JNeViS
Journal of Neurointervention and Stroke

Journal of Neurointervention and Stroke 2025;1(2):63-71
doi: 10.63937/jnevis-2025.12.16

Original Article

Enhancing Stroke Risk Prediction with Explainable AI: Leveraging Resampling and Machine Learning for Improved Accuracy

Minhazul Alam Mahin[1], Md. Mominul Islam[1]✉, MD. Zulfikar Alam[1], Arnob Dutta Pollob[1], Oxita Zaman[1]

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

Corresponding Author:
Md. Mominul Islam
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
Email: mominul.diu.cse@gmail.com

Received: September 19, 2025
Revised: October 8, 2025
Accepted: October 19, 2025
Published: December 2, 2025

Introduction: Stroke represents a significant global health concern, impacting millions worldwide and contributing substantially to morbidity and mortality. Early detection and accurate risk prediction remain critical for effective prevention strategies. **Objective:** This study aimed to improve stroke risk prediction by employing machine learning algorithms on health survey data to identify key predictors and enhance predictive performance. **Method:** A dataset derived from the National Health and Nutrition Examination Survey, comprising 4,603 participants, was utilized. The dataset exhibited class imbalance, with only 7.86% of individuals diagnosed with stroke. To address this imbalance, advanced resampling techniques, including SMOTE, SMOTETomek, and ADASYN, were applied. A range of tree-based algorithms was implemented, including Gradient Boosting, AdaBoost, XGBoost, and a Voting Classifier integrating Decision Tree, AdaBoost, and Gradient Boosting classifiers. Model evaluation included accuracy and AUC scores. Explainable Artificial Intelligence (XAI) analyses were conducted using SHAP (SHapley Additive exPlanations) to interpret feature importance. **Result**: The Gradient Boosting classifier, in conjunction with SMOTE, achieved the highest performance with an accuracy of 92% and an AUC score of 0.70. SHAP analysis identified age, general health condition, marital status, and BMI as the most influential predictors of stroke risk. **Conclusion:** This study underscores the essential need for ongoing advancements in early stroke detection methodologies. The findings highlight the transformative potential of machine learning and XAI in predictive healthcare, offering valuable insights for stroke prevention strategies.

Keywords: ADASYN, Class imbalance, Explainable AI (XAI), Resampling techniques SMOTE, SHAP, Stroke prediction

Highlights

- Advanced resampling techniques improved class balance in stroke datasets.
- Gradient Boosting with SMOTE reached 92% accuracy with SHAP interpretability.

Introduction

Stroke is a major global health issue that affects millions of people every year and causes high mortality rates and permanent impairments. Globally, 110 million people have experienced stroke, with 15 million new cases reported each year.[1,2] The impact is widespread, with 60% of strokes occurring in individuals aged <70 years, 38% in those aged <65 years, and 16% in those aged <50 years.[3,4] In the United States, a stroke occurs every 40 seconds, amounting to 795,000 cases annually, of which 610,000 are first-time incidents and

the remaining 185,000 are recurrent.[5,6] Alarmingly, a stroke-related death occurs every three and a half minutes in the U.S. alone.

Despite these staggering figures, up to 80% of strokes are preventable with increased public awareness and timely intervention. Unfortunately, only 38% of the population is familiar with major stroke symptoms, highlighting the urgent need for improved public education and early detection strategies.[7] Predicting stroke risk is crucial for reducing its global impact, particularly given the preventable nature of many stroke cases.

Machine learning has become an influential resource in the healthcare industry for recognizing patterns within complex data, which may assist in predicting strokes at an early stage. However, stroke datasets often suffer from class imbalances, where the number of stroke cases is disproportionately lower than the number of non-stroke cases, leading to biased predictions. Addressing this imbalance is key to developing models that are accurate and reliable for early diagnosis.

Although machine learning is increasingly being used to predict stroke, much of the current research fails to address class imbalance and rarely investigates how different resampling strategies interact with various machine learning models. Furthermore, few researchers have used interpretable AI tools to explain model predictions, which limits their application in real-world clinical decision-making.

The present study addresses these gaps by systematically evaluating modern resampling strategies (SMOTE, SMOTETomek, and ADASYN) in combination with tree-based ensemble classifiers (Gradient Boosting, AdaBoost, XGBoost, and a hard-voting ensemble) for stroke risk prediction in an imbalanced, population-based survey dataset. In addition, we incorporate explainable AI techniques, including feature importance analysis and SHAP, to quantify how demographic, lifestyle, and clinical variables contribute to individual predictions. Validating this framework on a large, publicly available dataset underscores the practical value of interpretable machine-learning–based stroke prediction tools for supporting timely clinical decision-making and early intervention.

