Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 2, April 2025, pp: 404-421; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received December 8, 2024; Revised February 10, 2025; Accepted March 1, 2025; date of publication March 25, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i2.698

How to cite: Luh Ayu Martini, Gede Angga Pradipta, and Roy Rudolf Huizen, "Analysis of the Impact of Data Oversampling on the Support Vector Machine Method for Stroke Disease Classification", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 2, pp. 404-421, April 2025.

Analysis of the Impact of Data Oversampling on the Support Vector Machine Method for Stroke Disease Classification

Luh Ayu Martini, Gede Angga Pradipta, and Roy Rudolf Huizen

Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia

Corresponding author: Gede Angga Pradipta (e-mail: angga_pradipta@stikom-bali.ac.id).

ABSTRACT Data imbalance is a critical challenge in the classification of medical data, particularly in stroke disease prediction, a life-threatening condition requiring immediate intervention. This imbalance arises due to the disproportionate number of non-stroke cases compared to stroke cases, which can lead to biased models favoring the majority class. Consequently, the model may struggle to correctly identify stroke cases, resulting in lower recall and an increased risk of misdiagnosis. This study evaluates the impact of various oversampling techniques, including Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE, SMOTE-Edited Nearest Neighbor (SMOTE-ENN), and SMOTE-Instance Prototypes Filtering (SMOTE-IPF), along with feature selection using Information Gain and Chi-Square, to assess their influence on model performance. Oversampling is utilized to address class imbalance by generating synthetic samples, thereby improving the representation of the minority class. Feature selection is employed to eliminate irrelevant or redundant features, enhancing both interpretability and computational efficiency. The dataset obtained from Kaggle, consists of 5,110 records and 12 features. Support Vector Machine (SVM) is used as the classification algorithm, with evaluations conducted on Linear, Radial Basis Function (RBF), and Polynomial kernels. Experimental results indicate that the highest performance is achieved by the combination of Borderline-SMOTE and the RBF kernel, yielding an accuracy of 96.86%, precision of 98.65%, recall of 94.99%, and an F1-score of 96.79%. This model outperforms others in stroke disease classification, demonstrating that the integration of oversampling techniques can effectively enhance prediction accuracy. Future research could focus on implementing deep learning-based models to further optimize stroke classification in the case of imbalanced data. These advancements are expected to enhance model performance, leading to a more effective and efficient approach for medical datasets.

INDEX TERMS Stroke, Machine Learning, Imbalanced Data, Oversampling, Feature Selection, Support Vector Machine.

I. INTRODUCTION

Stroke is a serious and life-threatening condition for those affected. Data from the Institute for Health Metrics and Evaluation (IHME) in 2019 shows that stroke is the leading cause of death in Indonesia (19.42% of total deaths). Stroke is defined by the World Health Organization (WHO) as a rapid or sudden clinical manifestation of focal brain function deficits that persist for 24 hours or more or result in death, with no apparent cause other than vascular factors. Classification can aid in identifying the characteristics of stroke in patients. Stroke can be diagnosed and classified using technologies such as machine learning [1].

Data imbalance is a major challenge in stroke disease classification, occurring when the number of samples in the majority class (patients without stroke) significantly exceeds those in the minority class (patients with stroke) [2]. This imbalance causes machine learning models to be biased toward the majority class, reducing their ability to classify the minority class effectively. Consequently, the model may yield a low recall for the minority class, increasing the risk of undetected stroke patients. To address data imbalance, oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) can generate synthetic data for the minority class. This allows the machine learning model to learn better from the minority class distribution without

disregarding the majority class [3]. In addition to SMOTE, several variant techniques focus on harder-to-classify samples, such as Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF [4]. Borderline-SMOTE generates synthetic data around the decision boundary between the majority and minority classes, while SMOTE-ENN combines SMOTE with the Edited Nearest Neighbor algorithm to remove overlapping or misclassified samples. SMOTE-IPF iteratively eliminates majority class samples contributing to imbalance, thereby improving class distribution.

Oversampling is crucial as it enhances the model's sensitivity to stroke patients, ultimately contributing to a more accurate diagnosis process. The effectiveness of oversampling techniques heavily depends on the characteristics of the dataset. In imbalanced datasets, oversampling can significantly improve model performance by addressing class imbalance, making the model more responsive to the minority class. However, in datasets that are already relatively balanced, oversampling may not provide additional benefits and could even lead to overfitting if the model becomes overly focused on the synthetic data generated. Therefore, it is essential to carefully assess the dataset's characteristics before deciding to apply oversampling techniques.

