Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 6, December 2025, Page. 5418-5434
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4760

Stroke Risk Prediction using Winsorizing Interquartile Range and Tree-Based Classification with Explainable Artificial Intelligence

Fitria Rahmadani[1], Wiharto*[2], Shaifudin Zuhdi[3]

[1,2,3]Informatics, Universitas Sebelas Maret, Indonesia

Email: [2]wiharto@staff.uns.ac.id

Received : May 21, 2025; Revised : Jun 4, 2025; Accepted : Jun 13, 2025; Published : Dec 22, 2025

Abstract

According to the Global Burden of Disease (GBD) Study, stroke is the third leading cause of death globally. Recognizing its signs early is crucial for both prevention and effective treatment. Although machine learning has made significant progress in predicting strokes, many current models operate like "black boxes", making them hard to interpret and often resulting in high error rates. This study aims to enhance prediction accuracy and interpretability in stroke risk detection by integrating Winsorizing Interquartile Range (IQR) for outlier management, a tree-based classification method, and Explainable Artificial Intelligence (XAI) techniques. The proposed approach applies Winsorizing Interquartile Range to handle extreme values while employing tree-based methods for prediction due to their superior performance in processing tabular data. Additionally, Explainable Artificial Intelligence techniques are utilized to improve model transparency and interpretability. Testing was conducted using the Cerebral Stroke Prediction-Imbalanced Dataset, comparing results with various existing models. The suggested approach demonstrated the lowest prediction error rates, achieving a False Positive Rate (FPR) of 15.74% and a False Negative Rate (FNR) of 8.56%. Additionally, it attained an accuracy of 84.39%, sensitivity of 91.43%, specificity of 84.26%, Area Under the Receiver Operating Characteristic Curve (AUROC) of 94.74%, and G-Mean of 87.76%, outperforming previous studies in stroke risk prediction. The combination of Winsorizing Interquartile Range, Random Under-Sampling, tree-based classification, and Explainable Artificial Intelligence techniques effectively enhances prediction accuracy and transparency, supporting early stroke detection with improved interpretability. This study contributes to medical informatics by integrating transparent predictive models suitable for decision support systems.

Keywords : *eXplainable Artificial Intelligence, Machine Learning, Stroke, Tree-based Method, Winsorizing IQR*

1. INTRODUCTION

Stroke occurs when blood flow to the brain is disrupted, either due to a blockage (ischemia) or bleeding in the brain's blood vessels [1]. This interruption leads to a sudden loss of brain function [2]. In 2021, the Global Burden of Disease (GBD) Study identified stroke as the third leading cause of death after ischemic heart disease and COVID-19 [3]. This statement is supported by data from the World Stroke Organization (WSO), which recorded more than 12.2 million new stroke cases in 2022, with an annual death toll reaching 3.3 million [4]. The high incidence and mortality rates of stroke highlight the importance of preventive measures, including early identification of stroke risk to reduce morbidity and mortality rates. Early identification of a person's potential risk of stroke can be achieved by recognizing the main symptoms and risk factors [5], [6]. A lack of public awareness regarding stroke symptoms and improper response to the condition can lead to delays in receiving medical treatment and increased prevalence of the disease [7], [8]. The most significant risk factors identified in stroke development include age, gender, family medical history, hypertension, smoking habits, excessive alcohol consumption, obesity, diabetes, heart disease, and lack of physical activity [9], [10], [11], [12].

Machine learning methods have been explored as a way to identify strokes early based on risk factors. Various studies have focused on improving the accuracy of stroke prediction using machine learning (ML) and deep learning (DL) techniques, as accurate predictions are essential for doctors to make well-informed clinical decisions [13]. Liu et al. [14] introduced a combination of Deep Neural Network (DNN) and Automated Hyperparameter Optimization (AutoHPO) to reduce prediction errors, achieving a FPR of 33.10% and a FNR of 19.10%. Meanwhile, Shih et al. [15] leveraged Transfer Learning (TL) alongside Deep Neural Networks (DNN) to enhance prediction accuracy, obtaining a FPR of 23% and a FNR of 14.40%. Despite their high accuracy, both models are considered "black-box" systems, meaning their decision-making processes lack transparency [16]. To address this, Kokkotis et al. [17] implemented SHapley Additive Explanations (SHAP) to provide insights into how a Multi-Layer Perceptron (MLP) model operates. However, the MLP model's overall performance remained relatively modest, with a FPR of 29.35% and a FNR of 18.60%.

While previous research has advanced stroke risk prediction using ML and DL methods [14], [15], [17], some obstacle still persist. The key concern is the relatively high prediction error rates, often caused by noise and outliers in the data [18]. Outlier detection is crucial in medical data prediction because it can reveal important information about a patient based on the physiological data provided [19]. If outliers are not properly addressed, they can disrupt the patterns learned by the model, leading to higher prediction errors [14]. However, employing effective outlier detection techniques can help mitigate these errors [20]. Another limitation of earlier studies is the use of black-box models like Deep Neural Networks and Multi-Layer Perceptron (MLP), which offer high accuracy but lack interpretability. This makes it difficult for doctors and healthcare professionals to understand the reasoning behind diagnostic outcomes [21].

This study aims to address these gaps by integrating Winsorizing Interquartile Range (IQR), tree-based models, and Explainable Artificial Intelligence (XAI) techniques to improve stroke risk prediction and its interpretability. Winsorizing IQR is capable of reducing the impact of outliers without losing valuable data [22]. Meanwhile, tree-based models are preferred because they effectively handle uninformative features and are more robust against outliers compared to deep learning models like MLP [23]. Furthermore, to enhance prediction transparency, this study will implement Explainable Artificial Intelligence (XAI) techniques, including SHAP (SHapley Additive exPlanations) [24], LIME (Local Interpretable Model-agnostic Explanations) [25], and PDP (Partial Dependence Plot) [26]. These XAI techniques will be compared to assess the consistency of their interpretations and determine the most effective and comprehensible method for explaining prediction results. This method aims to improve understanding of the risk factors associated with stroke prediction, thereby enabling healthcare professionals to make more accurate diagnoses and appropriate treatment decisions for patients.

2. METHOD

In this study, the stages carried out include data preprocessing, data sorting, data balancing on training data, model development, evaluation model, and interpretation model. The research stages are listed in Figure 1.

2.1. Data Retrieval

This study utilizes Cerebral Stroke Prediction-Imbalanced Dataset from Kaggle [27], which includes 43,400 patient records. Among these, 783 individuals have been diagnosed with stroke, making up roughly 1.8% of the dataset. Each record consists of 12 attributes, along with a target class labeled "stroke".