Objective

The research aims to improve stroke risk prediction by applying machine learning algorithms to an imbalanced dataset. Specifically, the study evaluates various resampling techniques and tree-based classifiers to identify the most effective combination for accurate and balanced stroke detection. Additionally, it employs explainable AI methods to clarify the key factors influencing stroke risk, thereby supporting clinical decision-making.

Method

The overall methodology shown in Figure 1 includes several stages of analysis. First, the dataset was pre-processed and standardized to ensure consistent feature scaling. The data was then split into training and testing sets. To address class imbalance, sampling techniques such as ADASYN, SMOTE, and SMOTETomek were applied. Various machine learning models, including Gradient Boosting, AdaBoost, XGBoost, and Voting classifiers, were trained and evaluated using metrics such as classification reports, confusion matrices, ROC curves, and AUC scores. Gradient Boosting was identified as the best-performing model. Finally, XAI techniques, including Feature Importance and SHAP, were implemented to interpret the model's predictions on the testing set.

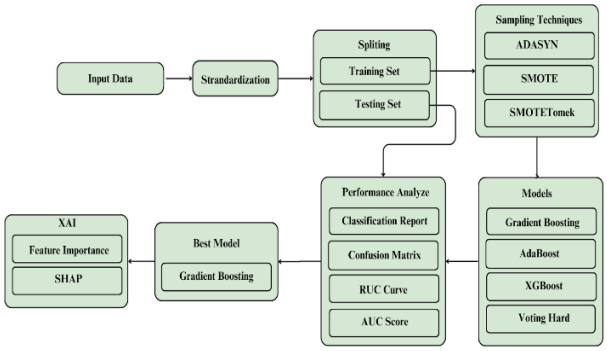

Figure 1. Research methodology diagram

Dataset

This study was based on a dataset that included 4603 participants obtained from the National Health and Nutrition Examination Survey.[8] These subjects met the inclusion and exclusion criteria defined for this analysis. Among them, 362 individuals (7.86%) were diagnosed with stroke, whereas 4,241 individuals (92.14%) were non-stroke patients, making the dataset highly imbalanced. The outcome variable, stroke, is a binary classification target, and the remaining columns represent predictors, including demographic, lifestyle, clinical, and dietary features. This dataset was collected from Mendeley Data, making it a rich source for analyzing stroke risk factors and predicting stroke occurrences using machine learning models. The details of the datasets are presented in Table 1.

Data preprocessing

During preprocessing, scaling was performed to standardize the continuous features so that they have a mean of zero and a standard deviation of one.

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

In this context, Z represents the standardized value, x denotes the original value, μ is the mean of the attribute, and σ corresponds to the standard deviation of the attribute.

Table 1. Dataset's Attribute

Research Variables	
Categorical Variables	
Demographics	Gender
	Race
	Marital status
Lifestyle Factors	Alcohol consumption
	Smoking status
	Sleep disorder
	Health Insurance
Health Conditions	General health condition
	Depression
	Diabetes
	Hypertension
	High cholesterol
	Coronary Heart Disease
Continuous Variables	
Physical Measurements	Age
	Body Mass Index
	Waist Circumference
	Systolic
	Diastolic blood pressure
Biochemical Markers	High-density lipoprotein (HDL)
	Low-density lipoprotein (LDL)
	Triglycerides
	Fasting Glucose
	Glycohemoglobin
Dietary Intake	Energy
	Protein
	Carbohydrate
	Dietary fiber
	Total fat
	Saturated
	Monounsaturated
	Polyunsaturated fatty acids
	Potassium
	Sodium
Activity Levels	Minutes sedentary activity
	Sleep time

Splitting

To ensure fair and reliable model testing, we divided the dataset into training and testing sets in an 80:20 ratio using stratified sampling methods. All models were trained and tested on the same stratified partitions to ensure comparability. Our dataset contained 4,603 records; in total, 4,241 were labeled as "no stroke" and 362 as "stroke," indicating that the data were highly imbalanced (approximately 92% vs. 8%). Stratification helped maintain this proportion in both the training and testing sets so that each set reflected the real-world distribution of cases. This is especially important in medical datasets, where class imbalance can lead to misleading results.

Sampling technique

To address the challenges of imbalanced datasets in stroke prediction, this study employed several effective sampling techniques: ADASYN, SMOTE, and SMOTETomek. ADASYN was used to create synthetic samples for the minority class, giving priority to instances that were difficult to classify, thereby enhancing the model's focus on challenging cases. In contrast, SMOTE generates synthetic samples to increase the F1-score for the minority class. Additionally, SMOTETomek combines the advantages of SMOTE with Tomek links to refine the majority class by removing ambiguous instances. These methods were systematically applied to ensure a balanced dataset, thereby improving model performance and enabling more effective learning from the data.

