

Performance Evaluation of Stroke Prediction Model on Healthcare Dataset with Imbalanced Data

Harprith Kaur Rajinder Singh¹, Afgha Jameel Aufan^{*2}

¹ Faculty of Data Science and Information Technology, INTI Interational University, Malaysia

²Department of Information Technology, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia

e-mail: harprith.randhawa@newinti.edu.my¹, afghajr12@gmail.com²

Article History	Abstract
Received : 06 January 2026 Revised : 22 February 2026 Accepted : 02 March 2026 Published : 28 March 2026 Keywords: <i>stroke, XGBoost, random forest, SVM, medical prediction</i>	This study is motivated by the growing need for machine learning-based prediction models for identifying stroke risk in medical datasets with severe class imbalance. This research compares three machine learning algorithms, namely Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), for predicting stroke occurrence using the Healthcare Stroke Dataset. Data preprocessing included missing value imputation, numerical feature standardization, and categorical feature encoding. The dataset was divided into training and testing sets using a stratified split. To address class imbalance, class weighting and Synthetic Minority Over-sampling Technique (SMOTE) were applied. Model performance was evaluated using macro precision, macro recall, and macro F1-score, which are appropriate for imbalanced medical data. XGBoost with class weighting achieved the best performance, yielding the highest macro F1-score of 0.59 while maintaining balanced stroke detection. Random Forest achieved high accuracy but showed limitations in detecting stroke cases, whereas SVM exhibited higher sensitivity but produced more false positives. These findings indicate that XGBoost with class weighting is the most suitable model and can serve as a baseline for developing clinical decision-support systems for early stroke risk detection. Future work includes hyperparameter optimization, additional clinical features, and larger datasets to improve model generalization.

^{*)}Corresponding author : afghajr12@gmail.com

1. INTRODUCTION

Stroke is a non-communicable disease and a leading cause of death and long-term disability worldwide. This condition occurs due to disrupted blood flow to the brain, causing permanent brain tissue damage if not treated promptly. Epidemiological data indicate that stroke incidence continues to rise along with increasing life expectancy, lifestyle changes, and growing prevalence of risk factors such as hypertension, diabetes, obesity, and other metabolic disorders. The impact of stroke extends beyond individual sufferers, imposing significant economic and social burdens on families and healthcare systems overall [1].

Prevention and early detection efforts are crucial considering that most stroke risk factors can be identified before clinical events occur. In conventional medical practice, stroke risk assessment typically relies on clinical experience and simple statistical analysis, which often inadequately capture complex relationships between clinical and demographic variables. With the increasing availability of large-scale health data, machine learning approaches have been widely adopted to support more accurate and objective disease risk prediction. Various machine learning algorithms have been employed in stroke prediction research, including Random Forest, XGBoost, and Support Vector Machine (SVM). Random Forest, an ensemble method based on bagging, effectively handles non-linear data and noise, making it suitable for highly variable clinical data. XGBoost, a boosting algorithm, offers advantages in iterative learning and regularization, proving effective for complex tabular data. Meanwhile, SVM, a

margin-based algorithm, is commonly used for medium-sized datasets due to its ability to establish optimal classification boundaries [2], [3].

Research by Nabila and Pamungkas demonstrated that Random Forest and XGBoost tend to yield more stable performance than SVM in stroke risk classification, particularly when datasets exhibit non-linear and multivariate characteristics [2]. Similar findings by Almuzadid and Subhiyakto indicated that ensemble methods produce more consistent results in stroke prediction compared to single models, especially for complex health datasets [3]. However, both studies emphasized that model performance is heavily influenced by class distribution in the dataset. A major challenge in machine learning-based stroke prediction is class imbalance, where the number of non-stroke patients substantially exceeds stroke patients. This condition causes models to bias toward the majority class, producing deceptively high accuracy while maintaining low stroke detection capability. Research by Melnykova confirmed that for imbalanced stroke datasets, evaluation metrics such as precision, recall, and F1-score, particularly macro F1-score, are far more representative than accuracy as they fairly assess model performance on minority classes [3].

Beyond class imbalance, preprocessing quality significantly determines prediction model success. Noor stated that preprocessing stages including missing value handling, categorical feature encoding, and numerical feature normalization are crucial factors that can enhance or diminish machine learning model performance in stroke prediction [4]. Hassan further noted that prediction errors for stroke cases carry far more serious clinical implications than errors for non-stroke cases, thus model evaluation should emphasize the ability to detect high-risk classes [5]. Although numerous studies have examined machine learning applications in stroke prediction, a research gap remains regarding the lack of consistent comparative analysis of Random Forest, XGBoost, and SVM within a unified experimental framework, particularly incorporating class imbalance handling and macro F1-score-based evaluation as primary metrics. Variations in preprocessing approaches, imbalance handling strategies, and evaluation metrics across previous studies render direct result comparisons difficult [6].

