Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 3 (2025) 646 - 658 e-ISSN: 2580-0760

Enhancing Stroke Prediction with Logistic Regression and Support Vector Machine Using Oversampling Techniques

Syamsul Risal[1*], Fajar Apriyadi[2], A. Sumardin[3], Andini Dani Achmad[4], Annisa Nurul Puteri[5]
[1,2,3] Department of Informatics, Universitas Teknologi Akba Makassar, Indonesia
[4] Department of Electrical Engineering, Universitas Hasanuddin, Indonesia
[5] Department of Computer and Network Engineering, Politeknik Negeri Ujung Pandang, Indonesia

[1]syamsulrisal22@mhs.unitama.ac.id*, [2]fajarap.0204@gmail.com, [3]andis@unitama.ac.id, [4]andini.achmad@unhas.ac.id, [5]annisanurulputeri@gmail.com

Abstract

Stroke is a significant health concern that can result in both death and disability, making the early identification of risk factors crucial. Previous studies on stroke prediction have been limited by inadequate handling of class imbalance, lack of comprehensive feature selection, and parameter optimization, with accuracy rates usually below 80%. This study compares the performance of Logistic Regression (LR) and Support Vector Machine (SVM) algorithms combined with different oversampling methods—SMOTE, Borderline-SMOTE, ADASYN, Random Over Sampling (ROS), and Random Under Sampling (RUS)—on a stroke prediction dataset. Correlation-based feature selection identified age, hypertension, and heart disease as significant predictors. GridSearchCV with 10-fold cross-validation was used for hyperparameter optimization, and performance was evaluated using precision, recall, accuracy, and ROC curves. The results showed that SVM significantly outperformed Logistic Regression across all sampling methods. SVM+ROS achieved the highest performance with perfect recall (100%), precision of 97.18%, and accuracy of 98.56% (AUC: 0.9857), whereas SVM + Borderline-SMOTE offered balanced performance with a recall of 94.99%, precision of 95.06%, and accuracy of 95.17% (AUC: 0.9512). LR + Borderline-SMOTE performed the best with an accuracy of 84.98% (AUC: 0.8503), significantly better than previous studies. This improved accuracy shows significant clinical benefits, potentially reducing missed stroke diagnoses by identifying thousands of additional at-risk patients in large-scale screening programs. Healthcare providers should consider implementing SVM with ROS in critical care settings, where potentially missed stroke cases have severe consequences. Simultaneously, SVM with Borderline-SMOTE may be more appropriate for resource-constrained environments.

Keywords: grid search cross-validation; logistic regression; machine learning; stroke disease; support vector machine

How to Cite: S. Risal, Fajar Apriyadi, A. Sumardin, Andini Dani Achmad, and Annisa Nurul Puteri, "Enhancing Stroke Prediction with Logistic Regression and Support Vector Machine Using Oversampling Techniques", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 646 - 658, Jun. 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i3.6431

Received: February 28, 2025
Accepted: June 8, 2025
Available Online: June 22, 2025

1. Introduction

Stroke remains one of the leading causes of death and disability worldwide [1]. This disease occurs when blood vessels in the brain are blocked or ruptured, reducing the blood supply to the brain and causing brain cell death [2]. Every year, more than 15 million people worldwide suffer from stroke [3]. Until now, there has been no proper treatment for stroke. However, early detection provides an opportunity to delay or even prevent its development from worsening.

Machine learning classification algorithms have demonstrated significant potential in the medical field, particularly in detecting various diseases, including stroke. This innovative approach enables the identification of patterns in patient data that can indicate certain risks. In the specific context of stroke detection, machine learning classification can assist medical personnel in identifying high-risk patients based on a range of factors and symptoms related to stroke, offering a promising outlook for the future of healthcare.

Previous studies have applied various classification algorithms to disease detection with promising results. For example, research [4] developed a heart disease prediction model using algorithms such as Random Forest, Decision Tree, and Neural Network, with an

accuracy of up to 90%. However, the study did not investigate logistic regression's effectiveness in cases of extreme data imbalance. In addition, researchers [5] conducted a comparative study between SVM, Logistic Regression, and Decision Tree for liver disease prediction, showing that SVM provided the highest accuracy of 85% but did not perform comprehensive hyperparameter optimization.

In the context of stroke prediction, several studies have shown mixed results. Research [6] used ensemble learning with the SMOTE method to address data imbalance. Their research performed well on Random Forest, which achieved 91% accuracy. However, the study showed that Logistic Regression only achieved 75% accuracy without an in-depth analysis of the causes of the low performance. Likewise, research [7] also built a stroke disease prediction model using the SMOTE method with several classification algorithms, including ensemble methods such as Stacking. In their research, Stacking showed excellent performance by achieving 98% accuracy, but logistic regression also showed less than optimal performance with an accuracy of only 79%.

The study [8] conducted further research by applying feature selection together with SMOTE, which increased accuracy for ensemble methods such as Random Forest (96%) but did not provide a significant increase for Logistic Regression, which remained at 79%. Meanwhile, the study [9] took a different approach by applying Random Under Sampling (ROS) to overcome data imbalance. In his study, he showed an increase in performance for Logistic Regression, which reached 78% compared to the study results [6], although it was still lower than the ensemble technique.

Previous studies have shown a gap in the optimization of LR algorithms for stroke prediction. Despite their superior interpretability and computational efficiency, these algorithms consistently outperformed ensemble algorithms. This lower performance was due to the significant data imbalance in the stroke datasets. The proportion of stroke cases is much smaller than that of non-stroke cases, making LR models biased towards the majority class. Although SMOTE helps balance classes, LR may not be able to handle the complexity of synthetic data or ensemble methods. In addition, LR Regression is a linear model with limitations in capturing complex and nonlinear relationships between variables (age, BMI, and avg_glucose_level). Stroke risk factors often interact in complex and non-linear ways, and ensemble methods such as Random Forest or Stacking can better capture these complex patterns.

No comprehensive study has compared the effect of sampling techniques and hyperparameter optimization on the performance of Logistic Regression in the context of stroke prediction. In addition, most previous studies have used only one type of oversampling method, namely SMOTE or under-sampling, without directly comparing different methods in the same experimental framework, which has a limited understanding of the relative effectiveness of these methods in improving the performance of basic classification algorithms for stroke prediction.

The novelty of this study lies in its comprehensive approach to optimizing base classifiers. It systematically compares multiple sampling techniques that have rarely been explored together in stroke prediction. Unlike previous studies that focused primarily on a single sampling method, this study directly compares five techniques—SMOTE, Borderline-SMOTE, ADASYN, Random Over Sampling, and Random Under Sampling-within the same experimental framework. Furthermore, this study addresses a critical gap in the existing literature by applying Correlation-based Feature Selection to identify the most relevant stroke predictors, combined with extensive hyperparameter tuning via GridSearchCV, an approach not applied in previous studies. This comprehensive approach provides a solid foundation for the research, reassuring the audience about the thoroughness of the study.

The study [10] used GridSearchCV and compared three feature selection techniques, including Information Gain (IG), Chi-square (Chi2), and Correlation-based Feature Selection (CFS), to improve the accuracy of the SVM and Random Forest algorithms in diagnosing heart disease. The results of the study showed that CFS was able to obtain the highest accuracy in the SVM and Random Forest algorithms with accuracies of 92.19% and 91.88%, respectively, which experienced an increase in accuracy of 10.88% for SVM, and Random Forest obtained an increase of 9.47%. GridSearchCV was also applied in research [11] using SVM and KNN algorithms in predicting stroke. In his research, he managed to improve the performance of both algorithms to achieve an accuracy of 94% for SVM and 95% for KNN, which previously only achieved an accuracy of 83% and 91%.