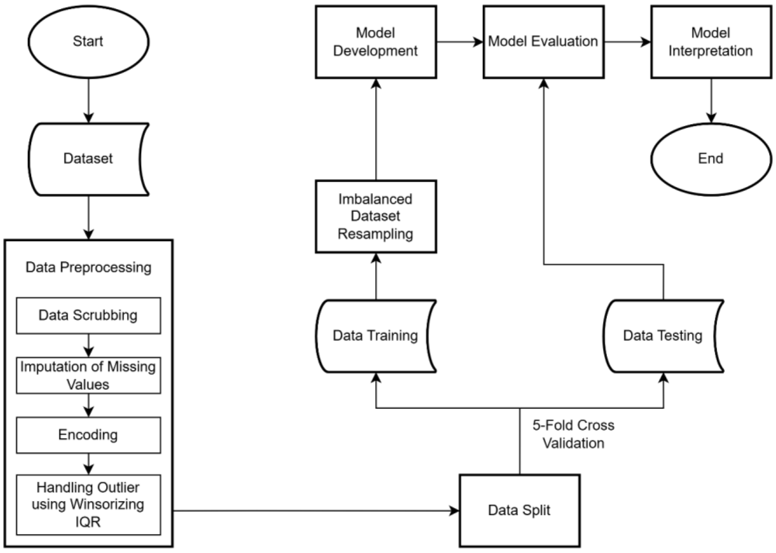


Figure 1. Research Stages

2.2. Data Preprocessing

Data preprocessing is conducted through several stages, including data cleaning, handling missing values, encoding, and outlier processing.

a. Data Cleaning: Removed irrelevant attribute, such as id. Additionally, the age attribute is converted from float to integer, and the gender attribute's "Other" category is removed to retain only Female and Male categories.
b. Handling Missing Values: Handling missing values involves imputing data for the bmi and smoking_status attributes. The bmi values are filled using the median, categorized by gender. Meanwhile, missing entries in smoking_status are replaced with "Unknown." After imputation, smoking_status consists of four categories: Unknown, formerly smoked, never smoked, and smokes.
c. Encoding: Categorical and numerical columns are separated. Since machine learning models cannot process categorical data directly, one-hot encoding is applied to categorical attributes.
d. Outlier Handling: Outliers are addressed using the Winsorizing Interquartile Range (IQR) method. This method replaces detected outlier values with the lower or upper quartile limits [22]. Visualization of the Interquartile Range (IQR) technique for outlier detection is shown in Figure 2.

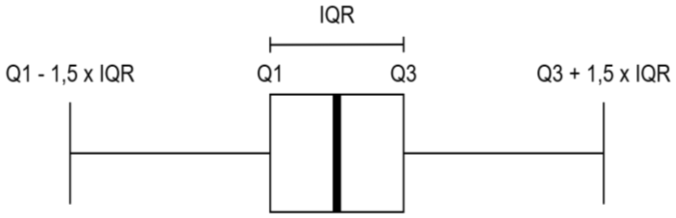


Figure 2. IQR Technique for Outlier Detection

2.3. Data Split

In this study, the dataset is split into training and testing subsets. Training data is used for model learning and construction. Testing data evaluates the model’s performance in predicting new data correctly. A 5-fold cross-validation technique is applied to ensure predictions are robust and prevent overfitting to specific classes.

2.4. Imbalanced Dataset Resampling

Imbalanced dataset resampling techniques aim to ensure a balanced dataset and minimize bias toward a particular class. Resampling methods are explained below:

a. RUS (Random Under-Sampling): Randomly decrease the number of samples from majority class [28]. A subset of these samples is chosen and combined with the minority class data to create a more balanced training dataset.

b. SMOTE (Synthetic Minority Over-sampling Technique): This method identifies K-Nearest Neighbors (KNN) for each minority class samples and creates synthetic data in proportion to the minority-majority ratio [29]. SMOTE is formulated in Equation 1 [30].

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \qquad (1)$$

x_{syn} is the new synthetic data, x_i represents an original minority class sample, x_{knn} is one of its nearest neighbors (from the same class), and δ is a randomly chosen value between 0 and 1.

c. ADASYN (Adaptive Synthetic Sampling): Generates synthetic data by considering noise, distribution, density, decision boundaries, and sample uncertainty [31].

2.5. Model Development

Model development integrates resampling techniques for imbalanced datasets with RF, GBTs, and XGBoost. Descriptions of these models are as follows:

a. Random Forest (RF): Predicts the target variable by aggregating results from multiple decision trees [32]. Predictions are based on the most frequently occurring class across different trees [33].

b. Gradient Boosting Trees (GBTs): This approach utilizes ensemble learning, where predictions are refined iteratively. By computing gradients of the loss function, the model gradually adjusts to correct errors and enhance accuracy [34] [35].

c. XGBoost: A highly effective classification model, enhancing decision tree performance by considering loss functions and regularization techniques [36].

Hyperparameter tuning is applied to determine the optimal parameter combinations for models with the lowest prediction error rates. Grid Search is used to explore the best hyperparameter configurations and is evaluated using cross-validation.

2.6. Model Evaluation

Model effectiveness is measured based on the confusion matrix to determine the level of success of a model in classifying test data. The confusion matrix consists of four main components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) is number of samples of class C_N that are correctly predicted by the model. True Negative (TN) is number of samples that are not class C_N and not predicted as C_N. False Positive (FP) is number of samples that are not class C_N, but predicted as C_N. False Negative (FN) is number of samples that should be included in C_N, but predicted as another class.

To evaluate the effectiveness of the stroke risk prediction method, this study employed various performance metrics. These include accuracy, as defined in Equation 2; sensitivity, described in Equation 3; specificity, outlined in Equation 4; the G-Mean metric, referenced in Equation 5; and the Area Under the Receiver Operating Characteristic Curve (AUROC). Accuracy represents the overall reliability of predictions, while sensitivity determines the ability to identify true positive cases. Specificity measures how effectively the model classifies negative instances. G-Mean evaluates the balance between sensitivity and specificity. AUROC evaluates the effectiveness of a classification model by measuring its performance across various threshold values [15].

$$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} \quad (2)$$

$$Sensitivity = \frac{TP(C_i)}{TP(C_i)+FN(C_i)} \quad (3)$$

$$Spesificity = \frac{TN(C_i)}{TN(C_i)+FP(C_i)} \quad (4)$$

$$GMean = \sqrt{Sensitivity\ (C_i) \times Spesificity\ (C_i)} \quad (5)$$

2.7. Model Interpretation

2.7.1. SHAP

Provides insights into feature contributions within predictions by applying game theory principles, specifically Shapley values, to measure the relative impact of each feature [37]. The Shapley value formula listed in Equation 6 [24].

$$\varphi_J(f,\boldsymbol{\tau}) = \sum_{V \subseteq M\{J\}} \frac{|V|!(|M|-|V|-1)!}{|M|!} \left[f_{V\cup\{J\}}\left(\boldsymbol{\tau}_{V\cup\{J\}}\right) - f_V(\boldsymbol{\tau}_V)\right] \quad (6)$$

$\varphi_J(f,\boldsymbol{\tau})$ represents contribution of Jth feature on the prediction outcome, f denotes prediction model, $\boldsymbol{\tau}$ denotes the actual input vector, $V \subseteq M\{J\}$ denotes the subset of features that do not contain the Jth feature, $f_V(\boldsymbol{\tau}_V)$ denotes the model prediction if only feature V is active, $\boldsymbol{\tau}_V$ denotes the input to only the subset V, and M is the total number of features.

2.7.2. LIME

LIME is a method that builds a locally interpretable model to approximate the behavior of the original model around a specific prediction point [25]. LIME method formula listed in Equation 7 [25].

$$\xi(\boldsymbol{\tau}) = argmin_{s\in S}\, L(Q,s,\pi_{\boldsymbol{\tau}}) + \vartheta(s) \quad (7)$$

$\xi(\boldsymbol{\tau})$ is a local explanation for instance $\boldsymbol{\tau}$, $s \in S$ is a local explanatory model from the set S which denotes the set of interpretable models. Q is the original (black-box) predictive model and $\pi_{\boldsymbol{\tau}}(\sigma)$ is a proximity weight between samples σ and $\boldsymbol{\tau}$ to determine how relevant the data σ is in providing a local explanation around $\boldsymbol{\tau}$. The function $L(Q,s,\pi_{\boldsymbol{\tau}})$ evaluates the difference in explanatory models according to the local behavior of the original model, while $\vartheta(s)$ measures the complexity of the explanatory model, which is used to maintain interpretability.

