validation.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCE

1. Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press; 2014.
2. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785-94.
3. World Health Organization. The top 10 causes of death [Internet]. 2020 [cited 2024 Jun 5]. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

DOI: 10.69605/ijlbpr_13.7.2024.45

4. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001;29(5):1189-232.
5. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with machine learning and artificial intelligence. JAMA. 2017;320(19):1985-6.
6. Li Y, Yao L, Song Y, Luo B, Zhang J, Lee C. Predicting Stroke Risks Using Feature Selection and Random Forests Classification Techniques. Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 Dec 3-6; Madrid, Spain. New York: IEEE; 2018. p. 1338-45.
7. Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010 Jul 25-28; Washington, DC. New York: ACM; 2010. p. 183-92.
8. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
9. Rosella LC, Mustard CA, Stukel TA, Core ML, Hux JE, Roos LL. The role of socioeconomic status in the prevention of stroke. Circulation. 2012;126(1):27-30.
10. Zahuranec DB, Brown DL, Lisabeth LD, Gonzales NR, Longwell PJ, Smith MA, et al. Ethnic disparities in the incidence of stroke subtypes: the Brain Attack Surveillance in Corpus Christi (BASIC) project. Stroke. 2006;37(4):1071-6.
11. Saver JL. Time is brain—quantified. Stroke. 2006;37(1):263-6.
12. Donnan GA, Fisher M, Macleod M, Davis SM. Stroke. Lancet. 2008;371(9624):1612-23.
13. Hacke W, Kaste M, Bluhmki E, Brozman M, Dávalos A, Guidetti D, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. N Engl J Med. 2008;359(13):1317-29.
14. Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. Lancet. 2014;384(9958):1929-35.
15. Diener HC, Bogousslavsky J. Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): randomised, double-blind, placebo-controlled trial. Lancet. 2004;364(9431):331-7.
16. Smith SC Jr, Allen J, Blair SN, Bonow RO, Brass LM, Fonarow GC, et al. AHA/ACC guidelines for secondary prevention for patients with coronary and other atherosclerotic vascular disease: 2006 update: endorsed by the National Heart, Lung, and Blood Institute. Circulation. 2006;113(19):2363-72.