Based on this background, the present study aims to compare the performance of Random Forest, XGBoost, and SVM algorithms in predicting stroke occurrence using the Healthcare Stroke Dataset, implementing class imbalance handling and macro F1-score-based evaluation. This research seeks to analyze each algorithm's performance in detecting stroke risk, evaluate the influence of class imbalance handling on minority class recognition capability, and determine the most effective and stable algorithm for health data applications [7]. The findings are expected to provide theoretical contributions to machine learning research in healthcare and offer practical benefits through stroke prediction model recommendations that can serve as foundations for clinical decision support system development, ultimately enhancing healthcare quality and early stroke prevention [8].

2. METHOD

This research method was designed to analyze the performance of three machine learning algorithms, namely Random Forest, XGBoost, and Support Vector Machine (SVM), in predicting stroke occurrence using the Healthcare Stroke Dataset. The research stages include understanding dataset characteristics, data preprocessing, model development, and model performance evaluation. All these stages were designed following the methodology commonly used in chronic disease prediction research based on medical data, particularly for datasets with class imbalance [9], [10].

2. 1. Dataset

The dataset used in this study is the Healthcare Stroke Dataset obtained from the Kaggle platform. This dataset has been widely utilized as a reference in medical classification research as it contains relevant patient clinical and demographic data for stroke risk analysis and has a tabular data structure suitable for applying machine learning algorithms [11]. The dataset consists of 5,110 samples, with 10 predictor features and 1 target feature representing stroke occurrence conditions in patients. All features in this dataset reflect parameters commonly used in stroke risk evaluation, encompassing demographic characteristics, medical history, health conditions, and patient physiological indicators. The use of such tabular clinical datasets has been extensively applied in chronic disease prediction

research due to their ability to comprehensively represent patient conditions and support data-driven model learning processes [12].

The target feature in this dataset is stroke, which represents stroke occurrence status based on medical diagnosis. The stroke variable has two classes: class 0 indicating patients who did not experience stroke and class 1 indicating patients who experienced stroke. The class distribution in this dataset is highly imbalanced, where the number of samples in the non-stroke class is substantially larger than the stroke class. Such class imbalance is a common characteristic in health datasets and can potentially cause model bias toward the majority class if not addressed with appropriate methods [9]. Therefore, model performance evaluation in this study was conducted using metrics capable of providing fair assessment of the minority class, namely macro F1-score, as recommended in disease classification research with imbalanced medical data [10]. A complete description of each variable used in the Healthcare Stroke Dataset is presented in Table 1.

Table 1. Feature Description of the Stroke Dataset

Features	Descriptions	Data Types
gender	Patient's gender	categorical
age	Patient's age	numerical
hypertension	History of hypertension (0 = no, 1 = yes)	binary
heart_disease	History of heart disease (0 = no, 1 = yes)	binary
ever_married	Patient's marital status	categorical
work_type	Patient's type of work	categorical
Residence_type	Type of residence (urban/rural)	categorical
avg_glucose_level	Average blood glucose level	numerical
bmi	Body mass index	numerical
smoking_status	Smoking habit status	categorical
stroke	Stroke occurrence status (target)	binary

2. 2. Preprocessing

The preprocessing stage was carried out to prepare the data to be suitable and optimal for processing by machine learning algorithms. Preprocessing is a crucial stage in chronic disease prediction research because data quality greatly influences the performance of classification models, especially in health data that has class imbalance and heterogeneous characteristics [13]. In this study, preprocessing includes handling missing values, transformation of categorical features, and scaling of numerical features.[14]

2. 2. 1. Handling Missing Values in Medical Features

In the Healthcare Stroke Dataset, missing values were found in the bmi feature. These missing values occurred due to incomplete data entry during the collection process and do not represent the actual physiological condition of patients. If these missing values are not addressed, the model may learn inaccurate patterns and cause decreased prediction performance, as reported in machine learning-based chronic disease prediction research [6], [10]. To overcome this problem, missing values in the bmi feature were replaced using the median value of the respective feature. The median was chosen because it is more robust to non-normal data distribution and the presence of outliers, which are commonly found in clinical data [8]. The median imputation method is also recommended in medical classification research because it maintains the statistical characteristics of the data without introducing significant bias. The imputation process is formulated as follows:

$$x'_i = \begin{cases} \text{median}(X), & \text{if } x_i \text{ is a missing value} \\ x_i, & \text{if } x_i \text{ is available} \end{cases}$$

with notation:

x_i : original value in the i -th sample

x'_i : value after imputation in the i -th sample

X : set of all values in the relevant feature

$\text{median}(X)$: median value of feature X

This imputation was consistently applied to all training data and testing data to maintain the validity of the model evaluation process.

