SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of XGBoost, Random Forest, and Logistic Regression Algorithms in Stroke Disease Classification

Lia Relita Sitompul [1)], Adli Abdillah Nababan [2)*], Mey Lestari Manihuruk [3)], Wildan Andika Ponsen [4)], Supriyandi [5)]
[1, 3, 4, 5)] Teknik Informatika, Universitas Prima Indonesia, Medan, Indonesia
[2)] Sistem Informasi, Universitas Prima Indonesia, Medan, Indonesia
[1)]liarelitaa21@gmail.com, [2)] adliabdillahnababan@unprimdn.ac.id, [3)] meilestarickp@gmail.com, [4)] wildanandika860@gmail.com,[5)] supriyandi445@gmail.com

Submitted :May 20, 2025 | **Accepted** : Jun 5, 2025 | **Published** : Jun 20, 2025

Abstract*:* Stroke remains one of the most critical global health challenges, ranking as the second leading cause of death and the third leading cause of disability worldwide. Early detection and precise classification of stroke risk are vital for enabling timely interventions and improving patient prognoses. This study aims to evaluate and compare the performance of three machine learning algorithms—Extreme Gradient Boosting (XGBoost), Random Forest, and Logistic Regression—for stroke classification, utilizing a dataset comprising 5,110 patient records containing 12 demographic, lifestyle, and clinical attributes. To overcome significant class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Model training and evaluation were conducted using 10-fold cross-validation, implemented in Python with the scikit-learn and imbalanced-learn libraries. Among the tested models, XGBoost demonstrated superior performance, achieving an accuracy of 95%, surpassing both Random Forest and Logistic Regression. Feature importance analysis identified age, average blood glucose level, and a history of heart disease as the most influential predictors of stroke risk. These findings underscore the potential of ensemble learning approaches, particularly XGBoost, in enhancing stroke risk prediction and highlight their prospective integration into clinical screening systems, especially within resource-constrained healthcare environments.

Keywords: Stroke, Machine Learning, XGBoost, Random Forest, Logistic Regression

INTRODUCTION

Stroke is a vascular disorder affecting the brain's blood vessels, characterized by reduced or obstructed blood flow and oxygen supply to the brain, which can result in tissue damage, neurological impairment, or even death. According to the World Health Organization (WHO), approximately 15 million people worldwide experience a stroke each year, resulting in 5 million deaths and another 5 million individuals living with permanent disabilities (Setyawan & Wakhidah, 2025). Globally, stroke ranks as the second leading cause of death and the third leading cause of disability, with approximately 70% of cases occurring in low- and middle-income countries. Moreover, about 87% of stroke-related deaths and disabilities are reported from these regions. Several risk factors contribute to the high mortality and morbidity rates associated with stroke, including gender, age, hypertension, heart disease, marital status, type of employment, residence type, average blood glucose levels, body mass index (BMI), and smoking habits (Luo et al., 2022). In recent years, machine learning (ML) techniques have gained increasing attention for their potential to analyze complex patterns within healthcare data and improve disease prediction, including stroke risk assessment. Studies have employed algorithms such as Extreme Gradient Boosting (XGBoost), Random Forest, and Logistic Regression to classify stroke risk factors and predict stroke events (Mridha et al., 2023).

These machine learning algorithms offer distinct advantages in medical classification tasks (Akmal et al., 2023). XGBoost is known for its ability to handle imbalanced datasets and generate highly accurate models through gradient boosting optimization. Random Forest improves classification accuracy by aggregating multiple decision trees while mitigating the risk of overfitting. Meanwhile, Logistic Regression, a classical statistical method, provides a simple yet effective probabilistic model for binary classification problems, offering interpretability in decision-making processes. However, despite the promising results of these algorithms in disease prediction, prior research has not comprehensively examined the trade-offs between these three classifiers specifically for stroke prediction using imbalanced medical datasets — a common issue in healthcare

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

analytics (Pratama et al., 2024). Although some studies have applied data balancing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), there remains a notable lack of systematic comparative analysis between ensemble methods (like XGBoost and Random Forest) and traditional statistical approaches (like Logistic Regression) under consistent experimental conditions . Additionally, limited research has been conducted to perform detailed feature importance analyses that reveal the most influential predictors for stroke diagnosis across different machine learning models (Fadmadika et al., 2024).

LITERATURE REVIEW

Stroke remains a major global health issue, affecting approximately 15 million individuals worldwide each year, of whom 5 million die and another 5 million suffer permanent disabilities. According to the World Health Organization (WHO), stroke ranks as the second leading cause of death and the third leading cause of disability globally, with around 70% of cases occurring in low- and middle-income countries. This alarming statistic highlights the urgent need for research aimed at identifying key risk factors associated with stroke, including age, gender, hypertension, heart disease, marital status, type of employment, blood glucose levels, body mass index (BMI), smoking status, and residential environment. In recent years, the application of machine learning (ML) algorithms has emerged as a promising approach for early detection and risk assessment of stroke. Among the widely adopted machine learning algorithms in medical classification tasks are Extreme Gradient Boosting (XGBoost), Random Forest, and Logistic Regression. XGBoost is a boosting algorithm known for its high accuracy and capability to handle imbalanced datasets effectively (Rice et al., 2024). It operates by incrementally building decision trees, where each subsequent tree is designed to correct the errors of the previous one. Despite its superior predictive performance, XGBoost can be computationally intensive and prone to overfitting if hyperparameters are not carefully optimized .

