ILKOM Scientific Journal Vol. 18, No. 1, April 2026, pp. 180-194
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779| | P-ISSN 2087-1716

SMOTE-Based Comparative Analysis of Machine Learning Models for Stroke Risk Prediction Using Imbalanced Healthcare Data

Ratu Mutiara Siregar [a,1] ; Budy Satria [b,2*] ; Sandi Fadilah [c,3,] ; Liga Mayola [d,4] ; Silky Safira [d,5]
[a] *Department of Information Systems and Technology, Institut Teknologi Sawit Indonesia, Jl. William Iskandar, Medan, 20226, Indonesia*
[b] *Department of Informatics, Universitas Andalas, Jl. Limau Manis, Padang, 25163, Indonesia*
[c] *Department of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, 02100 Padang Besar, Perlis, Malaysia*
[d] *Department of Information Systems, Universitas Putra Indonesia YPTK, Jl. Raya Lubuk Begalung, Padang, 25221, Indonesia*
[1] *ratu_ms@itsi.ac.id;* [2*] *budy.satria@it.unand.ac.id;* [3] *p5250078@student.umam.edu.my;* [4] *ligamayola@upiyptk.ac.id;* [5] *silkysafira@upiyptk.ac.id*
* *Corresponding author*

Article history: Received December 11, 2025; Revised February 10, 2026; Accepted April 01, 2026; Available online April 20, 2026

Abstract

Stroke remains one of the leading causes of mortality and long-term disability worldwide, with a significant burden in Indonesia. Early detection is crucial, as up to 90% of stroke cases are potentially preventable through timely intervention. However, predictive modeling for stroke risk is often challenged by imbalanced datasets, where non-stroke cases significantly outnumber stroke cases, potentially biasing classification models. This study aims to perform a systematic comparative evaluation of six machine learning algorithms Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost) for stroke risk prediction under imbalanced data conditions. The dataset consists of 5,110 patient records with 11 health-related features obtained from a publicly available healthcare dataset. Data preprocessing included anomaly removal, categorical encoding, feature scaling, and class balancing using the Synthetic Minority Oversampling Technique (SMOTE). Model evaluation was conducted using 5-fold cross-validation and assessed through accuracy, precision, recall, and F1-score metrics. The experimental results demonstrate that ensemble-based models outperform single classifiers. Random Forest achieved the highest mean accuracy of 97.12% (±0.42) with an F1-score of 0.96, followed closely by XGBoost with 96.85% (±0.51). Both models also exhibited superior recall performance, indicating improved minority class detection. The novelty of this study lies in the systematic evaluation of multiple machine learning models using SMOTE-based balancing and cross-validation on publicly available healthcare data, providing robust comparative insights for imbalanced medical classification problems.

Keywords: Stroke prediction, Machine Learning, SMOTE, Imbalanced Data, Random Forest, XGBoost

Introduction

Stroke is a major global health problem and remains one of the leading causes of mortality and long-term disability [1]. Rapid neuronal damage occurs during stroke events, emphasizing the urgency of early detection and intervention. In Indonesia, stroke prevalence continues to increase, placing a substantial economic and social burden on the healthcare system [2]. Although early screening can significantly reduce complications, many individuals remain unaware of early symptoms, resulting in delayed treatment.

With advancements in artificial intelligence, machine learning (ML) has emerged as a promising approach for disease risk prediction [3]. ML models can learn complex patterns from historical patient data and provide automated classification to support early diagnosis. However, medical datasets frequently suffer from class imbalance, where the number of negative cases significantly exceeds that of positive cases [4]. In stroke prediction datasets, non-stroke cases often dominate, leading to biased learning toward the majority class and poor minority detection performance [5].

Previous studies have reported promising results using ensemble-based algorithms, such as Random Forests and XGBoost, for medical classification tasks. Several works have also incorporated the Synthetic Minority Oversampling Technique (SMOTE) to mitigate imbalance issues [6]. However, many studies focus on limited model comparisons, lack systematic cross-validation, or do not explicitly analyze the impact of balancing strategies on model robustness[7]. Furthermore, comparative evaluations across multiple algorithms, using consistent preprocessing and validation frameworks, remain limited.

http://dx.doi.org/10.33096/ilkom.v18i1.3161.180-194

Therefore, this study aims to systematically compare six widely used machine learning algorithms for stroke risk prediction under imbalanced data conditions using SMOTE-based balancing and cross-validation. The main contributions of this study are:

- A comprehensive comparative evaluation of six machine learning algorithms under controlled preprocessing and validation settings.
- Implementation of SMOTE-based oversampling to address class imbalance in stroke prediction.
- Robust evaluation using 5-fold cross-validation and multiple performance metrics to ensure model reliability.
- Analytical discussion of ensemble learning effectiveness in imbalanced healthcare classification.

Several recent studies have demonstrated the effectiveness of machine learning for stroke prediction, particularly using ensemble learning and deep learning approaches [8], [9], [10], [11]. Ensemble methods such as Random Forests and XGBoost consistently achieve superior performance in medical classification tasks due to their robustness against overfitting and their ability to model nonlinear relationships [12], [13], [14], [15].

