IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 3, March, 2026, pp. 583-594
Journal homepage: http://iieta.org/journals/mmep

A Hybrid Sampling Approach for Classification of Imbalanced Health Data Using Synthetic Minority Oversampling Technique and Edited Nearest Neighbor

Fifin Ayu Mufarroha[1*], Eka Mala Sari Rochman[1], Aeri Rachmad[1], Fitriyatul Qomariyah[2], Yuli Panca Asmara[3]

[1] Department of Informatics, Faculty of Engineering, University of Trunodjoyo Madura, Bangkalan 69162, Indonesia
[2] Mathematics Education Program, Tarbiyah Faculty, Madura State Islamic University, Pamekasan 69371, Indonesia
[3] Faculty of Engineering and Quantity Surveying, INTI International University, Negeri Sembilan 71800, Malaysia

Corresponding Author Email: fifin.mufarroha@trunojoyo.ac.id

https://doi.org/10.18280/mmep.130313

Received: 7 December 2025
Revised: 7 March 2026
Accepted: 15 March 2026
Available online: 10 April 2026

Keywords:
human disease, stroke classification, imbalanced medical data, Random Forest, Synthetic Minority Oversampling Technique, SMOTE–ENN, minority class detection, innovation

ABSTRACT

Human disease remains a central challenge in global health, encompassing a wide spectrum of conditions that affect populations across all regions and socioeconomic levels. Stroke is a leading cause of death and disability worldwide, making early detection a crucial aspect of its prevention. Machine learning methods applied to medical record data have the potential to support this process, but they often face the problem of class imbalance. This condition can bias the model towards the majority class and reduce its ability to detect the most clinically critical stroke cases. This study evaluates the effectiveness of sampling techniques in improving minority-class detection using the Random Forest model. Three approaches were compared: Random Forest without sampling, Random Forest with the Synthetic Minority Oversampling Technique (SMOTE), and Random Forest with the hybrid SMOTE-Edited Nearest Neighbors (SMOTE-ENN). Model performance was evaluated using an independent test set with accuracy, sensitivity, specificity, precision, F1-score, area under the receiver operating characteristic curve (AUC), and Matthews correlation coefficient (MCC). The results show that Random Forest without sampling achieved high accuracy but failed to detect stroke cases effectively, with a near-zero sensitivity. The application of SMOTE and SMOTE-ENN significantly improved sensitivity, with the best performance achieved by SMOTE-ENN, reaching a sensitivity of up to 0.878 and an AUC of around 0.826 on the test set. This study contributes by providing a systematic and robust evaluation framework that emphasizes generalization performance and statistically validated model comparison, thereby representing a methodological innovation in addressing common pitfalls in imbalanced medical data analysis. These findings demonstrate that hybrid sampling provides a better balance between detecting minority cases and maintaining overall model stability, highlighting the effectiveness of SMOTE-ENN for improving stroke detection in imbalanced datasets.

1. INTRODUCTION

A stroke is a medical emergency caused by a disruption in blood flow to the brain. This condition can occur due to a blockage or rupture of a blood vessel, cutting off the supply of oxygen and nutrients to brain tissue. As a result, the affected brain can lose function due to damage to nerve cells. When oxygen and nutrients are not delivered, brain tissue begins to die, resulting in the cessation of functions controlled by that area [1]. As one of the most serious global health problems, stroke continues to show an increasing trend. Stroke is the leading cause of death and permanent disability in adults in middle- to high-income countries. Various risk factors, such as age, gender, genetics, smoking, and comorbidities like hypertension and diabetes, contribute to stroke [2, 3].

One way to identify the early risk of stroke is through early detection. This is crucial to help the public recognize risk early and reduce the prevalence of stroke-related deaths. This effort involves building predictive models capable of identifying the likelihood of stroke based on available medical data [4]. Early stroke detection using artificial intelligence (AI) approaches has shown potentially good results. However, the application of AI-based decision support systems in healthcare raises concerns. These concerns concern model bias, reliability, and ethical responsibility, particularly when trained on imbalanced medical data. These risks emphasize the need for transparent and ethically grounded AI models capable of robust generalization [5]. Therefore, the Random Forest method was chosen due to its reliability in diagnosing diseases, including stroke. As an ensemble learning algorithm, Random Forest works by combining several models to improve prediction performance. Previous research has shown that Random Forest provides high accuracy, reaching 96.67%, and outperforms the XGBoost method [6]. Another study in 2020

using fever symptom data from the University of California, Irvine (UCI) Machine Learning also showed that Random Forest produced the highest accuracy, at 84.30%, compared to Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and C4.5 [7]. However, the main challenge in the classification process is class imbalance, a condition where the amount of data in the majority class is much larger than the amount in the minority class [8, 9].

In the context of classification with datasets experiencing class imbalance, selecting the right resampling technique is very important to optimize model performance, especially for clinically crucial minority classes such as stroke cases. Sampling techniques such as oversampling, undersampling, and hybrid sampling are commonly used to balance class distributions [10]. Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique based on the K-Nearest Neighbor (KNN) algorithm that has been shown to improve accuracy in several studies, for example, in hepatitis C classification using Random Forest [11]. To address class imbalance in medical datasets, resampling techniques such as SMOTE and hybrid approaches like SMOTE-ENN have been widely adopted. These methods improve minority class representation and reduce noise, leading to more robust classification performance, particularly in tabular healthcare data [12]. However, overfitting can occur when using the SMOTE technique itself [13].

The Edited Nearest Neighbor (ENN) method is an undersampling technique that removes samples from the majority class that have labels different from those of their neighbors. ENN works in contrast to the SMOTE oversampling technique. This technique has been used in research using Logistic Model Tree Forest to predict steel plate damage, demonstrating an accuracy of up to 86.65% [14]. Another sampling technique, SMOTE-ENN, employs a hybrid resampling approach that combines two main strategies: synthetic oversampling with SMOTE and noisy sample cleaning using ENN. SMOTE-ENN not only improves the representation of the minority class but also removes inconsistent or ambiguous samples from both classes. The result of this technique is a cleaner feature space on which the model can be trained more effectively. In 2020, research was conducted on addressing class imbalance in medical datasets. This study compared oversampling, undersampling, and hybrid resampling methods. The study found that the Random Forest M-SMOTE-ENN (RFMSE) method performed best with an MCC of 99.0% and a specificity of 100% [15]. Other studies, such as those on the UCI Machine Learning heart failure dataset, demonstrated that Random Forest combined with SMOTE-ENN achieved 90% accuracy [16]. SMOTE-ENN was also applied to heart failure data from Shanxi Province, China, demonstrating improved diagnostic performance [17]. Furthermore, comparative studies across various classification domains have shown that this hybrid resampling method performs remarkably well in improving minority class separation and training data distribution without increasing overfitting. According to the study, the hybrid resampling method SMOTE-ENN combined with the stacking classifier consistently produced a Receiver Operating Characteristic–Area Under the Curve (ROC–AUC) of 99.6%, accuracy, and f1-score of up to 97.9% [18].

In addition to data balancing, feature selection is also important in improving classification accuracy. The purpose of feature selection is to eliminate irrelevant attributes or features, simplify the data structure, reduce the computational burden, and retain features that truly contribute to classification. This will indeed add to the research stage, but it is important to do so so that the model can find its best version in predicting classes. The main resampling methodology used in this study is the SMOTE-ENN method. SMOTE-ENN uses gain ratio as a feature selection method along with synthetic oversampling and noise sampling to improve class distribution and reduce uncertainty in the model's decision boundary. This combination allows the model to learn and process effectively. Random Forest was chosen as the classification model for predicting stroke due to its proven high performance [7, 19]. With this approach, it is hoped that this study can make a significant contribution to supporting early detection of stroke risk in imbalanced data.