Machine Learning Algorithms

In this study, we analyzed various machine learning algorithms for stroke prediction. The models used included Gradient Boosting (GB), AdaBoost (AB), XGBoost (XGB), and a Voting Classifier (hard voting) that combined the AdaBoostClassifier, DecisionTreeClassifier, and GradientBoostingClassifier. The algorithms were implemented on the dataset to assess their effectiveness in predicting stroke events, with their predictive performance evaluated in terms of accuracy, precision, recall, confusion matrix, and AUC score.

Gradient Boosting

Gradient Boosting, widely employed in supervised learning, is capable of handling both classification and regression problems. The method focuses on optimizing the loss function through an iterative process in which each new model is fitted to the residuals of the previous models.[9,10]

$$F_m(X) = F_{m-1}(X) + \eta \times f_m(X) \quad (2)$$

Applied to stroke prediction, Gradient Boosting sequentially builds decision trees that correct the errors of preceding ones, optimizing performance using a loss function. It achieved the best results among all models, indicating strong capability in capturing complex patterns in the data.

AdaBoost

AdaBoost has been effectively applied to classification tasks, particularly in medical domains such as diagnosis and prognosis. It operates by integrating multiple weak classifiers into a strong ensemble model capable of producing reliable predictions.[11,12]

$$F(x)=(\sum_{m=1}^{M} a_m h_m(x)) \quad (3)$$

XGBoost

The XGBoost algorithm is derived from the CART (Classification and Regression Tree) framework. It

constructs weak learners by selecting subsets of features, incrementally fits the residuals in stages, and ultimately integrates them into a robust predictive model.[13,14]

$$L_{xgb} = \sum_{i=1}^{N} \quad L\left(y_i, F(x_i)\right) + \sum_{m=1}^{M} \quad \Omega(h_m) \text{ (4)}$$

Voting Hard

The Voting Classifier aggregates predictions from multiple models to enhance overall accuracy. By employing a hard voting mechanism, it combines the outputs of AdaBoost, Decision Tree, and Gradient Boosting classifiers, leveraging their individual strengths to improve stroke prediction reliability.

Result

Across all models, SMOTE and SMOTETomek consistently produced the highest accuracy (92% for Gradient Boosting) and recall scores (up to 0.92). However, the precision was slightly lower when using these sampling techniques, which implies that while the models are good at detecting stroke cases, there may be a slightly higher rate of false-positive results. The Gradient Boosting emerges as the best-performing approach for this dataset, offering a balance between precision, recall, and overall accuracy.

It is clear from this analysis that sampling strategies are important for improving machine learning models' performance while working with unbalanced datasets. SMOTE and SMOTETomek generally yielded superior results, particularly for the Gradient Boosting models. The interplay between precision and recall demonstrates that these techniques are useful for recognizing stroke risk factors, which makes them appropriate for practical medical use where both sensitivity and specificity are essential.

Table 2 reports test-set performance for Gradient Boosting, AdaBoost, XGBoost, and a Voting classifier across ADASYN, SMOTE, and SMOTETomek. Table 3 reports the corresponding training-set results. SMOTE and SMOTETomek generally outperform ADASYN; Gradient Boosting provides the best balance of accuracy, precision, recall, and F1-score, while the Voting classifier is weaker. Training values are higher but preserve the same ordering as the test set, and small train–test gaps suggest limited overfitting under the shared stratified split.

Accuracy

Accuracy is the proportion of correctly classified samples to the total number of samples in the dataset. This reflects the overall effectiveness of the model in correctly predicting both positive and negative

Table 2. Results of all models on testing set

Model	Sampling	Accuracy	Precision	Recall	F1-score
Gradient Boosting	ADASYN	0.955	0.956	0.961	0.958
	SMOTE	0.956	0.966	0.970	0.968
	SMOTETomek	0.956	0.964	0.969	0.967
AdaBoost	ADASYN	0.922	0.955	0.944	0.949
	SMOTE	0.922	0.955	0.961	0.958
	SMOTETOmek	0.922	0.955	0.961	0.958
XGBoost	ADASYN	0.972	0.969	0.969	0.974
	SMOTE	0.972	0.979	0.979	0.979
	SMOTETomek	0.972	0.979	0.979	0.979
Voting Hard	ADASYN	0.961	0.955	0.960	0.957
	SMOTE	0.951	0.966	0.960	0.963
	SMOTETomek	0.951	0.966	0.960	0.963