2.7.3. Partial Dependence Plot (PDP)

Analyzes black-box models to demonstrate how predictions are influenced by specific features [35]. This method aims to measure the influence of one or more features on model predictions, by averaging the model predictions against all other feature combinations. In a prediction model $\hat{f}(X)$ consisting of a set of input features $\boldsymbol{X}$ where $\boldsymbol{X}_{\boldsymbol{\alpha}}$ is the feature being analyzed and $\boldsymbol{X}_{\boldsymbol{\beta}}$ is the marginalized feature, the partial dependence function for feature $\boldsymbol{X}_{\boldsymbol{\alpha}}$ is formulated in Equation 8 [26].

$$PD_\alpha(\boldsymbol{X}) = \int \hat{f}\left(\boldsymbol{X}_{\boldsymbol{\alpha}}, \boldsymbol{X}_{\boldsymbol{\beta}}\right) dP\left(\boldsymbol{X}_{\boldsymbol{\beta}}\right) \quad (8)$$

3. RESULT

This study focuses on predicting stroke risk by incorporating Winsorizing Interquartile Range (IQR), tree-based models, and Explainable Artificial Intelligence (XAI) methods. The goal is to enhance both the accuracy of stroke risk prediction and the interpretability of the results. This research utilizes Python and several machine learning libraries, including pandas, matplotlib, imblearn, and sklearn.

3.1. Data Pre-processing

3.1.1. Data Scrubbing

This initial phase involved a data scrubbing which includes removed id attribute as it does not provide meaningful information regarding the patient's condition, converted age attribute from float to integer, and eliminated "Other" category in the gender attribute to retain only Female and Male categories. The results of data scrubbing are shown in Table 1.

Table 1. Data Scrubbing Results

gender	age	hypertension	heart_ disease	ever_ married	work _type	Residence _type	avg _glucose _level	bmi	smoking _status	stroke
Male	3	0	0	No	children	Rural	95.12	18.0	NaN	0
Male	58	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
					…					
Female	64	1	0	Yes	Govt_job	Rural	228.43	NaN	smokes	0
Male	14	0	0	No	children	Urban	82.48	24.8	NaN	0

3.1.2. Missing Data Management

Missing values in the bmi attribute are imputed using the median value, categorized by gender. Gender is considered to anticipate potential differences in BMI calculations across different genders. Figure 3 shows that the BMI distribution across gender categories is skewed, making the median a more stable choice for imputation compared to the mean. The median BMI values are 27.5 for females and 28.1 for males, and these values are used for imputing missing BMI values based on gender. Additionally, missing values in smoking_status are imputed with the value "Unknown".

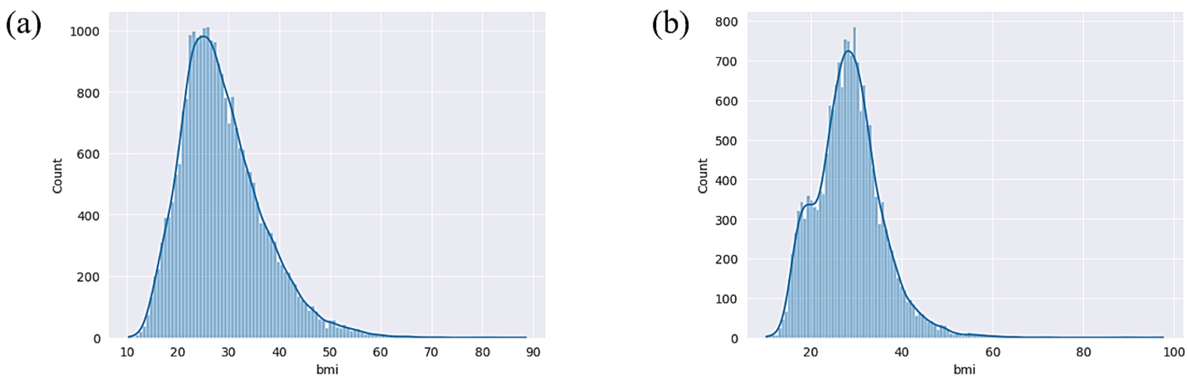


Figure 3. Distribution of BMI Attribute Data based on Class on Stroke Attribute:
(a) gender="Female"; (b) gender="Male"

3.1.3. One-Hot Encoding

Attributes like gender, ever_married, Residence_type, smoking_status, and work_type are categorical nominal variables, meaning they lack a natural order or ranking. One-hot encoding is performed to ensure the model does not mistakenly infer relationships between categories. The value 1 on the columns signifies the presence of a specific category, while 0 denotes its absence. The results of One-Hot Encoding are shown in Table 2.

Table 2. One-Hot Encoding Results

age	hypertension	heart _disease	avg_glucose _level	bmi	ever_married _No	…	smoking _status_never smoked	smoking _status _smokes	stroke
3	0	0	95.12	18.0	1	…	0	0	0
58	1	0	87.96	39.2	0		1	0	0
				…					
64	1	0	228.43	27.5	0	…	0	1	0
14	0	0	82.48	24.8	1		0	0	0

3.2. Handling Outliers using Winsorizing IQR

In this study, the Winsorizing Interquartile Range (IQR) method is used for outlier detection and processing. The first step involves checking the correlation between each attribute and the target variable (stroke) to determine which attributes require outlier treatment. The correlation of attributes to stroke is shown in Figure 4.

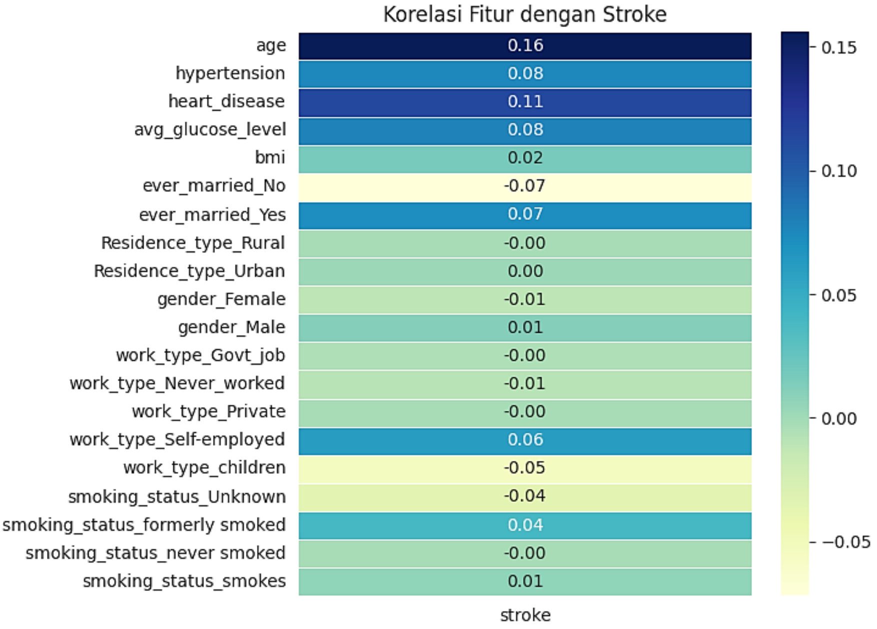


Figure 4. Features Correlation to Target Stroke Variables before Winsorizing IQR

Attributes with high correlation to stroke include age (0.16), heart_disease (0.11), hypertension (0.08), and avg_glucose_level (0.08), whereas low correlation attributes include smoking_status_formerly smoked (0.04), bmi (0.02), gender_Male (0.01), and smoking_status_smokes

(0.01). Since most low to high correlation attributes are categorical data, they do not contain extreme values or outliers. Therefore, Winsorizing IQR is only applied to numerical attributes, as they are more susceptible to outlier effects.