2. 2. 2. Categorical Feature Encoding

The stroke dataset has several categorical features, namely gender, ever_married, work_type, Residence_type, and smoking_status. Categorical features cannot be directly processed by numerical-based machine learning algorithms, so they need to be transformed into numerical form. In this study, categorical features were transformed using One-Hot Encoding, where each category is represented as a binary variable. This method was chosen because it does not assume an ordinal relationship between categories and has been widely used in medical classification research based on tabular data. This approach also helps the model capture the contribution of each category independently [9].

2. 2. 3. Feature Scaling

Standardization was performed to ensure that all numerical features were on a uniform scale. In the stroke dataset, numerical features such as age, avg_glucose_level, and bmi had significantly different value ranges. This scale difference can cause certain models, particularly distance-based algorithms like Support Vector Machine (SVM), to be biased toward features with larger values [7]. Therefore, all numerical features were standardized using StandardScaler, which transforms the data to have a mean of 0 and a standard deviation of 1. The standardization process is formulated as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

with notation:

z_i : standardized value

x_i : original feature value

μ : mean of the feature

σ : standard deviation of the feature

Standardization was applied to all numerical features to ensure that each feature has a balanced contribution in the model learning process. This step aligns with recommendations from previous studies showing that SVM performance is highly sensitive to changes in data scale, and that standardization can improve the stability and accuracy of medical classification models [7]. Furthermore, standardization also helps maintain learning consistency in ensemble algorithms such as Random Forest and XGBoost when combined with encoded features [8].

2. 2. 4. Training and Testing Data Split

The division of the dataset into a training set and a testing set was carried out to separate the data used in the model learning process from the data used to evaluate its performance. This stage aims to ensure that the built model is not only able to learn patterns from the training data but also has good generalization ability to new, unseen data [15]. In this study, the dataset was split with a ratio of 80% training data and 20% testing data. The 80:20 ratio was chosen because it is a standard practice in machine learning research in the health domain, where a larger amount of training data is needed for the model to learn clinical patterns more effectively, while the testing data remains sufficient to assess model performance objectively and reliably. This ratio is also considered stable for maintaining a balance between the learning process and evaluation on a medium-sized dataset [16].

Data splitting was performed using the stratified split technique, a data partitioning method that preserves the proportion of the target class in both subsets, i.e., the training data and testing data. The use of stratified split is crucial in this study because the stroke prediction dataset has significant class imbalance, where the number of non-stroke patient samples is much larger than that of stroke patients.

Without stratification, random data splitting could potentially result in unrepresentative class distributions in the testing data and lead to biased evaluation results [17]. Mathematically, the data split is formulated as follows:

$$D = D_{\text{train}} \cup D_{\text{test}}$$

$$|D_{\text{train}}| = 0.8 |D|, |D_{\text{test}}| = 0.2 |D|$$

with notation::

D : entire dataset

D_{train} : training data (for model learning process)

D_{test} : testing data (for model performance evaluation)

$|D|$: total number of samples in the dataset

$|D_{\text{train}}|$: number of samples in the training subset

$|D_{\text{test}}|$: number of samples in the testing subset

The application of stratified split ensures that the class proportions in the training data and testing data follow the class distribution in the original dataset, thereby making model evaluation more stable and unbiased toward the majority class. The dataset used consists of 5,110 patient records, with 10 predictor features and 1 target feature, namely the stroke variable. Initial exploration results showed that approximately 95% of the data were in the non-stroke class and only about 5% in the stroke class, which is a common characteristic in health datasets based on critical disease events.

Furthermore, initial data exploration also revealed missing values in the Body Mass Index (BMI) feature. These missing values were handled during the preprocessing stage using the median imputation method, as this method is more robust to non-normal data distribution and the presence of outliers in clinical data. The descriptive statistics results of the main numerical features after the imputation process are presented in Table 2.

Table 2. Descriptive Statistics of the Stroke Dataset

Features	Mean	Std	Min	25%	50%	75%	Max
age	43.23	22.61	0.08	25.0	45.0	61.0	82.0
avg_glucose_level	106.15	45.28	55.12	77.24	91.88	114.09	271.74
bmi	28.89	7.85	10.30	23.50	28.40	33.10	97.60

Table 2 shows that the numerical features have considerable value variation, reflecting the heterogeneity of patient clinical conditions. The age feature shows a wide age range, while avg_glucose_level and bmi have broad value distributions with the presence of extreme values. This condition confirms the importance of applying preprocessing techniques, such as missing value imputation and feature standardization, before the data is used in the machine learning model training process. Overall, data splitting with an 80:20 ratio using the stratified split technique, accompanied by adequate descriptive statistical analysis, is a crucial step to ensure that the training and evaluation process of the stroke prediction model in this study can be conducted fairly, accurately, and scientifically accountable.

2. 3. Machine Learning Methods

This study applied three machine learning algorithms that are widely used in chronic disease prediction, namely Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). These three algorithms were chosen because they have good capabilities in processing tabular clinical data that is multivariate, has non-linear relationships between features, and faces class imbalance problems commonly found in health data [4].