However, medical datasets are frequently affected by class imbalance, which may bias predictive performance toward majority classes [16], [17]. To address this challenge, oversampling techniques such as SMOTE have been widely adopted [18], [19], with empirical evidence suggesting improved minority class detection in healthcare applications [21], [22]. Despite these advancements, systematic comparative evaluations under consistent preprocessing and validation frameworks remain limited. This study addresses this gap by integrating SMOTE-based balancing with cross-validated comparative modelling.

Methods

This study compares six methods: Logistic Regression, Random Forest, Decision Tree, Naive Bayes, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). The purpose of this study is to find the best method for predicting stroke.

A. Data Sources

Data is a very important part of a study, because the type and characteristics of the data will influence the selection of an appropriate method to solve the problem being studied. To produce a model capable of accurately learning patterns, we need to select high-quality, large-scale, relevant data.

In addition to quality, data must also be obtained legally, as illegal data collection could expose researchers to legal liability. Based on these considerations, the data used in this study is public and was obtained from *kaggle.com* using the search term *"stroke prediction dataset*" (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset). This dataset consists of 5,110 patient records covering 12 attributes, including ID, gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose, BMI, smoking status, and stroke.

Table 1. Attribute Description

Attribute Name	Data Type	Description
Id	Int	For administrative purposes
Gender	Object	Patient gender (Male, Female, Other)
Age	Float	Patient age (in years)
Hypertension	Int	History of hypertension (1 = yes, 0 = no)
Heart_disease	Int	History of heart disease (1 = yes, 0 = no)
Ever_married	Object	Marital status patient (Yes, No)
Work_type	Object	Type of employment patient (Private, Self-employed, Govt_job, Children, Never worked)
Residence_type	object	Place of residence(Urban, Rural)
Average_glucose_level	float	Average blood glucose level blood glucose level (mg/dL)
bmi	float	Body mass index (kg/m²)
Smoking_status	object	Smoking status (formerly smoked, never smoked, smokes, Unknown)
stroke	Int	Target variable: Has the patient Ever had a stroke (1 = yes, 0 = no)

Table 1 showed attribute description data is used to predict whether a patient is at risk of stroke, based on these features. This dataset has been widely used in previous studies to develop early-detection systems for stroke, due to its comprehensive nature and its representativeness of commonly encountered medical conditions.

This data is binary, with the target column (stroke) taking values 0 (no stroke) or 1 (stroke), making it well-suited for classification modeling.

B. Research Flow

This research flow comprises several systematic stages, from data collection to model evaluation. The steps are described as follows:

1. Data Collection and Understanding

   The dataset used in this study was obtained from a publicly available healthcare dataset hosted on Kaggle. It contains 5,110 patient records with 11 predictive attributes and one binary target variable (stroke: 1 = stroke, 0 = no stroke).

2. Data Preprocessing

   Data preprocessing is a series of steps performed to prepare raw data before it is used in a machine learning model[8] . Data preprocessing involved:

   - Removal of missing BMI values (201 records)
   - Removal of anomalous gender category (“Other”)
   - Categorical encoding:
     Binary features: Label Encoding
     Multi-class features: One-Hot Encoding
   - Feature scaling using StandardScaler.
   - Train-test split (80:20, random_state=42)

   The Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set to generate synthetic minority samples and balance class distribution. To ensure robustness and reduce variance bias, 5-fold cross-validation was implemented during model evaluation.

3. Modeling

   In the modeling stage, we will apply and evaluate the performance of six machine learning classification algorithms. Each model is trained on balanced, transformed training data. The algorithms to be used in this study are Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Model Configuration Hyperparameter settings:

   - Logistic Regression: solver = lbfgs, max_iter = 1000
   - Random Forest: n_estimators = 100, max_depth = None, random_state = 42
   - Decision Tree: criterion = gini, max_depth = None
   - Naïve Bayes: GaussianNB (default parameters)
   - SVM: kernel = rbf, C = 1, gamma = scale
   - XGBoost: n_estimators = 100, learning_rate = 0.1, max_depth = 6

4. Model Evaluation

   The evaluation stage measures how well the machine learning model predicts stroke risk using the prepared test data. One of the evaluation methods used is the confusion matrix, which is a table that presents a comparison between the actual labels and the predicted results of the classification system[23] .

   The confusion matrix contains four main components, namely:

   - True Positive (TP): the number of stroke cases that are correctly detected as strokes.
   - True Negative (TN): the number of non-stroke cases that were correctly detected as non-stroke.
   - False Positive (FP): non-stroke cases that are incorrectly classified as strokes,

- False Negative (FN): stroke cases that are incorrectly classified as non-stroke.

This matrix is used to calculate various evaluation metrics, such as accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Accuracy measures the proportion of correct predictions against the entire data set. However, in classification cases with unbalanced data such as the data in this study, accuracy alone is not sufficient to measure the overall performance of the model.

Therefore, other metrics are also used, such as:

a. Recall

Recall indicates the model's ability to correctly detect positives (strokes). Recall allows us to determine how many of all patients who actually had a stroke were successfully detected. A high recall value means that the model rarely misses stroke cases.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

b. Precision

Precision measures how many stroke predictions are actually positive. Here, we will see how many of all patients who actually had a stroke were successfully detected. Precision is important in the context of health to minimize false diagnoses of healthy patients.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

c. F1-Score

The F1-score is the harmonic mean of precision and recall, suitable for use when the data is imbalanced between the two.

$$\text{F1} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

Particular emphasis was placed on recall and F1-score due to their importance in imbalanced medical classification.