Based on the gap analysis, this study aims to improve the performance of Logistic Regression in stroke prediction with a more comprehensive approach. This study proposes a combination of techniques that have not been fully explored in previous studies in predicting stroke, namely: the application of Correlation-based Feature Selection to identify the most relevant features for stroke, a comparison of various over-sampling methods (SMOTE, Borderline-SMOTE, ADASYN, and Random Over Sampling) and under-sampling methods (Random Under Sampling) to overcome data imbalance, and the use of GridSearchCV for model parameter optimization.

The selection of Logistic Regression (LR) and Support Vector Machine (SVM) as the main algorithms in this study is a strategic and reasonable choice. The logistic regression model is versatile, has strong interpretation, and has been used to describe phenomena in various medical and non-medical research fields [12]. The

study [13] that detected heart disease using LR obtained an accuracy of 91.65%, and in the study [14] also predicted heart disease using LR successfully obtained an accuracy of 92.30%. Both studies showed good performance for LR, instilling confidence in its potential to better predict stroke disease through proper processing. In addition, LR offers significant clinical value due to its high interpretability, allowing health practitioners to understand the relative contribution of each risk factor in predicting stroke through its coefficients. This interpretability is especially important in medical settings where transparency in decision making is critical. The computational efficiency of LR makes it suitable for real-time clinical applications with limited computing resources.

In contrast, SVM excels in handling complex non-linear relationships between variables through its kernel function, capturing subtle patterns in stroke risk factors that linear models may miss. Furthermore, SVM's resilience to overfitting, especially in high-dimensional feature spaces resulting from encoding categorical variables, makes it invaluable for medical datasets with many risk factors. SVM also performed well in studies [10] and [11], achieving accuracies of around 90%, but in studies [6] and [9], SVM was only able to achieve accuracies of 81%, making it an ideal algorithm to compare with Logistic Regression to determine the most effective approach for stroke prediction.

The main objective of this study is to improve the performance of Logistic Regression in stroke prediction, which has consistently performed poorly in previous studies. In addition, this study aims to compare optimized Logistic Regression models with Support Vector Machines to determine the most effective approach for stroke prediction under various sampling conditions. The researchers expect that the application of correlation-based feature selection, exploring various alternative sampling techniques, and optimizing hyperparameters through GridSearchCV will significantly improve the predictive accuracy of both algorithms, especially improving the performance of Logistic Regression beyond the 75-79% accuracy range reported in previous studies.

The contents of this research paper are structured as follows. Section 2 discusses the method and description of the analysis using Logistic Regression with selection and hyperparameter tuning features. Section 3 contains the results and discussion of this research. Finally, Section 4 contains the conclusions of the research.

2. Methods

This study uses Logistic Regression (LR) and Support Vector Machine (SVM) classification methods for stroke prediction. Figure 1 outlines the research methodology, which begins with data acquisition from Kaggle, followed by data preprocessing, addressing class synchronization, data verification, hyperparameter tuning, classification using Logistic Regression and Support Vector Machine, and finally, model performance evaluation.

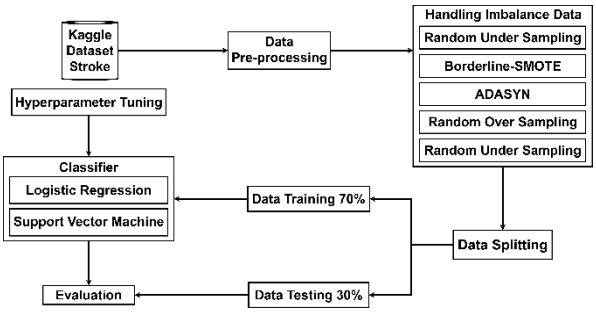

Figure 1. Research Method

2.1 Dataset Description

The dataset used in this study comes from the dataset created by Fedesoriano in 2020, uploaded on Kaggle [15]. This dataset has a total of 5110 data and 12 features. The available dataset features can be seen in Table 1.

Table 1. Dataset Description

Feature	Description
Id	Unique identifier for each patient
gender	Patient gender (Male, Female, and Other)
age	Age of the patient
hypertension	Whether the patient has hypertension or not (0 if no, 1 if yes)
heart_disease	Whether the patient has heart disease or not (0 if no, 1 if yes)
ever_married	Whether the patient is married or not (Yes, No)
work_type	Patient's occupation type (children, Govt_job, Private, Self-employed, Never_worked)
Residence_type	Type of patient residence (Urban, Rural)
avg_glucose_level	The average value of glucose levels in the patient's blood
bmi	BMI value in patients
smoking_status	Smoking status of patients (formerly smoked, never smoked, smokes, Unknown)
Stroke	Stroke diagnosis status of the patient (0 if not stroke, 1 if stroke)

2.2 Data Preprocessing

Before testing the model, the data must go through a preprocessing stage. Preprocessing is a stage that is useful for preparing the data to be used so that the model built can work optimally and effectively. In the preprocessing stage, data cleaning is needed to clean the data from missing values. In addition, in the preprocessing stage, the data will be cleaned from irrelevant or redundant data using the feature selection method [16].

Correlation-Based Feature Selection is used in this study to select relevant and influential features in the dataset. CFS (Correlation-based Feature Selection) is a multivariate filter feature selection that works by selecting features based on correlation by measuring the relationship between two variables; Irrelevant features will be ignored because they do not have a high

correlation value [10], [17]. After going through Feature Selection, Feature Encoding is performed on the dataset using One-Hot Encoding. It is a popular encoding method to utilize when processing datasets containing categorical variables [18]. In one-hot encoding, the original feature vector is expanded into a multidimensional matrix, with the matrix dimension being the number of states in this feature and each dimension representing a particular state; This processing results in only one dimension of the feature matrix being expressed for a particular state (usually '1′), and all other state dimensions are zero [19]. This method leads to a significant increase in the number of features in the dataset.

In addition, a normalization stage is carried out on data with continuous numeric variable types such as age, avg_glucose_level, and BMI. Data normalization is the process of scaling attribute values into smaller ranges with equal weights; The new scale of data attribute values can help classification performance because it can remove features with high noise and low relevance [20].

2.3. Handling Imbalance Data

The dataset used in this study has a total of 5110 data, with a very unbalanced distribution: 249 data (4.87%) are stroke class while 4861 data (95.13%) are non-stroke class, as shown in Figure 2. This extreme imbalance can cause the classification model to tend to be biased towards the majority class, so it must be overcome first to build an optimal and effective model.

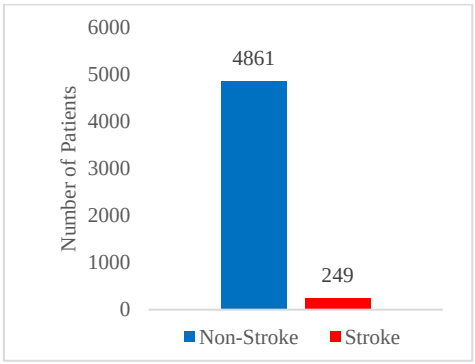

Figure 2. Amount of Data Distribution

The imbalance of main approaches can address the imbalance in data:

Over sampling: This method balances the data by increasing the number of samples of the minority class (in this case, stroke patients). Oversampling copies or creates new data similar to existing stroke cases so the number is close to non-stroke cases.