2. 3. 1. Random Forest

Random Forest is an ensemble learning algorithm based on the bagging technique (bootstrap aggregating) that builds a number of decision trees independently. Each tree is trained using data

samples randomly selected through a bootstrap sampling process. At the prediction stage, each tree produces a classification decision, and the final model outcome is determined by majority voting from all trees in the ensemble [3], [8]. This approach allows Random Forest to increase model stability and reduce the risk of overfitting, particularly in clinical data containing noise and variation between individuals. Additionally, at each node splitting process, Random Forest only considers a randomly selected subset of features. This strategy increases diversity among trees and makes the model more robust to feature correlations that often appear in medical data. Mathematically, the Random Forest prediction for classification problems can be formulated as follows:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_T(x))$$

with notation:

$\hat{y}$: final prediction result

h_t : prediction from the t-th tree

T: number of trees in the ensemble

mode: the most frequently occurring value

Random Forest has several advantages, including the ability to handle high-dimensional data, relative resistance to outliers, and effectiveness in modeling non-linear relationships without requiring specific data distribution assumptions. However, the main weakness of Random Forest lies in its high model complexity, as the use of many trees can increase computational requirements and inference time compared to single decision tree-based methods [8].

2. 3. 2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm based on boosting that builds models sequentially by adding new decision trees to correct prediction errors from previous models. This approach enables XGBoost to iteratively minimize prediction errors by focusing learning on samples that were difficult to classify in previous iterations. Due to its ability to capture non-linear relationships and complex interactions between features, XGBoost is widely used in disease prediction research based on tabular clinical data [6], [9]. One of the main advantages of XGBoost compared to conventional boosting methods is the application of L1 and L2 regularization, which explicitly controls model complexity. This regularization plays an important role in reducing the risk of overfitting, particularly in medical data that has high dimensionality, feature correlations, and class imbalance. Therefore, XGBoost is considered highly suitable for handling health data with complex structures and high variability. Mathematically, XGBoost optimizes an objective function consisting of a loss function and a regularization component, which is formulated as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k)$$

Notation:

$\mathcal{L}^{(t)}$: objective function value at the t-th iteration

$l(Y_i, \hat{Y}_i^{(t)})$: loss function that measures the difference between the actual label Y_i and the model prediction

n: number of data samples

f_k : k-th decision tree added in the boosting process

$\Omega(f_k)$: regularization function that measures the complexity of the tree

The regularization function in XGBoost is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

with notation:

T: number of leaves in the tree

w_j : weight on the j -th leaf
 γ and λ : regularization parameters

The XGBoost model is built using a gradient boosting approach, where at each iteration a new tree is added that minimizes the objective function through first and second-order gradient approximation. This approach enables a faster and more stable optimization process compared to traditional boosting methods [18]. The advantages of XGBoost include high computational efficiency, strong regularization capabilities, and consistent performance across various chronic disease prediction studies and medical classification tasks. However, XGBoost has the weakness of being sensitive to hyperparameter selection, thus requiring a careful tuning process to prevent the model from overfitting and to achieve optimal performance [19].

2. 3. 3. Support Vector Machine

Support Vector Machine (SVM) is a margin-based classification algorithm that aims to find an optimal hyperplane capable of separating two classes with maximum distance (margin). This approach is designed to produce decision boundaries with high generalization ability, making it effective for complex and noisy medical data classification problems [20]. In the case of clinical data with non-linear relationships between features, SVM utilizes kernel functions to map data from a low-dimensional input space to a higher-dimensional feature space. One of the most commonly used kernel functions is the Radial Basis Function (RBF), due to its ability to capture non-linear patterns and complex interactions between clinical variables without requiring explicit knowledge of the transformation form [21]. The main concept of SVM is to maximize the margin between two classes while minimizing classification errors. The SVM learning process is formulated as the following optimization problem:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

This objective function aims to minimize model complexity by reducing the weight norm $\|\mathbf{w}\|^2$, while simultaneously maximizing the margin and controlling the number of classification errors through parameter C . A value of C that is too large can cause the model to become overly complex and prone to overfitting, whereas a value that is too small can reduce the model's ability to optimally separate classes. SVM has several advantages, including being effective on high-dimensional data, performing well on relatively small datasets, and having a strong theoretical foundation in statistical learning. However, SVM also has limitations, namely being sensitive to feature scaling and kernel parameter selection. Therefore, preprocessing steps such as feature standardization and appropriate kernel parameter determination are crucial to ensure that SVM achieves optimal performance on medical data [22].