Figure 1 is an illustration of the research flow we will conduct:

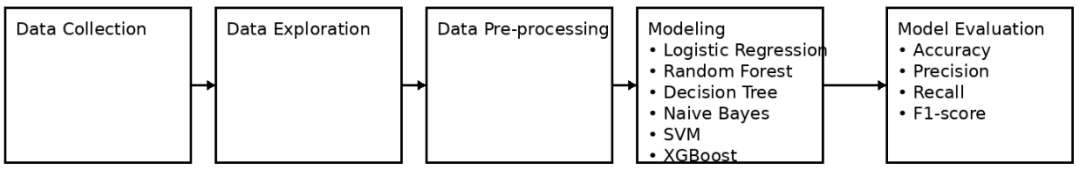


Figure 1. Research Flow

C. Results and Discussion

1. Data Collection

The data used in this study was sourced from the Kaggle platform under the name Stroke prediction Dataset. This dataset consists of 5110 data sets covering 11 health attributes and 1 target column (stroke). Each row represents one individual, complete with information such as age, gender, history of hypertension, blood glucose levels, smoking status, and whether the individual has ever had a stroke. Additionally, an id attribute was added as a unique identifier, but it was not used in the model training process.

This dataset is public and has been widely used in various medical prediction studies. The data is categorical and numerical, so adjustments are needed before it can be used in machine learning modeling.

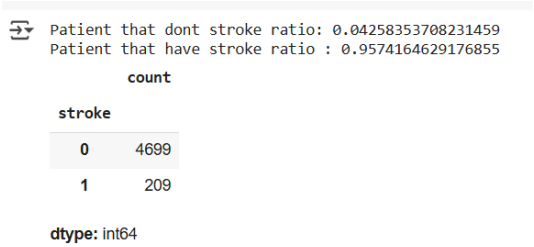
This dataset is used because it represents a binary classification problem commonly found in the health sector, particularly in the early detection of stroke.

2. Data Exploration

The data exploration stage was conducted to understand the basic characteristics of the dataset before the modeling process was carried out.

Some of the exploratory findings obtained include:

- From the 5,110 patient data available, it was found that the class distribution in the target (stroke) was very unbalanced, As shown in **Figure 2**, the distribution of the target variable is highly imbalanced, with approximately 95% of instances belonging to the non-stroke class and only 5% to the stroke class. This became a major concern in selecting the data balancing technique at the preprocessing stage.



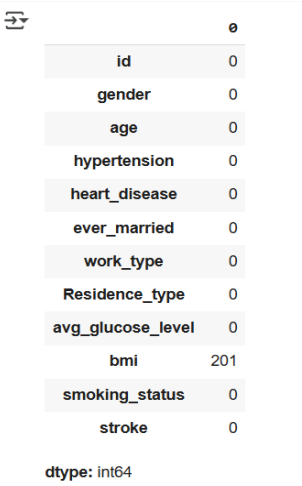
```
Patient that dont stroke ratio: 0.04258353708231459
Patient that have stroke ratio : 0.9574164629176855
```

stroke	count
0	4699
1	209

dtype: int64

Figure 2. Target Data Distribution

- **Figure 3** presents the missing value analysis, indicating that the BMI (Body Mass Index) attribute contains 201 missing entries. These incomplete records were removed to maintain data integrity prior to model training.



	0
id	0
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0

dtype: int64

Figure 3. Missing Value

- gender column has unique values that do not match. The gender column should have two unique values, but the data has three unique values. As illustrated in **Figure 4**, the gender attribute includes an anomalous category ("Other") with only a single occurrence. This outlier was excluded to prevent potential bias during encoding and model training.

	0
id	5110
gender	3
age	104
hypertension	2
heart_disease	2
ever_married	2
work_type	5
Residence_type	2
avg_glucose_level	3979
bmi	418
smoking_status	4
stroke	2

Figure 4. Unique Data Values

- Figure 5 shows the distribution of the average glucose level, where several outliers are detected. Although extreme values are present, they were retained after evaluation to preserve clinically relevant information.

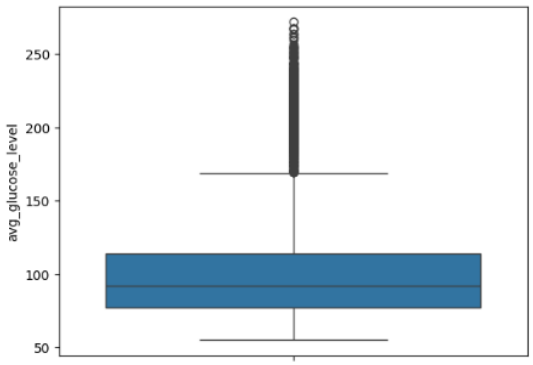


Figure 5. Glucose Level Outliers

The correlation heatmap presented in Figure 6 indicates that no pair of features exhibits extremely high correlation, suggesting minimal multicollinearity among predictors. Therefore, all relevant features were retained for modeling.