Under sampling: This method takes the opposite approach by reducing the number of samples from the majority class (non-stroke) until the number is comparable to the minority class.

This study prioritizes the oversampling method based on two primary considerations. First, research [9] showed that under sampling produces lower accuracy than oversampling in stroke prediction. Second, this dataset has extreme imbalance (249 stroke data vs 4861 non-stroke data), so under sampling would result in only 498 total samples, which risks missing important information from the majority class and reducing the model's generalization ability.

However, to provide a comprehensive analysis and fill the gap in the literature, this study also includes a comparison with the under-sampling method. The methods used in this study are:

SMOTE (Synthetic Minority Over-sampling Technique): The most frequently used oversampling method [21]. SMOTE creates synthetic data for the minority class based on its nearest neighbors. SMOTE is included in this study as a basis for comparison with previous studies.

Borderline-SMOTE: An enhanced version of SMOTE that focuses on creating synthetic samples along the border between two classes [22]. This method is more selective in choosing data to be oversampled, especially for minority samples in the vulnerable area or border between two classes (0 or 1), because these points are more susceptible to misclassification by the model.

ADASYN (Adaptive Synthetic Sampling): This method like SMOTE, generates synthetic data for the minority class [23]. The difference is that ADASYN focuses on minority samples that are more difficult to classify [24]. The data generation process uses linear interpolation between minority samples and randomly selected minority neighbors.

Random Over Sampling (ROS): A simple approach that randomly duplicates minority class samples until class balance is achieved [25], [26]. Despite its simplicity, ROS provides a valuable foundation and can perform well when combined with a robust classification algorithm.

Random Under Sampling (RUS): An under-sampling method that randomly removes samples from the majority class until balance is achieved with the minority class. Despite the risk of information loss, RUS can provide important comparative insights to oversampling approaches.

Applying these five methods to a stroke dataset allows for a comprehensive analysis of the relative effectiveness of different approaches in dealing with data imbalance. This is important for determining the optimal strategy to improve the performance of basic classification algorithms such as Logistic Regression and Support Vector Machine for stroke prediction.

2.4 Dataset Bias Analysis

The stroke prediction dataset used in this study exhibits significant data representation and balance issues that can significantly impact the performance of predictive models. Extreme class imbalance, with stroke cases comprising only 4.87% (249) of the dataset compared

to 95.13% (4861) of non-stroke cases, creates a fundamental challenge for developing a reliable model. This imbalance naturally biases the algorithm towards the majority class, potentially resulting in misleading accuracy metrics and inadequate sensitivity to stroke risk patterns.

Demographic features of the dataset reveal concerning distributional biases. The bimodal age distribution peaking at ages 45-60 and 80+ years suggests potential sampling bias that may not correctly represent the full spectrum of age-related stroke risk. The occupational distribution heavily favors individuals in private sector jobs, introducing occupational bias despite relatively balanced urban-rural representation.

Health-related variables exhibit additional problematic patterns. The BMI distribution that is predominantly between 25-30 suggests the underrepresentation of individuals with higher BMI values. Conversely, missing values imputed to the mean reduce variance and potentially obscure the relationship between extreme BMI values and stroke risk. Similarly, the bimodal distribution of glucose levels may cause the model to treat glucose as a binary rather than a continuous risk factor. The underrepresentation of hypertension and heart disease cases in the dataset may reduce the model's sensitivity to these critical stroke risk factors.

The various sampling techniques used to address class imbalance introduce their own biases. Synthetic data generation methods such as SMOTE create artificial cases through interpolation that may not accurately reflect the real-world presentation of stroke, especially for categorical features or multimodal distributions. Borderline-SMOTE's focus on decision boundary cases may amplify noise or outliers, while ADASYN risks overemphasizing unusual stroke cases. Random Over Sampling (ROS) duplicates existing minority cases without introducing new information, potentially leading to overfitting. Conversely, Random Under Sampling (RUS) reduces the model's exposure to non-stroke variation, potentially missing important information about differences between stroke and non-stroke cases.

Models trained on such resampled data may perform satisfactorily on similarly sampled test data but suffer significant degradation when applied to real-world populations with natural distributions. The fundamental representational imbalances risk creating models that primarily learn patterns from dominant groups while performing inadequately for minority populations—precisely those for whom accurate stroke prediction could be most critical.

2.5 Data Splitting

Data splitting is one of the important steps in the machine learning process that aims to ensure that the model built cannot only learn patterns from training data but also generalize to new data that has never been seen before. This study divides the dataset into two parts: training data and testing data. Training data is used to train the model and recognize patterns in the data. In contrast, testing data objectively measures the model's performance on entirely new data independent of training.

This process is critical to prevent overfitting, which is when the model focuses too much on training data, so it cannot perform well on new data. In addition, data splitting allows for a more accurate evaluation of the model's ability to handle real situations. Ensuring proper data division, the built model can be more reliable in predicting data outside the training sample. The data in this study was divided into 70% for training data and 30% for testing data.

2.6 Classifier Logistic Regression

Logistic Regression (LR) is one of the statistical analysis methods used to model the relationship between categorical dependent variables (responses) with one or more independent variables (predictors) in the form of categorical or continuous data so that it is possible to perform classification analysis and also allows to provide information about variables that have a significant influence [27].

Logistic Regression is one of the most popular supervised learning machine learning algorithms. LR can process large amounts of data at high speed because it requires less computing capacity, such as memory and processing power [28]. The Logistic Regression model is stated in Equation 1 [29].

$$Logit\ (P) = \ln\left(\frac{P}{1-P}\right) \quad (1)$$

This equation is the core of Logistic Regression, which transforms probability (P) into logit values. This transformation makes it possible to model the relationship between predictor variables (stroke risk factors) and the probability of stroke using a linear function. The P value ranges between 0 and 1 (0% to 100% probability), while the logit can range from -∞ to +∞, allowing for more flexible modeling. In the context of this study, Equation 2 shows how the probability of stroke can be predicted from a linear combination of risk factors.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = n_0 + n_1x_1 + n_2x_2 + \cdots + n_nx_n \quad (2)$$

p is the probability of a patient having a stroke, x_1, $x_2 \ldots, x_n$ are predictor variables such as age, hypertension, and heart disease, n_0 is a constant (intercept), and $n_1, n_2, \ldots, n_n$ are coefficients that indicate how strong the influence of each risk factor is. The higher the coefficient value for a risk factor, the greater its influence on the probability of stroke.