In this dataset, only three attributes–age, avg_glucose_level, and bmi–are continuous requiring an outlier check within stroke categories, as shown in Figure 5.

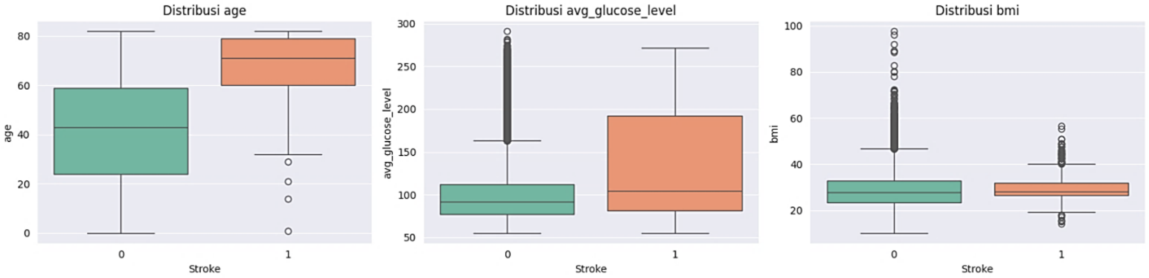


Figure 5. Data Distribution before Winsorizing IQR

Figure 5 shows that the age attribute exhibits several outliers among stroke patients, the avg_glucose_level attribute shows outliers among those without a stroke, and the bmi attribute displays outliers across both groups. Winsorizing IQR is applied to these three attributes to correct outliers. The method replaces extreme values with the upper or lower quartile limits. Following the application of Winsorizing, the data distribution becomes more compact, effectively minimizing extreme values across all three attributes, as depicted in Figure 6.

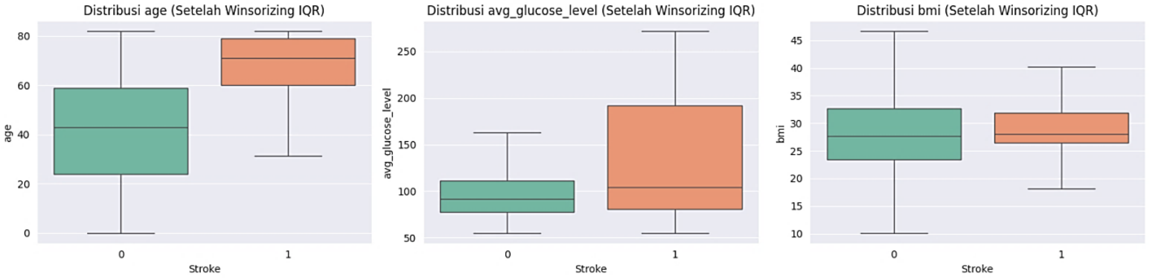


Figure 6. Data Distribution after Winsorizing IQR

In attribute age (stroke class = 1), outliers below the lower quartile are replaced with 31.5. Attribute avg_glucose_level (stroke class = 0), outliers above the upper quartile are replaced with 162.92 mg/dL. Attribute bmi (stroke class = 0), outliers above the upper quartile are replaced with 46.65. Attribute bmi (stroke class = 1), outliers below 18.15 or above 40.15 are replaced accordingly. These adjustments significantly reduce extreme values in avg_glucose_level and bmi, while age distribution remains relatively unchanged as most data falls within a reasonable range. The change in correlation after Winsorizing IQR is shown in Figure 7.

A stronger correlation was observed between stroke and avg_glucose_level increasing from 0.08 to 0.13, indicating that the previous extreme values of avg_glucose_level were obscuring the true relationship between the feature and the target. The remaining attributes showed minimal changes in correlation, indicating that outliers did not significantly impact stroke prediction.

Vol. 6, No. 6, December 2025, Page. 5418-5434
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4760

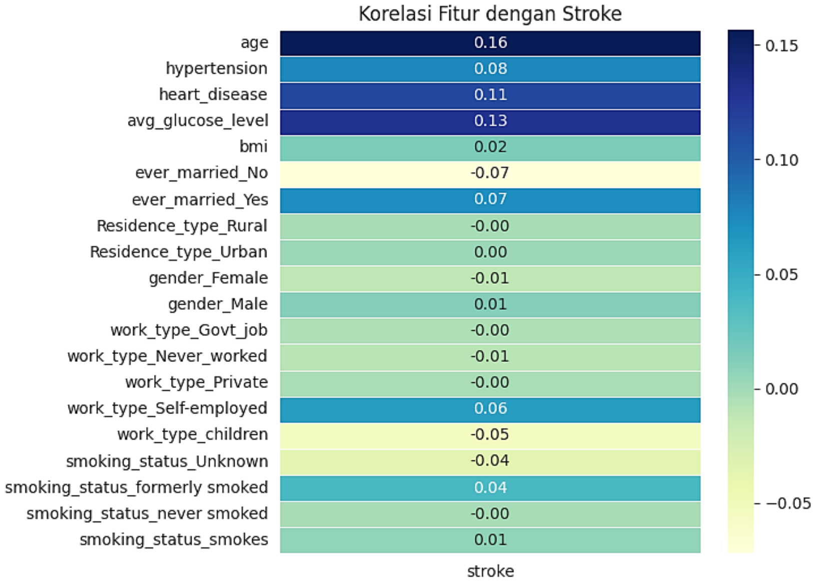


Figure 7. Features Correlation to Target Stroke Variables after Winsorizing IQR

3.3. Comparison of Resampling Methods with Tree-based Models

The comparison of imbalanced dataset resampling methods and tree-based classification models is shown in Table 3. The evaluation metric used is sensitivity (recall). Sensitivity is critical in medical applications as it measures how effectively the model detects disease cases (positives). False Negatives (misclassifying sick patients as healthy) can delay treatment, making sensitivity the most prioritized metric. High sensitivity allows the model to identify individuals at risk of stroke, enabling early medical intervention.

Table 3. Experimental Results per Class

Model	Imbalanced Dataset Resampling	Class	Sensitivity
Random Forest	RUS	0	**0.817913**
		1	**0.825061**
	SMOTE	0	0.999155
		1	0.330786
	ADASYN	0	0.999061
		1	0.238870
Gradient Boosting Trees	RUS	0	**0.825635**
		1	**0.859587**
	SMOTE	0	0.998568
		1	0.332068
	ADASYN	0	0.991856
		1	0.362739
XGBoost	RUS	0	**0.806905**
		1	**0.831488**
	SMOTE	0	0.998287
		1	0.325674
	ADASYN	0	0.997864
		1	0.325682

SMOTE and ADASYN yield imbalanced sensitivity between stroke (positive class = 1) and non-stroke (negative class = 0) predictions. False Negative rates remain high because synthetic samples from SMOTE and ADASYN tend to overlap, making stroke detection challenging. RUS achieves balanced sensitivity between stroke (1) and non-stroke (0) predictions. RUS also produces higher sensitivity for stroke cases (1) compared to SMOTE and ADASYN, making it more effective in classifying stroke risks accurately. Thus, RUS is selected as the preferred resampling technique for handling imbalanced stroke data. The sensitivity results across RF, GBTs, and XGBoost are relatively similar. To enhance performance and identify the most effective model, hyperparameter tuning is employed for optimization.