Random Forest, another popular ensemble learning method, constructs multiple decision trees using a bootstrap aggregating (bagging) technique to improve classification accuracy and reduce overfitting risks. By combining the predictions of several decision trees, Random Forest generates a more robust and reliable model compared to individual trees. However, one limitation of Random Forest is reduced model interpretability due to the complexity of aggregating numerous decision trees, which can hinder clinical adoption in settings where decision transparency is crucial f(Ruescas-Nicolau et al., 2021). In contrast, Logistic Regression represents a traditional linear classification algorithm widely used in clinical decision-making applications due to its simplicity and ease of interpretation. This model estimates the probability of a binary event based on predictor variables and is computationally efficient. Nevertheless, Logistic Regression is less effective when dealing with complex, non-linear relationships among variables, which can limit its predictive power in intricate medical datasets (Dhar et al., 2023).

Several studies have demonstrated the effectiveness of these algorithms in stroke prediction tasks. (Mochurad et al., 2025), in a study published in BMC Medical Informatics and Decision Making, integrated XGBoost with optimized Principal Component Analysis (PCA) and explainable artificial intelligence techniques, achieving accuracies of 95% and 98% on two different stroke datasets. The study emphasized the importance of combining high predictive accuracy with model interpretability to enhance clinical applicability. Similarly, (Shobayo et al., 2023) evaluated multiple classifiers and reported that Random Forest outperformed Decision Tree and Logistic Regression, achieving a macro F1 score of 94%, with age and BMI identified as the most influential predictors of stroke incidence. Despite these advancements, few comparative studies have systematically evaluated the trade-offs between ensemble methods such as XGBoost and Random Forest and traditional models like Logistic Regression within the context of imbalanced stroke datasets. Moreover, limited research has incorporated comprehensive feature importance analyses to identify the most significant predictors contributing to stroke risk across different algorithms. This gap highlights the necessity for further empirical investigation, as addressed in this study (Chakraborty et al., 2024).

METHOD

Data Collection

The dataset utilized in this study was sourced from the Kaggle platform, specifically the "Stroke Prediction Dataset," comprising 5,110 patient records with 12 attributes. These features include gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status (Chen et al., 2023). The target variable represents stroke occurrence, labeled as 0 for no stroke and 1 for stroke. Initial exploratory data analysis revealed a significant class imbalance, with only 249 stroke cases (4.9%) compared to 4,861 non-stroke cases (95.1%). Missing data was observed in the BMI attribute and was addressed by imputing the missing values with the mean of the respective column . Categorical variables, including gender, marital status (ever_married), work type, residence type, and smoking status, were transformed into numerical format using Label Encoding to facilitate model processing (Ilham et al., 2024).

Meanwhile, numerical features such as age, average glucose level, and BMI were normalized via Min-Max Scaling to bring all feature values within a uniform range of 0 to 1, thereby preventing potential biases during

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

model training. Essential data visualizations were generated using Python's matplotlib and seaborn libraries to examine distribution patterns and relationships between features (Saleem et al., 2024). Key findings from these visualizations include an increased prevalence of stroke in patients over the age of 50, a female predominance of 58.6%, and weak to moderate correlations among several features as demonstrated by a correlation heatmap. Additionally, the distribution of blood glucose levels indicated that stroke cases were more frequent among patients with elevated glucose levels exceeding normal thresholds. These insights played a vital role in guiding feature selection and enhancing model interpretability, while avoiding the inclusion of excessive graphical details that could complicate the analysis.

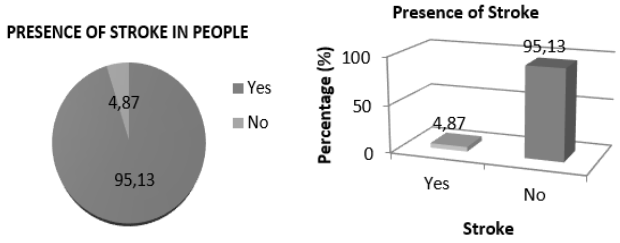

Figure 1. Distribution of Target Samples from Original Dataset

Table 1. Description of Dataset Attributes

Nama Fitur	Deskripsi
Gender	Male" , "Female" or "Other"
Age	Patient'S age in years
Hypertension	0: Not hypertensive, 1: Hypertensive
Heart_Disease	0: No heart disease, 1: Has heart disease
Ever_Married	Ever-married status (encoded)
Work_Type	Type of patient's occupation (encoded)
Residence_Type	Place of residence (encoded: 0 = Urban, 1 = Rural)
Avg_Glucose_Level	Average blood glucose level
Bmi	Body Mass Index
Smoking_Status	Patient's smoking status (encoded)
Stroke	Target: 0 = Did not have a stroke, 1 = Had a stroke

Data Prepocessing

In this study, data preprocessing was carried out to ensure the quality and suitability of the dataset for training and testing stroke prediction models. The preprocessing steps aimed to clean the data by addressing invalid and missing values, converting categorical variables into numerical form, normalizing numerical features, and preparing balanced datasets for machine learning algorithms (Rahman et al., 2023). Initially, the dataset was examined for duplicates and missing data. Missing values were identified in the BMI attribute and subsequently imputed using the mean value of the corresponding column. Categorical variables including gender, ever_married, work_type, residence_type, and smoking_status were transformed using Label Encoding, enabling the machine learning models to process these variables numerically (Agustiningsih et al., 2023). Following data cleaning, numerical features—age, average glucose level, and BMI—were normalized using the Min-Max Scaling method to rescale values between 0 and 1, ensuring balanced feature ranges and preventing bias during model training.