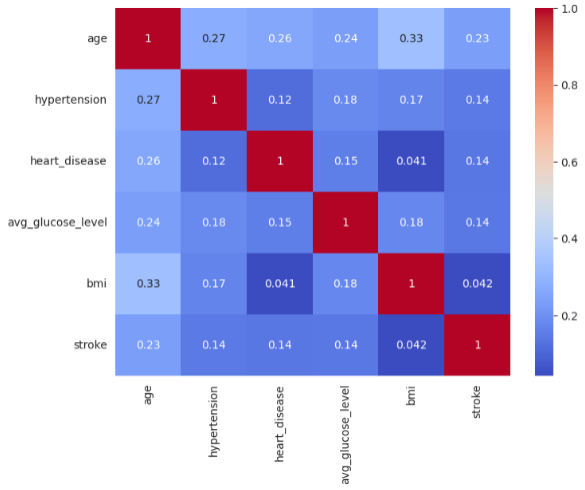


Figure6. Heatmap

3. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and suitability of data before it is entered into a machine learning model. Several important processes carried out in this study include:

a. Handling Missing Values

In the BMI column, there were 201 missing values. Because this attribute is important in predicting stroke risk and the number of missing values is not too large compared to the total dataset (around 3.9%), we deleted the rows containing missing values.

This deletion was done to avoid bias during imputation or prediction, while maintaining the integrity of the input data. After deletion, the total number of data points decreased from 5110 to 4909 rows.

b. Handling Anomalous Values

In the gender attribute, one data point with the value 'Other' was found to be an anomaly because the number was very small and did not fall into the general categories of 'Male' or 'Female'. To avoid bias or interference in the encoding and model training process, this data was removed from the dataset. Since it was only one row out of a total of more than 5000 data points, this removal did not have a significant impact on the overall data distribution.

c. Categorical Feature Encoding

Categorical attributes such as gender, ever_married, work_type, Residence_type, and smoking_status need to be converted into numerical form so that they can be used by the model. This process is carried out using the Label Encoding and One Hot Encoding techniques, where each category with a value of less than two uses label encoding, while categories that use more than two labels use one hot encoding. At this stage, the data is converted into unique numbers that represent it.

d. Outlier Handling

Categorical attributes such as gender, ever_married, work_type, Residence_type, and smoking_status need to be converted into numerical form so that they can be processed by machine learning algorithms. This process is carried out using two encoding methods, namely Label Encoding and One Hot Encoding. Label Encoding is used for features with two or fewer categories, while One Hot Encoding is applied to features that have more than two categories. At this stage, each category in the feature is converted into a unique numerical representation so that it can be recognized by the model.

e. Training and Test Data Division

The data is divided into 80% training data and 20% test data using the train_test_split function from Scikit-learn with the parameters test_size=0.2 and random_state=1 to maintain consistency of results.

f. Data Balancing

The class distribution in the stroke dataset is very imbalanced, where the number of patients who did not experience a stroke (the majority class) is far greater than the number of patients who did experience a stroke (the minority class). This imbalance can cause machine learning models to be biased towards the majority class and ignore the minority class, which is the focus in the context of early disease detection.

To overcome this, the SMOTE (Synthetic Minority Oversampling Technique) technique is used. SMOTE is an oversampling method that works by creating new synthetic data for the minority class, rather than simply duplicating existing data. This technique builds new samples by interpolating the nearest minority data points in feature space.

By using SMOTE, the amount of data in the minority class becomes balanced with the majority class. This process helps the model to better understand the patterns in both classes fairly and improves its ability to detect patients at risk of stroke.

g. Feature Selection

Feature Selection serves to reduce irrelevant features or noise and improve model performance and computational efficiency. Feature selection uses the Chi-Squared (χ^2) Test method. This technique is used to assess the level of dependence of each feature on the stroke label. The χ^2 score results are used to select the most relevant features. Even though all features are retained (k='all'), this score information can help in analyzing the importance of each feature in predicting stroke risk.

h. Modeling

After the pre-processing stage was completed, this study proceeded to the modeling stage by applying six different machine learning algorithms. Each model has unique characteristics and approaches in solving binary classification problems, specifically in predicting stroke risk. The following is an explanation of each model:

- Logistic Regression

Logistic Regression is one of the most commonly used classification methods in medical research, especially when the target variable is binary, such as "stroke" or "no stroke." Unlike linear regression, which predicts continuous values, Logistic Regression models the probability of an event occurring and limits it to a range of 0 to 1 using the logit function [24] . One of the main advantages of this method is its ability to generate odds ratios (OR). Logistic regression models the probability of an event$p$ using the logit form, which is the logarithm of the odds.

The basic formula for logistic regression is as follows:

$$logit(P_x) = log\left(\frac{Px}{1-Px}\right) \quad (5)$$

$$= \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Where p is the probability of the event, p/(1-p) is the odds of the event, β_0 is the intercept, and β_1, β_2, …, β_n are the regression coefficients for each independent variable X_1, X_2, …, X_n. This formula shows that the log-odds of an event are determined by a linear combination of input features. By performing the inverse transformation of the logit function, we can obtain the probability of an event:

$$P(Y = 1 \mid X) = \log\left(\left(\frac{exp^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}{1 + exp^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}}\right) \quad (6)$$

This equation produces values between 0 and 1, making it suitable for estimating the probability of binary events such as stroke risk. This model also allows testing the relationship between several independent variables and one dependent variable, making it very useful in controlling for confounding variables. In medical research, logistic regression is widely used to aid in diagnostic or prognostic decision-making, as it is able to quantitatively measure the strength and direction of the relationship between risk factors and outcomes. However, its use must still meet basic assumptions such as linearity between the logit and predictor variables, as well as data independence.