Several studies have explored stroke disease classification and the application of oversampling techniques to imbalanced datasets. Priyanka Bathla and Rajneesh Kumar [2] compared five classification algorithms: Naive Bayes, Support Vector Machine (SVM), Random Forest, Adaptive Boosting, and XGBoost, while implementing SMOTE to handle class imbalance. Their findings indicated that the combination of Feature Importance and Random Forest achieved the highest accuracy, whereas SVM with Feature Importance yielded a lower accuracy of 76.84%. This study highlights the importance of thoroughly evaluating oversampling techniques and feature selection to enhance SVM model performance. Chowdhury *et al.* [3] compared oversampling techniques like ENN, SMOTE-N, SMOTE-Tomek, and SMOTE-ENN on an imbalanced BRFSS dataset. The results showed that ENN with Gradient Boosting achieved the highest recall in detecting the minority class, while SMOTE-Tomek produced the best accuracy (74.2%). Additionally, Katerina Iscra *et al.* [5] used SMOTENC and other oversampling techniques on a stroke patient dataset with epileptiform EEG patterns, finding the best performance with Naive Bayes (72%) and Neural Networks (74%). Sushila Paliwal *et al*. [6] applied various classifiers, including Logistic Regression, Decision Tree, Random Forest, SVM, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Voting Classifier, for stroke disease classification. The SVM model achieved an accuracy of 84.06% using SMOTE and 83.13% using SMOTENC. Ashrafuzzaman *et al.* used a Convolutional Neural Network (CNN) model for stroke disease classification and achieved the highest accuracy of 95.5% [7]. Another study [8] combined clustering techniques (k-means) with classifiers to enhance stroke severity prediction, finding that the k-means with Artificial Neural Network (ANN) model yielded the best results, with 89% sensitivity, 89% specificity, and 90% accuracy, and an AUC-ROC of 96%.

He *et al.* developed and validated a prediction model for ischemic stroke patient discharge outcomes using machine learning, with Random Forest performing best, achieving an AUC of 90.3% [9]. Another study compared pre-processed and non-pre-processed datasets for stroke risk prediction using K-Nearest Neighbor (KNN), Decision Tree, and SVM, with Decision Tree achieving the highest accuracy (92.05%) and precision (96%) on pre-processed data [10]. A study on stroke classification using feature extraction techniques such as PCA, FA, and FPCA combined with machine learning algorithms (Logistic Regression, Random Forest, KNN, SVM, and Gradient Boosting) on the Kaggle Stroke Prediction Dataset (5,110 records, 12 features) found that Random Forest with PCA-FA achieved the best performance, with an accuracy of 92.55%, precision of 90.53%, and recall of 94.35% [11]. A study [12] focused on Random Forest for stroke classification using Mutual Information feature selection, achieving a classification accuracy of 77.90%.

Machine learning methods can be combined with feature selection to achieve better results [13], [14]. This study also evaluates the feature selection techniques Information Gain and Chi-Square to assess their impact on model performance. Information Gain measures the contribution of each feature in reducing data uncertainty by determining the best attributes through entropy calculation [15], while Chi-Square evaluates the relationship between features and the target class [16]. The purpose of feature selection is to analyze whether these techniques enhance classification performance or if their impact is insignificant for the given dataset. Support Vector Machine (SVM) is chosen as the classification algorithm due to its strong generalization ability, making it effective for classification, regression, and clustering tasks [17]. SVM operates by finding an optimal hyperplane that separates classes within the dataset and offers flexibility in using various kernel types, such as Linear, Radial Basis Function (RBF), and Polynomial [18]. Selecting the appropriate kernel is crucial for improving model performance, particularly in imbalanced datasets. Although several studies have discussed the implementation of SMOTE and other oversampling techniques on imbalanced datasets, a comprehensive analysis of their impact on stroke disease classification performance using Support Vector Machine (SVM) with Linear, Radial Basis Function (RBF), and Polynomial kernels remains limited and requires further exploration.

Previous studies have examined the use of oversampling techniques in stroke classification. Priyanka Bathla and Rajneesh Kumar [2] compared SMOTE with an SVM model, but their study did not explore the impact of different SVM kernel types. Similarly, Chowdhury *et al*. [3] tested SMOTE-Tomek but did not discuss the application of oversampling techniques with different SVM kernels. Therefore, this study addresses this research gap by analyzing the effects of various oversampling techniques on the performance of SVM with different kernels in a single experiment. By exploring the interaction between oversampling methods and SVM kernel

types, this research aims to provide deeper insights into the optimal combination that can enhance model performance when handling imbalanced datasets.

The primary issue addressed in this study is how various oversampling techniques, including SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF, influence the performance of SVM models on imbalanced datasets. The main focus of this research is to evaluate the impact of oversampling techniques on stroke classification using three different SVM kernels: Linear, RBF, and Polynomial. Additionally, this study will assess the effectiveness of feature selection methods, namely Information Gain and Chi-Square, in improving accuracy, precision, recall, and F1-score on an imbalanced dataset.