The dataset was then split into training and testing subsets with an 80:20 ratio, employing stratified sampling to maintain proportional class distributions across both sets. To address the significant class imbalance between stroke and non-stroke cases in the training data, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to synthetically generate minority class samples, thereby improving class balance and reducing model bias toward the majority class (Aulia et al., 2024). Subsequently, feature selection was performed to retain the most relevant attributes influencing stroke prediction, including gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. The ID attribute was excluded as it lacked predictive relevance (Suhliyyah et al., 2023).

Following preprocessing, three machine learning algorithms—Random Forest, XGBoost, and Logistic Regression—were implemented using Python's scikit-learn library within the Jupyter Notebook environment. Models were initially trained with default hyperparameters to establish baseline performance. The training data

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

had undergone SMOTE to mitigate class imbalance, which significantly influenced model outcomes. Among the models, XGBoost delivered superior performance, attributed to its gradient boosting mechanism that iteratively corrects prediction errors and its inherent ability to manage imbalanced data through class weighting (Djaya et al., 2021). Random Forest also exhibited strong results by leveraging an ensemble of decision trees to produce stable and accurate predictions. Conversely, Logistic Regression demonstrated limitations, particularly in handling non-linear relationships and the synthetic noise introduced by SMOTE, which reduced its predictive efficacy.

Model evaluation utilized multiple metrics derived from confusion matrices, including accuracy, precision, recall, and F1-score, with additional cross-validation to ensure robustness and minimize overfitting risks. Emphasis was placed not only on overall accuracy but also on the ability to correctly classify stroke cases, given the clinical importance of minimizing false negatives. Advanced analysis included feature importance assessment using the XGBoost model to identify key predictors of stroke, providing valuable clinical insights to support risk assessment and medical decision-making. Through these comprehensive preprocessing and modeling steps, this study developed an optimal and reliable stroke prediction system aimed at enhancing early stroke detection efforts based on data-driven methodologies.

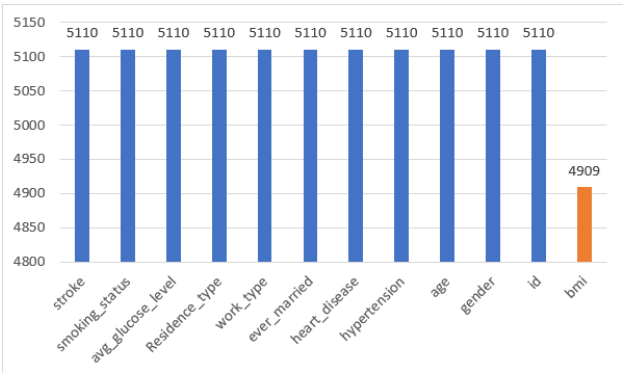

Figure 2. Null Values Visualitation

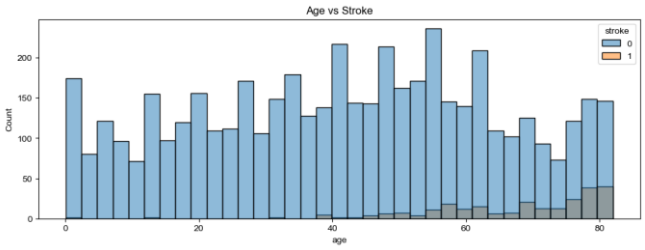

Figure 3. Patient Age Distribution with Respect to Stroke Incidence

Stroke incidence is widely distributed across various age groups, as illustrated in Figure 3, with cases documented from infancy to those aged over 80 years. The data demonstrate that the majority of stroke occurrences are concentrated in individuals aged 50 to 80 years, a pattern that reflects the increased vulnerability associated with aging. Conversely, stroke occurrence among individuals under 40 years old is relatively low compared to older age groups. This distribution highlights that advanced age is a significant risk factor for stroke, although cases in younger populations, including children, do occur. These findings align with established epidemiological evidence identifying age as a critical determinant in stroke risk.

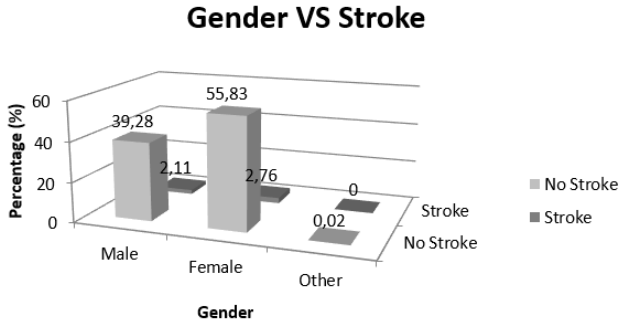

Figure 4. Proportional Gender Distribution Using a Pie Chart

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In many epidemiological studies, gender differences are evident in stroke prevalence. Within this dataset, as illustrated in Figure 4, female patients account for a larger proportion (58.60%) compared to males (41.40%), indicating a greater representation of females among recorded stroke cases. When examining stroke cases specifically, 2.76% of the total patients were female stroke cases, while male stroke cases comprised 2.11%. Among patients without stroke, females represented 55.83%, males 39.28%, and a small fraction (0.02%) identified as other genders. These results suggest that the incidence of stroke is slightly higher in females than in males within this dataset.