2. METHODOLOGY

The entire research process is illustrated in Figure 1. The process begins with partitioning the dataset into training and testing sets to facilitate model development and evaluation. This separation allows the model to be trained and subsequently assessed on unseen data to ensure its generalization capability. Following the data partitioning stage, a series of preprocessing steps was carried out on the training data to enhance data quality and suitability for modeling. These steps included data selection, categorical coding, missing value imputation using the KNN method, outlier handling, and data normalization. The purpose of this stage is to improve feature representation and ensure comparability across variables. The transformation parameters obtained from this stage were then consistently applied to the testing data to maintain alignment between both datasets. After preprocessing, feature selection was conducted using the gain ratio to identify the most relevant and influential features. To address class imbalance, a hybrid sampling approach combining SMOTE and ENN was applied to the selected training features. In the classification stage, the Random Forest method is implemented to perform the classification task. The trained model was subsequently evaluated using the testing data to assess its performance on unseen samples. Finally, the model was used to generate class predictions.

2.1 Data imputation

Incomplete data found or missing values in a dataset can be processed using three methods: (1) deletion, (2) learning without handling missing values, and (3) data imputation. Data imputation is a technique used to replace any missing values using statistical methods, such as the mean and mode. Machine learning-based techniques can also be used [20]. K-NN Imputer is a simple and effective imputation method for handling missing values, making it frequently used for various prediction problems. The advantage of the KNN imputation method is its ability to predict two types of data: discrete data (using the mode) and continuous data (using the mean). The K-NN Imputer replaces missing values by analyzing nearest neighbors. The Euclidean distance matrix is then implemented by calculating two points in multidimensional space [21]. The distance is computed using Eqs. (1) and (2).

$$D(x,y) = \sqrt{w \times \sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

$$w = \frac{total\ features}{the\ number\ of\ features\ that\ have\ a\ value} \quad (2)$$

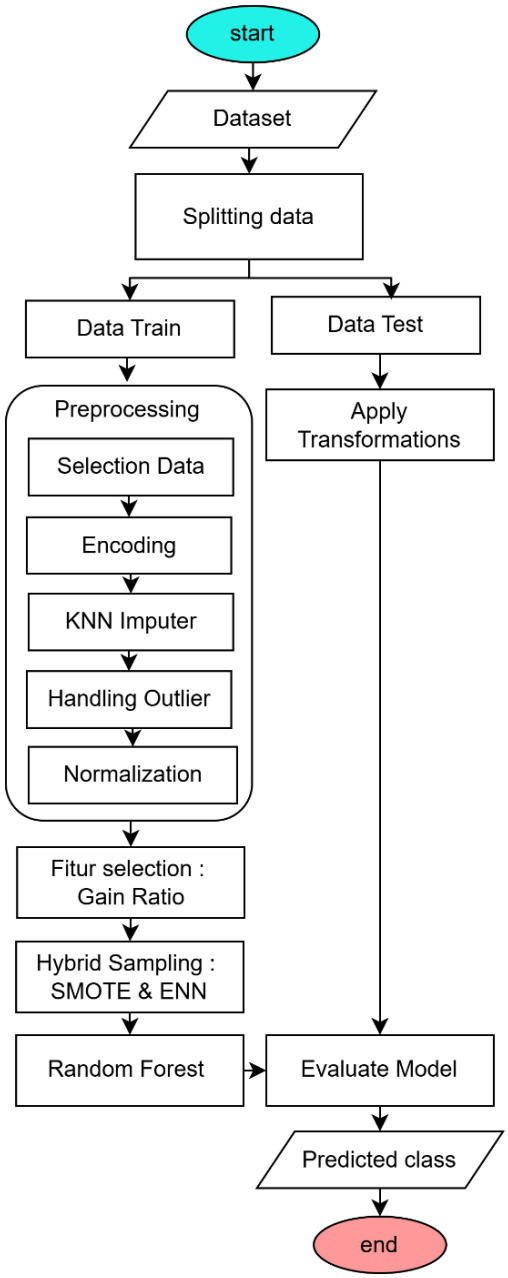


Figure 1. SMOTE-ENN Random Forest flowchart
Note: KNN: K-Nearest Neighbor, SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors.

2.2 Undersampling

Undersampling is a technique used to address class imbalance by reducing the number of majority samples. ENN is one such undersampling approach, which uses the nearest neighbor rule. The ENN works by eliminating samples whose class differs from the majority class of their nearest neighbors. The primary goal of this algorithm is to eliminate the majority of data that indicates noise [22]. The ENN algorithm uses the nearest neighbor rule by identifying the closest samples from each majority sample based on the distance between the samples. Then, it continues by determining whether the majority sample is noise or not by checking whether its label is consistent with the sample. If the label of the majority sample is inconsistent with the majority label of its neighboring samples, then the sample is considered noise and is removed from the dataset.

The nearest neighbor rule (KNN) of sample S_i is the number of samples in dataset S whose distance from S_i is smaller than or equal to the distance between S_i and its k-th nearest neighbor, which can be seen in Eq. (3).

$$KNN(S_i, k) = \{S_j \in S \mid dist\,(S_j, S_i) \leq dist\,(S_i', S_i)\} \quad (3)$$

where, $KNN(S_i, k)$ represents the set of nearest neighbors of a sample S_i in a dataset S. The parameter k is the number of nearest neighbors used in the classification process. The entire dataset is denoted as S. Each data point in S is represented as S_i, which is a particular sample that is being evaluated or whose class is to be predicted. Meanwhile, S_j is another sample in the dataset that is one of the nearest neighbors of S_i. In some cases, nearest neighbors can also be ordered by their distance. Therefore, S_i' is used to denote other nearest neighbors. The calculation of the proximity between samples is determined by the distance function. The notation $dist(S_j, S_i)$ indicates the distance between two samples, namely S_j and S_i, while $dist(S_i', S_i)$ indicates the distance between other nearest neighbor samples and S_i.

2.3 Oversampling

Unlike undersampling, oversampling does not handle the majority of samples, but rather, this approach is used to increase samples within the minority class. SMOTE, or the abbreviation for Synthetic Minority Over-Sampling Technique, is a part of the oversampling technique. This approach is carried out by creating new minority samples by using linear interpolation on randomly selected neighboring samples. This is done to improve the ability to recognize minority samples. SMOTE works by randomly selecting a group of x_i, long with its neighbors. New data sample examples can be generated using the following Eq. (4).

$$x_{new} = x_i + (x_i - x_i')\delta \quad (4)$$

where, x_{new} represents new synthetic data generated through the oversampling process. Meanwhile, x_i is the original sample from the minority class and is used as a reference point. x_i' is one of x_i's nearest neighbors, randomly selected from the KNN search. The variable δ is a random parameter with a value between 0 and 1 that functions as an interpolation factor.

2.4 Hybrid sampling

Hybrid sampling is an approach that combines under-sampling and over-sampling methods. This method is used to reduce the number of majority samples and increase the number of minority samples. Conceptually, SMOTE creates additional samples from the minority class to correct underrepresentation, but oversampling alone may introduce noise and synthetic data points that overlap with the majority class. ENN is then used to delete occurrences in ambiguous decision areas, both in the majority and minority classes. This combination enhances class separability while reducing ambiguity at the decision boundary, allowing models such as Random Forest to properly identify and separate minority classes [23]. The operational flow of the SMOTE-ENN procedure is illustrated in Figure 2 and Algorithm 1, which

provides a visual representation of the oversampling and noise removal stages. The stages of SMOTE-ENN are as follows.