2.7 Classifier Support Vector Machine

The Support Vector Machine (SVM) algorithm is a supervised learning method that separates data into different categories by maximizing the margin between the classes [30]. In solving classification problems

using datasets that cannot be separated linearly, this algorithm uses a kernel function (kernel trick) to map data into a high-dimensional feature space to obtain a hyperplane that separates the data into two classes [31]. Some kernels often used in SVM are polynomial, Radial Basis Function (RBF), and sigmoid, which have been proven to improve SVM performance. This kernel function uses Equations 3, 4, and 5 to produce a hyperplane in the classification process [32].

$$K_{Polynomial} = (yx^T x_i + r)^p, y > 0 \quad (3)$$

This polynomial kernel allows SVM to handle non-linear relationships with a degree of complexity p. In practice, this function helps the model identify complex patterns, such as how a combination of certain risk factors, such as age and blood pressure, disproportionately increases the risk of stroke.

$$K_{RBF} = \exp(-y||x_i - x||^2), y > 0 \quad (4)$$

The Radial Basis Function (RBF) kernel effectively handles very complex data. The parameter y (gamma) controls how much influence one data sample has on another. For stroke prediction, RBF allows the model to identify "high-risk areas" in the feature space where certain combinations of risk factors are strongly associated with stroke occurrence.

$$K_{Sigmoid} = \tanh(yx^T x_i + r) \quad (5)$$

The sigmoid kernel mimics the activation function in a neural network. The parameter c is a bias that shifts the curve. In the context of stroke prediction, this function can capture the threshold relationship where stroke risk increases sharply when a risk factor reaches a certain level.

2.8 Hyperparameter Tuning

Hyperparameter tuning is finding the best combination of hyperparameters in the model to be built to produce optimal and effective performance. In addition, hyperparameter tuning functions so that the model that is built can learn data patterns effectively so that it can avoid overfitting or underfitting where if the combination of hyperparameters is wrong, it will cause the model to become too complex (overfitting) or too simple (underfitting).

Grid Search Cross-Validation or GridSearchCV is one method that is often used for hyperparameter tuning. GridSearchCV is a technique that helps find the parameters that produce the best performance for a particular model [33]. With GridSearchCV, various sets of hyperparameters will be tested individually in a grid, allowing for a structured and consistent assessment of model performance. This process ensures that the best combination is selected based on the model evaluation results to optimize the model properly to produce more accurate and reliable predictions.

Previous studies have primarily ignored systematic hyperparameter tuning, which has the potential to contribute to less-than-optimal performance. This study applies GridSearchCV with cross-validation 10 to systematically optimize model parameters. The parameters used in this study are shown in Table 2. Optimization of these parameters addresses significant gaps in previous research and allows for a fair comparison between algorithms in their optimal configurations.

Table 2 Parameters used in LR and SVM

Logistic Regression	Support Vector Machine
Regularization strength (C)	Kernel type
Class weights	Regularization parameter (C)
Maximum iterations	Kernel coefficient (gamma)
Tolerance (tol)	Tolerance (tol)

While GridSearchCV offers a systematic approach to hyperparameter engineering, it introduces significant computational overhead that must be considered when unpacking the feasibility of model implementation. Without GridSearchCV, both models run in less than 2 seconds, but with GridSearchCV and oversampling, Logistic Regression takes 1–2 minutes, while SVM takes 13–15 minutes.

These differences in execution time reflect the characteristics of the respective algorithms. Logistic Regression uses more straightforward optimization, while SVM solves complex optimization problems with more intensive kernel calculations. An extreme case occurs when using a polynomial kernel with oversampling, increasing the SVM execution time to 232 minutes, well above the RBF and sigmoid kernels (13–15 minutes). Conversely, undersampling results in shorter computational times for both models (Logistic Regression <10 seconds, SVM 16–25 seconds) due to the reduced amount of data processed.

This information is important for practical applications in clinical settings, where the balance between accuracy and computational efficiency is critical. Although GridSearchCV improves model performance, its computational cost must be carefully considered, especially for applications requiring rapid development or frequent retraining.

2.9 Performance Evaluation

After testing, the model is evaluated to calculate the error rate made during testing. Performance evaluation of the model is critical to understand how well the prediction model works on the test data. Confusion Matrix is one of the most frequently used methods to evaluate the performance of the classification model. Confusion Matrix consists of four parts, namely True Positive (TP), True Negative (TN), False Positive (FN), and False Positive (FP), as shown in Table 3.

The value of True Positive (TP) is positive data that is predicted as positive data by the model. At the same time, True Negative (TN) is positive data predicted as unfavorable. The value of False Positive (FP) is harmful data that is predicted as positive data by the model. At the same time, false negative (FN) data is positive but is predicted to be negative. The value of False Negative

(TN) and False Negative (FN) is the error value or mistake the model makes in making predictions.

Table 3. Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Other evaluation matrices, such as accuracy, precision, and recall, can then be calculated from the confusion matrix. Accuracy will measure how many predictions are correct among all predictions made by the model. Accuracy is obtained using Equation 6.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision shows how many of the optimistic predictions are positive. Precision is stated in Equation 7.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Recall shows how much positive data is found. Recall is defined in Equation 8.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

The ROC curve is a visual representation utilized to evaluate the classifier's performance; In the ROC curve, The Sensitivity or True Positive Rate (TPR), which represents the proportion of correct predictions classified as true and the False Positive Rate (FPR), which represents to the ratio of incorrect predictions classified as correct predictions [34]. The ROC curve is used to visually depict the model's capacity to accurately classify stroke risk by evaluating the ratio of accurate and inaccurate predictions. TPR and FPR are defined in Equations 9 and 10 [35].

$$TPR = \frac{TP}{TP+FN} \quad (9)$$

$$FPR = \frac{FP}{FP+TP} \quad (10)$$

Then there is an important index of ROC, namely AUC, which is the value of the area between the ROC curve and the abscissa; A better credit risk score is indicated by a higher value of the AUC (Area Under the Curve), which ranges from 0 to 1 [35]. AUC is stated in Equation 11.

$$AUC = \frac{1+TPR-FPR}{2} \quad (10)$$

3. Results and Discussions

This study aims to improve the performance of Logistic Regression and Support Vector Machine through the use of Correlation-based Feature Selection (CFS), oversampling techniques, and GridSearchCV. Model performance evaluation is carried out using Confusion Matrix to calculate the level of error made during testing. The data used comes from Kaggle and data that has gone through the data preprocessing stage is divided into 70% training data and 30% test data.

3.1 Data Preprocessing

After the initial data set inspection, the id feature was removed as it did not affect the model performance. It handled 201 missing values in the BMI feature by imputing the mean value. The Correlation-Based Feature Selection technique produced 9 relevant features with the top ranking of the threshold of 0.01, as shown in Table 4.

Table 4. Feature Selection Results

Feature	Correlation Coefficients
Age	0.245257
Hypertension	0.134914
Heart_disease	0.131945
Ever_married	0.127904
Work_type	0.108340
Residence_type	0.038971
Avg_glucose_level	0.032316
Bmi	0.015458
Smoking_status	0.015458
Stroke (output)	(output)

Through correlation-based feature selection, this study not only validates the primacy of age, hypertension, and heart disease as major predictors of stroke but also sheds new light on their relative importance (age: 0.245257, hypertension: 0.134914, heart disease: 0.131945). These quantitative rankings provide evidence-based prioritization for clinical risk assessment, offering a deeper understanding of stroke prediction. The study highlights the relative contribution of additional factors such as marital status (0.127904) and occupation (0.108340) and validates the importance of considering socioeconomic and lifestyle factors in stroke risk assessment. This support for a more holistic approach reassures clinicians about the validity of their current practices.

Furthermore, this study shows that although mean glucose level (0.032316) and BMI (0.015458) are considered important in clinical practice, their predictive power is much lower than that of cardiovascular factors. These findings may help clinicians appropriately consider these factors in their risk assessment. The performance improvements achieved through feature selection underscore the importance of targeted clinical assessment. By focusing on the most predictive factors rather than collecting extensive patient data, clinicians can enhance the efficiency and accuracy of their risk stratification, empowering them to make more informed decisions in their practice.