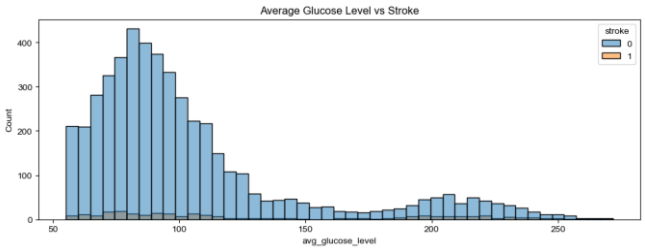

Figure 5. Distribution of Average Blood Glucose Levels in Relation to Stroke Status

The distribution of patients' average blood glucose levels based on stroke status, as presented in Figure 5, indicates that most individuals have glucose levels ranging between 70 and 130 mg/dL. This range represents the typical blood glucose profile observed in the majority of patients within the dataset.Stroke cases (marked as label 1 in gray) are more frequently observed among patients with elevated blood glucose levels, although their number remains relatively small compared to the overall non-stroke population. This pattern suggests that higher blood glucose levels may be a contributing risk factor for stroke, warranting further investigation through classification analysis to validate this association.

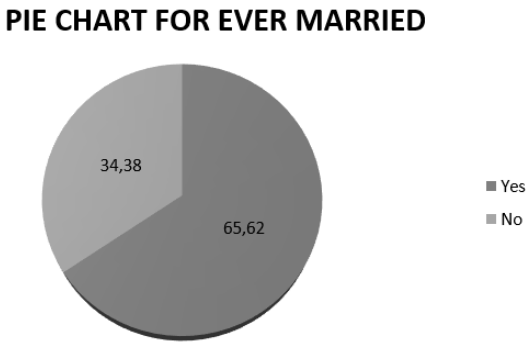

Figure 6. Pie Chart Ever Married

A review of respondent characteristics based on marital status, depicted in Figure 6, indicates that 65.62% of the participants are married, whereas 34.38% have never been married. This pattern highlights the dominance of married individuals within the dataset.

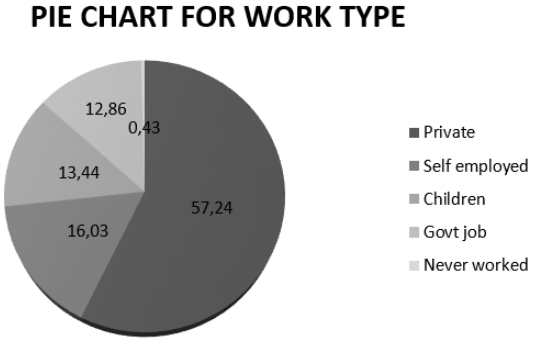

Figure 7. Pie Chart Work Type

Occupational classification among respondents reveals that most individuals, as illustrated in Figure 7, work in the private sector (57.24%). Other occupations include self-employment (16.03%), children (13.44%), government employees (12.86%), and a minor proportion of individuals who have never been employed (0.43%). This distribution highlights the predominance of private-sector employment within the dataset.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

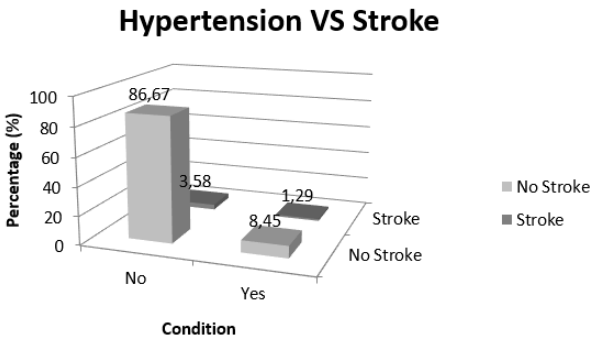

Figure 8. Hypertension vs Stroke

Blood pressure status remains a critical variable in stroke research. According to Figure 8, 86.67% of patients without a stroke history were normotensive. Surprisingly, in stroke patients, a slightly higher proportion (3.58%) had no history of hypertension than those with hypertension (1.29%), indicating that stroke can still occur in the absence of elevated blood pressure and suggesting a multifactorial etiology. While this diagram provides an overview of the distribution, additional statistical and classification analysis is required to accurately assess the influence of hypertension as a risk factor for stroke.

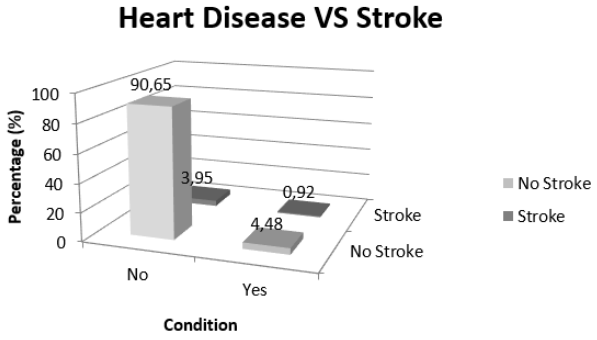

Figure 9. Heart Disease

Heart disease is commonly associated with increased stroke risk. However, the data in Figure 9 indicate that most respondents without heart disease (90.65%) were also free from stroke. Of those affected by stroke, a greater proportion (4.48%) had comorbid heart disease, although the overall prevalence of individuals experiencing both conditions was relatively low at 0.92%.

This distribution indicates that although most stroke cases occur in individuals without heart disease, the presence of heart disease appears to increase the likelihood of stroke. This finding supports existing clinical knowledge that heart disease is a contributing risk factor for stroke, warranting further investigation through comprehensive statistical analysis.

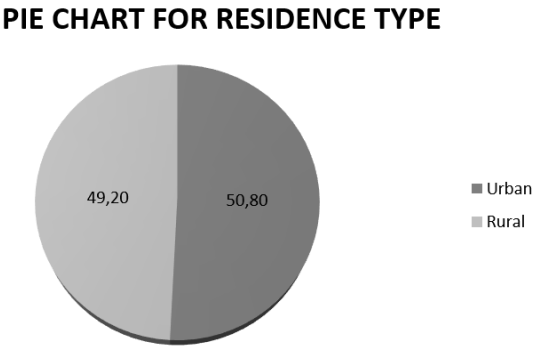

Figure 10. Residence Type

Residence type is a relevant factor in health outcomes research. According to Figure 10, the dataset demonstrates an evenly distributed population between urban (50.80%) and rural (49.20%) areas. This equilibrium

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

provides a representative basis for comparing health characteristics and risk factors across different residential environments.