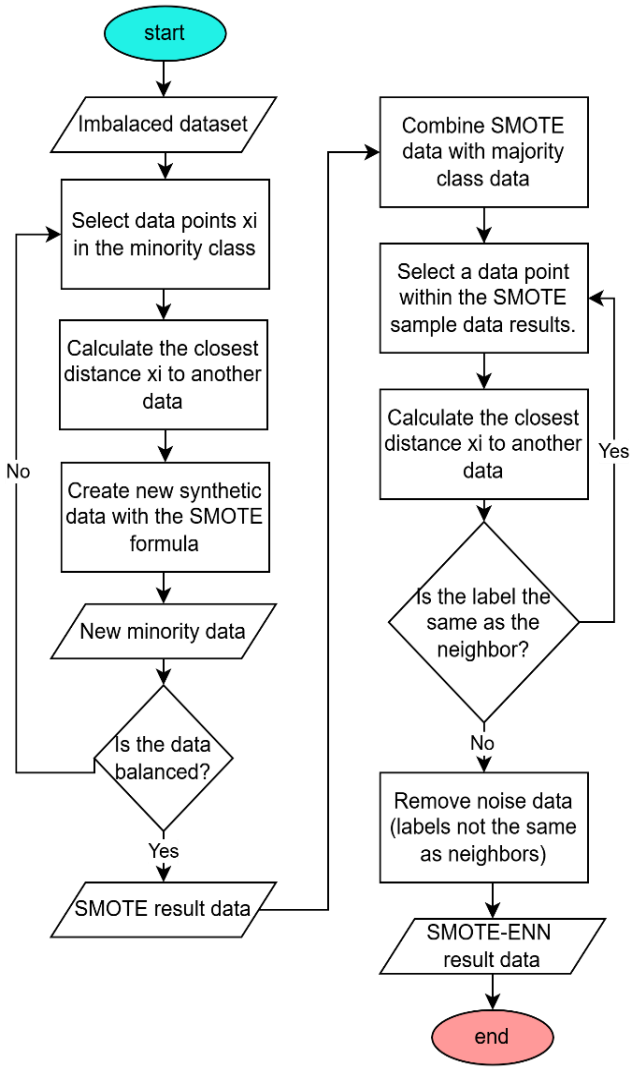


Figure 2. SMOTE-ENN flow
Note: KNN: K-Nearest Neighbor, SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors.

Algorithm 1. Hybrid SMOTE–ENN Sampling Procedure

Input: Dataset: X, minority sample $x_{i_{min}}$ with $i = 1,2,\dots,N$, majority sample $x_{j_{maj}}$, with $j = 1,2,\dots,M$.

Output: Data sampling results

Steps:

1. Set the oversampling rate (IR) based on the degree of sample imbalance.
2. For each $l = 1,2\dots,N$,
   - Calculate the distance $x_{i_{min}}$ to all minority samples using the Euclidean distance, and obtain the k_1 nearest neighbor samples $x_{ik_{1min}}$.
3. For each $l = 1,2\dots,$ IR,
   - For each $x_{i_{min}}$, randomly select a number of k_1, nearest neighbor samples, assuming the selected nearest neighbor is $x_{ik_{1min}}$.
   - For each randomly selected nearest neighbor $x_{ik_{1min}}$, synthesize a new minority sample x_{new} from the minority sample $x_{i_{min}}$ according to Eq. (1).
   - Add a newly synthesized minority sample x_{new} to the original minority sample.
4. For each $l = 1,2\dots,$ IR,
   - For each $x_{j_{maj}}$, calculate the distance between $x_{j_{maj}}$ and the majority sample using the Euclidean distance, and obtain the k2 nearest neighbor samples $x_{ik_{2maj}}$.
   - For each majority sample $x_{j_{maj}}$, select three nearest neighbor samples from its k2 nearest neighbors, assuming the nearest neighbor samples selected are $x_{ik_{2maj}}$.
   - For each majority sample $x_{j_{maj}}$ determine whether it is a noise sample. If it is a noise sample, then delete $x_{j_{maj}}$.
   - Remove the noise sample from the majority sample.

2.5 Feature selection

Gain ratio is a modified method of feature selection using the information gain method, which reduces the level of bias [24]. In feature selection using the gain ratio, improvements are made by considering the intrinsic information of the attributes. Gain ratio also has the ability to correct data instability, making it more suitable for two-class numerical data. This approach is simple, resulting in faster computation. In this work, the gain ratio was used to choose features on the training data before the sampling procedure. This was done to optimize data representation before sampling. Furthermore, the sampling method became more efficient and focused on relevant information. The following steps can be used to calculate the gain ratio:

Step 1: Calculate entropy using Eq. (5)

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i \times \log_2\ p_i \tag{5}$$

Step 2: Calculate the information gain using Eq. (6)

$$Information\ Gain\ (S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{6}$$

Step 3: Calculate SplitInfo using Eq. (7)

$$SplitInfo_A(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|} \tag{7}$$

Step 4: Calculate the gain ratio using Eq. (8)

$$Gain\ ratio\ (A) = \frac{Information\ Gain\ (A)}{Splitinfo\ (A)} \tag{8}$$

where, S denotes the dataset, p_i represents the proportion of data in the i-th category, n denotes the number of categories in the data set, A represents the attribute or feature to be evaluated, S_i denotes a subset of the dataset, $|S_i|$ denotes the number of data points in subset S_i, $|S|$ denotes the number of data in subset.

2.6 Random Forest classification

Random Forest is an ensemble classification method that creates a forest of decision trees. The majority vote of all decision trees is used to determine the class of the input data. The Random Forest method is capable of producing relatively low error rates with good classification performance [25, 26]. There are several stages in the Random Forest algorithm, namely*:*

- Take n random data samples from the dataset.
- Use these sample data to build the i-th tree up to k iterations.
- Repeat steps (1) and (2) k times to form a forest containing the previously constructed classification trees. Each classification tree produces one vote and one class.
- The final classification result is determined by taking the majority vote of the k votes formed.

3. RESULTS AND DISCUSSION

3.1 Experimental data

The dataset used in this study is stroke prediction data. This dataset comes from Oluwafemi Emmanuel Zachariah's thesis on "Stroke Prediction with Demographic and Behavioral Data Using the Random Forest Algorithm" from Sheffield Hallam University. It was compiled from health records from various hospitals in Bangladesh, which can be accessed through the following link. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/ [27].

The dataset comprises 5,110 records with 11 features and one binary label: 1 (stroke) and 0 (non-stroke), with 249 stroke and 4,861 non-stroke instances, and 201 missing values. Three key challenges are addressed in this study: the severe class imbalance between stroke and non-stroke cases, the presence of missing data requiring imputation, and the need for feature selection to identify variables with meaningful predictive impact on model accuracy.

3.2 Preprocessing result

The preprocessing stage aims to improve data quality and ensure that the dataset is suitable for subsequent analysis. Initially, a data selection process was conducted by removing irrelevant attributes, such as the "id" column, and eliminating records with invalid categorical values, specifically the “other” category in the gender variable. Next, one-hot encoding was applied to convert categorical features into binary values of 1 and 0. For example, the gender feature, which contained two categories (male and female), was split into two binary feature columns. As a result of this step, the original 11 features were expanded to 22 features, including 'age', 'avg_glucose_level', 'bmi', 'smoking_status_never smoked', 'hypertension_1', 'work_type_Private', 'Residence_type_Rural', 'Residence_type_Urban', 'ever_married_Yes', 'smoking_status_smokes', 'smoking_status_Unknown', 'gender_Female', 'gender_Male', 'work_type_Self-employed', 'heart_disease_0', 'heart_disease_1', 'hypertension_0', 'work_type_children', 'smoking_status_formerly smoked', 'ever_married_No', 'work_type_Never_worked', 'work_type_Govt_job'. Outlier detection and removal were then performed using the Interquartile Range (IQR) method. The Body Mass Index (BMI) and avg_glucose_level features were found to contain outliers, and data rows with values outside the IQR range were removed, reducing the overall dataset size.

An inspection of missing values revealed that, among all features, only the BMI column contained 201 missing entries. These were imputed using the KNN Imputation method with $k = 5$, enabling estimation based on the similarity of neighboring data points. The final preprocessing step involved normalization using the MinMax Scaler, which rescaled all feature values to the range [0, 1] by subtracting the minimum value of each feature and dividing by the difference between its maximum and minimum values. Table 1 summarizes the changes in data size across each preprocessing step.

Table 1. Information about data size changes

Data History	Data Update	Stroke	Non-Stroke	Data Deleted	Reason
Original data	5.110	249	4,861	–	Initial dataset
Drop Gender “other”	5.109	249	4,860	1	Invalid gender data
Eliminate outlier BMI	4.999	247	4,752	110	BMI outlier using the IQR method
Eliminate Outlier avg_glucose_level	4.400	165	4,235	599	avg_glucose_level outlier using the IQR method

Note: BMI: Body Mass Index, IQR: Interquartile Range.