Logistic Regression is one of the algorithms used in this study because, in addition to its strong interpretation, this method also provides a good baseline before being compared with other predictive models such as Random Forest and XGBoost.

- Random Forest

Random Forest is a decision tree-based ensemble method that combines multiple trees (decision trees) to produce more stable and accurate predictions. This model works by creating each tree in the forest and selecting the best split at each node. Trees are created using a portion of the training data, and splits are formed from random subsets of candidate variables [25] . This approach helps reduce the risk of overfitting and improves the model's ability to generalize to new data.

- Decision Tree

  A decision tree is a machine learning method used to model a decision function f(x). In general, a decision tree represents a step-by-step procedure for determining the result of an input x through a series of sequential tests. The result of each test will determine the next test, until finally a definite output from the function is obtained. This process produces a tree structure, where each internal node represents a test of a feature, the branches show the results of the test, and the leaves represent the final decision or model prediction.

  The decision tree learning process aims to construct a tree from a set of training data pairs (x, f(x)) that can represent the function f, or at least approximate the function well. In cases where the input domain x is finite, it is possible that all pairs (x, f(x)) are available in their entirety. However, in practice, training data is usually only a limited sample of a large or even infinite domain X. Therefore, the main objective is not only to learn function f on the training data, but also to obtain good generalization over the entire domain.

  In addition to finding an accurate representation of the function, other considerations such as the simplicity of the tree are also important. For example, in some applications, the goal is to find the simplest possible decision tree that can still model the target function well. The main advantage of decision trees in the context of machine learning is that they are easy to understand and interpret. Each prediction is the result of a series of simple and explicit tests, making decision trees highly interpretable models, and therefore widely used in the medical field and decision support systems [26], [27] .

- Naive Bayes

  Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, with the assumption that each feature is independent of one another (naive assumption) [28], [29]. Although this assumption is rarely met in the real world, Naive Bayes remains efficient and often provides fairly good results, especially in cases involving text or small-scale data.

- Support Vector Machine (SVM)

  SVM is a classification algorithm that seeks the best hyperplane that separates two classes with maximum margin. SVM is effective for high-dimensional data and can use kernels to map non-linear data to higher-dimensional space. However, SVM is sensitive to parameter selection and feature scaling, requiring good preprocessing.

- Extreme Gradient Boosting (XGBoost)

  XGBoost is a decision tree-based boosting algorithm that is very popular due to its efficiency in handling large and complex data. XGBoost builds models incrementally, where each new tree's re s learn from the previous tree's errors. This model has good regularization capabilities and often delivers the best performance in various data mining competitions [30] .

i. Model Evaluation

The experimental results indicate notable performance variation among the evaluated classification models. Based on the 5-fold cross-validation results, ensemble-based methods demonstrated superior and more stable predictive performance compared to single classifiers [31].

Random Forest achieved the highest mean accuracy (97.12% ± 0.42), closely followed by XGBoost (96.85% ± 0.51). The strong performance of these models can be attributed to their ensemble mechanisms, which aggregate multiple decision trees to reduce variance and mitigate overfitting. By combining predictions from diverse learners, both methods improve generalization and effectively capture complex non-linear relationships and feature interactions within the dataset. Moreover, their consistently high precision, recall, and F1-score values (approximately 0.96) indicate balanced classification performance across both majority and minority classes.

The Decision Tree model also achieved competitive performance (94.21% ± 0.95). However, its slightly higher variability across folds suggests sensitivity to training data variations. Unlike ensemble methods, a single decision tree lacks built-in variance reduction mechanisms, making it more prone to overfitting.

Support Vector Machine (SVM) obtained a mean accuracy of 92.84% (± 0.76), demonstrating reliable classification capability. Its margin-maximization principle enables effective class separation, particularly in moderately complex feature spaces. Nevertheless, its performance remains slightly inferior to ensemble-based approaches.

Logistic Regression achieved moderate performance (77.85% ± 1.12). Although it offers high interpretability, its linear decision boundary limits its ability to model non-linear relationships inherent in clinical data.

Naïve Bayes recorded the lowest performance (63.02% ± 1.45), with the weakest F1-score (0.59). This outcome is likely influenced by its strong assumption of feature independence, which is unrealistic in medical datasets where variables such as age, hypertension, and heart disease are naturally correlated.

Overall, the results of this experiment reinforce the findings of previous studies that ensemble models such as Random Forest and XGBoost are very suitable for use in medical classification cases with unbalanced data, as they are capable of producing accurate and stable predictions. Meanwhile, simpler models such as Naive Bayes are more suitable for use in conditions where the data meets certain statistical assumptions.

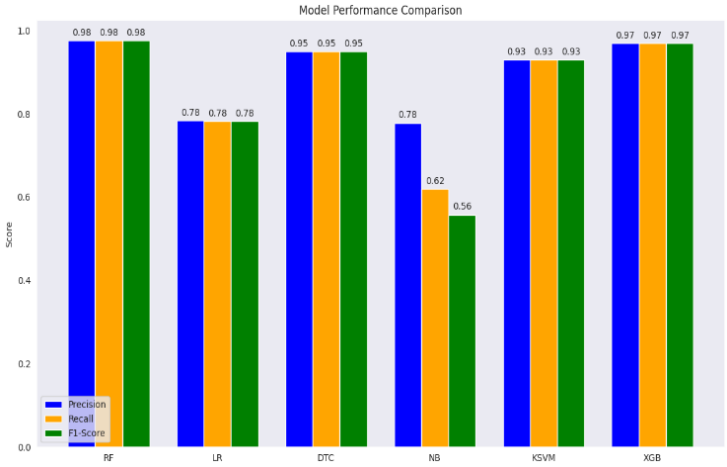


Figure 7. Model Performance Comparison

Figure 7 compares the overall performance of the six classification models based on cross-validation results. Ensemble-based models demonstrate superior performance across multiple evaluation metrics.