Although various studies have applied oversampling techniques and SVM in stroke classification, several research gaps remain to be explored further. First, previous studies primarily used SMOTE without comparing other oversampling techniques, such as Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF, which may have different effects on stroke classification performance. Second, no study has systematically evaluated how different oversampling techniques influence the performance of various SVM kernels. The choice of kernel is crucial for the effectiveness of oversampling, as each kernel handles imbalanced data distributions differently. Third, prior research has shown that feature selection methods like PCA and RFE reduce the accuracy of SVM models. However, it remains unexplored whether Information Gain and Chi-Square feature selection can enhance model performance when combined with oversampling techniques.

Based on these research gaps, this study contributes by evaluating four oversampling techniques, SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF to address data imbalance in stroke classification. This study also compares the performance of the Support Vector Machine (SVM) algorithm with three different kernel types: Linear, Radial Basis Function (RBF), and Polynomial, which have not been extensively explored in previous literature. Most prior studies have only examined one or two oversampling techniques and have not simultaneously compared all SVM kernels. Some previous studies have reported suboptimal results, with low accuracy for SVM algorithms on imbalanced datasets. Therefore, this study aims to make a significant contribution to improving the accuracy and effectiveness of stroke diagnosis through the application of machine learning techniques, particularly SVM. Additionally, this study integrates feature selection using Information Gain and Chi-Square to evaluate the impact of feature selection on model performance when dealing with imbalanced stroke datasets. No prior study has comprehensively compared multiple oversampling techniques with different SVM kernels within a single experiment for stroke classification. Thus, this research is expected to provide new insights into selecting the optimal strategy for handling imbalanced datasets.

This study aims to analyze the impact of oversampling techniques on the performance of the SVM algorithm with Linear, RBF, and Polynomial kernels on imbalanced stroke datasets. This research also compares models using oversampling techniques with models that do not understand their effects on accuracy, precision, recall, and F1-score.

This study makes a novel contribution by evaluating four oversampling techniques SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF to address data imbalance in stroke classification. Additionally, it compares the performance of the Support Vector Machine (SVM) algorithm with three different kernel types: Linear, Radial Basis Function (RBF), and Polynomial, which have not been extensively explored in previous studies. Prior research has primarily focused on one or two oversampling techniques and has not systematically compared all SVM kernels within a single study. Moreover, previous studies have reported suboptimal results, with low accuracy for SVM models on imbalanced datasets. Therefore, this research aims to make a significant contribution to improving the accuracy and effectiveness of stroke diagnosis through the application of machine learning techniques, particularly SVM. Furthermore, this study integrates feature selection methods, namely Information Gain and Chi-Square, to assess their impact on model performance in handling imbalanced stroke datasets.

II. METHODOLOGY

The research framework is designed to analyze and compare the impact of various oversampling techniques and feature selection methods on the performance of the Support Vector Machine (SVM) algorithm in stroke disease classification on an imbalanced dataset, as illustrated in Figure 1. This study employs a structured methodology to assess the effects of oversampling techniques, feature selection methods, and different SVM kernels on the classification performance of stroke in an imbalanced dataset.

The Support Vector Machine (SVM) was selected as the primary classification algorithm in this study due to its effectiveness in handling limited and imbalanced datasets, as well as its strong generalization capabilities. Additionally, SVM offers flexibility in utilizing various kernel types, allowing the model to adapt class separation based on complex data distribution patterns. SVM also excels in constructing optimal decision boundaries by maximizing the margin between classes, which is particularly crucial for datasets with class imbalances.

Although some prior studies have reported that SVM performs worse than other methods, these findings are often influenced by suboptimal kernel selection and imbalance-handling techniques. Therefore, this study focuses on analyzing the impact of kernel selection and oversampling techniques on model performance rather than comparing different algorithms. This approach enables a more in-depth evaluation of SVM optimization in stroke classification.

Several previous studies have evaluated SVM for stroke classification, but the results vary, indicating that kernel selection and oversampling techniques significantly influence model performance. Priyanka Bathla and Rajneesh Kumar [2] reported that SVM with SMOTE achieved only 74.72% accuracy, whereas Hanqing Zhang [27] found that SVM without oversampling reached 79.20% accuracy, albeit with

lower precision than recall, suggesting a bias toward the majority class. Conversely, other studies have demonstrated high accuracy with SVM, such as Aakanshi Gupta *et al*. [29], who achieved 95.04% accuracy, and Windy Junita Sari *et al*. [30], who reached 94.11%. However, these studies did not explicitly examine the impact of oversampling techniques on model performance, leaving the effect of oversampling methods on different SVM kernels for stroke classification largely unexplored.