3.4. Hyperparameter Tuning

Hyperparameter tuning is performed on RUS combined with Random Forest, Gradient Boosting Trees, and XGBoost. The GridSearchCV library from scikit-learn is used to find the optimal parameter combination for better predictions. The results of hyperparameter tuning are shown in Table 4.

Table 4. Hyperparameter Tuning

Model	Optimized Parameter	Optimal Parameter	Evaluation

Random Forest	n_estimators: [100, 300, 500]
	max_depth: [3, 5, 7]
	min_samples_split: [2, 10, 20]
	min_samples_leaf: [1, 5, 10]
	max_features: ['sqrt', 'log2']
		n_estimators = 500
		max_depth = 7
		min_samples_split = 20
		min_samples_leaf = 1
		max_features = 'sqrt'
			Accuracy = 0.8071
			Sensitivity = 0.8467
			Specificity = 0.8064
			AUROC = 0.9086
			G-Mean = 0.8261
			FPR = 0.1936
			FNR = 0.1533

Gradient Boosting Trees	n_estimators: [100, 300, 500]
	learning_rate: [0.001, 0.01, 0.1]
	max_depth: [3, 5, 7]
	subsample: [0.7, 0.8, 1.0]
	min_samples_split: [2, 10, 20]
		n_estimators = 300
		learning_rate = 0.1
		max_depth = 3
		subsample = 0.7
		min_samples_split = 10
			Accuracy = 0.8439
			Sensitivity = 0.9144
			Specificity = 0.8426
			AUROC = 0.9474
			G-Mean = 0.8776
			FPR = 0.1574
			FNR = 0.0856

XGBoost	n_estimator: [100, 300, 500]
	learning_rate: [0.001, 0.01, 0.1]
	max_depth: [3, 5, 7]
	subsample: [0.7, 0.8, 1.0]
	colsample_bytree: [0.7, 0.8, 1.0]
	min_child_weight: [1, 5, 10]
		n_estimator = 100
		learning_rate = 0.1
		max_depth = 3
		subsample = 0.7
		colsample_bytree = 0.8
		min_child_weight = 1
			Accuracy = 0.8244
			Sensitivity = 0.8479
			Specificity = 0.8240
			AUROC = 0.9203
			G-Mean = 0.8357
			FPR = 0.1760
			FNR = 0.1521

Table 4 shows that Gradient Boosting Trees outperforms Random Forest and XGBoost across all metrics. Additionally, combination RUS and Gradient Boosting Trees shows improved sensitivity after hyperparameter tuning, confirming that optimization enhances classification model performance and reduces prediction errors. Hyperparameter tuning helps determine the best model parameters, providing valuable insights for evaluating the overall model improvement.

3.5. Model Interpretation

3.5.1. Feature Importance

Figure 8 illustrates the feature importance results from the Gradient Boosting Trees model. The three attributes with the highest contribution to predictions are age, avg_glucose_level, and bmi. Age and avg_glucose_level contribute 0.381902 and 0.378323, respectively, indicating they are the most important attributes in stroke predictions. BMI contributes 0.144196, which, while lower than age and avg_glucose_level, still supports the overall predictions. Other attributes contribute less than 0.05, meaning their impact on prediction is minimal.

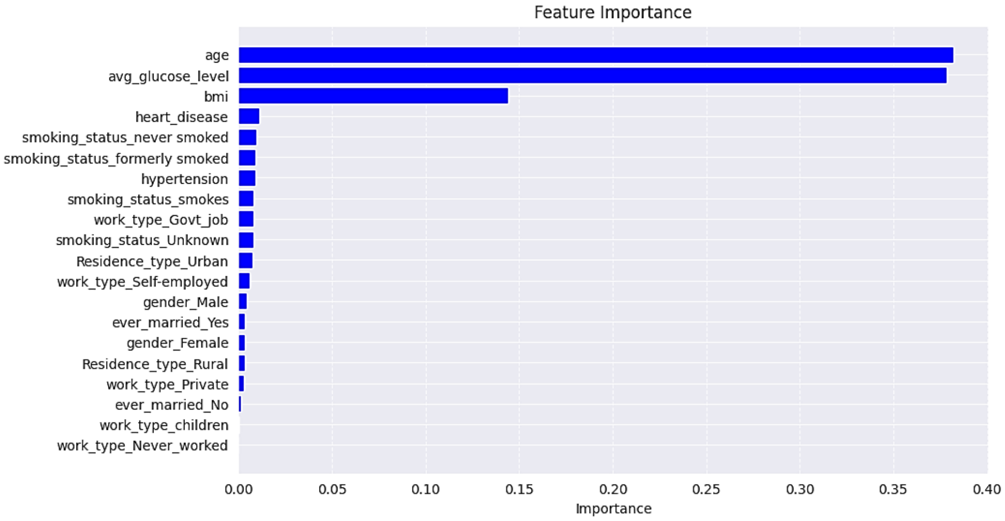


Figure 8. Feature Importance

3.5.2. SHAP (Shapley Additive Explanations)

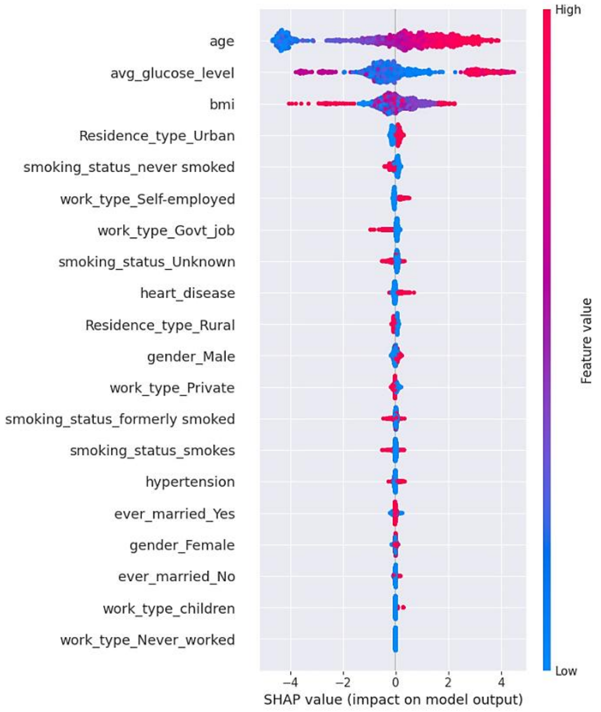


Figure 9. Global Interpretation of SHAP

Figure 9 highlights that the key factors influencing stroke predictions are age, average glucose level, and BMI. Older individuals have a higher likelihood of stroke, while younger individuals have a lower stroke risk. Higher glucose levels increase stroke vulnerability compared to lower glucose levels. Higher BMI contributes to stroke risk, though not as significantly as age and glucose levels.