3.3 Gain ratio feature selection result

Feature selection using the gain ratio was performed exclusively on the training data to enhance data representation and improve the efficiency of subsequent modeling stages. The features processed comprised the 22 features resulting from the preprocessing stage, from which the most informative subset was identified during the model-building process. By retaining only the most relevant characteristics, this step enhances model performance and computational efficiency. The feature ranking from highest to lowest is presented in Table 2. As shown in Table 2, 20 features were retained based on gain ratio rankings, selecting all features with a score greater than zero, as features with a zero score contribute no discriminative information. To further evaluate its effectiveness, a comparative experiment between models using the full feature set and the selected feature subset is presented in Section 3.5.1. This comparison aims to determine whether the gain ratio-based feature selection contributes to improved model performance.

3.4 Synthetic Minority Oversampling Technique–Edited Nearest Neighbors sampling data

Class imbalance is a critical issue in this study, as the number of non-stroke instances significantly exceeds the number of stroke cases. Such an imbalance can bias the classification model toward the majority class, reducing its

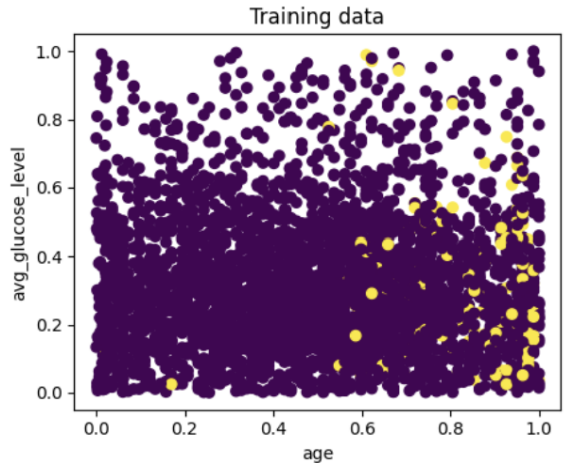
ability to correctly identify stroke instances. Therefore, resampling techniques were applied to the training data to address this issue. Figure 3 presents a visual comparison of the original data, data resampled with SMOTE, and data resampled with SMOTE-ENN. The yellow dots represent the distribution of stroke strokes, while the purple dots represent the non-stroke class. A comparison of the training data and the sampling results can be seen in Table 3.

The data distribution based on two labels is visualized in Figure 3, where the stroke class is represented by yellow dots and the non-stroke class by purple dots. The predominance of non-stroke data over stroke data indicates a substantial class imbalance in the original data. The stroke class distribution tends to be dispersed and concentrated in specific regions of the feature space as a result of this circumstance. The distribution of the stroke class becomes more uniform and proportionate to the non-stroke class with the use of the resampling techniques, specifically SMOTE and SMOTE-ENN, with changes in the parameter k. By creating synthetic samples around the closest neighbors, SMOTE increases the density of stroke data while expanding the representation of minority classes without removing the global characteristics of the data. Meanwhile, SMOTE-ENN produces a more organized data distribution by both increasing the amount of stroke samples and cleaning the data by eliminating possibly noisy samples. This difference indicates that SMOTE-ENN prioritizes data quality over quantity, which may contribute to improved generalization performance.

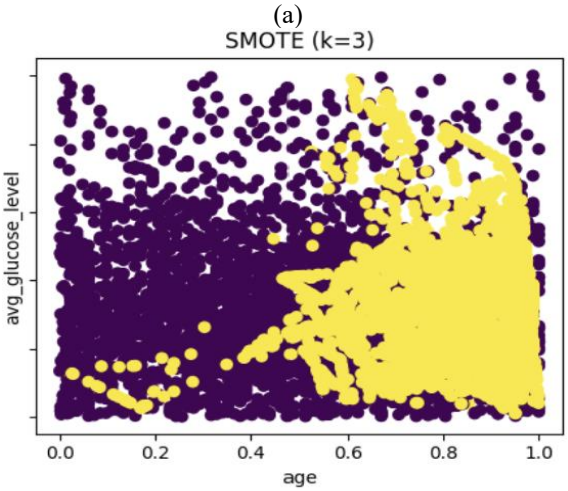
The dataset was partitioned into distinct subsets for training and evaluation. The training data was used to identify optimal classification parameters and model configurations. Following preprocessing, a total of 4,400 clean data records were available for analysis. These were divided into training and testing sets at an 80:20 ratio, yielding 3,520 records for model training and 880 records for performance evaluation. To prevent data leakage and ensure impartial model evaluation, all subsequent data sampling and balancing procedures were applied exclusively to the training data. The numerical class distribution at each stage is summarized in Table 3.

Table 2. Results of the ranking gain ratio

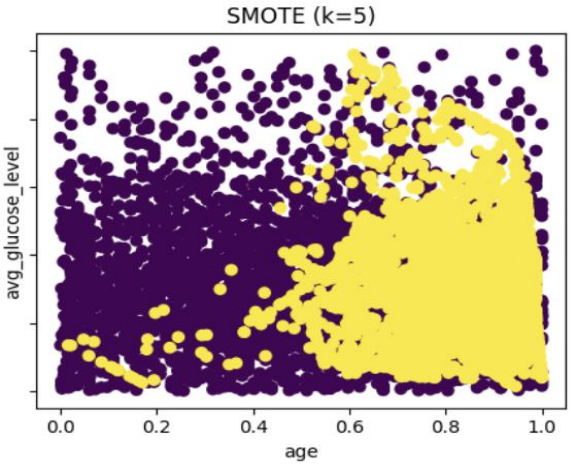
Features	Gain Ratio
age	1.350980
avg_glucose_level	1.246309
bmi	1.006987
smoking_status_never smoked	0.082929
hypertension_1	0.077050
work_type_Private	0.071030
Residence_type_Rural	0.067175
Residence_type_Urban	0.055853
ever_married_Yes	0.050687
smoking_status_smokes	0.047261
smoking_status_Unknown	0.043266
gender_Female	0.041857
gender_Male	0.038171
work_type_Self-employed	0.034052
heart_disease_0	0.027995
heart_disease_1	0.026448
hypertension_0	0.024594
work_type_children	0.016957
smoking_status_formerly smoked	0.014704
ever_married_No	0.010142
work_type_Never_worked	0.000000
work_type_Govt_job	0.000000



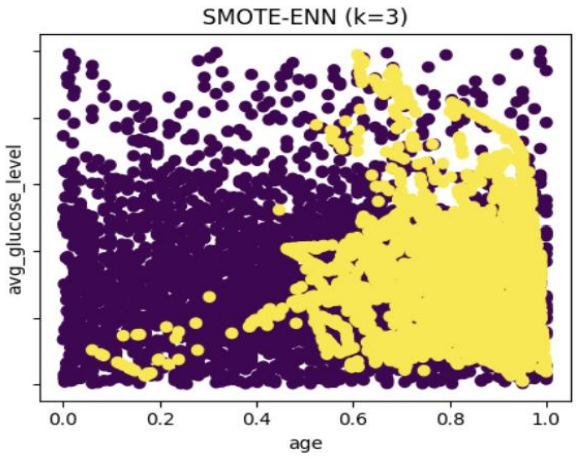
(a)



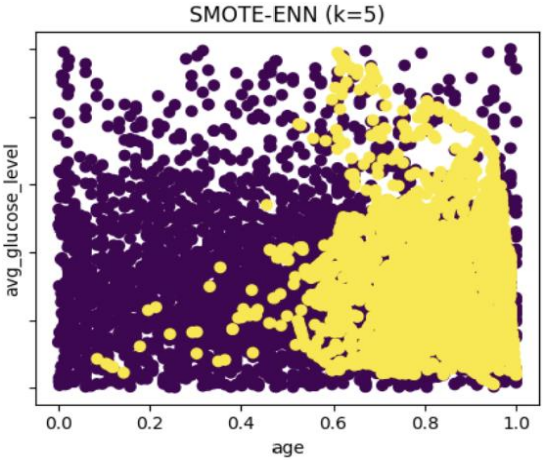
(b)



(c)



(d)



(e)

Figure 3. Data distribution: (a) Training data, (b) SMOTE k = 3, (c) SMOTE k = 5, (d) SMOTE-ENN k = 3, (e) SMOTE-ENN k = 5

Note: KNN: K-Nearest Neighbor, SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors.