The research process comprises several key stages, including data preprocessing, the application of oversampling techniques, feature selection, SVM model development, and comprehensive model evaluation. During the data preprocessing stage, missing values are addressed, categorical variables are encoded, and feature scaling is performed to ensure consistency in model training. Oversampling techniques are employed to mitigate class imbalance, thereby enhancing the model's capability to accurately identify stroke cases. Feature selection is conducted to identify the most relevant attributes contributing to stroke prediction, reducing computational complexity, and improving model interpretability. The SVM model is then developed utilizing different kernel functions to examine their impact on classification performance. Finally, the model's effectiveness is assessed using comprehensive performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. The findings from this study are expected to provide insights into the optimal combination of oversampling techniques and SVM kernel functions for stroke classification. The research flow is presented in FIGURE 1.

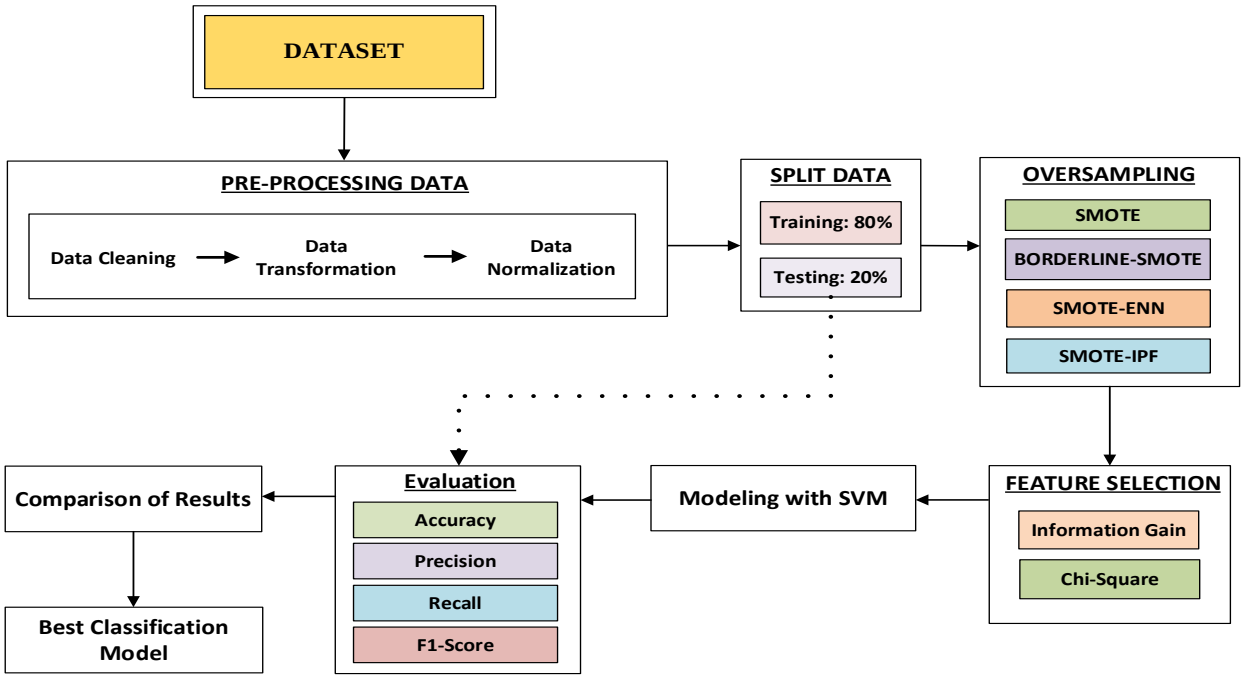

FIGURE 1. Methodology for Stroke Classification using SVM, Oversampling, and Feature Selection

A. DATASET

This study utilizes a public dataset from Kaggle named healthcare-dataset-stroke-data, comprising 5,110 records and 12 features, where one feature serves as the label or target. A detailed description of the dataset features is provided in Table 1.

TABLE 1
Attributes of dataset

No	Fitur	Data Type
1	Id	Numerical
2	Gender	Categorical
3	Age	Numerical
4	Hypertension	Categorical
5	Heart_disease	Categorical
6	Ever_married	Categorical
7	Work_type	Categorical
8	Residence_type	Categorical
9	Avg_glucose_level	Numerical
10	Bmi	Numerical
11	Smoking_status	Categorical
12	Stroke	Categorical

B. PRE-PROCESSING DATA

The data pre-processing steps include data cleaning, normalization, and transformation. The dataset cleaning process is carried out to remove inconsistent data, normalization is applied to standardize the scale of data values, and data transformation is performed to enhance the quality of features used as input for the model. Additionally, categorical features are converted into numerical formats to ensure their compatibility with subsequent processes.

1) MISSING VALUE

Incomplete data can significantly affect the performance of machine learning models. To address missing values, SimpleImputer was employed with the mean strategy to fill in the missing values in the BMI column. This technique was selected to maintain the distribution of the numerical data without introducing significant bias.