2. 4. Model Evaluation

2. 4. 1. Accuracy

Accuracy is a metric that measures the proportion of correct predictions made by the model relative to the total number of samples tested. This metric is often used as an initial indicator of classification model performance due to its simplicity and ease of interpretation. Mathematically, accuracy is formulated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

with notation:

TP (True Positive): number of stroke samples correctly predicted as stroke

TN (True Negative): number of non-stroke samples correctly predicted as non-stroke

FP (False Positive): number of non-stroke samples incorrectly predicted as stroke

FN (False Negative): number of stroke samples incorrectly predicted as non-stroke

Accuracy indicates the overall proportion of correct predictions made by the model. However, in the context of highly imbalanced medical datasets, such as in stroke prediction where the number of non-stroke patients is far greater than stroke patients, the accuracy metric can become less informative. A model can achieve high accuracy simply by predicting all samples as the majority class, even though it fails to detect any stroke cases at all [24]. Therefore, in health research, accuracy is not used as the sole evaluation metric but is analyzed alongside other metrics such as Precision, Recall, and F1-Score, particularly in the form of macro-average, to ensure that the model is not biased toward the majority class and maintains good capability in detecting the minority class, which is clinically more critical [25].

2. 4. 2. Precision

Precision is an evaluation metric that measures the proportion of positive predictions that are truly positive out of all positive predictions generated by the model. In the context of stroke prediction research, precision represents the model's ability to avoid false positive prediction errors, i.e., conditions when the model classifies a patient as having a stroke when in fact the patient did not experience a stroke. Mathematically, precision is formulated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A high precision indicates that the model is not reckless in assigning positive labels, so that most of the stroke predictions generated are indeed actual stroke cases. This is crucial in the medical context, because incorrect positive predictions can cause psychological anxiety, unnecessary follow-up examinations, and waste of healthcare resources [10]. In health research, several studies emphasize that precision is a critical metric, particularly for diseases requiring further clinical follow-up after initial screening results. Therefore, precision needs to be analyzed together with other metrics such as recall and F1-score to achieve a balance between the ability to detect stroke cases and minimizing misdiagnosis [26].

2. 4. 3. Recall

Recall, also known as sensitivity, is an evaluation metric that measures the proportion of positive samples correctly detected by the model out of all samples that truly belong to the positive class. In the context of stroke prediction research, recall describes the model's ability to identify patients who actually experienced a stroke. Mathematically, recall is formulated as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall plays a very important role in the medical field because it is directly related to the risk of undetected patients who are actually sick. A low recall value indicates that many stroke cases are still missed by the model (false negative), which in clinical practice can have serious impacts on patient safety due to delays in medical treatment [1], [10]. High recall indicates that the model is able to capture most stroke cases present in the data, making it highly relevant as a primary metric in disease screening systems. Several studies in the health field emphasize that, in the case of chronic and critical diseases such as stroke, false negative errors are considered more dangerous than false positives, because they can cause patients not to receive timely diagnosis and necessary interventions [1], [9]. Therefore, in this study, recall is analyzed together with precision and F1-score metrics to ensure that the model is not only sensitive to the stroke class but also maintains a balance between case detection and overall prediction accuracy.

2. 4. 4. F1-Score

F1-Score is an evaluation metric that combines precision and recall into a single measure through the harmonic mean approach. This metric is designed to provide a balanced assessment between

the model's ability to avoid false positives and minimize false negatives, making it highly relevant for use in medical datasets with class imbalance. Mathematically, F1-Score is formulated as follows:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In the context of stroke prediction research, the F1-Score provides a comprehensive overview of model performance because it considers two crucial aspects of medical diagnosis: the accuracy of identifying stroke patients and the ability to detect as many existing stroke cases as possible. This is important because a model with high precision but low recall still has the potential to miss many stroke patients, whereas a model with high recall but low precision can generate many false alarms [5], [8]. Several studies in the field of medical machine learning affirm that the F1-Score is more representative than accuracy in imbalanced data conditions, because accuracy can give the impression of good performance even when the model fails to adequately recognize the minority class [5], [9]. Therefore, the F1-Score is often used as the primary metric in evaluating prediction systems for chronic and high-risk diseases such as stroke.

In this study, the F1-Score, particularly the macro F1-score, is used as the main basis for comparing the performance of Random Forest, XGBoost, and SVM. The use of the macro F1-score enables a fair evaluation of each class by giving equal weight to both the majority and minority classes, so that the evaluation results are not biased toward the more dominant non-stroke class [8], [10].