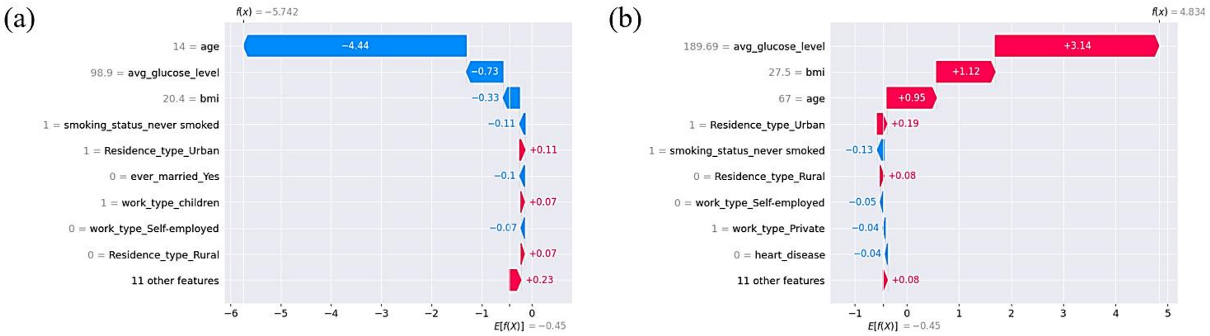


Figure 10. Local Interpretation of SHAP: (a) stroke (class = 0) ; (b) stroke (class = 1)

Figure 10 (a) reveals that young age significantly reduces stroke probability on this sample. Normal glucose levels and ideal BMI further lower stroke risk. Additional factors such as non-smoking status, unmarried status, and non-self-employment have a mi nor impact on reducing stroke risk. Figure 10 (b) highlights that high glucose levels are the strongest contributors to stroke probability on this sample. Obesity (high BMI) and old age further increase stroke risk. Urban residence has a minor impact on stroke risk.

3.5.3. LIME (Local Interpretable Model-agnostic Explanations)

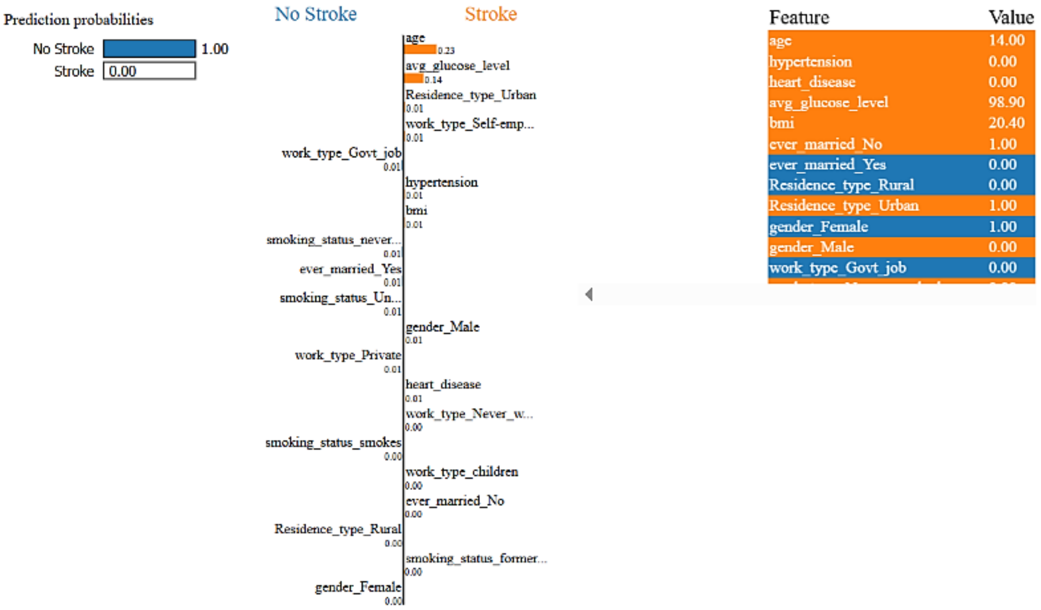
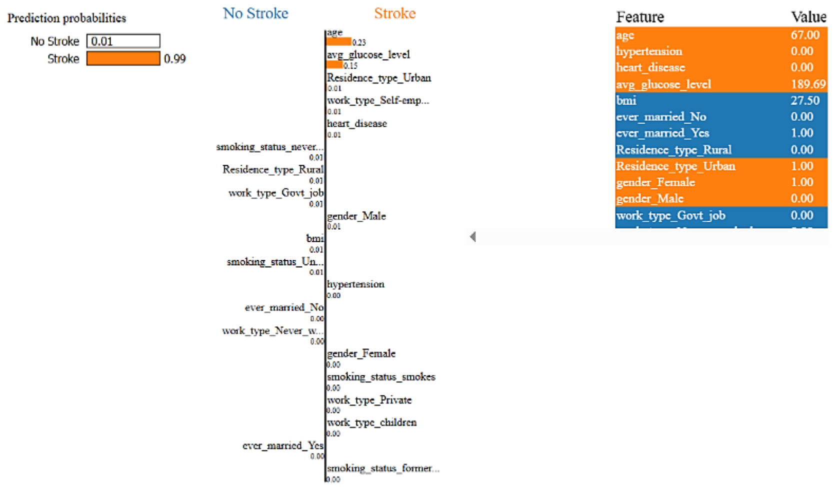


Figure 11. Local Interpretation using LIME on Sample Stroke (class = 0)

Figure 11 shows a model prediction of 100% probability of no stroke for an individual, determined by key attributes such as age, avg_glucose_level, and health history. Specifically, the person's young age (14 years) is a major factor in lowering stroke risk, while their normal blood glucose level (98.9 mg/dL) further supports a low probability of stroke. Additionally, the absence of hypertension and heart disease contributes to the prediction, albeit with a minor impact. Other attributes, such as residence type, work type, smoking status, marital status, and gender, have very little to no influence on the prediction outcome.

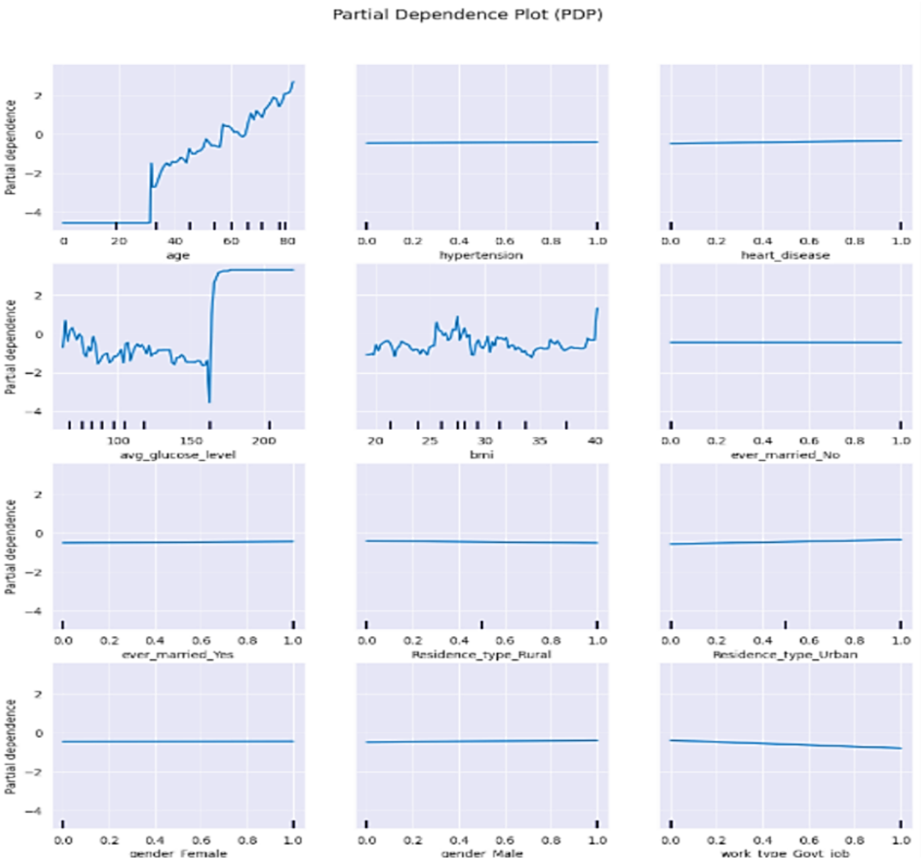


Feature	Value
age	67.00
hypertension	0.00
heart_disease	0.00
avg_glucose_level	189.69
bmi	27.50
ever_married_No	0.00
ever_married_Yes	1.00
Residence_type_Rural	0.00
Residence_type_Urban	1.00
gender_Female	1.00
gender_Male	0.00
work_type_Govt_job	0.00

Figure 12. Local Interpretation using LIME on Sample Stroke (class = 1)

Figure 12 shows a model prediction of 99% probability of stroke for another individual, primarily influenced by advanced age and elevated glucose levels. In this case, the age of 67 years significantly increases stroke risk, while a high avg_glucose_level (189.69 mg/dL) further intensifies the likelihood of stroke. Additionally, factors like urban residence and self-employment contribute positively to the probability of stroke, although their impact is relatively minor compared to age and glucose levels.