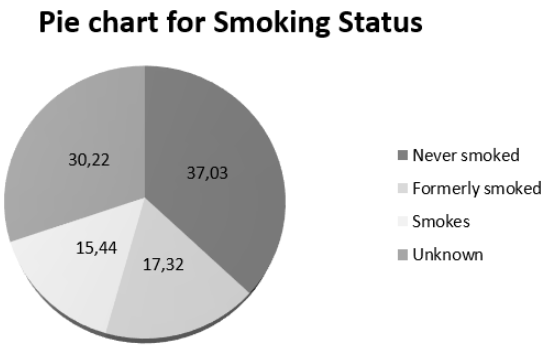

Figure 11. Smooking Status

An analysis of smoking habits within the dataset reveals that 37.03% of respondents have never engaged in smoking, as presented in Figure 11. This category represents the largest group among the total study participants. This is followed by the 'Unknown' category at 30.22%, which represents individuals whose smoking status was not recorded. Meanwhile, 'Formerly smoked' accounts for 17.32%, and the 'Smokes' category has the smallest proportion at 15.44%. This distribution highlights that the majority of individuals in the dataset either never smoked or had unrecorded smoking status, while a smaller proportion consists of active and former smokers.

The full experimental pipeline from data preprocessing to classification evaluation is summarized in Figure 12.

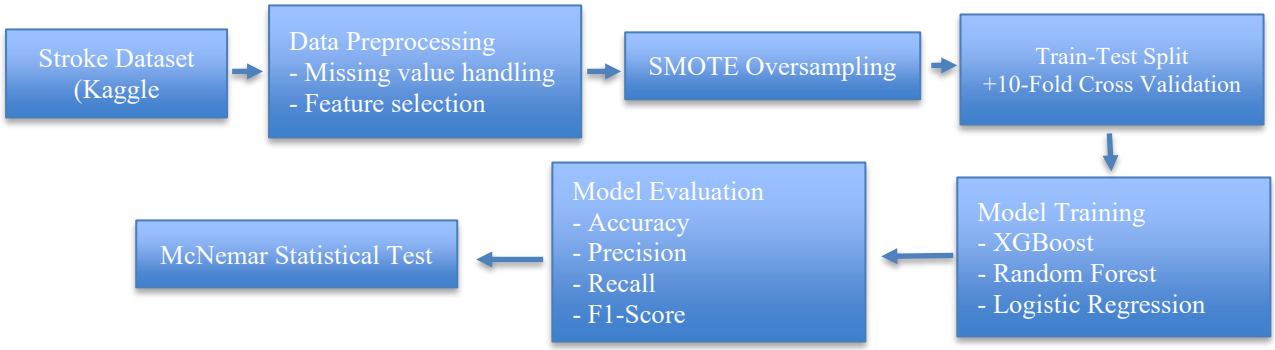

Figure 12. Workflow of the Machine Learning-Based Stroke Classification Experiment

Based on the experimental flow described above, this study proceeded to evaluate the classification performance of each algorithm using multiple quantitative metrics

RESULT

This study utilized a comprehensive set of evaluation metrics to assess the performance of three machine learning algorithms for stroke classification. The metrics included accuracy, measuring the overall correctness of predictions; precision, indicating the ability to correctly identify positive stroke cases; recall (sensitivity), reflecting the ability to capture all true positive cases; and F1-score, the harmonic mean of precision and recall, balancing both false positives and false negatives.

To complement these quantitative measures, confusion matrices were generated to visualize prediction patterns, providing insight into true positive, false positive, true negative, and false negative rates. Additionally, detailed classification reports were produced, summarizing the performance metrics for each class individually, thereby offering a granular understanding of model effectiveness across stroke and non-stroke categories.

Table 2. Most Common Machine Learning Evaluation Metrics

		Predicted Values		
		True	**False**	
Actual	**True**	True Positive (TP)	False Negative (FN) Type 1 Error	Accurracy=$\frac{TP+TN}{TP+TN+FP+FN}$

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	False	False Positive (FP) Type 1 Error	True Negative (TN)	Specificity=$\frac{TN}{TN+FP}$
		Precision$\frac{TP}{TP+TN}$		Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$ F1=$\frac{2\ X\ Precision\ x\ Recall}{Precision+Recall}$

The comparative performance evaluation of three machine learning algorithms — XGBoost, Random Forest, and Logistic Regression — is summarized in Table 3. XGBoost demonstrated the highest accuracy at 95%, followed closely by Random Forest with 94%, while Logistic Regression yielded the lowest performance with an accuracy of 82%. In terms of precision, XGBoost obtained values of 0.96 for the non-stroke class and 0.95 for the stroke class, while its recall values reached 0.95 for non-stroke and 0.96 for stroke.

Random Forest closely followed, achieving precision scores of 0.95 (non-stroke) and 0.93 (stroke), with corresponding recall values of 0.93 (non-stroke) and 0.95 (stroke). In contrast, Logistic Regression demonstrated the weakest performance, with precision values of 0.83 (non-stroke) and 0.81 (stroke), alongside recall values of 0.80 (non-stroke) and 0.83 (stroke). These results indicate that ensemble-based algorithms, particularly XGBoost, are more effective for stroke classification tasks, especially when handling imbalanced datasets.