2. 4. 5. Macro F1-Score (For Imbalanced Data)

Macro F1-Score is a variant of the F1-Score calculated by determining the F1-Score for each class separately, then taking the arithmetic average of all these values without considering the proportion of sample sizes in each class. This approach ensures that each class has an equal contribution to the evaluation process, regardless of the imbalanced data distribution. Mathematically, Macro F1-Score is formulated as follows:

$$\text{Macro F1} = \frac{1}{K} \sum_{i=1}^K \text{F1}_i$$

In the context of this study, Macro F1-Score was selected as one of the primary evaluation metrics because the Healthcare Stroke Dataset is highly imbalanced, with non-stroke cases substantially outnumbering stroke cases. Unlike accuracy, Macro F1-Score assigns equal importance to each class, thereby providing a more balanced evaluation of the model's ability to classify both the majority (non-stroke) and minority (stroke) classes. This characteristic makes Macro F1-Score particularly appropriate for evaluating classification models developed using imbalanced medical datasets[13]. Various studies in the field of chronic disease classification, including diabetes and chronic kidney disease, recommend the use of Macro F1-Score as the primary metric in imbalanced data conditions due to its ability to fairly reflect model performance across all classes, particularly in detecting disease cases that are relatively few in number [5], [12].

3. RESULTS AND DISCUSSION

3. 1. Results

The dataset used in this study is the Healthcare Stroke Dataset, consisting of 5,110 samples, with 10 predictor features and 1 target feature, namely stroke. The stroke variable has two classes: class 0 (non-stroke) and class 1 (stroke). Based on initial exploration, the class distribution in the dataset is highly imbalanced, where the number of non-stroke patients is far more dominant compared to stroke patients. After the preprocessing stage was completed, the dataset was divided into training data and testing data using the stratified split method with an 80% training data and 20% testing data ratio. The use of stratified split aims to maintain the class proportions in both subsets so that model evaluation can be conducted fairly and representatively according to the actual data conditions. The class distribution in the testing data is shown in Table 3.

Table 3. Outcome Class Distribution in Training & Testing Data

Dataset	Total Samples	Class 0 (Non-Stroke)	Class 1 (Stroke)	Class 0 Proportion	Class 1 Proportion
y_test	1,022	972	50	95.11%	4.89%

Based on Table 3, it can be seen that the dataset has a very high class imbalance, where the non-stroke class dominates more than 95% of the total testing data. This condition has the potential to cause the classification model to be biased toward the majority class if only evaluated using accuracy metrics. Therefore, this study uses the Macro F1-Score as the main metric to assess model performance more fairly across both classes, particularly the minority class (stroke).

Table 4. Performance Comparison Before and After Applying Class Weighting and SMOTE

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Random Forest (Class Weight)	0,95	0,48	0,50	0,49
SVM (Class Weight)	0,76	0,55	0,71	0,53
XGBoost (Class Weight)	0,94	0,62	0,58	0,59
Random Forest (SMOTE)	0,93	0,59	0,57	0,58
SVM (SMOTE)	0,81	0,54	0,63	0,54
XGBoost (SMOTE)	0,93	0,57	0,55	0,56

Based on the results in Table 4, Random Forest with class weighting produced the highest accuracy value but failed to detect stroke cases entirely, as indicated by the stroke class recall value of 0.00. This confirms that high accuracy does not always reflect good model performance on imbalanced medical datasets. The SVM model with class weighting showed a significant improvement in detecting the stroke class, with relatively high recall. However, the low precision indicates that there were still many false positives, so the model's performance was not yet optimal. The XGBoost model with class weighting provided the most balanced performance, demonstrated by the highest Macro F1-Score of 0.59. This model was able to maintain a balance between prediction precision and the ability to detect stroke cases, without drastically sacrificing accuracy.

In the SMOTE approach, all models showed improved ability to detect the minority class compared to Random Forest without balancing, but they were unable to surpass the performance of XGBoost with class weighting. This indicates that on a highly imbalanced stroke dataset, the class weighting approach with XGBoost is more effective than SMOTE oversampling.

3. 2. Discussion

This study aims to identify a machine learning model capable of providing the best performance in predicting stroke occurrence based on the Healthcare Stroke Dataset. After going through the stages of data preprocessing, stratified data splitting, model training, and evaluation using macro average-based metrics, the results obtained showed significant performance differences among the tested algorithms. The test results indicated that Random Forest with class weighting produced a very high accuracy value. However, further analysis revealed that this model failed to detect all stroke cases, characterized by very low recall and F1-score values for the stroke class. This condition confirms that in datasets with extreme class imbalance, high accuracy does not always reflect good model performance in a medical context, as the model tends to be biased toward the majority class (non-stroke).

The Support Vector Machine (SVM) model demonstrated a better ability to detect the stroke class, particularly when applied with class weighting or SMOTE. The relatively higher recall value for the stroke class indicates that SVM is quite sensitive in recognizing stroke cases. However, the low precision indicates a high number of false positives, making the model potentially prone to providing excessive stroke alerts and less than ideal for clinical implementation without additional filtering. Meanwhile, XGBoost with class weighting showed the most balanced performance compared to the other models. This model was able to maintain high accuracy while also achieving the highest Macro F1-Score, indicating a balance between the ability to detect the majority class and the minority class. XGBoost's advantage lies in its boosting mechanism, which is capable of learning from previous model errors, and its regularization support that helps control overfitting, making it more adaptive to complex and highly imbalanced medical datasets such as the stroke data.