3) DATA TRANSFORMATION

In the data transformation process, numerical features such as age, hypertension, heart_disease, avg_glucose_level, and bmi are handled using imputation to address missing values. Categorical features like gender, ever_married, work_type, Residence_type, and smoking_status are transformed into binary columns using One-Hot Encoding. For example, the gender column is split into two columns, representing 'Male' and 'Female'. This transformation prepares the dataset for model training with numerical representations suitable for machine learning algorithms.

2) DATA NORMALIZATION

Normalization transforms numerical features into a uniform scale, ensuring that no single feature dominates the model training process. In this study, the MinMaxScaler method is applied to rescale the data within the range of 0 to 1. Normalization is performed on the age, avg_glucose_level, and bmi features. This process is particularly important for machine learning algorithms such as SVM, which are sensitive to feature scaling.

C. OVERSAMPLING DATA

Oversampling is a technique aimed at addressing class imbalance by enhancing the representation of the minority class to improve model performance. This approach works by generating synthetic samples or duplicating existing ones from the minority class, thereby increasing its presence in the dataset. The distribution of data before applying SMOTE can be observed in FIGURE 2, while the balanced data after the SMOTE process is presented in FIGURE 3.

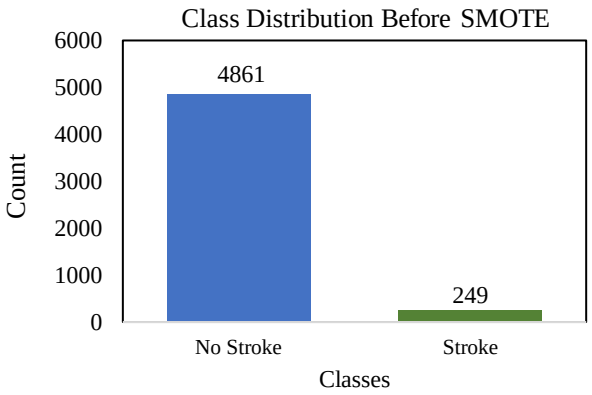

FIGURE 2. Imbalanced Data Before SMOTE

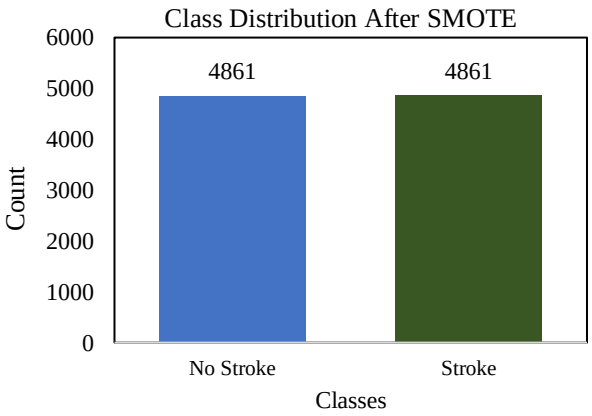

FIGURE 3. Balance Data After SMOTE

The dataset utilized in this study comprises two classes: "stroke" and "no-stroke," with a highly imbalanced distribution where the "no-stroke" class significantly outnumbers the "stroke" class. Before applying SMOTE, the dataset contained only 249 instances in the "stroke" class compared to 4,861 instances in the "no-stroke" class. To address this imbalance, the study employs various oversampling techniques, including the Synthetic Minority Over-Sampling Technique (SMOTE), Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF.

1) SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

SMOTE (Synthetic Minority Over-Sampling Technique) is a technique used to address the problem of data imbalance in datasets. SMOTE generates synthetic samples for the minority class by interpolating existing data, thereby increasing the number of samples in the minority class and enabling the model to better recognize patterns associated with that class [19]. The formula for this can be represented as shown in Eq. (1).

$$x_{new} = x_p + rand\,(0,1).\,(x_q - x_p) \quad (1)$$

Where x_p represents an instance from the minority class, x_q is one of the kk-nearest neighbors of x_p (also from the minority class), $rand\ (0,1)$ is a random number between 0 and 1 that determines the magnitude of the difference between x_q and x_p and x_{new} is the synthetic sample generated.

2) BORDERLINE-SMOTE

Borderline-SMOTE is a variant of SMOTE that focuses on samples located near the decision boundary between the majority and minority classes. This technique aims to enhance the model's ability to differentiate between classes that are more challenging to separate [20]. Borderline-SMOTE performs oversampling exclusively on samples situated near the boundary between the majority and minority classes, referred to as borderline samples. A sample is considered borderline if the number of majority class samples among its kk-nearest neighbors exceeds half of the total neighbors. The formula for this can be represented as shown in Eq. (2).

$$n \geq x_p\ \frac{k}{2} \quad (2)$$

Where n represents the number of majority class (negative) samples included in the k-nearest neighbors (kNN) of a minority sample, and k denotes the total number of nearest neighbors used to calculate the distance, typically set to k=5 [21]. After identifying the borderline samples, the SMOTE process is applied to generate synthetic data for the minority class.