Table 3. Results of all models on training set

Model	Sampling	Accuracy	Precision	Recall	F1-score
Gradient Boosting	ADASYN	0.912	0.862	0.910	0.885
	SMOTE	**0.922**	**0.891**	**0.922**	**0.906**
	SMOTETomek	0.921	0.891	0.922	0.906
AdaBoost	ADASYN	0.895	0.870	0.910	0.889
	SMOTE	0.884	0.871	0.880	0.876
	SMOTETomek	0.884	0.871	0.880	0.876
XGBoost	ADASYN	0.912	0.865	0.910	0.876
	SMOTE	0.912	0.865	0.910	0.887
	SMOTETomek	0.912	0.865	0.910	0.887
Voting Hard	ADASYN	0.920	0.873	0.902	0.887
	SMOTE	0.881	0.871	0.912	0.891
	SMOTETomek	0.881	0.871	0.912	0.891

categories. The formula for calculating accuracy is shown here.[15]

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision

Precision, in the context of classification models, refers to the proportion of true positives (TP) among all instances predicted as positive. It measures the model's accuracy in identifying only relevant cases.[16] The formula for calculating precision is:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

Recall

Recall evaluates a model's ability to correctly detect positive cases among all actual positives in the dataset.[17] The formula for calculating recall is:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

doi: 10.63937/jnevis-2025.12.16

F1-score

Defined as the harmonic average of precision and recall, the F1-score serves as a consolidated indicator that maintains equilibrium between the two. This measure is
especially valuable in situations with imbalanced data, since it incorporates the influence of both false positives and false negatives.[18] The F1-score is determined using the following equation:

$$\text{F 1-score} = 2\times \frac{Precision\times Recall}{Precision+Recall} \quad (8)$$

Matrix Analysis

The confusion matrix in Figure 2 evaluates the Gradient Boosting model with SMOTE, showing 837 true negatives and only 8 false positives, but also 72 false negatives and just 4 true positives, indicating difficulty in correctly identifying stroke cases. The ROC curve in Figure 2 yields an AUC of 0.7094, reflecting moderate discrimination between stroke and non-stroke classes and highlighting the need to further improve model sensitivity.

Feature Importance

In machine learning, feature importance measures how much each input variable influences a model's predictions. By assessing the importance of each feature, practitioners can enhance the model interpretability, improve performance, and streamline the feature selection processes.[19] The feature importance analysis of stroke risk prediction using the Gradient Boosting model is illustrated in Figure 3, which displays the top predictors identified by the model. The most significant predictor was age, reflecting established medical knowledge that stroke risk escalates with age due to physiological changes and increased comorbidities. Following age, health conditions emerged as a critical factor, highlighting the impact of chronic illnesses such as diabetes and hypertension on stroke risk. Smoking status and High Cholesterol levels are also prominent predictors, reinforcing their established associations with cardiovascular diseases and atherosclerosis. Furthermore, marital status and race enhanced the ability to predict outcomes, highlighting the role of social support and health inequalities. Other notable factors include BMI and sleep time, both of which are linked to various stroke risk factors. Overall, this analysis underscores the necessity of incorporating a holistic view of both physiological and lifestyle factors in stroke risk assessment, enhanced by the balanced dataset achieved through SMOTE.