After data cleaning and feature selection, one-hot encoding was applied to the categorical features, creating binary variables for each category, as shown in Figure 3. At the same time, normalization was performed on continuous numeric features to ensure consistent scaling and reduce the impact of outliers, as shown in Figure 4.

ever_married_No	ever_married_Yes	work_type_Govt_job	work_type_Never_worked	work_type_Private
0	1	0	0	1
0	1	0	0	0
0	1	0	0	1
0	1	0	0	1
0	1	0	0	0

Figure 3. Sample One-Hot Encoding Dataset

age	avg_glucose_level	bmi
1.051434	2.706375	1.005086
0.786070	2.121559	-0.098981
1.626390	-0.005028	0.472536
0.255342	1.437358	0.719327
1.582163	1.501184	-0.631531

Figure 4. Sample Normalization Dataset

3.2 Imbalance Data

The results of over sampling and under sampling are shown in Figure 5.

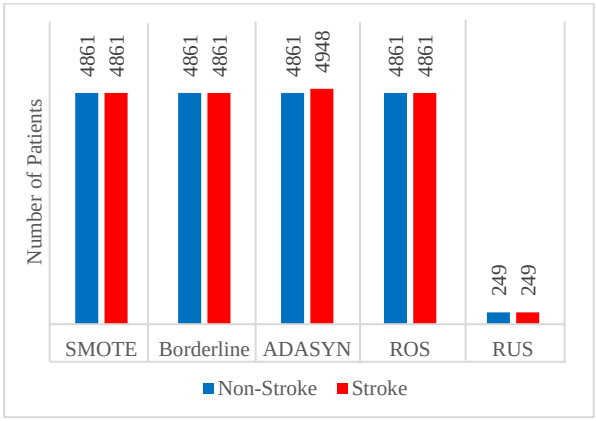

Figure 5. Amount of Data Distribution After Sampling Method

Figure 5 shows that SMOTE, Borderline-SMOTE, and ROS (Random Over Sampling) produce the same number of samples for both classes, which is 4861, indicating a perfect balance between the Non-Stroke and Stroke classes. SMOTE creates balanced datasets in each class by generating synthetic samples that maintain statistical properties while avoiding simple duplication. However, it may produce unrealistic samples if feature correlations are complex or create class overlapping.

Borderline SMOTE similarly achieves balance but focuses specifically on border regions between classes where classification is most challenging, making it more effective for complex boundaries despite being computationally expensive and potentially struggling with sparse minority classes. Random Over-Sampling (ROS) offers simplicity with minimal computational overhead and preserves original data points without modification but risks overfitting by creating exact duplicates without adding new information.

ADASYN (Adaptive Synthetic Sampling) takes a different approach by adaptively generating more synthetic data for difficult-to-learn minority samples, resulting in slightly more stroke samples (4948) than non-stroke (4861), which helps with complex decision boundaries but may amplify noise and create regional imbalances. Finally, Random Under-Sampling (RUS) drastically reduces the dataset to only 249 samples per class, which significantly reduces training time and simplifies the decision boundary but discards potentially valuable majority class information and can lead to underfitting and increased susceptibility to outliers.

The selection of the proper sampling method depends on the dataset's characteristics. In this study, ADASYN shows a slightly different approach by generating non-exact distributions for both classes, which may be more adaptive to the learning difficulty in the dataset. Furthermore, data splitting is performed with 70% training data and 30% test data.

3.3 Hyperparameter Tuning

Hyperparameter tuning is performed using GridSearchCV with 10-fold cross-validation, which performs an exhaustive search on various parameter combinations to find the optimal configuration. With 10-fold cross-validation, this process becomes more robust because the data will be divided into 10 parts (folds) that are used alternately as validation data. Table 5 shows the best parameter results obtained by the Logistic Regression algorithm.

Table 5 The Best Parameters Logistic Regression

	Tol	C	Class_ weight	Max_i ter	Score
SMOTE	1e-6	10	None	100	0.837
Borderline-SMOTE	1e-4	10	Balanced	100	0.861
ADASYN	1e-5	100	Balanced	200	0.835
Random Over Sampling	1e-4	0.1	None	100	0.776
Random Under Sampling	1e-4	0.1	None	100	0.756

Table 5 shows the parameter optimization results for the Logistic Regression model applied to five different sampling methods. For the Borderline-SMOTE method, the model achieved the best performance with a score of 0.861 using a tolerance (Tol) of 1e-4, a regularization parameter (C) of 10, class_weight balanced, and a max_iter of 100, indicating that Borderline-SMOTE performed better than other methods in balancing the dataset.

SMOTE and ADASYN also performed competitively, with scores of 0.837 and 0.835, respectively. SMOTE was optimal with a tighter tolerance (1e-6) and C=10,

while ADASYN required a higher C value (100) and a larger max_iter (200), indicating higher complexity in the data it generated.

Random Over Sampling and Random Under Sampling performed lower, with scores of 0.776 and 0.756. Both methods achieve optimal results with identical parameters (Tol=1e-4, C=0.1, class_weight=None, max_iter=100). A smaller C value (0.1) indicates that stronger regularization is needed to prevent overfitting the data generated by these methods. The class_weight parameter has a significant effect, with "Balanced" giving the best results for Borderline-SMOTE and ADASYN. At the same time, the other methods do not require additional weighting, indicating that both methods generate data distributions that benefit from adjusting the class weights in the classification algorithm. Furthermore, Table 6 shows the best parameter results the Support Vector Machine obtained.

Table 6. The Best Parameters Support Vector Machine

	Kernel	C	Gamma	Tol	Score
SMOTE	RBF	100	1	1e-3	0.935
Borderline-SMOTE	RBF	1	10	1e-3	0.954
ADASYN	RBF	100	1	1e-3	0.967
Random Over Sampling	RBF	1	10	1e-3	0.992
Random Under Sampling	RBF	1	Scale	1e-3	0.787

Table 6 shows that the RBF kernel is more dominant than the other kernels, indicating that the data is not linearly separable and requires transformation to a higher dimension. The tolerance (Tol) value is consistent across all methods at 1e-3, which determines the stopping criterion for the optimization algorithm.

Random Over Sampling scored the highest at 0.992, with parameters C=1 and gamma=10. This lower C value indicates that the model emphasizes wider margins for better generalization. A high gamma indicates that the model focuses on data points that are closer together, thus creating a more complex and flexible decision boundary.

ADASYN came in second with a score of 0.967, using C=100 and gamma=1. A higher C value indicates that the model emphasizes minimizing the classification error, while a lower gamma creates a slightly smoother decision boundary. Borderline-SMOTE scored 0.954 with C=1 and gamma=10, while conventional SMOTE scored 0.935 with C=100 and gamma=1. These differences reflect the differences in the data distribution characteristics produced by each method.

Random Under Sampling showed the lowest performance with a score of 0.787, using C=1 and gamma="scale". The use of gamma="scale" indicates that the gamma value is automatically determined based on the data's number of features and variance, which may be suboptimal for the smaller datasets generated by this method.

The 10-fold cross-validation used in GridSearchCV ensures that the selected parameters have good generalization ability and are robust to variations in the data, which reduces the risk of overfitting and provides higher confidence in the model's performance when applied to new data.