Table 3. Summary of dataset composition and class distribution at each data processing stage

Model	Amount Data	Stroke	Non-Stroke
Original data	5,110	249	4,861
Preprocessed data	4,400	165	4,235
Training data	3,520	132	3,388
Testing data	880	33	847
SMOTE (k = 3)	6,776	3,388	3,388
SMOTE-ENN (k = 3)	6,622	3,234	3,388
SMOTE (k = 5)	6,776	3,388	3,388
SMOTE-ENN (k = 5)	6,391	3,003	3,388

Note: SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors.

3.5 Classification result

To examine the effect of model complexity on classification performance, the Random Forest algorithm was evaluated under various hyperparameter configurations. The maximum depth parameter is set at 4, 5, and none, and the number of estimators utilized is modified into three scenarios set at 5, 10, and 15 trees. Each decision tree uses entropy as its splitting criterion to maximize information gain at each node. Three data schemes were used to train and evaluate all hyperparameter combinations: the baseline Random Forest on imbalanced data, Random Forest with SMOTE oversampling, and Random Forest with hybrid SMOTE-ENN sampling. This systematic approach aimed to comprehensively assess the combined influence of hyperparameter settings and data balancing strategies on model performance.

3.5.1 Result of feature selection (gain ratio)

Feature selection using the gain ratio was performed exclusively on the training data to avoid data leakage. The evaluation of this feature selection step was focused on the Random Forest model, as it serves as the baseline approach in this study, while the Random Forest with SMOTE (RFS) and Random Forest with SMOTE-ENN (RFSE) models already incorporate feature selection mechanisms within their respective frameworks. The selected subset of features was then applied to both training and test data. To assess the effectiveness of this feature selection step, model performance was compared between the full feature set and the selected feature subset using the independent test data.

The results demonstrate that gain ratio-based feature selection did not improve classification performance relative to the full feature set (Table 4), where the reported results represent the best performance obtained from each respective configuration. This suggests that Random Forest is inherently robust to irrelevant or less informative features due to its ensemble-based mechanism, allowing the model to maintain performance even without explicit feature selection. Although the gain ratio reduces dimensionality, it may have inadvertently removed features with residual predictive value, resulting in a marginal performance decline. Nevertheless, its inclusion remains methodologically important to evaluate the impact of feature selection on the model. Additionally, despite achieving an accuracy of up to 0.9625, a sensitivity of 0.000 indicates complete failure to identify stroke cases, reflecting a severe class imbalance problem and confirming that accuracy alone is an inadequate evaluation criterion. Supplementary metrics such as sensitivity, F1-score, and AUC are needed to provide a more reliable evaluation, especially for minority classes.

Table 4. Comparison of classification results using the full feature set and gain ratio selection

Model	n_estimators	max_depth	Feature set	Accuracy	Sensitivity	Specificity	F1-Score	AUC
Random Forest	5	4.0	Full	0.9625	0.0000	1.0000	0.0000	0.8361
Random Forest	5	4.0	Selected	0.9625	0.0000	1.0000	0.0000	0.8079

Note: AUC: The area under the receiver operating characteristic curve.

3.5.2 Model performance on training data

This subsection examines model behavior during the training phase to characterize learning patterns under different class imbalance handling strategies. All models were trained using features selected via the gain ratio method, which was applied to reduce feature redundancy and assess the influence of feature selection on model learning. Training performance does not necessarily reflect generalization ability, particularly in imbalanced datasets. First, evaluate the training model by applying the random forest method, and the results for all scenarios are shown in Table 5.

In this scenario, Random Forest demonstrated that high accuracy did not reflect the model's ability to detect the minority class (stroke). This was evident in several model configurations with accuracy values above 96%, which nonetheless yielded low sensitivity. The model achieved 96.25% accuracy but only 0.76% sensitivity with n_estimators = 5 and max_depth = 4. Despite its high accuracy, the model was almost completely biased toward the majority class, even though its specificity value approached 100%. A similar pattern was observed in the configurations n_estimators = 5 and max_depth = 5, as well as n_estimators = 15 and max_depth = 4, where sensitivity was only 1.52% and 0%, respectively. In another configuration, the model failed to

detect stroke cases at all, although accuracy remained high at 96.25%. This reinforces the point that accuracy alone should not be used as a primary indicator of model performance on imbalanced datasets. Conversely, significant performance improvements began to be seen when max_depth was set to none. This allows the decision tree to grow deeper and capture complex patterns in the minority data. With n_estimators = 5 and max_depth = None, the sensitivity increased to 78.79% with an F1 score of 0.8814 and an MCC of 0.8840. The best configuration was obtained with n_estimators = 15 and max_depth = None, yielding a sensitivity of 87.88%; however, this substantial increase in model capacity suggests potential overfitting.

In the second scenario, the Random Forest model was evaluated in combination with SMOTE oversampling. Synthetic samples were generated for the minority class, resulting in a balanced class distribution between stroke and non-stroke instances. The results for all configurations are presented in Table 6.

Table 5. Performance evaluation of Random Forest (RF)

Model	n_estimators	max_depth	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	MCC
RF	5	4	0.9625	0.0076	0.9997	0.50	0.0149	0.8538	0.0580
RF	5	5	0.9631	0.0152	1.0000	1.00	0.0299	0.8929	0.1208
RF	5	None	0.9920	0.7879	1.0000	1.00	0.8814	0.9973	0.8840
RF	15	4	0.9625	0.0000	1.0000	0.00	0.0000	0.8669	0.0000
RF	15	5	0.9636	0.0303	1.0000	1.00	0.0588	0.9061	0.1709
RF	15	None	0.9955	0.8788	1.0000	1.00	0.9355	0.9999	0.9352

Note: AUC: The area under the receiver operating characteristic curve, MCC: Matthews correlation coefficient.

Table 6. Performance evaluation of Random Forest with SMOTE

Model	k	n_estimators	max_depth	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	MCC
RFS	3	5	4	0.7929	0.9298	0.6561	0.7300	0.8179	0.8899	0.6091
RFS	3	5	5	0.8264	0.9283	0.7246	0.7712	0.8425	0.9100	0.6669
RFS	3	5	None	0.9963	0.9985	0.9941	0.9941	0.9963	0.9999	0.9926
RFS	3	15	4	0.7953	0.9348	0.6558	0.7309	0.8204	0.9066	0.6150
RFS	3	15	5	0.8229	0.9109	0.7349	0.7746	0.8372	0.9148	0.6560
RFS	3	15	None	0.9996	0.9997	0.9994	0.9994	0.9996	1.0000	0.9991
RFS	5	5	4	0.8179	0.9324	0.7034	0.7586	0.8366	0.8941	0.6531
RFS	5	5	5	0.8301	0.9519	0.7084	0.7655	0.8486	0.9147	0.6808
RFS	5	5	None	0.9948	0.9976	0.9920	0.9921	0.9948	0.9999	0.9897
RFS	5	15	4	0.7987	0.9543	0.6432	0.7278	0.8258	0.9087	0.6286
RFS	5	15	5	0.8124	0.9448	0.6800	0.7470	0.8344	0.9246	0.6480
RFS	5	15	None	0.9990	0.9991	0.9988	0.9988	0.9990	1.0000	0.9979

Note: SMOTE: Synthetic Minority Oversampling Technique, AUC: The area under the receiver operating characteristic curve, MCC: Matthews correlation coefficient, RFS: Random Forest with SMOTE.