Table 3. Accuracy, Precision, Recall, and F-1 Score Respectively a,b,c, and d

A. Precision

Model		XGBoost	Random Forest	Logistic Regression
Precision	0	0.96	0.95	0.83
	1	0.95	0.93	0.81

B. Recall

Model		XGBoost	Random Forest	Logistic Regression
Recall	0	0.95	0.93	0.80
	1	0.96	0.95	0.83

C. F1-Score

Model		XGBoost	Random Forest	Logistic Regression
F1-Score	0	0.95	0.94	0.81
	1	0.95	0.94	0.82

D. Accuracy

Model	XGBoost	Random Forest	Logistic Regression
Accuracy	0.95	0.94	0.82

Evaluation metrics for the performance of each machine learning model, including accuracy (a), precision (b), recall (c), and F1-score (d), are displayed in Table 3. The classification was performed on a binary outcome, where '0' corresponds to 'Healthy' and '1' indicates 'Stroke'.

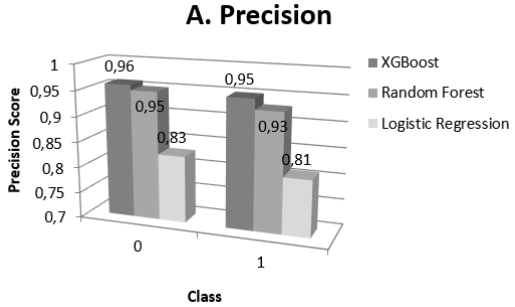

Figure 12. Precision comparison of XGBoost, Random Forest, and Logistic Regression.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The precision values achieved by each classification model, as shown in Figure 12, indicate that XGBoost obtained the highest precision for both classes, with 0.96 for class 0 (No) and 0.95 for class 1 (Yes). Random Forest also demonstrated strong performance, achieving precision values of 0.95 for class 0 and 0.93 for class 1. In contrast, Logistic Regression recorded the lowest precision, with 0.83 for class 0 and 0.81 for class 1.

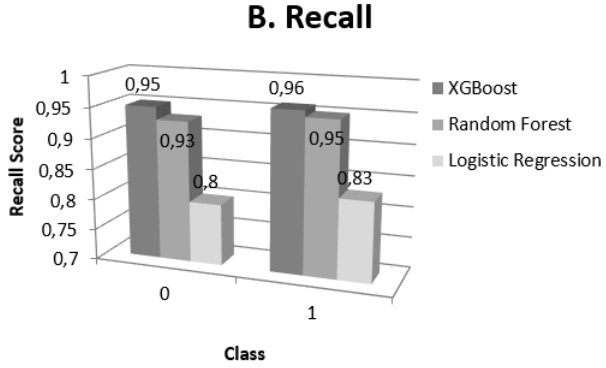

Figure 13. Recall comparison of XGBoost, Random Forest, and Logistic Regression.

Among the three models evaluated, XGBoost consistently produced the highest recall scores, obtaining 0.95 for class 0 (No) and 0.96 for class 1 (Yes), as presented in Figure 13. Random Forest followed with recall values of 0.93 and 0.95 for class 0 and class 1, respectively. Meanwhile, Logistic Regression showed lower recall performance, with 0.80 for class 0 and 0.83 for class 1.

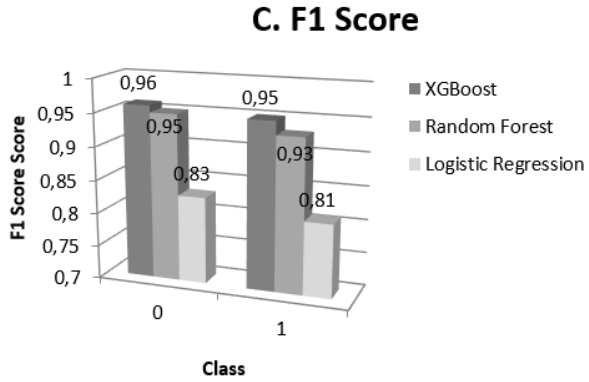

Figure 14. F1 Score comparison of XGBoost, Random Forest, and Logistic Regression.

The F1-score, which balances precision and recall, highlights the superior classification performance of XGBoost in this study. As shown in Figure 14, XGBoost achieved an F1-score of 0.95 for both class 0 (No) and class 1 (Yes). Random Forest followed with consistent scores of 0.94 for both classes, while Logistic Regression recorded the lowest F1-score performance, with 0.81 for class 0 and 0.82 for class 1.

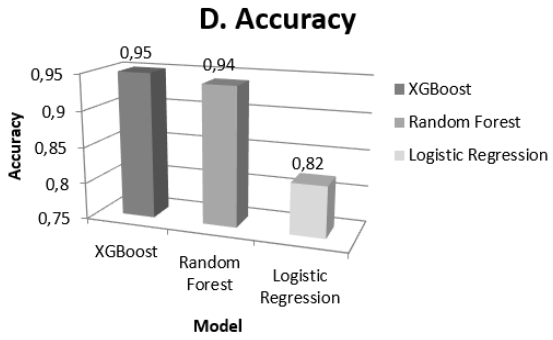

Figure 15. Accuracy comparison of XGBoost, Random Forest, and Logistic Regression.

Accuracy is a crucial indicator of a model's predictive capability. According to the results in Figure 15, XGBoost outperformed the other models by achieving an accuracy of 0.95, followed by Random Forest at 0.94. Logistic

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14794
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Regression displayed the weakest accuracy performance at 0.82. These results confirm the superior predictive ability of XGBoost in stroke classification tasks.