From these results, the selection of oversampling methods and model parameters highly depends on the algorithm used. Logistic Regression benefits more from the Borderline-SMOTE method, which adjusts the weights of the minority class more adaptively. At the same time, SVM shows the best performance with Random Over Sampling, which maintains the original characteristics of the data. Therefore, in dealing with imbalanced datasets, it is important to consider the nature of the model used and tune the parameters optimally to achieve the best results.

3.4 Model Evaluation Performance

This study uses the Logistic Regression and Support Vector Machine algorithm models to predict stroke disease using five sampling methods and the best parameters previously obtained through GridSearchCV with 10-fold cross-validation. Each method is evaluated using three main metrics: Recall, Precision, and Accuracy, while the ROC curve serves as a visual representation for performance evaluation. The performance of Logistic Regression (LR) is shown in Figures 6 and 7.

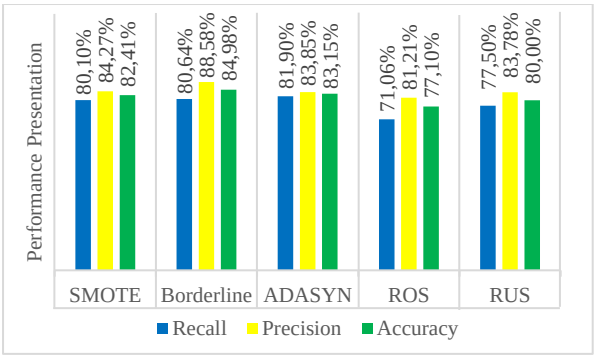

Figure 6. Performance Presentation Logistic Regression

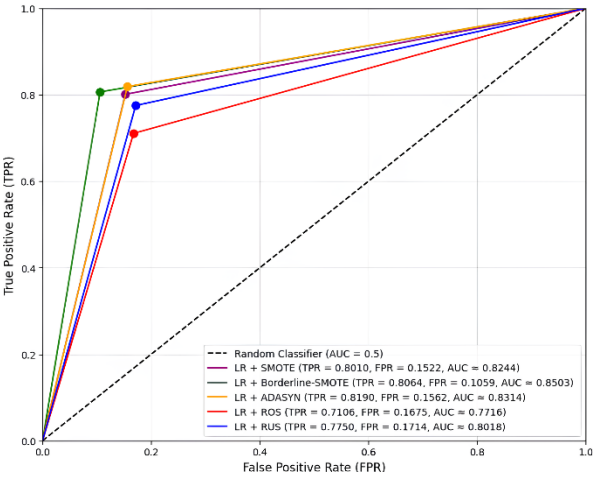

Figure 7. ROC Curve of Logistic Regression

Figures 6 and 7 show the performance of the Logistic Regression model on the stroke dataset with the five different standard sampling methods. LR + Borderline-SMOTE showed the strongest overall performance with the highest accuracy (84.98%), highest precision (88.58%), and Recall of 80.64%, indicating that the borderline approach's focus on hard-to-tell examples near class boundaries appears to produce robust and effective results, especially in precision. The ROC curves confirm this assessment, showing an AUC of 0.8503, the highest among all models, with a good balance between TPR (0.8064) and FPR (0.1059).

LR + ADASYN followed closely with the highest Recall (81.90%), strong precision (83.85%), and solid accuracy (83.15%), indicating that the adaptive synthetic sampling approach effectively identified stroke cases while maintaining good overall classification performance. Its ROC curve shows a TPR of 0.8190, FPR of 0.1562, and AUC of 0.8314, making it the second-best performer in terms of AUC, although its Precision and Accuracy are slightly lower than Borderline-SMOTE.

LR + SMOTE shows a competitive performance with a recall of 80.10%, a precision of 84.27%, and an accuracy of 82.41%. Its ROC curve shows a TPR of 0.8010, an FPR of 0.1522, and an AUC of 0.8244, making it the third-best performer in terms of AUC.

LR + RUS (Random Under-Sampling) shows a fairly good performance with a recall of 77.50%, a precision of 83.78%, and an accuracy of 80.00% despite the drastic decrease in the dataset size. This efficiency is reflected in the ROC metrics (TPR: 0.7750, FPR: 0.1714, AUC: 0.8018), indicating that most of the information in the class is likely to be redundant.

LR + ROS (Random Over-Sampling) showed the weakest performance among these techniques, with the lowest Recall (71.06%), precision (81.21%), and accuracy (77.10%). This poor performance is consistent with the ROC curve showing a TPR of 0.7675, the highest FPR of 0.1985, and the lowest AUC of 0.7718, which is most likely due to the overfitting of the minority duplicate samples.

These results indicate that choosing the right sampling method can significantly affect the performance of the Logistic Regression model in the context of stroke disease classification. Logistic Regression combined with Borderline-SMOTE and ADASYN proved to be more effective in handling class cohesion than other methods. The choice between Borderline-SMOTE and ADASYN will ultimately depend on clinical priorities: Borderline-SMOTE excels in minimizing false positives, while ADASYN maximizes the identification of true stroke cases.

Then, the performance of the Support Vector Machine (SVM) is shown in Figures 8 and 9.

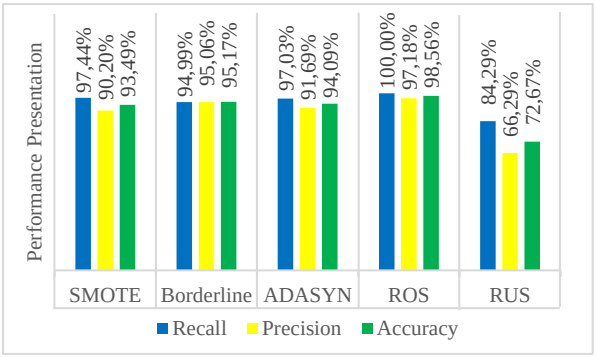

Figure 8. Performance Presentation Support Vector Machine

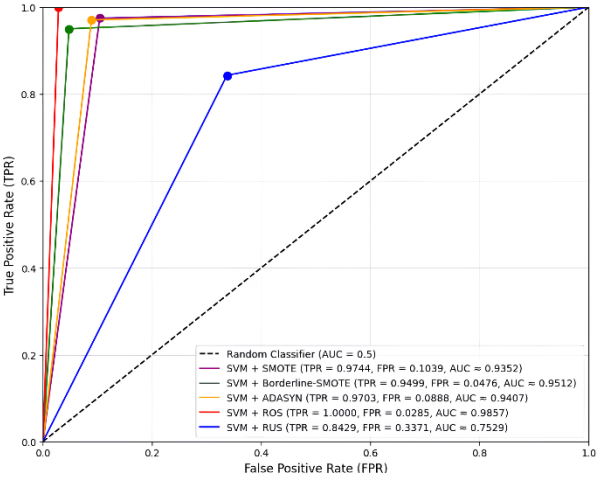

Figure 9. ROC Curve of Support Vector Machine

Figures 8 and 9 show the performance of the Logistic Regression model on the stroke dataset with the five different standard sampling methods. The SVM + ROS (Random Over-Sampling) model achieved perfect recall (100%) with high precision (97.18%) and excellent accuracy (98.56%). This performance is reflected in its ROC metric, which showed a TPR of 1.0000, a very low FPR of 0.0285, and an impressive AUC of 0.9857. The combination of perfect sensitivity and excellent specificity makes this approach very important.