The application of SMOTE to the three algorithms generally improved the models' ability to detect the stroke class compared to models without balancing. However, in this study, the SMOTE

approach was not able to surpass the performance of XGBoost with class weighting, indicating that class weighting is more effective than synthetic oversampling for stroke datasets with a very high imbalance ratio. Overall, the results of this study confirm that the selection of algorithms and strategies for handling class imbalance significantly influences the quality of stroke prediction. The use of the Macro F1-Score metric proved to be more representative than accuracy in evaluating model performance on medical datasets with a very small minority class.

4. CONCLUSION

This study compared the performance of three machine learning algorithms, namely Random Forest, XGBoost, and Support Vector Machine (SVM), for predicting stroke occurrence using the Healthcare Stroke Dataset. The proposed framework included data preprocessing, class imbalance handling, model training, and performance evaluation using Accuracy, Precision, Recall, and Macro F1-score. The experimental results demonstrated that class imbalance significantly affected the classification performance, particularly in detecting stroke cases as the minority class. Among the evaluated models, XGBoost with the Class Weighting approach achieved the best overall performance, obtaining the highest Macro F1-score of 0.59 while maintaining a more balanced trade-off between precision and recall than the other algorithms. Random Forest also produced competitive results but was less effective in recognizing the minority class, whereas Support Vector Machine (SVM) showed the lowest performance due to its sensitivity to class imbalance and feature distribution. The findings indicate that appropriate class imbalance handling plays a crucial role in improving stroke prediction performance. In this study, the Class Weighting strategy proved more effective than the baseline approach in enhancing the recognition of stroke cases without sacrificing overall classification performance. These results demonstrate that XGBoost combined with Class Weighting is the most suitable model for stroke prediction using the Healthcare Stroke Dataset and can serve as a promising foundation for developing clinical decision support systems for early stroke risk assessment. Future research should investigate more comprehensive hyperparameter optimization, advanced imbalance-handling techniques such as ADASYN and hybrid sampling methods, and larger, multicenter clinical datasets to improve model generalization and predictive performance. Furthermore, integrating the proposed model into real-world clinical decision support systems and validating it using external datasets would enhance its practical applicability in healthcare settings.

ACKNOWLEDGMENTS

The authors sincerely thank Universitas Amikom Purwokerto, particularly the Department of Information Technology, and all parties who provided support, data, and references for this research.

BIBLIOGRAPHY

- [1] 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,” *Lancet Neurol.*, vol. 23, no. 10, pp. 973–1003, Oct. 2024, doi: 10.1016/S1474-4422(24)00369-7.
- [2] X. Huang *et al.*, “Novel Insights on Establishing Machine Learning-Based Stroke Prediction Models Among Hypertensive Adults,” *Front. Cardiovasc. Med.*, vol. 9, May 2022, doi: 10.3389/fcvm.2022.901240.
- [3] F. Asadi, M. Rahimi, A. H. Daeeshini, and A. Paghe, “The most efficient machine learning algorithms in stroke prediction: A systematic review,” *Health Sci. Rep.*, vol. 7, no. 10, Oct. 2024, doi: 10.1002/hsr2.70062.
- [4] R. Palomo-Gómez *et al.*, “Prenatal Stress as a Risk Factor for Maternal–Foetal Morbidity: A Longitudinal Study,” *Healthcare*, vol. 12, no. 3, p. 312, Jan. 2024, doi: 10.3390/healthcare12030312.
- [5] G. Jia and G. Jin, “The prediction and feature importance analysis of stroke based on the machine learning algorithm,” *Applied and Computational Engineering*, vol. 18, no. 1, pp. 225–229, Oct. 2023, doi: 10.54254/2755-2721/18/20230994.