To rigorously assess the statistical significance of performance differences among the evaluated models, McNemar's test was conducted. The comparison between XGBoost and Random Forest yielded a p-value of 0.032 ($p < 0.05$), indicating a statistically significant difference in their classification performance. Furthermore, both XGBoost and Random Forest demonstrated statistically significant improvements over Logistic Regression, with p-values less than 0.001, thereby confirming the superior predictive capabilities of ensemble learning methods in stroke classification.

Figure 13 presents the confusion matrices for all three models. XGBoost exhibited the highest true positive rate coupled with a minimal number of false negatives, underscoring its efficacy in accurately identifying stroke cases. Random Forest displayed a similar trend, albeit with a slightly increased false negative rate. In contrast, Logistic Regression produced the greatest number of misclassifications across both positive and negative classes, particularly struggling to correctly identify positive stroke cases.

These findings not only highlight the enhanced accuracy of ensemble models but also emphasize the critical clinical value of minimizing false negatives in stroke prediction.

DISCUSSTION

The superior performance of XGBoost in this study can be attributed to several intrinsic algorithmic advantages. First, its gradient boosting framework allows iterative correction of prediction errors, enhancing overall model accuracy. Second, the presence of built-in regularization mechanisms (both L1 and L2) effectively mitigates overfitting—a common concern in medical data analysis. Third, XGBoost's capability to automatically handle missing values and model feature interactions without explicit transformation makes it highly suitable for complex healthcare datasets, where non-linear and interdependent relationships among variables frequently occur.

In contrast, Logistic Regression demonstrated the lowest predictive performance with an accuracy of 82%, primarily due to its inherent linear assumption, which restricts its ability to capture intricate interactions between stroke risk factors. Although the application of SMOTE helped address class imbalance, Logistic Regression remained biased toward the majority class and failed to accommodate the non-linear feature dependencies inherent in stroke-related datasets.

The clinical significance of these findings is noteworthy. XGBoost's low false negative rate (4%), compared to 20% in Logistic Regression, implies a substantial reduction in undetected stroke cases. This has direct implications for clinical decision-making, particularly in resource-constrained healthcare environments, where timely and accurate diagnoses are critical. Reducing missed diagnoses can facilitate earlier interventions and ultimately improve patient survival and quality of care.

These results are consistent with existing literature, which consistently reports that ensemble learning methods outperform traditional linear models in medical classification tasks, often yielding 10–15% improvements in predictive accuracy. The alignment of this study's findings with prior research reinforces the generalizability of ensemble models like XGBoost and Random Forest across diverse clinical datasets and conditions.

Furthermore, XGBoost exhibited a high sensitivity of 96%, making it an ideal candidate for deployment in clinical screening systems, where minimizing false negatives is paramount. In medical diagnostics, especially for high-risk conditions like stroke, the cost of missing a true case far outweighs the cost of a false alarm. However, this improvement in accuracy and sensitivity comes at the expense of model interpretability, a critical factor for adoption in clinical practice.

Therefore, while the technical efficacy of XGBoost is evident, future research must address the trade-off between performance and interpretability. The integration of Explainable AI (XAI) techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), is essential to enhance model transparency and gain clinician trust. Achieving a balance between predictive strength and interpretability is crucial to ensure that such advanced models can be effectively and ethically implemented in real-world clinical settings

CONCLUSION

This study has systematically evaluated and compared the perfsormance of three machine learning algorithms—XGBoost, Random Forest, and Logistic Regression—for the classification of stroke disease. The experimental results clearly indicate that XGBoost outperformed the other models with a classification accuracy of 95%, followed by Random Forest (94%) and Logistic Regression (82%). The implementation of Synthetic Minority Over-sampling Technique (SMOTE) effectively mitigated the issue of class imbalance, contributing to enhanced model reliability and robustness.

Through feature importance analysis, this research identified key predictors of stroke: age, average blood glucose level, and history of heart disease, all of which align with established medical and clinical literature. The findings highlight the superior capability of ensemble learning methods in modeling complex, nonlinear interactions

commonly found in healthcare data. These characteristics support their integration into Clinical Decision Support Systems (CDSS), which can facilitate earlier detection, more accurate diagnosis, and improved patient outcomes.

Despite the promising outcomes, several limitations must be acknowledged. The research relied solely on a single public dataset from Kaggle, limiting the generalizability of the findings. Although SMOTE was used to address class imbalance, inherent skewness in the data remained. Furthermore, the study lacked external validation across diverse populations and healthcare institutions. The absence of temporal data precluded longitudinal analysis of risk factor progression. Additionally, the dataset had a limited feature set, omitting critical lifestyle-related variables such as diet, physical activity, and family medical history.

REFERENCES

Agustiningsih, A., Findawati, Y., & Alnarus Kautsar, I. (2023). Classification of vocational high school graduates' ability in industry using extreme gradient boosting (XGBoost), random forest, and logistic regression. *Jurnal Teknik Informatika (Jutif)*, *4*(4), 977–985. https://doi.org/10.52436/1.jutif.2023.4.4.945

Akmal, K., Faqih, A., & Dikananda, F. (2023). Perbandingan metode algoritma naïve bayes dan k-nearest neighbors untuk klasifikasi penyakit stroke. *Jurnal Mahasiswa Teknik Informatika, 7*(1), 470–477. https://doi.org/10.36040/jati.v7i1.6367

Aulia, Y., Andriyansyah, A., Suharjito, S., & Nensi, S. W. (2024). Analisis prediksi stroke dengan membandingkan tiga metode klasifikasi decision tree, naïve bayes, dan random forest. *Jurnal Ilmu Komputer dan Informatika*, *3*(2), 89–98. https://doi.org/10.54082/jiki.90