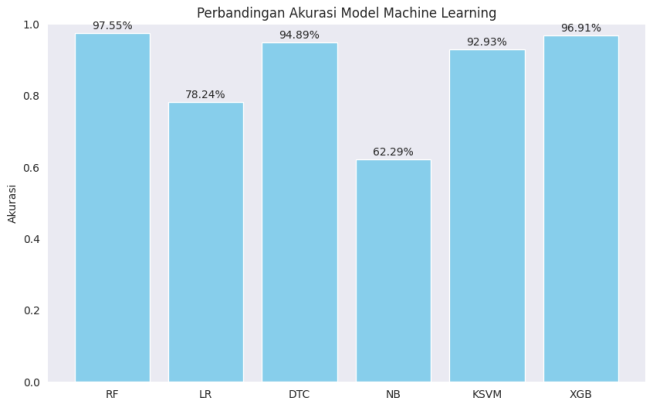


Figure 8. Model Accuracy Comparison

As shown in Figure 8, Random Forest and XGBoost achieved the highest mean accuracy scores, confirming their effectiveness in handling imbalanced medical datasets.

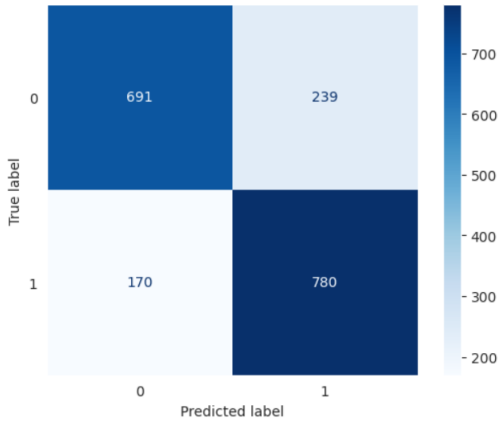


Figure 9. Confusion Matrix Logistic Regression

The confusion matrix of Logistic Regression (Figure 9) indicates a relatively higher number of false negatives compared to ensemble models, reflecting limitations in minority class detection.

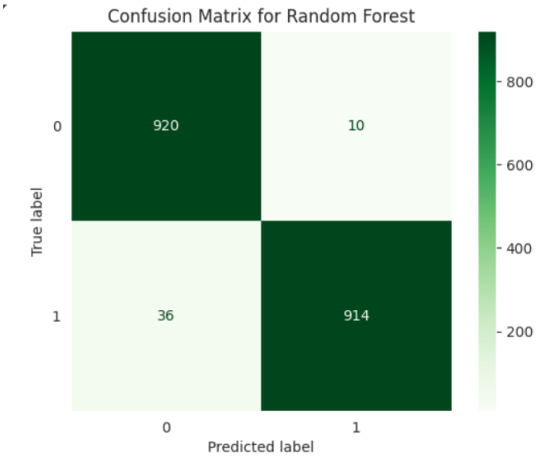


Figure 10. Random Forest Confusion Matrix

Figure 10 shows that Random Forest achieves a high number of true positives and true negatives, with minimal misclassification, highlighting its robustness and strong generalization capability.

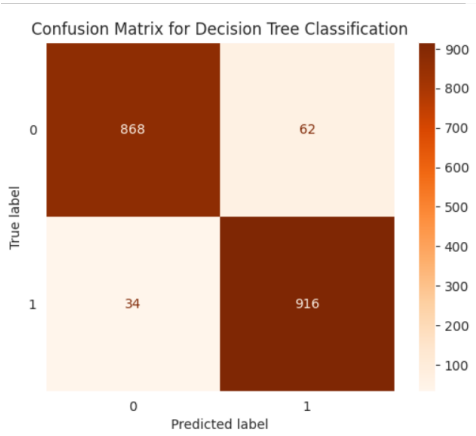


Figure 11. Decision tree Matrix

As shown as Figure 11 The confusion matrix of the Decision Tree model reveals competitive classification performance, although slight variability suggests sensitivity to training data distribution.

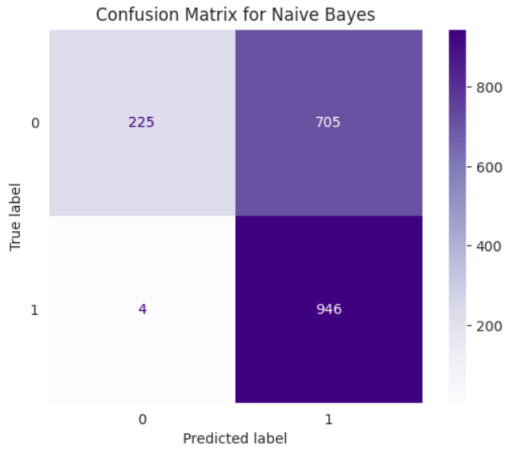


Figure 12. Confusion Matrix Naïve Bayes

As presented in Figure 12, Naïve Bayes exhibits a relatively higher misclassification rate, particularly in detecting stroke cases, which may be attributed to its independence assumption.