- [6] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/j.health.2022.100116.
- [7] E. M. Z. Akay, A. Hilbert, B. G. Carlisle, V. I. Madai, M. A. Mutke, and D. Frey, "Artificial Intelligence for Clinical Decision Support in Acute Ischemic Stroke: A Systematic Review," *Stroke*, vol. 54, no. 6, pp. 1505–1516, Jun. 2023, doi: 10.1161/STROKEAHA.122.041442.
- [8] A. Byna, M. M. Lakulu, I. Y. Panessai, and Nurhaeni, "Machine Learning-Based Stroke Prediction: A Critical Analysis," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 5, pp. 1609–1618, Oct. 2024, doi: 10.18517/ijaseit.14.5.19527.
- [9] K. Moulaei, L. Afshari, R. Moulaei, B. Sabet, S. M. Mousavi, and M. R. Afrash, "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models," *Sci. Rep.*, vol. 14, no. 1, p. 31392, Dec. 2024, doi: 10.1038/s41598-024-82931-5.
- [10] N. Melnykova, Y. Patereha, S. Skopivskyi, M. Farion, S. Fedushko, and K. Drohomiretska, "Machine learning for stroke prediction using imbalanced data," *Sci. Rep.*, vol. 15, no. 1, p. 33773, Sep. 2025, doi: 10.1038/s41598-025-01855-w.
- [11] I. G. Ivanov, Y. Kumchev, and V. J. Hooper, "An Optimization Precise Model of Stroke Data to Improve Stroke Prediction," *Algorithms*, vol. 16, no. 9, p. 417, Sep. 2023, doi: 10.3390/a16090417.
- [12] L. Fast *et al.*, "Machine learning-based prediction of clinical outcomes after first-ever ischemic stroke," *Front. Neurol.*, vol. 14, Feb. 2023, doi: 10.3389/fneur.2023.1114360.
- [13] P. Chakraborty *et al.*, "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," *BMC Bioinformatics*, vol. 25, no. 1, p. 329, Oct. 2024, doi: 10.1186/s12859-024-05866-8.
- [14] S. Zhi *et al.*, "An exploration on the machine-learning-based stroke prediction model," *Front. Neurol.*, vol. 15, Apr. 2024, doi: 10.3389/fneur.2024.1372431.
- [15] V. Padimi, V. S. Telu, and D. D. Ningombam, "Performance analysis and comparison of various machine learning algorithms for early stroke prediction," *ETRI Journal*, vol. 45, no. 6, pp. 1007–1021, Dec. 2023, doi: 10.4218/etrij.2022-0271.
- [16] S. Lolak, J. Attia, G. J. McKay, and A. Thakkinstant, "Comparing Explainable Machine Learning Approaches With Traditional Statistical Methods for Evaluating Stroke Risk Models: Retrospective Cohort Study," *JMIR Cardio*, vol. 7, p. e47736, Jul. 2023, doi: 10.2196/47736.
- [17] A. Zignoli, K. Skroce, D. J. Lipman, and H. C. Zisser, "Personalized nutrition and machine-learning: Exploring the scope of continuous glucose monitoring in healthy individuals in uncontrolled settings," *Biomed. Signal Process. Control*, vol. 90, p. 105809, Apr. 2024, doi: 10.1016/j.bspc.2023.105809.
- [18] Md. J. Raihan, Md. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 6263, Apr. 2023, doi: 10.1038/s41598-023-33525-0.
- [19] S. Friedrich *et al.*, "Regularization approaches in clinical biostatistics: A review of methods and their applications," *Stat. Methods Med. Res.*, vol. 32, no. 2, pp. 425–440, Feb. 2023, doi: 10.1177/09622802221133557.
- [20] S. Abdelfattah *et al.*, "Lightweight Multi-Class Support Vector Machine-Based Medical Diagnosis System with Privacy Preservation," *Sensors*, vol. 23, no. 22, p. 9033, Nov. 2023, doi: 10.3390/s23229033.
- [21] J. Wang, F. He, and S. Sun, "Construction of a new smooth support vector machine model and its application in heart disease diagnosis," *PLoS One*, vol. 18, no. 2, p. e0280804, Feb. 2023, doi: 10.1371/journal.pone.0280804.
- [22] S. Abdelfattah *et al.*, "Lightweight Multi-Class Support Vector Machine-Based Medical Diagnosis System with Privacy Preservation," *Sensors*, vol. 23, no. 22, p. 9033, Nov. 2023, doi: 10.3390/s23229033.
- [23] N. Melnykova, Y. Patereha, S. Skopivskyi, M. Farion, S. Fedushko, and K. Drohomiretska, "Machine learning for stroke prediction using imbalanced data," *Sci. Rep.*, vol. 15, no. 1, p. 33773, Sep. 2025, doi: 10.1038/s41598-025-01855-w.

-
- [24] F. H. Pereira, A. H. de A. Melani, F. N. Kashiwagi, T. G. da Rosa, U. S. dos Santos, and G. F. M. de Souza, "Imperfect Preventive Maintenance Optimization with Variable Age Reduction Factor and Independent Intervention Level," *Applied Sciences*, vol. 13, no. 18, p. 10210, Sep. 2023, doi: 10.3390/app131810210.
- [25] M. A. Hertz *et al.*, "Urine Flow Cytometry and Dipstick Analysis in Diagnosing Bacteriuria and Urinary Tract Infections among Adults in the Emergency Department—A Diagnostic Accuracy Trial," *Diagnostics*, vol. 14, no. 4, p. 412, Feb. 2024, doi: 10.3390/diagnostics14040412.
- [26] M. A. Hertz *et al.*, "Urine Flow Cytometry and Dipstick Analysis in Diagnosing Bacteriuria and Urinary Tract Infections among Adults in the Emergency Department—A Diagnostic Accuracy Trial," *Diagnostics*, vol. 14, no. 4, p. 412, Feb. 2024, doi: 10.3390/diagnostics14040412.