3.5.4. PDP (Partial Dependence Plot)



Vol. 6, No. 6, December 2025, Page. 5418-5434
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4760

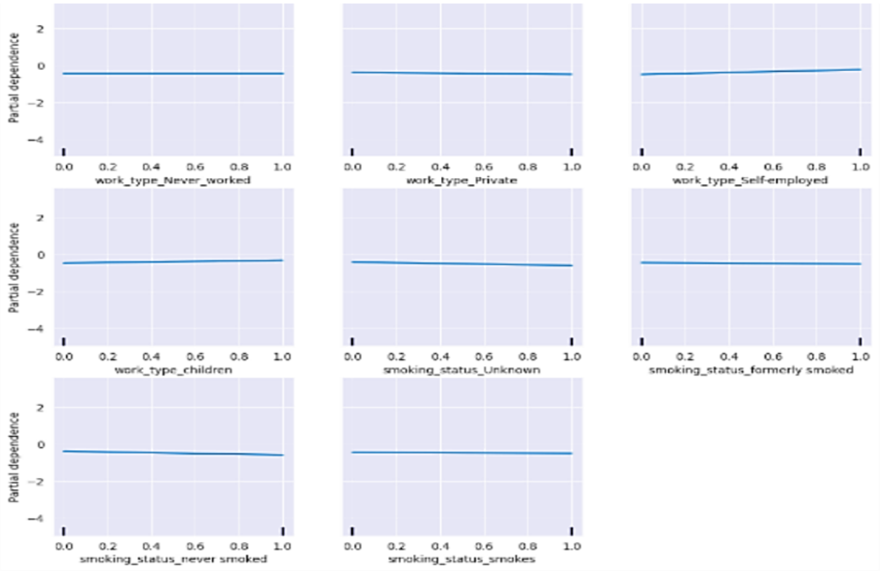


Figure 13. Global Interpretation of PDP

Figure 13 identifies three attributes with strong influences on stroke predictions are age, avg_glucose_level, and bmi. Stroke risk sharply increases after age 40 and is extremely high after age 60, confirming age as a primary risk factor. Partial dependence graphs show a drastic rise in stroke probability for glucose levels exceeding 150 mg/dL, reinforcing its significant impact on stroke risk. BMI effects are more fluctuating, with higher BMI values (>35) showing mild increases in stroke risk. However, BMI's influence is not as strong as age and glucose levels.

At the local interpretation level, differences arise between SHAP and LIME, particularly in ranking feature importance. However, both methods agree that age and avg_glucose_level contribute the most to stroke predictions. Thus, feature importance, SHAP, LIME, and PDP consistently identify age, avg_glucose_level, and BMI as the key predictive attributes for stroke risk.

4. DISCUSSIONS

The effectiveness of the proposed approach—combining Winsorizing IQR, RUS, and Gradient Boosting Trees (GBTs)—was evaluated against previous studies, including AutoHPO with DNN [14], RUS with Multi-Layer Perceptron (MLP) [17], ADASYN with Neural Network - Random Forest (NN-RF) [38], and RUS with TL-DNN [15]. These approaches were evaluated using the Cerebral Stroke Prediction-Imbalanced Dataset, with their performance metrics summarized in Table 5.

Table 5. Metric Comparison with Previous Research.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUROC (%)	G-Mean (%)	FPR (%)	FNR (%)
AutoHPO with DNN [14]	71.6	67.4	32.60	-	46.9	33.10	19.1
RUS with MLP [17]	71	81.4	70.65	82.14	75.83	29.35	18.6
ADASYN with NN-RF [38]	84	-	-	86	-	-	-
RUS with TL-DNN [15]	80.5	85.6	77	80	81.20	23	14.4
Winsorizing IQR with RUS and GBTs (Proposed)	**84.39**	**91.4**	**84.26**	**94.74**	**87.76**	**15.74**	**8.56**

Integrating advanced techniques such as outlier handling, data balancing, an optimized classification process, and targeted hyperparameter tuning, this study presents a robust approach for predicting stroke risk that markedly outperforms previous models. The proposed method achieved an accuracy of 84.39%, a sensitivity of 91.44% (ensuring most true stroke cases are correctly identified), a specificity of 84.26% (accurately classifying non-stroke cases), an AUROC of 94.74% (demonstrating excellent discriminative ability across thresholds), and a G-Mean of 87.76%. Additionally, it recorded a false positive rate (FPR) of 15.74% and a false negative rate (FNR) of 8.56%. Notably, compared to the RUS-TL-DNN model developed by Shih et al. [15], the approach reduced the FPR by 7.24% and the FNR by 5.84%, underscoring its enhanced reliability in clinical contexts.

Moreover, the research leverages explainable AI (XAI) techniques to provide deeper insights into the model's predictions by highlighting critical factors such as age, average glucose level, and BMI. These key predictive factors not only validate the findings of previous studies like those by Kokkotis et al. [17] but also underscore their central role in stroke risk assessment. By integrating transparent and interpretable predictive models tailored for decision support systems, this research makes a significant contribution to both medical informatics and interpretable machine learning, ultimately fostering more trustworthy and informed decision-making in clinical practice.

5. CONCLUSION

This study proposes the Winsorizing IQR technique combined with Random Undersampling (RUS) for data balancing and Gradient Boosting Trees for classification. The combination of these methods achieved the highest evaluation metrics compared to other approaches, with the results achieved 84.39% accuracy, 91.44% sensitivity, 84.26% specificity, 94.74% AUROC, 87.76% G-Mean, 15.74% FPR, and 8.56% FNR. The findings indicate that applying Winsorizing IQR in data preprocessing enhances data quality by reducing outlier influence, positively impacting classification model performance. Additionally, the combination of RUS and Gradient Boosting Trees proved more effective in classifying imbalanced stroke data compared to SMOTE or ADASYN. Moreover, the use of eXplainable Artificial Intelligence (XAI) techniques—such as SHAP, LIME, and PDP—offers valuable insights into the global risk factors affecting stroke diagnosis. These methods highlight age, blood glucose level, and BMI as the most significant predictive attributes. This research makes a significant contribution to the field of medical informatics and interpretable machine learning by integrating transparent predictive models suitable for decision support systems. The study enhances our understanding of how to develop reliable and comprehensible tools in medical informatics. The integration of robust data preprocessing methods with cutting-edge classification algorithms not only improves predictive performance but also fosters trust among healthcare professionals by elucidating the underlying decision-making process. Looking forward, future studies should consider to explore alternative imbalanced dataset resampling techniques that can achieve a more optimal balance—such as a combination of undersampling and oversampling. This approach aims to reduce the bias introduced by the majority class while preserving as much valuable information as possible.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] I. Surakka *et al.*, “Multi-ancestry meta-analysis identifies 5 novel loci for ischemic stroke and reveals heterogeneity of effects between sexes and ancestries,” *Cell Genomics*, vol. 3, no. 8, Aug. 2023, doi: 10.1016/j.xgen.2023.100345.
[2] B. Hum *et al.*, “Unveiling the evolving landscape of stroke care costs: A time-driven analysis,”

Journal of Stroke and Cerebrovascular Diseases, vol. 33, no. 6, p. 107663, 2024, doi: 10.1016/j.jstrokecerebrovasdis.2024.107663.