Table 7. Performance evaluation of Random Forest with SMOTE-ENN

Model	k	n_estimators	max_depth	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	MCC
RFSE	3	5	4	0.8055	0.8757	0.7385	0.7617	0.8147	0.8798	0.6187
RFSE	3	5	5	0.8162	0.8751	0.7600	0.7768	0.8230	0.9041	0.6382
RFSE	3	5	None	0.9968	0.9985	0.9953	0.9951	0.9968	0.9999	0.9937
RFSE	3	15	4	0.8188	0.9066	0.7349	0.7655	0.8301	0.9058	0.6494
RFSE	3	15	5	0.8173	0.8788	0.7586	0.7765	0.8245	0.9148	0.6407
RFSE	3	15	None	0.9998	1.0000	0.9997	0.9997	0.9998	1.0000	0.9997
RFSE	5	5	4	0.8208	0.9207	0.7323	0.7530	0.8285	0.8873	0.6592
RFSE	5	5	5	0.8305	0.9321	0.7406	0.7610	0.8379	0.9222	0.6792
RFSE	5	5	None	0.9970	0.9980	0.9962	0.9957	0.9968	0.9999	0.9940
RFSE	5	15	4	0.7927	0.9610	0.6434	0.7049	0.8133	0.9149	0.6288
RFSE	5	15	5	0.8360	0.9417	0.7423	0.7641	0.8437	0.9309	0.6915
RFSE	5	15	None	0.9997	0.9997	0.9997	0.9997	0.9997	1.0000	0.9994

Note: SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors, AUC: The area under the receiver operating characteristic curve, MCC: Matthews correlation coefficient, RFSE: Random Forest with SMOTE-ENN.

The application of SMOTE to the Random Forest model substantially improved performance characteristics relative to the baseline RF scenario with imbalanced data. In contrast to the RF results, all RFS configurations demonstrated substantial improvements in sensitivity for the stroke class, indicating that the model was more effective in identifying stroke cases and considerably reduced false-negative errors. Accuracy ranged from 79.29% to 83.01% in configurations with max_depth values of 4 and 5. Although these configurations achieved relatively lower accuracy than the baseline RF, they produced high sensitivity values exceeding 92%, reflecting a more realistic, appropriate trade-off, given that failure to detect a stroke case carries far more serious consequences than false-positive errors. F1-scores in this configuration ranged from 0.8179 to 0.8486, while MCC increased to 0.6808, indicating a better balance between performance on the majority and minority classes. The most significant performance improvement was observed when max_depth is set to none, both at n_estimators = 5 and 15. At k = 3 and n_estimators = 15 in the RFS configuration, the

model achieved 99.96% accuracy, 99.97% sensitivity, F1-score 0.9996, AUC 1.00, and MCC 0.9991. A comparable performance pattern was observed at k = 5, with MCC reaching 0.9979, confirming the model's stability against SMOTE parameter variations. Overall, SMOTE significantly improves sensitivity across all configurations; however, several configurations exhibited near-perfect training performance, suggesting potential overfitting to synthetic samples rather than generalizable patterns.

In the third scenario, the Random Forest model was trained using data resampled with the SMOTE-ENN hybrid approach. SMOTE-ENN addresses class imbalance while effectively controlling noise by combining synthetic oversampling via SMOTE with noise removal via ENN. The results for each configuration are presented in Table 7.

The SMOTE-ENN approach further refines the data by combining oversampling with noise reduction, producing a more structured training distribution and an improved balance between sensitivity and specificity. In configurations with max_depth values of 4 and 5, RFSE demonstrated accuracy ranging from 79.27% to 83.60%, with sensitivity increasing to 96.10% at k = 5, n_estimators = 15, and max_depth = 4. This indicates the model's excellent ability to identify stroke cases, although accompanied by a decrease in specificity due to the data's more aggressive detection of minority classes. The F1-score in this configuration ranged from 0.8133 to 0.8437, while the MCC increased to 0.6915, reflecting a stronger classification balance compared to Random Forest without imbalance treatment. Optimal performance of RFSE is achieved when max_depth is set to None, both for k = 3 and k = 5. In the RFSE configuration with k = 3 and n_estimators = 15, the model achieved 99.98% accuracy, 100% sensitivity, an F1-score of 0.9998, an AUC of 1.00, and an MCC of 0.9997. Although this configuration yielded highly balanced performance in terms of both sensitivity and specificity, the near-perfect training results indicate a substantial risk of overfitting, particularly when deep trees are employed.

Model performance evaluation was conducted by comparing the best-performing configurations across all approaches, as illustrated in Figure 4. The selection of the best model for each approach was based on a balanced assessment of evaluation metrics relevant to imbalanced data classification, specifically sensitivity, F1-score, and AUC. Experimental results demonstrated that Random Forest produced high accuracy but exhibited substantial weaknesses in detecting stroke cases, with a sensitivity of only 0.8788. This finding confirms that high accuracy does not automatically reflect the model's ability to identify minority classes in imbalanced data. The application of SMOTE significantly improved stroke detection performance. The best SMOTE-based model achieved a sensitivity of 0.9997, an F1-score of 0.9996, and an AUC of 1.0000, with only one false negative error, indicating improved discriminatory ability and more balanced class predictions. The SMOTE-ENN approach yielded optimal and stable results, with the best model achieving a sensitivity of 1.0000, an F1-score of 0.9997, and an AUC of 1.0000 without producing false negatives. The combination of oversampling and data cleaning proved effective in reducing class bias while increasing prediction reliability.

3.5.3 Generalization performance on test data

Model performance on the independent test set was evaluated to assess the generalization capability of each proposed approach. Unlike training performance, which reflects a model's ability to fit observed data, test set evaluation provides a more reliable estimate of generalization, particularly in imbalanced classification problems. Three approaches were systematically compared using the same parameter combination: RF, RFS, and RFSE. The performance of each configuration is illustrated in Figures 5-7, while the comparison of the best-performing models is summarized in Table 8.