3) SMOTE-ENN

SMOTE-ENN combines the SMOTE technique with the Edited Nearest Neighbor (ENN) algorithm to eliminate overlapping or misclassified samples after oversampling, resulting in a cleaner dataset and enabling the model to learn more effectively [5]. After oversampling with SMOTE, the dataset is refined using ENN, which removes the majority

class samples deemed misclassified or irrelevant based on distance and nearest neighbor voting. The formula is presented in Eq. (3).

$$\text{ENN(D)} = \{x_i \in D | \, y_i \neq majority(kNN\,(x_i))\} \quad (3)$$

D represents the original dataset, where x_i is a sample within the dataset and y_i denotes the class label of sample x_i. kNN is the k-Nearest Neighbors function used to determine the majority class of the nearest neighbors of x_i. Misclassified samples from the majority class are removed, resulting in a cleaner dataset [22].

3) SMOTE-IPF

SMOTE-IPF combines SMOTE with Iterative Proportional Fitting (IPF) to iteratively address the class imbalance by removing majority class samples that influence the class distribution and generating synthetic samples for the minority class. The steps involved in SMOTE-IPF [23] are as follows:

Determining KNN (K-Nearest Neighbors): Similar to SMOTE, SMOTE-IPF utilizes KNN to identify the nearest neighbors for each sample in the minority class.

a) Minority Sample Selection: Specific samples from the minority class are selected to generate synthetic samples.
b) Iterative Process: SMOTE-IPF performs iterations for each selected minority sample. During each iteration, the sample is used to create synthetic data, and the synthetic class distribution is updated based on the actual distributions of the minority and majority classes.
c) Distribution Update: The synthetic class distribution is updated to reflect the original distribution of the minority and majority classes. This step ensures that the synthetic samples are not only concentrated in dense areas of the feature space but also mirror the actual distribution of the minority class.

D. FEATURE SELECTION

In this study, the evaluation was conducted on models with and without feature selection. Feature selection was performed using two distinct methods, each with different functions and objectives.

1) Information Gain, this method ranks each feature to identify the most relevant ones that exhibit a strong relationship with other features [24]. The formula is presented in Eq. (4).

$$Entropy\,(L) = \sum_{i}^{c} -P_i \, log_2 \, P_i \quad (4)$$

Where c represents the total number of classification classes, P_i is the proportion of samples belonging to the class i. Once the Entropy value is obtained, the Information Gain can be calculated using the formula shown in Eq. (5).

$$Gain\,(L, f) = Entropy\,(L) - \sum_{v=1}^{v} \frac{|L_v|}{|L|} \, Entropy\,(L^v) \quad (5)$$

where $Gain\,(L, f)$ represents the Gain value of the feature f, v denotes a possible value of the feature f, values (f) is the set of all possible values of f, $|L_v|$ is the number of examples in the subset L^v, $|L|$ is the total number of data samples, and Entropy (L^v) is the entropy of the examples with value L^v.

2) Chi-Square aims to eliminate less significant features without compromising the overall accuracy of the model. The purpose of this feature selection method is to measure the relevance and influence of each feature on the classification outcome, thereby enhancing the efficiency and performance of the developed model. The formula is shown in Eq. (6).

$$x_c^2 = \frac{\sum (O_i - E_i)^2}{E_i} \quad (6)$$

where c represents the degrees of freedom for each variable, O_i represents the observed value in cell i, and E_i represents the expected value in cell i [25].

E. MODELING SVM

Support Vector Machine (SVM) is a robust machine learning algorithm widely applied for both classification and regression tasks. The primary goal of SVM is to find the optimal hyperplane that effectively separates two classes in the feature space by maximizing the margin between them. This hyperplane acts as a decision boundary, enabling the prediction of classes for new data points. By transforming data, SVM constructs an ideal hyperplane that minimizes classification risk while maintaining high generalization capabilities [25], [18]. The mathematical formulation of the hyperplane is presented in Eq. (7).

$$g(x) = w^T x + b \quad (7)$$

Where w^T represents an n-dimensional vector, and b is the bias term.

SVM employs several kernels to map data into higher-dimensional feature spaces, making class separation more straightforward. Kernel functions enable SVM to operate efficiently in complex or even infinite-dimensional feature spaces without explicitly calculating transformations into higher-dimensional spaces. The linear kernel is used when the analyzed data is linearly separable. The polynomial kernel transforms input data into a higher-dimensional space. The RBF (Radial Basis Function) kernel is utilized to classify data that is not linearly separable [18]. The formula is presented in Eq. (8) – (10).