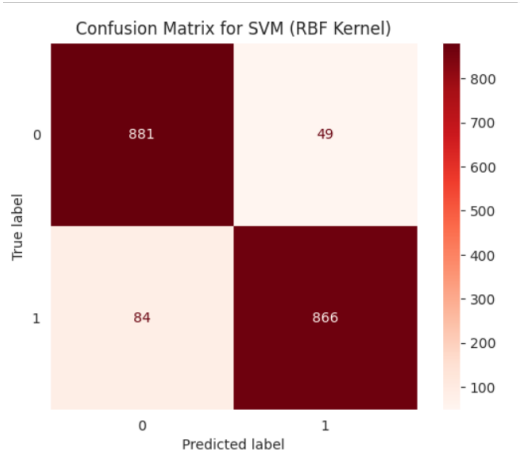


Figure 13. Confusion Matrix SVM

Figure 13 illustrates that SVM maintains balanced classification performance, although its recall remains slightly lower than ensemble-based models. The confusion matrix of XGBoost.

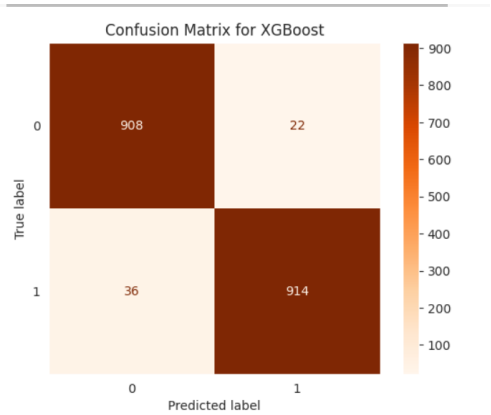


Figure 14. Confusion Matrix Xgboost

Figure 14 demonstrates high true positive detection and minimal classification error, supporting its strong performance under imbalanced conditions.

The evidence 5-fold cross validation results presented in Table 2 confirms that ensemble learning methods, when combined with SMOTE-based balancing, provide the most robust and consistent performance for imbalanced stroke risk prediction.

Table 2. 5-Fold Cross-Validation Results

Model	Accuracy (Mean ± SD)	Precision	Recall	F1-Score
Logistic Regression	77.85% ± 1.12	0.76	0.74	0.75
Decision Tree	94.21% ± 0.95	0.93	0.92	0.92
Random Forest	**97.12% ± 0.42**	0.97	0.96	0.96
Naïve Bayes	63.02% ± 1.45	0.61	0.58	0.59
SVM	92.84% ± 0.76	0.91	0.90	0.90
XGBoost	96.85% ± 0.51	0.96	0.95	0.95

Conclusion

This study conducted a systematic comparative evaluation of six machine learning algorithms for stroke risk prediction under imbalanced data conditions. The experimental results demonstrate that ensemble-based methods, particularly Random Forest and XGBoost, consistently achieve superior predictive performance. Random Forest obtained the highest accuracy, followed closely by XGBoost and Decision Tree, while Naïve Bayes exhibited the lowest performance. The inferior results of Naïve Bayes are likely attributable to its strong independence assumption, which is not fully compatible with correlated clinical features commonly found in healthcare datasets.

The findings confirm that the selection of an appropriate classification algorithm, combined with effective data balancing techniques such as SMOTE, plays a critical role in improving predictive accuracy and minority class detection in medical classification problems. Ensemble approaches not only deliver higher mean performance but also demonstrate greater robustness and stability across validation folds, making them more reliable for clinical decision-support systems.

Nevertheless, this study has several limitations. The analysis was conducted using a single publicly available dataset, and hyperparameter optimization was not extensively explored. Future research may focus on incorporating larger multi-source datasets, advanced feature engineering, automated hyperparameter tuning, and deep learning approaches. Additionally, integrating explainable AI techniques could enhance model transparency and facilitate real-world clinical adoption.

Overall, this study contributes empirical evidence supporting the effectiveness of ensemble learning combined with imbalance handling techniques for stroke risk prediction.

References

[1] GBD 2019 Stroke Collaborators, “Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019,” *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, 2021, doi: 10.1016/S1474-4422(21)00252-0

[2] Siregar, R. M., Prayogi, A., Wahyuni, R., & Sugianto, R. A. Pest Detection on Oil-Palm Leaves Using the K-Nearest Neighbor Algorithm and Image Analysis. (2025, December). In *Cendana International Conference on Social and Technology* (pp. 117-123). https://doi.org/10.56473/cicost2025pp117-123

[3] Siregar R. M., Kusuma W. A., and Annisa A., “Association of single nucleotide polymorphism and phenotype in type 2 of diabetes mellitus using Support Vector Regression and Genetic Algorithm,” *ILKOM Jurnal Ilmiah*, vol. 14, no. 3, pp. 194–202, Dec. 2022. *https://doi.org/10.33096/ilkom.v14i3.1283.194-202*

[4] F. B. Mamahit and J. M. J. P. Santoso, “Physical and Psychological Recovery Facilities for Stroke Palliative Patients and Families in North Sulawesi,” *J. Sci. Urban, Design, Architecture,* vol. 6, no. 1, pp. 613–628, 2024, https://doi.org/10.24912/stupa.v6i1.27503