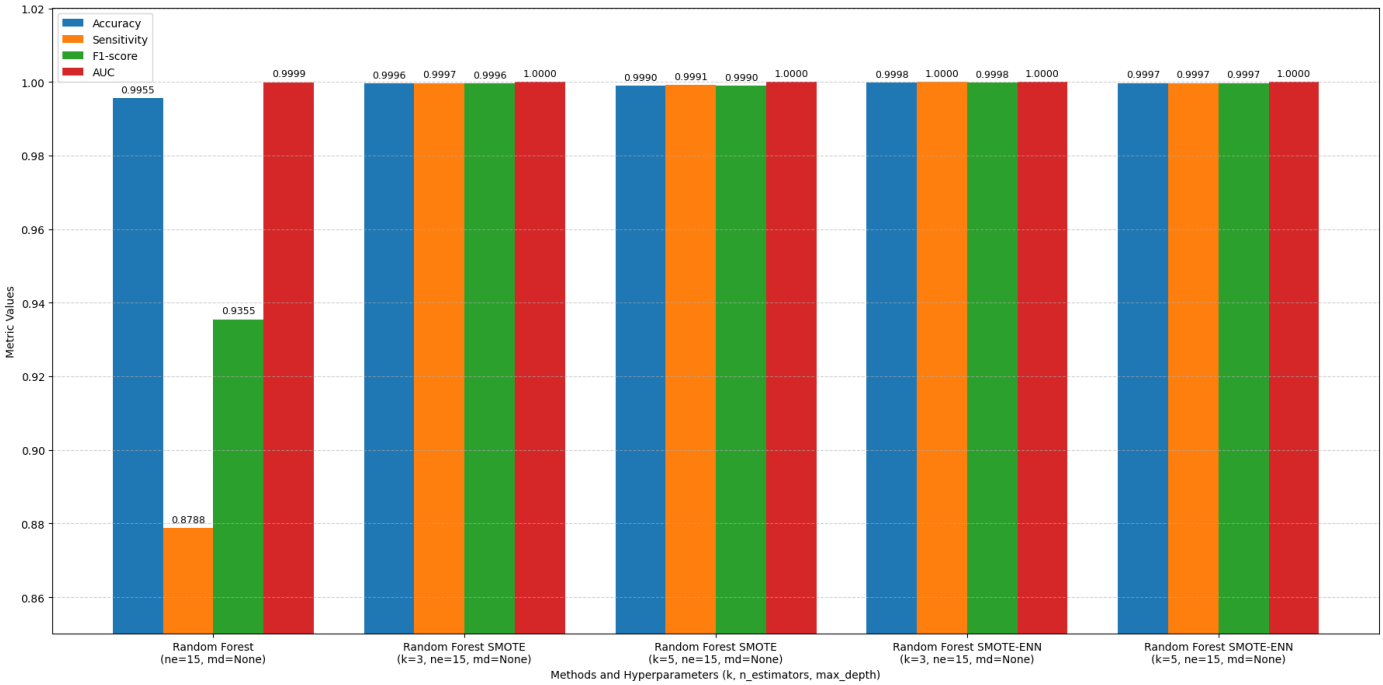


Figure 4. Model performance comparison by method and hyperparameters

Table 8. Best model comparison on the test set

Model	k	n_estimators	max_depth	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC	MCC
RF	–	5	–	0.9580	0.0909	0.9917	0.3000	0.1395	0.6933	0.1481
RFS	3	5	5	0.7080	0.7879	0.7048	0.0942	0.1683	0.8116	0.2018
RFS	5	5	–	0.9205	0.3333	0.9433	0.1864	0.2391	0.7611	0.2102
RFSE	3	15	4	0.7261	0.8788	0.7202	0.1090	0.1940	0.8261	0.2478
RFSE	5	5	–	0.9318	0.2121	0.9599	0.1707	0.1892	0.7299	0.1550

Note: AUC: The area under the receiver operating characteristic curve, MCC: Matthews correlation coefficient, RF: Random Forest, RFS: Random Forest with SMOTE, RFSE: Random Forest with SMOTE-ENN.

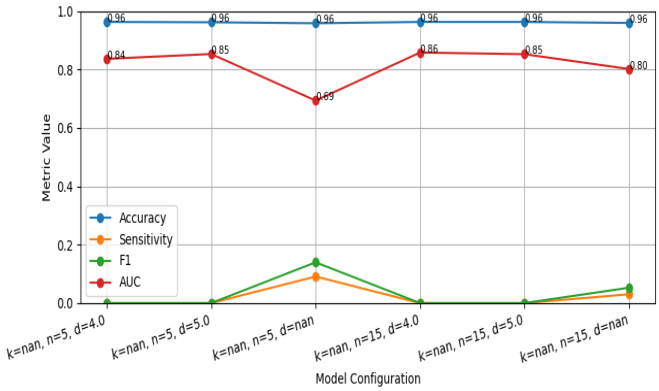


Figure 5. Performance of the Random Forest model on the independent test dataset

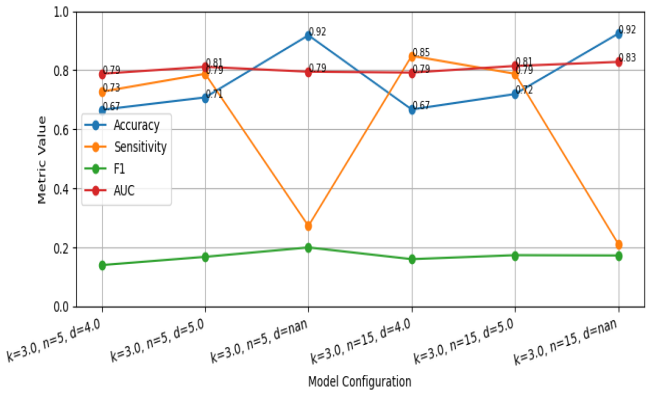


Figure 6. Performance of Random Forest with SMOTE model on the independent test dataset
Note: SMOTE: Synthetic Minority Oversampling Technique.

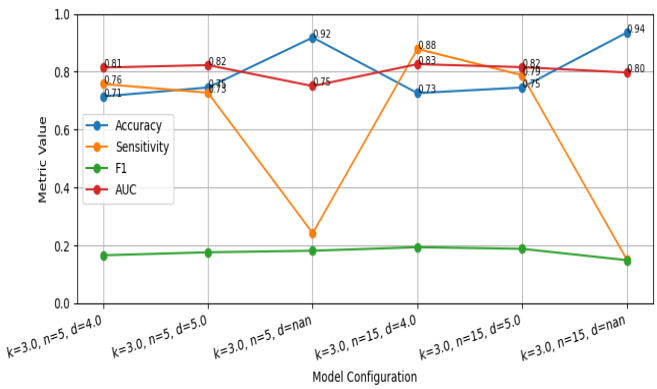


Figure 7. Performance of Random Forest with SMOTE-ENN model on the independent test dataset
Note: SMOTE: Synthetic Minority Oversampling Technique, ENN: Edited Nearest Neighbors.

The test results showed that the Random Forest model on the original data tended to produce high accuracy, reaching 0.96, but consistently failed to detect the minority class (stroke), as evidenced by near-zero or zero sensitivity values across most configurations. This confirms that the model is heavily biased toward the majority class, rendering the accuracy insufficient as the sole evaluative criterion in imbalanced settings. Even in the best-performing configuration (max_depth = None), sensitivity reached only 0.09 with a correspondingly low F1-score, indicating the model's limited capacity to capture discriminative patterns in the minority class.

The application of SMOTE substantially improved model sensitivity on the test set, with values ranging from 0.72 to 0.85 across several configurations, demonstrating that oversampling enhanced minority class representation and improved stroke detection capability. However, this gain in sensitivity was accompanied by a notable reduction in specificity and precision, reflecting an increased rate of false positives. In configurations with higher model complexity (max_depth = None), a marked decline in sensitivity was observed, suggesting that the model had overfit the synthetic training data and consequently lost generalization ability on the test set. The SMOTE-ENN approach yielded more balanced performance than pure SMOTE. In its best configuration (k = 3, n_estimators = 15, max_depth = 4), the model achieved a sensitivity of 0.8788, an AUC of 0.8261, and the highest MCC among all evaluated methods. Relative to SMOTE, SMOTE-ENN produced a more stable prediction distribution with a better balance between sensitivity and specificity. This is due to the ENN mechanism that removes ambiguous or noisy samples after oversampling, resulting in more representative training data and less bias towards synthetic data.

The experimental results revealed a substantial performance gap between training and testing phases, particularly for models trained on resampled data. Near-perfect scores (AUC close to 1.0) were observed during training but were not sustained on the independent test set. This discrepancy indicates overfitting, where the model captures not only underlying data patterns but also noise and synthetic structures introduced during the resampling process. In particular, the use of SMOTE and SMOTE-ENN may have produced overly optimistic training performance due to the increased homogeneity of synthetic samples. Consequently, generalization ability is more accurately reflected by the test set result, which showed more moderate sensitivity and AUC values. These findings underscore the importance of relying on independent test evaluation rather than training metrics when assessing model effectiveness in imbalanced classification tasks. Notably, the SMOTE-ENN approach exhibited a more controlled performance decline compared to SMOTE, indicating superior generalization ability. To verify that observed performance differences between models were

statistically significant rather than due to chance, McNemar's test was applied to the prediction results on the test data.

3.5.4 Statistical comparison using the McNemar test

To validate whether the performance differences between the models were statistically significant, a McNemar test was performed on the prediction results from the test data. McNemar's test is specifically designed to compare two classification models based on paired predictions on the same dataset, making it well-suited for evaluating whether observed differences in classification errors are statistically meaningful.

The McNemar test results confirmed that all pairwise model comparisons yielded statistically significant differences (Table 9). The comparison between RF and RFS produced a test statistic of 213.63 with a p-value of 2.21×10^{-48}, while the comparison between RF and RFSE yielded a test statistic of 161.09 with a p-value of 6.55×10^{-37}. Both results indicate that the application of resampling techniques, whether SMOTE or SMOTE-ENN, produced meaningful performance improvements over the unbalanced baseline. The comparison between RFS and RFSE further revealed a significant difference, with a test statistic of 35.15 and a p-value of 3.05×10^{-9}, confirming that the hybrid SMOTE-ENN approach is statistically superior to pure SMOTE on this dataset. This finding is consistent with earlier evaluations, in which SMOTE-ENN demonstrated a more favorable balance between sensitivity and specificity and more stable generalization performance.