1) Linear Kernel $K(x_i, x_j) = x_i^T x_j$ (8)

2) RBF Kernel $K(x_i, x_j) = (x_i^T x_j + 1)^{\,d}$ (9)

3) Polynomial $K(x_i, x_j) = \exp\,(-\sigma ||x_i - x_j||^{\,2})$ (10)

Where x_i, x_j are the feature vectors of two data points being compared, T denotes the transpose to simplify the dot product operation, d represents the polynomial order in the Polynomial Kernel, and σ is the kernel width or scale in the Gaussian Kernel.

In this study, 39 models were evaluated, derived from the combination of four main scenarios, three types of SVM kernels, and various oversampling techniques. The four main

scenarios in this research consist of: (1) models without feature selection and without oversampling (baseline model), (2) models without feature selection with oversampling, (3) models with Information Gain feature selection and oversampling, and (4) models with Chi-Square feature selection and oversampling. Each scenario was assessed using three types of SVM kernels: Linear, RBF, and Polynomial to determine the impact of each kernel on classification performance. For scenarios that applied oversampling (scenarios 2, 3, and 4), four oversampling techniques were used: SMOTE, Borderline SMOTE, SMOTE ENN, and SMOTE IPF. In the scenario without oversampling (scenario 1), only three models were evaluated, as models were tested solely with the three SVM kernels. Meanwhile, in scenarios that applied oversampling (scenarios 2, 3, and 4), each combination of an oversampling technique with an SVM kernel resulted in 12 models (3 kernels × 4 oversampling techniques). Thus, the combination of the four scenarios, three types of kernels, and four oversampling techniques resulted in a total of 39 models evaluated in this study. This evaluation aims to identify the best combination of feature selection methods, oversampling techniques, and SVM kernel types to enhance stroke disease classification performance.

F. CONFUSION MATRIX

This study compares the effectiveness of oversampling techniques, including SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF, in addressing data imbalance issues using evaluation metrics such as accuracy, precision, recall, and F1-score, while also analyzing their impact on the performance of Support Vector Machine (SVM) models with various kernel types, such as linear, polynomial, and RBF, to identify the most effective combination for stroke disease classification.

To ensure a reliable evaluation of model performance, this study employs 5-fold cross-validation. The dataset is randomly partitioned into five equal subsets (folds), where each fold is used once as the test set, while the remaining four folds are used for training. This process is repeated five times, ensuring that every data point is used for both training and testing.

The formulas for each evaluation metric are presented in Eq. (11) – (14) [26].

1) $$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)} \quad (11)$$

2) $$Precision = \frac{TP}{(TP + FP)} \quad (12)$$

3) $$Recall = \frac{TP}{(TP + FN)} \quad (13)$$

4) $$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (14)$$

True Positive (TP) refers to the number of cases correctly predicted as positive, while True Negative (TN) represents the number of cases correctly predicted as negative. False Positive (FP) indicates the number of cases predicted as positive but are actually negative, whereas False Negative (FN) refers to the number of cases predicted as negative but are actually positive.

III. RESULT

This study utilizes a stroke disease dataset characterized by data imbalance. The dataset is split into 80% training data and 20% testing data using the stratified splitting method to ensure balanced class distribution. The preprocessing steps include data cleaning, data transformation, and normalization of numerical features. To address data imbalance, four oversampling techniques are evaluated: SMOTE, Borderline-SMOTE, SMOTEENN, and SMOTE-IPF.

Additionally, the study assesses the impact of feature selection techniques, namely Information Gain and Chi-Square, on the dataset. Experiments are conducted using SVM with three different kernels: Linear, RBF, and Polynomial. Four main testing scenarios are considered: (1) models without oversampling and feature selection, (2) models with oversampling but without feature selection, (3) models with oversampling and feature selection using Information Gain, and (4) models with oversampling and feature selection using Chi-Square. For each of the four main models, experiments are carried out with the three SVM kernels, resulting in 39 testing scenarios. This comprehensive setup enables an in-depth analysis of the effects of feature selection and oversampling techniques on model performance. The evaluation metrics employed include Accuracy, Precision, Recall, and F1-score, which are compared to identify the model with the best performance.

A. MODEL WITHOUT FEATURE SELECTION AND WITHOUT OVERSAMPLING

This model serves as a baseline for comparison. In this configuration, the original dataset is utilized without applying any feature selection or oversampling techniques. The primary objective is to evaluate the model's performance using the raw dataset, identifying potential limitations caused by data imbalance or irrelevant features.

By analyzing the outcomes without oversampling, this study examines the extent to which class imbalance affects key performance metrics, including accuracy, precision, recall, and F1-score in the Support Vector Machine (SVM) model. If the baseline model exhibits poor performance in recognizing the minority class, oversampling techniques are expected to enhance predictive balance.