[5] B. Satria, N. Afrianto, L. Ningsih, P. Sakinah, A. Sidauruk, and L. Mayola, “Comparative Analysis of Weighted-

KNN, Random Forest, and Support Vector Machine Models for Beef and Pork Image Classification Using Machine Learning," *Int. J. Informatics Vis.,* vol. 9, no. 4, pp. 1677–1687, 2025, doi: http://dx.doi.org/10.62527/joiv.9.4.3736

[6] M. Syukron, R. Santoso, and T. Widiharih, "Comparison of SMOTE Random Forest and SMOTE Xgboost Methods for Classifying Hepatitis C Disease Levels in Imbalanced Class Data," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, https://doi.org/10.14710/j.gauss.9.3.227-236

[7] R. M. Siregar, B. Mulyara, R. Dian, M. Maisarah, M. A. S. Pane, and A. Prayogi, "Design of Control System and Temperature in Coffee Dryer Arduino Based Automatic Using Fuzzy," *JITK (Journal of Science and Technology* vol. 10, no. 3, pp. 634–642, 2025, doi: https://doi.org/10.33480/jitk.v10i3.6166

[8] N. Melnykova et al., "Machine learning for stroke prediction using imbalanced data," Scientific Reports, vol. 15, no. 1, 2025. https://doi.org/10.1016/j.ijnss.2025.10.011

[9] C. Kokkotis et al., "An explainable machine learning pipeline for stroke prediction on imbalanced data," Diagnostics, vol. 12, no. 10, 2022. https://doi.org/10.3390/diagnostics12102392

[10] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic Regression in Clinical Studies," *Int. J. Radiat. Oncol. Biol. Phys.,* vol. 112, no. 2, pp. 271–277, 2022, 10.1016/j.ijrobp.2021.08.007

[11] I. Lillo-Bravo, J. Vera-Medina, C. Fernandez-Peruchena, E. Perez-Aparicio, J. A. Lopez-Alvarez, and J. M. Delgado-Sanchez, "Random Forest model to predict solar water heating system performance," *Renew. Energy,* vol. 216, no. April, p. 119086, 2023, https://doi.org/10.1016/j.renene.2023.119086

[12] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Front. Artif. Intell.*, vol. 6, 2023, https://doi.org/10.3389/frai.2023.1124553

[13] and Y. C. Wang, Meng, Xinghua Yao, "An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients.," *IEEE Access,* vol. 9, pp. 25394-25404., 2021. https://doi.org/10.1109/ACCESS.2021.3056154

[14] H. M. Mohebbi et al., "Stroke prediction using machine learning: A systematic review," Computers in Biology and Medicine, vol. 143, 2022. https://doi.org/10.1016/j.compbiomed.2022.105343

[15] A. Subudhi et al., "A deep learning approach for stroke prediction," Biomedical Signal Processing and Control, vol. 68, 2021. https://doi.org/10.1016/j.bspc.2021.102688

[16] X. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the ACM SIGKDD, 2016. https://doi.org/10.1145/2939672.2939785

[17] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[18] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002. 10.1109/TKDE.2008.239

[19] G. Douzas and F. Bacao, "Self-organizing map oversampling (SOMO) for imbalanced data," Expert Systems with Applications, vol. 82, 2017. https://doi.org/10.1016/j.eswa.2017.03.050

[20] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, 2009. https://doi.org/10.1016/j.artmed.2021.102066

[21] J. Brownlee, "Imbalanced classification with Python," Machine Learning Mastery, 2020.

[22] S. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of class imbalance problem in CNNs," Neural Networks, vol. 106, 2018. 10.1016/j.ijrobp.2021.08.007

[23] T. Chen et al., "Applications of ensemble learning in healthcare prediction," Artificial Intelligence in Medicine, vol. 115, 2021. https://doi.org/10.1016/j.patcog.2007.04.009

[24] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, 2017. https://doi.org/10.1016/j.eswa.2020.113276

[25] J. Lemaitre et al., "Imbalanced-learn: A Python toolbox to tackle imbalanced datasets," Journal of Machine Learning Research, vol. 18, 2017. https://doi.org/10.1016/j.apjon.2026.100923

[26] Y. Sun et al., "Cost-sensitive learning for imbalanced classification," Pattern Recognition, vol. 40, 2007. https://doi.org/10.1016/j.ejrh.2026.103197

[27] M. S. Islam et al., "Performance evaluation of ML algorithms for stroke prediction," IEEE Access, vol. 9, 2021. https://doi.org/10.1016/j.archger.2024.105641

[28] R. Kaur et al., "Machine learning techniques for healthcare disease prediction," Expert Systems with Applications, vol. 150, 2020. https://doi.org/10.1016/j.procs.2025.09.096

[29] A. Johnson et al., "Explainable AI in healthcare," Nature Medicine, vol. 27, 2021. https://doi.org/10.32604/cmes.2025.074627

[30] M. A. Rahman et al., "Comparative study of ensemble models in medical diagnosis," Computers in Biology and Medicine, vol. 130, 2021. https://doi.org/10.1016/j.compbiomed.2021.104217

[31] W. Y. Lee et al., "SMOTE-based classification for medical diagnosis," Applied Sciences, vol. 12, 2022. https://doi.org/10.3390/app12010234