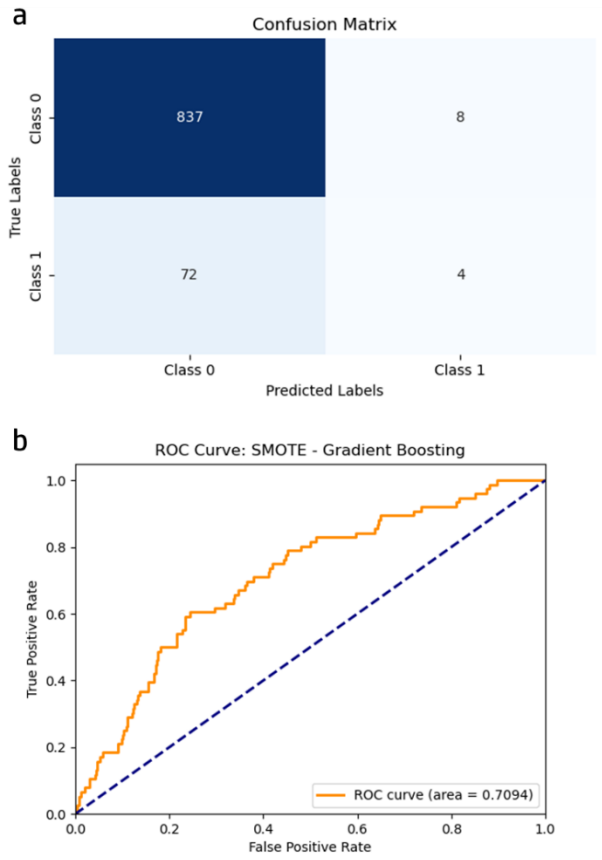

Figure 2. Confusion matrix (a) and ROC (b) of the proposed model

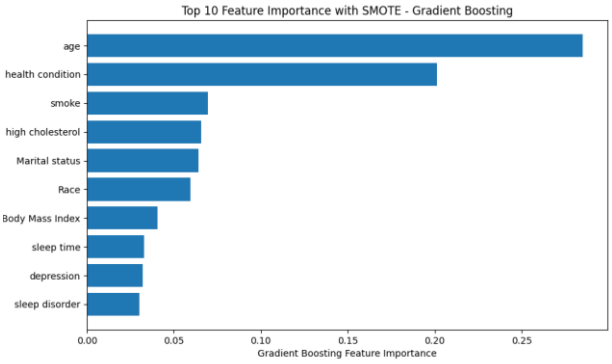

Figure 3. Feature importance of gradient boost classifier

SHAP Analysis

SHAP is an approach that explains machine learning predictions by measuring how much each feature contributes to the outcome. It provides both global and local explanations, showing the extent to which each feature influences a specific prediction.[20] The SHAP analysis, illustrated in Figure 4, provides a comprehensive understanding of the influence of features on stroke predictions made by the Gradient Boosting model. The SHAP summary plot revealed that General Health Condition exhibited the highest SHAP value range, indicating its strong impact on stroke risk,

doi: 10.63937/jnevis-2025.12.16

particularly as high values correlated with chronic health issues such as heart disease and diabetes. Age is another critical predictor, where an increased age significantly raises the likelihood of stroke, reinforcing its established role as a risk factor. Marital Status displayed a variable influence, potentially reflecting lifestyle factors and social support that affect health outcomes. Additionally, BMI and Sleep Time are relevant predictors, with a higher BMI associated with obesity-related risks and sleep patterns affecting overall health. The model also captures the influence of Race, Smoking, and High Cholesterol, all of which align with known cardiovascular risk factors. Taken together, SHAP shows that physiological and behavioral variables are key to stroke-risk evaluation, boosting transparency and aligning with established medical evidence.

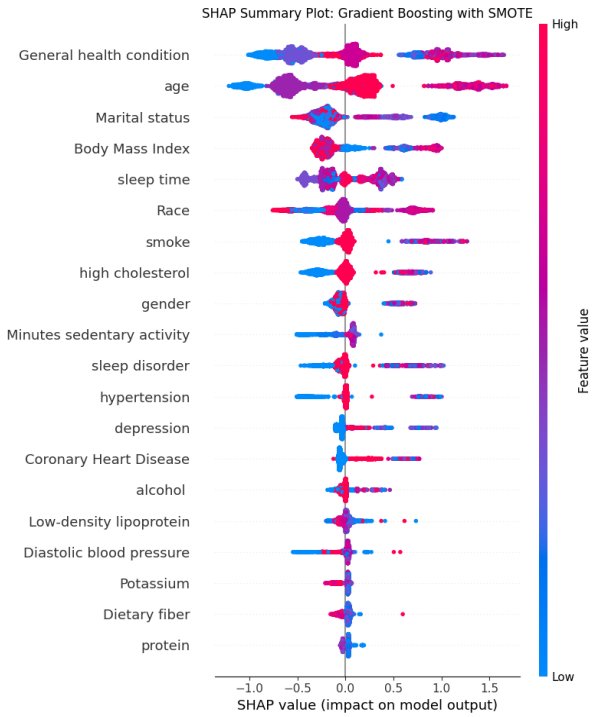

Figure 4. SHAP summary plot

Discussion

A study by Alageel, Nojood, et al. focuses on improving stroke prediction by analyzing electronic health records and identifying key risk factors such as age, average glucose level, heart disease, and hypertension. The research employed a Kaggle dataset and experimented with seven algorithms, including Naive Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbors, Decision Tree, Stacking, and Majority Voting. The dataset was first balanced using sub-sampling to address the issue of class imbalance in stroke occurrences. Following data preprocessing, the algorithms were assessed according to their accuracy, F1 score, recall, and precision. The Naive Bayes classifier achieved the lowest accuracy at 86%, while the other algorithms performed similarly well, with accuracies of approximately 96%, F1 scores of 0.98, precision of 0.97, and perfect recall.[21]