SVM + Borderline-SMOTE delivered an excellent balanced performance with a recall of 94.99%, precision of 95.06%, and accuracy of 95.17%. Its ROC curve showed a TPR of 0.9499, a very low FPR of 0.0476, and a high AUC of 0.9512, indicating excellent discrimination ability while maintaining a good balance between sensitivity and specificity, making it the second best in terms of AUC.

SVM + SMOTE performed impressively with 97.44% recall, 90.20% precision, and 93.49% accuracy. The corresponding ROC metrics (TPR: 0.9744, FPR: 0.1039, AUC: 0.9352) confirmed its strong performance, albeit with a slightly higher false positive rate than some alternatives.

SVM + ADASYN showed similar strength with 97.03% recall, 91.69% precision, and 94.09% accuracy. Its ROC values (TPR: 0.9703, FPR: 0.0888, AUC:

0.9407) position it as a highly effective approach, balancing high sensitivity with reasonable specificity.

SVM + RUS (Random Under-Sampling) shows significantly weaker performance with 84.29% recall, 66.29% precision, and 72.67% accuracy. This significant performance gap is reflected in the ROC metrics (TPR: 0.8429, FPR: 0.3371, AUC: 0.7529), indicating that the information loss due to the reduced random under-sampling has a significant impact on the classification ability of SVM.

Among the sampling methods, Random Over-Sampling (ROS) with SVM gave the best results with perfect recall and the highest AUC (0.9857), making it the optimal choice for cases where detecting all stroke patients is crucial. However, Borderline-SMOTE with SVM offers a better balance between recall (94.99%) and precision (95.06%) with a very low FPR (0.0476), making it more suitable for resource-constrained environments.

Although the perfect recall (100%) achieved by the SVM + ROS model initially seems ideal for stroke prediction, it requires careful consideration. Perfect recall means that all true stroke cases are identified, which is critical for life-threatening conditions where a missed diagnosis can have severe consequences. In a medical context, the perfect recall of SVM + ROS is invaluable for life-threatening conditions such as stroke. However, it must be weighed against the clinical implications of small numbers of false positives. Despite its relatively high precision (97.18%), the model will produce some false positives that may lead to unnecessary diagnostic procedures, specialist consultations, and preventive interventions for patients who are not actually at risk.

These false-positive stroke risk predictions can cause significant anxiety and stress for patients and their families, potentially leading to decreased quality of life and increased healthcare utilization for anxiety-related problems. Even with a low FPR of 0.0285, when applied to a large population, this can result in a large number of false positives, potentially overwhelming specialized stroke care resources. The economic impact of false-positive investigations must be weighed against the benefits of capturing every case of true stroke risk, especially in resource-constrained healthcare settings. Furthermore, stroke prevention often involves other interventions with their risk profiles, and administering these treatments to false-positive cases exposes patients to unnecessary risk. The choice between perfect recall and a more balanced performance metric ultimately depends on the specific clinical context, resource availability, and the relative costs of false negatives versus false positives in a given healthcare setting.

These results show that the Support Vector Machine model outperforms Logistic Regression. The superior performance of SVM can be attributed to its ability to handle non-linear relationships and create more complex decision boundaries, which is very beneficial for the stroke disease classification task. The significant performance improvement with Random Over Sampling for the Support Vector Machine is particularly noteworthy, as this method showed the weakest performance with Logistic Regression, underscoring the importance of considering sampling methods and classification algorithms as an integrated system when developing predictive models for imbalanced datasets.

3.5 Comparison Logistic Regression with Previous Studies

The comparison between the current study and previous studies on stroke prediction using Logistic Regression (LR) with various sampling methods is shown in Table 7. The results show significant improvement in the current study, which can be attributed to several factors, including parameter optimization using GridSearchCV and the practical application of the Feature Selection (FS) technique.

Table 7. Result Comparison With Previous Studies

Machine Learning Models	Accuracy	References
LR + SMOTE	75%	[6]
LR + SMOTE	79%	[7]
LR + SMOTE	79%	[8]
LR + RUS	78%	[9]
LR + FS + GridSearchCV + SMOTE	82.42%	This research
LR + FS + GridSearchCV + Borderline-SMOTE	84.98%	This research
LR + FS + GridSearchCV + ADASYN	83.15%	This research
LR + FS + GridSearchCV + ROS	77.10%	This research
LR + FS + GridSearchCV + RUS	80.00%	This research

The current study used correlation-based feature selection, as shown in Table 3, which identified the most relevant features for stroke prediction. Age emerged as the most significant predictor with a correlation coefficient of 0.24527, followed by hypertension (0.13914) and heart disease (0.13194). Other important features included marital status, occupation, residence type, mean glucose level, BMI, and smoking status.

By focusing on these highly correlated features, Logistic Regression achieved superior performance compared to previous studies. The combination of LR + Borderline-SMOTE achieved the highest accuracy of 84.98%, substantially improving over the reported best result of 79%. Similarly, LR + ADASYN (83.15%), LR + SMOTE (82.42%), and LR + RUS (80.00%) all outperformed previous implementations by a significant margin.

The feature selection process significantly improved model performance by reducing dimensionality and focusing on the most predictive variables. This approach minimizes noise and potential overfitting from less relevant features. The correlation-based selection method provides a clear ranking of the importance of features, allowing the model to

concentrate on the relationships that most strongly indicate stroke risk.

The accuracy improvement can be attributed to several factors: (1) the application of Feature Selection techniques that select the most relevant and influential features with stroke, (2) the application of advanced sampling techniques such as Borderline-SMOTE and ADASYN, and (3) thorough parameter optimization through GridSearchCV with 10-fold cross-validation.

These results demonstrate the importance of choosing the proper sampling method, optimizing parameters, and applying effective feature selection techniques when dealing with medical datasets. The current study shows that a comprehensive approach that addresses data imbalance, parameter tuning, and feature relevance can significantly improve the accuracy of stroke prediction models.

4. Conclusions

This study revealed a breakthrough in stroke prediction accuracy by showing that the borderline-smote method, combined with logistics regression, could achieve an unprecedented accuracy of 84.98% with an AUC value of 0.8503. This is far beyond the previous study, which only reached an accuracy of 79%. The Support Vector Machine method, which is combined with the Random Over Sampling technique, also gives extraordinary results, with a perfect recall value of 100% and accuracy of 95.56% with an AUC value of 0.9857, which is almost perfect. These results underline the important role of selecting features and adjusting parameters in developing predictive models.

While the results of this study are promising, the model's applicability to a broader and more diverse population is currently limited due to potential biases in the training data set. The data may not fully represent the global population's demographic, genetic, and environmental variations, which could affect the model's accuracy and reliability. Therefore, before the stroke prediction model can be widely applied in clinical practice, it must undergo rigorous external validation on independent data sets from various populations and clinical environments. This validation process ensures the model's reliability and applicability in diverse healthcare scenarios.

The stroke prediction model presented in this study has the potential to serve as a powerful tool for making clinical decisions in diverse healthcare environments. However, its successful implementation is contingent on its seamless integration with local clinical knowledge and considering unique population factors. This study lays a robust methodological foundation for developing an improved stroke prediction model. However, it is crucial to conduct further research and validation in more diverse populations before its widespread application.