[3] V. L. Feigin *et al.*, “Global , regional , and national burden of stroke and its risk factors , 1990 – 2021 : a systematic analysis for the Global Burden of Disease Study 2021,” vol. 23, no. 10, pp. 973–1003, 2024, doi: 10.1016/s1474-4422(24)00369-7.

[4] V. L. Feigin *et al.*, “World Stroke Organization (WSO): Global Stroke Fact Sheet 2022,” *International Journal of Stroke*, vol. 17, no. 1, pp. 18–29, 2022, doi: 10.1177/17474930211065917.

[5] T. Vu *et al.*, “Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study,” *Journal of Cardiovascular Development and Disease*, vol. 11, no. 7, 2024, doi: 10.3390/jcdd11070207.

[6] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, “Stroke Risk Factors, Genetics, and Prevention,” *Physiology & behavior*, vol. 176, no. 1, pp. 100–106, 2017, doi: 10.1177/0022146515594631.Marriage.

[7] M. L. De Mélo Silva Júnior, N. C. D. S. Menezes, and M. V. D. S. Vilanova, “Recognition, reaction, risk factors and adequate knowledge of stroke: A Brazilian populational survey,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 32, no. 8, p. 107228, 2023, doi: 10.1016/j.jstrokecerebrovasdis.2023.107228.

[8] J. Attakorah, K. B. Mensah, P. Yamoah, V. Bangalee, and F. Oosthuizen, “Awareness of stroke, its signs, and risk factors: A cross-sectional population-based survey in Ghana,” *Health Science Reports*, vol. 7, no. 6, 2024, doi: 10.1002/hsr2.2179.

[9] T. Zuo, F. Li, X. Zhang, F. Hu, L. Huang, and W. Jia, “Stroke classification based on deep reinforcement learning over stroke screening imbalanced data,” *Computers and Electrical Engineering*, vol. 114, Mar. 2024, doi: 10.1016/j.compeleceng.2023.109069.

[10] C. Fernandez-Lozano *et al.*, “Random forest-based prediction of stroke outcome,” *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-89434-7.

[11] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/7633381.

[12] X. Zhu *et al.*, “Effect of the number of unhealthy lifestyles in middle-aged and elderly people on hypertension and the first occurrence of ischemic stroke after the disease,” *Frontiers in Cardiovascular Medicine*, vol. 10, no. May, pp. 1–9, 2023, doi: 10.3389/fcvm.2023.1152423.

[13] S. Yalçın and H. Vural, “Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks,” *Computers in Biology and Medicine*, vol. 149, Oct. 2022, doi: 10.1016/j.compbiomed.2022.105941.

[14] T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” *Artificial Intelligence in Medicine*, vol. 101, no. August, p. 101723, 2019, doi: 10.1016/j.artmed.2019.101723.

[15] D. H. Shih, Y. H. Wu, T. W. Wu, H. Y. Chu, and M. H. Shih, “Stroke Prediction Using Deep Learning and Transfer Learning Approaches,” *IEEE Access*, vol. 12, no. June, pp. 130091–130104, 2024, doi: 10.1109/ACCESS.2024.3429157.

[16] S. Ali *et al.*, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[17] C. Kokkotis *et al.*, “An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data,” *Diagnostics*, pp. 153–157, 2022, doi: 10.3390/diagnostics12102392.

[18] J. W. Osborne and A. Overbay, “The power of outliers (and why researchers should ALWAYS check for them),” *Practical Assessment, Research and Evaluation*, vol. 9, no. 6, 2004, doi: 10.7275/qf69-7k43.

[19] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, and S. Silvia, “A survey on outlier explanations,” *The VLDB Journal*, vol. 31, no. 5, pp. 977–1008, 2022, doi: 10.1007/s00778-021-00721-1.

[20] H. P. Vinutha, B. Poornima, and B. M. Sagar, “Detection of outliers using interquartile range technique from intrusion dataset,” *Advances in Intelligent Systems and Computing*, vol. 701, pp. 511–518, 2018, doi: 10.1007/978-981-10-7563-6_53.

[21] M. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, “Explainable AI approaches in deep learning: Advancements, applications and challenges,” *Computers and Electrical Engineering*, vol. 117, Jul. 2024, doi: 10.1016/j.compeleceng.2024.109246.

[22] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, “An outliers detection and elimination framework in classification task of data mining,” *Decision Analytics Journal*, vol. 6, no. May 2022, 2023, doi: 10.1016/j.dajour.2023.100164.

[23] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?,” *Advances in Neural Information Processing Systems*, vol. 35, no. NeurIPS, pp. 507–520, 2022, doi: 10.48550/arXiv.2207.08815.

[24] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*, 2017. doi: 10.48550/arXiv.1705.07874.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier,” *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[26] G. Szepannek and K. Lübke, “How much do we see? On the explainability of partial dependence plots for credit risk scoring,” *Argumenta Oeconomica*, vol. 2023, no. 1, pp. 137–150, 2023, doi: 10.15611/aoe.2023.1.07.

[27] T. Liu, W. Fan, and C. Wu, “Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets,” *Mendeley Data*, 2019, doi: 10.17632/x8ygrw87jw.1.

[28] Z. Sun, W. Ying, W. Zhang, and S. Gong, “Undersampling method based on minority class density for imbalanced data,” *Expert Systems with Applications*, vol. 249, no. February, 2024, doi: 10.1016/j.eswa.2024.123328.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, pp. 321–357, 2002, doi: 10.1613/jair.953.

[30] H. Sain and S. W. Purnami, “Combine Sampling Support Vector Machine for Imbalanced Data Classification,” in *Procedia Computer Science*, Elsevier, 2015, pp. 59–66. doi: 10.1016/j.procs.2015.12.105.

[31] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[32] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[33] F. R. Aszhari, Z. Rustam, F. Subroto, and A. S. Semendawai, “Classification of thalassemia data using random forest algorithm,” in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1490/1/012050.

[34] K. Omari, “Phishing Detection using Gradient Boosting Classifier,” in *3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023) Phishing*, 2023, pp. 120–127. doi: 10.1016/j.procs.2023.12.067.

[35] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[36] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[37] M. T. Hosain, J. R. Jim, M. F. Mridha, and M. M. Kabir, “Explainable AI approaches in deep learning: Advancements, applications and challenges,” *Computers and Electrical Engineering*, vol. 117, p. 109246, 2024, doi: 10.1016/j.compeleceng.2024.109246.

[38] V. S. Elangovan, R. Devarajan, O. I. Khalaf, M. S. Sharif, and W. Elmedany, “Analyzing an Imbalanced Stroke Prediction Dataset Using Machine Learning Techniques,” *Karbala International Journal of Modern Science*, vol. 10, no. 2, pp. 246–259, 2024, doi: 10.33640/2405-609X.3355.