Uddin Emon et al. focused on the early prediction of stroke using machine learning models that incorporated a range of health-related indicators, including hypertension, BMI, history of heart disease, glucose levels, smoking habits, past stroke incidents, and age. For model training, ten different classifiers were employed, including Logistic Regression, SGD, Decision Tree, AdaBoost, Gaussian Naive Bayes, QDA, Multi-Layer Perceptron, K-Nearest Neighbors, Gradient Boosting, and XGBoost. The results from these base models were combined using a weighted voting approach, leading to an overall accuracy of 97%. The weighted voting classifier outperformed individual models in both accuracy and AUC, while minimizing false positive and false negative rates. The study concluded that this model is highly effective in stroke prediction and can be a valuable tool for physicians and patients in detecting potential strokes early.[22]

The study by Hassan, Ahmad, et al. focuses on addressing the challenges of early stroke detection, particularly missing and imbalanced data. It employs three methods for imputing missing data and applies the Synthetic Minority Oversampling Technique to achieve balance in the dataset. The study starts with a foundational model and advances to more sophisticated models, applying k-fold cross-validation on both imbalanced and balanced datasets. Important factors influencing stroke risk include age, BMI, blood glucose levels, presence of heart disease, high blood pressure, and marital status. It also presents a Dense Stacking Ensemble (DSE) model that integrates well-tuned advanced models, with the top-performing model acting as the meta-classifier. The DSE model achieved an accuracy of more than 96% and recorded an AUC score of 83.94% on imbalanced datasets, while achieving 98.92% on balanced datasets. The results demonstrate the DSE model's superior performance compared to previous research, highlighting its potential for early stroke detection and improved patient outcomes.[23]

Dritsas et al. focus on using machine learning to develop models for long-term stroke risk prediction. The study highlights the importance of early symptom recognition to improve stroke prediction and health outcomes. The main contribution is a combining technique that integrates several models, resulting in enhanced performance across different metrics such as AUC, precision, recall, F-measure, and accuracy. The results from the experiment indicate that the stacking classifier surpasses other models, achieving an AUC of 98.9%, with precision, recall, and F-measure scores of 97.4%, along with an overall accuracy of 98%.[24]

doi: 10.63937/jnevis-2025.12.16

Sailasya, et al. focus on predicting brain stroke using various machine learning models, addressing the gap in research on brain stroke risk prediction. Physiological features were used to construct models employing Logistic Regression, Decision Tree, Random Forest, KNN, SVM, and Naive Bayes classifiers. Among the various models, the Naive Bayes model attained the highest accuracy at around 82%, positioning it as the top performer in forecasting the probability of stroke. The study highlights the importance of using machine learning to enhance stroke prediction and diagnosis.[25]

Rahim, Abd Mizwar A., et al. conducted a study aimed at improving stroke prediction accuracy using the XGBoost algorithm. Stroke, the second most deadly disease, occurs when a blood vessel ruptures, cutting off oxygen supply to parts of the brain. The study addresses the challenge of low accuracy in previous healthcare models by applying XGBoost, which was trained on a dataset split 70/30 into training and test sets. The model achieved a high accuracy of 96%, significantly improving prediction performance compared to previous studies, making it a more reliable tool for predicting stroke cases.[26]

Sundaram .M, Sathya, et al, focuses on using machine learning to predict stroke occurrence by analyzing physiological parameters. The study employed four machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, and Voting Classifier—to develop models for accurate stroke prediction. Of these, the Random Forest model demonstrated the highest accuracy, reaching around 96%. The study used an open-access stroke prediction dataset and showed that the accuracy of these models is significantly higher than in previous studies, highlighting their reliability. Extensive model comparisons confirmed their robustness, and the analysis supports the effectiveness of the proposed approach for early stroke prediction.[27]

Biswas, Nitish, et al. conducted a separate study employing machine learning to predict strokes, tackling the issue of imbalanced data using Random Over Sampling. The study evaluated eleven classifiers, including SVM, Random Forest, KNN, Decision Tree, Naive Bayes, Voting Classifier, AdaBoost, Gradient Boosting, Multi-Layer Perceptron, and Nearest Centroid. Prior to data balancing, ten classifiers recorded an accuracy of over 90%, while after balancing, four classifiers surpassed an accuracy of 96%. Adjusting hyperparameters and utilizing cross-validation further enhanced the outcomes. The Support Vector Machine achieved an impressive accuracy of 99.99%, accompanied by comparably high values for recall, precision, and F1-measure. The Random Forest model followed, achieving an accuracy of 99.87%. Additionally, the study developed user-friendly web and mobile applications based on the most accurate model.[28]

Guhdar, Mohammed, et al. focus on predicting the early onset of stroke using a Logistic Regression model. To improve the model's effectiveness, techniques like SMOTE were utilized to balance the class distribution, along with feature selection and methods for managing outliers. The main risk factors considered include elevated blood pressure, body weight, cardiovascular conditions, blood glucose levels, smoking habits, previous strokes, and age. Compared to five other studies using Logistic Regression and the same dataset, this approach achieved the highest F1 score and AUC, with an accuracy of 86%. This predictive model has significant potential for early stroke diagnosis and application in clinical practice.[29]