Future research must focus on external validation by engaging health professionals and using clinical datasets from various regions or hospitals with diverse ethnicities, age groups, and socioeconomic conditions. Researchers should also explore the integration of additional clinical variables, such as medical history, family stroke history, and lifestyle factors, to enhance prediction accuracy. Furthermore, future research should investigate the potential benefits of merging various sampling techniques (hybrid sampling methods). For instance, testing a combination of Smote with Tomek Links or Smote with Edited Nearest Neighbors could yield improved results.

References

[1] T. N. Rochmah, I. T. Rahmawati, M. Dahlui, W. Budiarto, and N. Bilqis, “Economic burden of stroke disease: A systematic review,” Int. J. Environ. Res. Public Health, vol. 18, no. 14, 2021, doi: 10.3390/ijerph18147552.

[2] U. N. Wisesty, T. A. B. Wirayuda, F. Sthevanie, and R. Rismala, “Analysis of Data and Feature Processing on Stroke Prediction using Wide Range Machine Learning Model,” J. Online Inform., vol. 9, no. 1, pp. 29–40, 2024, doi: 10.15575/join.v9i1.1249.

[3] B. W. Negasa, T. W. Wotale, M. E. Lelisho, L. K. Debusho, K. Sisay, and W. Gezimu, “Modeling Survival Time to Death among Stroke Patients at Jimma University Medical Center, Southwest Ethiopia: A Retrospective Cohort Study,” Stroke Res. Treat., vol. 2023, 2023, doi: 10.1155/2023/1557133.

[4] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[5] V. R. Modhugu and S. Ponnusamy, “Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree,” Asian J. Res. Comput. Sci., vol. 17, no. 6, pp. 188–201, 2024, doi: 10.9734/ajrcos/2024/v17i6467.

[6] S. Ghanipour and S. Yousefzadeh Boroujeni, “Stroke Prediction with Logistic Regression and assessing it using Confusion Matrix,” no. October, 2022, [Online]. Available: https://www.researchgate.net/publication/364359247

[7] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” Sensors, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[8] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/7633381.

[9] G. Sailasya and G. L. A. Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.

[10] E. Utami, “Enhanced Heart Disease Diagnosis Using Machine Learning Algorithms: A Comparison of Feature Selection,” vol. 9, no. 2, pp. 385–392, 2025, doi: 10.29207/resti.v9i2.6175.

[11] W. Aprilliandhika and F. Fauzi Abdulloh, “Comparison of K-Nearest Neighbor and Support Vector Machine Algorithm Optimization With Grid Search Cv on Stroke Prediction,” vol. 5, no. 4, pp. 991–1000, 2024, doi: 10.52436/1.jutif.2024.5.4.1951.

[12] 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, doi: 10.1016/j.ijrobp.2021.08.007.

[13] J. Premsmith and H. Ketmaneechairat, “A predictive model for heart disease detection using data mining techniques,” J. Adv. Inf. Technol., vol. 12, no. 1, pp. 14–20, 2021, doi: 10.12720/jait.12.1.14-20.

[14] C. Gupta, A. Saha, N. V. S. Reddy, and U. D. Acharya, “Cardiac Disease Prediction using Supervised Machine Learning Techniques,” J. Phys. Conf. Ser., vol. 2161, no. 1, 2022, doi: 10.1088/1742-6596/2161/1/012013.

[15] Fedesoriano, "Stroke Prediction Dataset." 2020. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

[16] A. N. Puteri, A. Arizal, and A. D. Achmad, "Feature Selection Correlation-Based on Bank Telemarketing Customer Predictions for Time Deposits," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 20, no. 2, pp. 335–342, 2021, doi: 10.30812/matrik.v20i2.1183.

[17] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," Sci. Rep., vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-49962-w.

[18] T. Al-shehari and R. A. Alsowail, "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques," Entropy, vol. 23, no. 10, 2021, doi: 10.3390/e23101258.

[19] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?," Emerg. Mark. Financ. Trade, vol. 58, no. 2, pp. 472–482, 2020, doi: 10.1080/1540496X.2020.1825935.

[20] G. A. B. Suryanegara, Adiwijaya, and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[21] M. K. Rezki, M. I. Mazdadi, F. Indriani, Muliadi, T. H. Saragih, and V. A. Athavale, "Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine," J. Electron. Electromed. Eng. Med. Informatics, vol. 6, no. 4, pp. 343–354, 2024, doi: 10.35882/jeeemi v6i4.434.

[22] Y. Sun et al., "Borderline SMOTE Algorithm and Feature Selectiom-Based Network Anomalies Detection Strategy," Energies, vol. 15, no. 13, 2022, doi: 10.3390/en15134751.

[23] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," Technologies, vol. 13, no. 3, pp. 1–40, 2025, doi: 10.3390/technologies13030088.

[24] G. Ahmed et al., "DAD-Net: Classification of Alzheimer's Disease Using ADASYN Oversampling Technique and Optimized Neural Network," Molecules, vol. 27, no. 7085, pp. 1–21, 2022, doi: 10.3390/molecules27207085.

[25] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," Inf., vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[26] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," Int. J. Artif. Intell. Res., vol. 4, no. 2, pp. 86–94, 2020, doi: 10.29099/ijair.v4i2.152.

[27] S. Annas, A. Aswi, M. Abdy, and B. Poerwanto, "Stroke Classification Model using Logistic Regression," J. Phys. Conf. Ser., vol. 2123, no. 1, 2021, doi: 10.1088/1742-6596/2123/1/012016.

[28] Y. Dani and M. A. Ginting, "Classification of Predicting Customer Ad Clicks Using Logistic Regression and k-Nearest Neighbors," Int. J. Informatics Vis., vol. 7, no. 1, pp. 98–104, 2023, doi: 10.30630/joiv.7.1.1017.

[29] R. A. Khan et al., "A novel framework for classification of two-class motor imagery EEG signals using logistic regression classification algorithm," PLoS One, vol. 18, no. 9 September, pp. 1–18, 2023, doi: 10.1371/journal.pone.0276133.

[30] M. M. Siregar, R. Hizria, and D. Pardede, "Perbandingan Kinerja Kernel SVM dalam Klasifikasi Kategori Kanker Kulit Menggunakan Transfer Learning," vol. 4, no. 1, pp. 83–90, 2024, doi: 10.47709/dsi.v4i1.4665.

[31] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Inf., vol. 15, no. 4, 2024, doi: 10.3390/info15040235.

[32] E. Winarno, W. Hadikurniawati, A. Septiarini, and H. Hamdani, "Analysis of color features performance using support vector machine with multi-kernel for batik classification," Int. J. Adv. Intell. Informatics, vol. 8, no. 2, pp. 151–164, 2022, doi: 10.26555/ijain.v8i2.821.

[33] M. R. Siregar, D. Hartama, I. Engineering, S. Program, I. Systems, and S. Program, "OPTIMIZING THE KNN ALGORITHM FOR CLASSIFYING CHRONIC," vol. 10, no. 3, pp. 680–689, 2025, doi: 10.33480/jitk.v10i3.6214.

[34] F. A. Nasution, S. Saadah, and P. E. Yunanto, "Credit Risk Detection in Peer-to-Peer Lending Using CatBoost," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 7, no. 5, pp. 1056–1062, 2023, doi: 10.29207/resti.v7i5.5139.

[35] N. Hafidz and D. Yanti Liliana, "Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 2, pp. 213–219, 2021, doi: 10.29207/resti.v5i2.2960.