Table 9. Statistical comparison between methods

Comparison	Test Statistic	P-Value
RF vs. RFS	213.6338	2.21×10^{-48}
RF vs. RFSE	161.0865	6.55×10^{-37}
RFS vs. RFSE	35.1486	3.05×10^{-9}

Note: RF: Random Forest, RFS: Random Forest with SMOTE, RFSE: Random Forest with SMOTE-ENN.

4. CONCLUSIONS

This study demonstrates that applying Random Forest without handling class imbalance results in high accuracy but fails to effectively detect stroke cases, as indicated by extremely low sensitivity and a high number of false negatives. The use of sampling techniques, particularly SMOTE and SMOTE-ENN, significantly improves the model's ability to recognize minority classes. Based on evaluation using independent test data, SMOTE-ENN provides the best overall performance, achieving higher sensitivity and a more balanced trade-off between sensitivity and specificity compared to other approaches.

Although near-perfect performance was observed during training on resampled data, the results on the test set reveal a noticeable performance gap, indicating potential overfitting. This highlights the importance of evaluating models using unseen data to obtain a realistic assessment of their predictive capability. In addition to the performance improvements, this study contributes by providing a systematic comparison of sampling strategies within a consistent experimental framework, including evaluations on independent test data and statistical validation using McNemar's test. The findings also highlight the importance of distinguishing between training and testing performance, as near-perfect results on resampled training data do not necessarily imply strong generalization ability. This study demonstrates that integrating hybrid sampling techniques such as SMOTE-ENN with Random Forest can serve as an effective and robust approach for addressing imbalanced medical datasets, particularly in improving the detection of clinically critical cases such as stroke, while maintaining reliable generalization performance.

REFERENCES

[1] Lutz, P.L., Nilsson, G.E., Prentice, H.M. (2002). The Brain Without Oxygen: Causes of Failure-Physiological and Molecular Mechanisms for Survival. Springer Dordrecht. https://doi.org/10.1007/0-306-48197-9

[2] Murphy, S.J., Werring, D.J. (2020). Stroke: Causes and clinical features. Medicine, 48(9): 561-566. https://doi.org/10.1016/j.mpmed.2020.06.002

[3] Krishnamurthi, R.V., Feigin, V.L. (2022). Global burden of stroke. In Stroke, pp. 163-178.e2. https://doi.org/10.1016/B978-0-323-69424-7.00014-4

[4] Amann, J. (2022). Machine learning in stroke medicine: Opportunities and challenges for risk prediction and prevention. In Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues, pp. 57-71. https://doi.org/10.1007/978-3-030-74188-4_5

[5] Pokhrel, L., Kumar, A., Garg, P., Anand, N., Singh, N. (2025). AI and IoT in global health: Ethical lessons from pandemic response. In Development and Management of Eco-Conscious IoT Medical Devices, pp. 367-394. https://doi.org/10.4018/979-8-3373-4134-7.ch013

[6] Chheda, V.S., Kapadia, S.K., Lakhani, B.K., Sonawane, P. (2021). Stroke prediction using machine learning. International Journal of Advances in Engineering and Management, 3(6): 985-992.

[7] Devi, R.D.H., Sreevalli, P., Keerthana, K., Prathyusha, P., Asia, M. (2020). Prediction of diseases using random forest classification algorithm. Zeichen Journal, 6(5): 19-26.

[8] Napierala, K., Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems, 46: 563-597. https://doi.org/10.1007/s10844-015-0368-1

[9] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., Japkowicz, N. (2024). The class imbalance problem in deep learning. Machine Learning, 113: 4845-4901. https://doi.org/10.1007/s10994-022-06268-8

[10] Faran, J., Triayudi, A. (2024). Analysis of the effectiveness of polynomial fit SMOTE mesh on imbalance dataset for bank customer churn prediction with XGBoost and Bayesian optimization. Jurnal Teknik Informatika, 5(3): 661-667. https://doi.org/10.52436/1.jutif.2024.5.3.1284

[11] Lilhore, U.K., Manoharan, P., Sandhu, J.K., Simaiya, S., et al. (2023). Hybrid model for precise hepatitis-C classification using improved random forest and SVM method. Scientific Reports, 13: 12473. https://doi.org/10.1038/s41598-023-36605-3

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

[13] Ramentol, E., Caballero, Y., Bello, R., Herrera, F.

(2012). Smote-rs b*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information Systems, 33: 245-265. https://doi.org/10.1007/s10115-011-0465-6

[14] Ghasemkhani, B., Yilmaz, R., Birant, D., Kut, R.A. (2023). Logistic model tree forest for steel plates faults prediction. Machines, 11(7): 679. https://doi.org/10.3390/machines11070679

[15] Xu, Z.Z., Shen, D.R., Nie, T.Z., Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data. Journal of Biomedical Informatics, 107: 103465. https://doi.org/10.1016/j.jbi.2020.103465

[16] Nishat, M.M., Faisal, F., Ratul, I.J., Al-Monsur, A., Ar-Rafi, A.M., Nasrullah, S.M., Reza, M.T., Khan, M.R.H. (2022). A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. Scientific Programming, 2022: 3649406. https://doi.org/10.1155/2022/3649406

[17] Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Li, C.H., Han, Q.H., Zhang, Y.B. (2021). Improving risk identification of adverse outcomes in chronic heart failure using SMOTE+ ENN and machine learning. Risk Management and Healthcare Policy, 14: 2453-2463. https://doi.org/10.2147/RMHP.S310295

[18] Kumar, S., Kumari, R., Gosain, A. (2025). Hybrid resampling for enhanced multiclass classification. Discover Computing, 28: 306. https://doi.org/10.1007/s10791-025-09841-6

[19] Wei, Z.M., Li, M.Q., Zhang, C.H., Miao, J.L., Wang, W.M., Fan, H. (2024). Machine learning–based predictive model for post-stroke dementia. BMC Medical Informatics and Decision Making, 24: 334. https://doi.org/10.1186/s12911-024-02752-4

[20] Liu, C.H., Tsai, C.F., Sue, K.L., Huang, M.W. (2020). The feature selection effect on missing value imputation of medical datasets. Applied Sciences, 10(7): 2344. https://doi.org/10.3390/app10072344

[21] Chen, X.Y., Aljrees, T., Umer, M., Saidani, O., Almuqren, L., Mzoughi, O., Ishaq, A., Ashraf, I. (2023). Cervical cancer detection using K nearest neighbor imputer and stacked ensemble learning model. Digital Health, 9: 20552076231203800. https://doi.org/10.1177/20552076231203802

[22] Nizam-Ozogur, H., Orman, Z. (2024). A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. Expert Systems, 41(8): e13596. https://doi.org/10.1111/exsy.13596

[23] Yang, F.Y., Wang, K., Sun, L.S., Zhai, M.J., Song, J.J., Wang, H. (2022). A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. BMC Medical Informatics and Decision Making, 22(1): 344. https://doi.org/10.1186/s12911-022-02075-2

[24] Ghasemi, F., Neysiani, B.S., Nematbakhsh, N. (2020). Feature selection in pre-diagnosis heart coronary artery disease detection: A heuristic approach for feature selection based on information gain ratio and Gini index. In 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, pp. 27-32. https://doi.org/10.1109/ICWR49608.2020.9122285

[25] Speiser, J.L., Miller, M.E., Tooze, J., Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134: 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

[26] Ansori, N., Rachmad, A., Rochman, E.M.S., Fauzan, H.B., Asmara, Y.P. (2024). Corn stalk disease classification using random forest combination of extraction features. Communications in Mathematical Biology and Neuroscience, 2024: 19. https://doi.org/10.28919/cmbn/8404

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