Mezher, Mohammad A. proposed an enhanced version of the Genetic Folding algorithm to predict strokes using patient symptoms, comparing its performance with several machine learning techniques applied in this research, including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and SVM. The proposed Minimal Genetic Folding approach, developed using the Stroke Prediction dataset from Kaggle, utilizes minimal kernel operators. The model attained an accuracy of 83.2%, surpassing Logistic Regression by 4.2%, Naive Bayes by 1.2%, and Decision Tree by 17.2%, and matching the performance of the Support Vector Machine. Additionally, the model demonstrated a 7% improvement in the area under the curve, making it a more reliable predictor of stroke compared to previous methods.[30]

From a clinical perspective, the proposed Gradient Boosting framework could be integrated into electronic health record systems as a decision-support tool to flag patients at high risk of stroke, enabling timely intervention and lifestyle management. The explainable AI (SHAP) component allows clinicians to visualize key risk factors such as age, BMI, and hypertension, promoting transparent communication and informed decision-making. However, successful integration will require regulatory validation, data privacy safeguards, and interoperability with existing electronic health record standards.

Conclusion

The results of this study provide strong evidence for the effectiveness of machine learning methods in forecasting stroke risk. By effectively addressing class imbalance through advanced resampling techniques, we significantly enhanced the predictive performance of various models. The Gradient Boosting classifier proved to be the most effective approach for early stroke detection, achieving an accuracy of 92% and an AUC of 0.70. Our feature importance analysis indicated that age, general health condition, marital status, and BMI were

paramount predictors, corroborating established medical literature on stroke risk factors. Additionally, the SHAP analysis offered valuable insights into feature contributions, illustrating the influence of both physiological and lifestyle-related factors in stroke risk assessment. These findings not only contribute to the existing body of knowledge but also highlight the importance of integrating machine learning into clinical practice to improve patient outcomes. Future research should focus on refining these models and exploring additional variables to improve accuracy, while incorporating fairness analyses to ensure equitable performance across diverse populations.

Acknowledgement
None.

Conflict of Interest
All authors have no conflict of interest.

Ethic consideration
This research utilized a de-identified, publicly accessible dataset from the National Health and Nutrition Examination Survey. Because the dataset was anonymized and had already received approval from the Institutional Review Board of the National Center for Health Statistics at the U.S. Centers for Disease Control and Prevention, no additional ethical approval or informed consent from participants was required for this secondary analysis.

Funding
None

Author contribution
Minhazul Alam Mahin: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing-Original Draft, Writing-Review and Editing. **Md. Mominul Islam:** Formal Analysis, Methodology, Project Administration, Validation, Writing-Review and Editing. **MD Zulfikar Alam:** Data Curation, Investigation, Resources, Validation, Writing-Review and Editing. **Arnob Dutta Pollob:** Formal Analysis, Investigation, Software, Validation, Writing-Review and Editing. **Oxita Zaman:** Data Curation, Investigation, Project Administration, Supervision, Validation, Writing-Review and Editing.

References

1. McLaren. Stroke in 2024: By the Numbers [Internet]. 2024. https://www.mclaren.org/main/news/stroke-in-2024-by-the-numbers-4449
2. Golubnitschaja O, Potuznik P, Polivka J, Pesta M, Kaverina O, Pieper CC, et al. Ischemic stroke of unclear aetiology: a case-by-case analysis and call for a multi-professional predictive, preventive and personalised approach. EPMA Journal. 2022;13(4):535–45. DOI: 10.1007/s13167-022-00307-z.
3. Centracare. Strokes By the Numbers [Internet]. 2023. https://www.centracare.com/articles-stories/ strokes-by-the-numbers/
4. World Stroke Organization. Impact of Stroke [Internet]. 2025. https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke
5. Centers for Disease Control and Prevention. Stroke facts [Internet]. 2024. https://www.cdc.gov/stroke/data-research/facts-stats/index.html
6. NIH. How many people are affected by/at risk for stroke? [Internet]. 2016 https://www.nichd.nih.gov/health/topics/stroke/conditioninfo/risk
7. Correction to: Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. Circulation. 2023;148(4). DOI: 10.1161/cir.0000000000001167.
8. Wang Ping. Imbalanced Data-based Prediction and Risk Factor Analysis of Stroke. 2024. DOI: 10.17632/xggs239bnw.1.
9. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics. 2013;7(21). DOI: 10.3389/fnbot.2013.00021.
10. Lyashevska O, Malone F, MacCarthy E, Fiehler J, Buhk JH, Morris L. Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data. Statistical Methods in Medical Research. 2020;30(3):916–25. DOI: 10.1177/0962280220980484.
11. P. Nandal, Malik S. Leveraging AdaBoost and CatBoost to Classify the Likelihood of Brain Stroke. Journal of Scientific Research. 2024;16(3):637–46. DOI: 10.3329/jsr.v16i3.67891.
12. Hornyák O, Iantovics LB. AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics. Mathematics. 2023;11(8):1801. DOI: 10.3390/math11081801.
13. Rui C, Zhang S, Li J, Guo D, Zhang W, Wang X, et al. A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm. BMC Medical Informatics and Decision Making. 2023;23(1). DOI: 10.1186/s12911-023-02140-4.
14. Chang W, Ji X, Xiao Y, Zhang Y, Chen B, Liu H, et al. Prediction of Hypertension Outcomes Based on