Chakraborty, P., Bandyopadhyay, A., Sahu, P. P., Burman, A., Mallik, S., Alsubaie, N., Abbas, M., Alqahtani, M. S., & Soufiene, B. O. (2024). Predicting stroke occurrences: A stacked machine learning approach with feature selection and data preprocessing. *BMC Bioinformatics*, *25*(1), 329. https://doi.org/10.1186/s12859-024-05866-8

Chen, R., Zhang, S., Li, J., Guo, D., Zhang, W., Wang, X., Tian, D., Qu, Z., & Wang, X. (2023). 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*, *23*(1), 49. https://doi.org/10.1186/s12911-023-02140-4

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis—Improving explainability and trust. *IEEE Transactions on Technology and Society*, *4*(1), 68–75. https://doi.org/10.1109/tts.2023.3234203

Djaya, A. M. F. M., Sjattar, E. L., & Majid, A. (2021). Risk stratification schemes dalam mendeteksi stroke pada pasien atrial fibrillation. *Jurnal Kesehatan Masyarakat, 11*(2), 164–171. https://doi.org/10.56338/pjkm.v11i2.2143

Fadmadika, F., Handayani, H. H., Mudzakir, T. Al, & Indra, J. (2024). Pengaruh SMOTE terhadap performa algoritma random forest dan algoritma gradient boosting dalam memprediksi penyakit stroke. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *7*(2), 837–846. https://doi.org/10.37600/tekinkom.v7i2.1575

Ilham, M. A. R., Hunaifi, I., & Dirja, B. T. (2024). Effect of MLC901 on red cell distribution width (RDW) in acute ischemic stroke: Literature review. *Jurnal Biologi Tropis*, *24*(2), 431–440. https://doi.org/10.29303/jbt.v24i2.6833

Luo, J., Tang, X., Li, F., Wen, H., Wang, L., Ge, S., Tang, C., Xu, N., & Lu, L. (2022). Cigarette smoking and risk of different pathologic types of stroke: A systematic review and dose-response meta-analysis. *Frontiers in Neurology*, *12*, 1–10. https://doi.org/10.3389/fneur.2021.772373

Mochurad, L., Babii, V., Boliubash, Y., & Mochurad, Y. (2025). Improving stroke risk prediction by integrating XGBoost, optimized principal component analysis, and explainable artificial intelligence. *BMC Medical Informatics and Decision Making*, *25*(1), 63. https://doi.org/10.1186/s12911-025-02894-z

Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M. M., & Mridha, M. F. (2023). Automated stroke prediction using machine learning: An explainable and exploratory study with a web application for early intervention. *IEEE Access*, *11*, 52288–52308. https://doi.org/10.1109/ACCESS.2023.3278273

Pratama, R., Siregar, A. M., Lestari, S. A. P., & Faisal, S. (2024). Implementation of diabetes prediction model using random forest algorithm, k-nearest neighbor, and logistic regression. *Jurnal Teknik Informatika (Jutif)*, *5*(4), 1165–1174. https://doi.org/10.52436/1.jutif.2024.5.4.2593

Rahman, S., Hasan, M., & Sarkar, A. K. (2023). Prediction of brain stroke using machine learning algorithms and deep neural network techniques. *European Journal of Electrical Engineering and Computer Science*, *7*(1), 23–30. https://doi.org/10.24018/ejece.2023.7.1.483

Rice, H., de Villiers, L., Scarica, R., Bocquet, A. L., Dargan, K., & Barthe, T. (2024). Health budget implications of mechanical thrombectomy for acute ischaemic stroke in Australia. *Journal of Medical Imaging and Radiation Oncology, 68*(5), 564–569. https://doi.org/10.1111/1754-9485.13652

Ruescas-Nicolau, M. A., Sánchez-Sánchez, M. L., Cortés-Amador, S., Pérez-Alenda, S., Arnal-Gómez, A., Climent-Toledo, A., & Carrasco, J. J. (2021). Validity of the international physical activity questionnaire

e-ISSN : 2541-2019
p-ISSN : 2541-044X

long form for assessing physical activity and sedentary behavior in subjects with chronic stroke. *International Journal of Environmental Research and Public Health*, *18*(9), 1–16. https://doi.org/10.3390/ijerph18094729

Saleem, M. A., Javeed, A., Akarathanawat, W., Chutinet, A., Suwanwela, N. C., Asdornwised, W., Chaitusaney, S., Deelertpaiboon, S., Srisiri, W., Benjapolakul, W., & Kaewplung, P. (2024). Innovations in stroke identification: A machine learning-based diagnostic model using neuroimages. *IEEE Access*, *12*, 35754–35764. https://doi.org/10.1109/ACCESS.2024.3369673

Setyawan, N. H., & Wakhidah, N. (2025). Analisis perbandingan metode logistic regression, random forest, gradient boosting untuk prediksi diabetes. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *10*(1), 150–162. https://doi.org/10.29100/jipi.v10i1.5743

Shobayo, O., Zachariah, O., Odusami, M. O., & Ogunleye, B. (2023). Prediction of stroke disease with demographic and behavioural data using random forest algorithm. *Analytics*, *2*(3), 604–617. https://doi.org/10.3390/analytics2030034

Suhliyyah, Handayani, H. H., & Baihaqi, K. A. (2023). Implementasi algoritma logistic regression untuk klasifikasi penyakit stroke. *Syntax: Jurnal Informatika, 12*(1), 15–23. https://doi.org/10.35706/syji.v12i01.8329