Gain Sequence Forward Tabu Search Feature Selection and XGBoost. Diagnostics. 2021;11(5):792. DOI: 10.3390/diagnostics11050792.
15. Sheela Lavanya J M, Subbulakshmi P. Unveiling the potential of machine learning approaches in predicting the emergence of stroke at its onset: a predicting framework. Scientific Reports. 2024;14(1). DOI: 10.1038/s41598-024-70354-1.
16. Asadi F, Rahimi M, Daeechini AH, Paghe A. The most efficient machine learning algorithms in stroke prediction: A systematic review. Health Science Reports.2024;7(10). DOI: 10.1002/hsr2.70062.
17. 1.Yin Q, Ye X, Huang B, Qin L, Ye X, Wang J. Stroke Risk Prediction: Comparing Different Sampling Algorithms. International Journal of Advanced Computer Science and Applications. 2023;14(6). DOI: 10.14569/ijacsa.2023.01406115.
18. 1.Li X, Bian D, Yu J, Li M, Zhao D. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. BMC Medical Informatics and Decision Making. 2019;19(1). DOI: 10.1186/s12911-019-0998-2.
19. Ewald FK, Bothmann L, Wright MN, Bischl B, Casalicchio G, König G. A Guide to Feature Importance Methods for Scientific Inference. Communications in Computer and Information Science. 2024;440–64. DOI: 10.1007/978-3-031-63797-1_22.
20. Hu L, Wang K. Computing SHAP Efficiently Using Model Structure Information. arXiv (Cornell University). 2023. DOI: 10.48550/arxiv.2309.02417.
21. Alageel N, Alharbi R, Alharbi R, Alsayil M, Alharbi LA. Using Machine Learning Algorithm as a Method for Improving Stroke Prediction. International Journal of Advanced Computer Science and Applications. 2023;14(4). DOI: 10.14569/ijacsa.2023.0140481.
22. Emon MU, Keya MS, Meghla TI, Rahman MdM, Mamun MSA, Kaiser MS. Performance Analysis of Machine Learning Approaches in Stroke Prediction. IEEE Xplore. 2020. p. 1464–9. DOI: 10.1109/ICECA49313.2020.9297525
23. Hassan A, Gulzar Ahmad S, Ullah Munir E, Ali Khan I, Ramzan N. Predictive modelling and identification of key risk factors for stroke using machine learning. Scientific Reports. 2024;14(1):11498. DOI: 10.1038/s41598-024-61665-4.
24. Dritsas E, Trigka M. Stroke Risk Prediction with Machine Learning Techniques. Sensors. 2022;22(13):4670. DOI: 10.3390/s22134670.
25. Sailasya G, Kumari GLA. Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications [Internet]. 2021;12(6). DOI: 10.14569/ijacsa.2021.0120662.
26. Rahim AMA, Sunyoto A, Arief MR. Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm. MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer.2022;21(3):595–606. DOI: 10.30812/matrik.v21i3.1666.
27. Sundaram .M S, K Pavithra, V Poojasree. STROKE PREDICTION USING MACHINE LEARNING. IARJSET. 2022;9(6). DOI: 10.17148/iarjset.2022.9620.
28. Biswas N, Uddin KMM, Rikta ST, Dey SK. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics. 2022;2:100116. DOI: 10.1016/j.health.2022.100116.
29. Guhdar M, Ismail Melhum A, Luqman Ibrahim A. Optimizing Accuracy of Stroke Prediction Using Logistic Regression. Journal of Technology and Informatics. 2023;4(2):41–7. DOI: 10.37802/joti.v4i2.278.
30. Mezher MA. Genetic Folding (GF) Algorithm with Minimal Kernel Operators to Predict Stroke Patients. Applied Artificial Intelligence. 2022;36(1). DOI: 10.1080/08839514.2022.2151179