The results provide a reference point for evaluating the impact of advanced techniques in subsequent models, ensuring that any observed performance improvements stem from the applied preprocessing methods. The findings for the model without feature selection and oversampling are presented in TABLE 2.

TABLE 2
Model without feature selection and without oversampling

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
No Oversampling + Linear	81.80	17.62	74.00	28.46
No Oversampling + RBF	95.11	100	00.00	00.00
No Oversampling + Polynomial	95.11	100	00.00	00.00

TABLE 2 presents the evaluation results for data without the use of oversampling or feature selection techniques in stroke disease classification. The linear kernel achieved an accuracy of 81.80%, precision of 17.62%, recall of 74.00%, and F1-score of 28.46%. In contrast, both the RBF and polynomial kernels attained the same high accuracy of 95.11%, with 100% precision. However, these two kernels exhibited recall and F1-score values of 0.00%, indicating their inability to detect the minority class (stroke).

The recall of 0% without oversampling occurs because the model fails to identify any instances of the stroke class. This issue arises due to data imbalance, where the number of majority class samples (non-stroke) significantly outweighs the minority class (stroke). Consequently, the model primarily learns patterns from the majority class, leading it to classify nearly all instances as non-stroke. While this strategy may yield high accuracy, it compromises the model's ability to recognize stroke cases, which is the primary objective of the classification task.

These findings emphasize the critical importance of employing oversampling techniques to address data imbalance. By increasing the number of minority class samples, oversampling enables the model to learn patterns from both classes more effectively, thereby enhancing recall and F1-score. Without oversampling, the model tends to exhibit bias toward the majority class, resulting in suboptimal classification performance in detecting stroke cases.

B. MODEL USING OVERSAMPLING AND FEATURE SELECTION ON SVM KERNEL

This section examines the impact of combining oversampling techniques SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF with feature selection methods, namely Information Gain and Chi-Square, in SVM classification for stroke prediction. The evaluation results presented in TABLE 3 demonstrate that incorporating oversampling methods and feature selection substantially improves the performance of SVM models across all kernel types: Linear, RBF, and Polynomial.

Among all tested models, the Borderline-SMOTE with the RBF kernel yielded the highest classification performance, achieving an accuracy of 96.86%, precision of 99.14%, recall of 94.55%, and F1-score of 96.79%. These results indicate that Borderline-SMOTE is particularly effective in addressing the class imbalance issue, as it significantly enhances recall and F1-score. Notably, recall is a critical metric in medical diagnosis, as it measures the model's ability to correctly identify stroke patients, reducing the risk of false negatives. The increase in recall suggests that the model successfully learns from minority class samples without overfitting to the majority class.

For the Linear kernel, applying SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF resulted in moderate improvements, with accuracy values consistently around 96%. However, precision was generally lower compared to the RBF kernel, indicating a potential trade-off between correctly identifying positive cases and minimizing false positives. Meanwhile, the Polynomial kernel exhibited higher sensitivity to minority class samples when combined with SMOTE and Borderline-SMOTE, but its performance remained inferior to the RBF kernel. This suggests that the RBF kernel is more effective in capturing complex patterns in the dataset, particularly after the application of oversampling techniques.

Feature selection using Information Gain (IG) and Chi-Square (Chi2) produced varying effects across different kernels. While some models benefited from feature selection, particularly in terms of recall and F1-score, the overall trend suggests that applying oversampling without feature selection tends to yield more stable and optimal performance, especially with the RBF kernel. This may be due to the fact that oversampling techniques already mitigate the class imbalance issue, reducing the need for additional feature filtering. Moreover, in some cases, feature selection may inadvertently remove features that contribute to the discrimination of stroke cases, leading to slight performance degradation.

The Borderline-SMOTE technique, particularly when paired with the RBF kernel, consistently outperforms other models in terms of overall classification metrics. This combination effectively enhances the model's sensitivity to the minority class while maintaining high precision and accuracy. Evaluation based on accuracy, precision, recall, and F1-score demonstrates that the Borderline-SMOTE with the RBF kernel is the most robust and reliable approach for stroke classification. These findings reinforce the importance of selecting the appropriate oversampling strategy and kernel function in SVM classification, particularly in medical datasets where class imbalance is a prevalent issue. Furthermore, this approach improves the model's stability across different data distributions, ensuring consistent performance. It also effectively balances the trade-off between recall and precision, leading to more optimal classification results. As a result, Borderline-SMOTE with the RBF kernel stands out as a highly recommended strategy for developing accurate and reliable stroke prediction models.

TABLE 3
Model using oversampling and feature selection on SVM kernel

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BORDERLINE SMOTE + RBF	**96.86**	**99.14**	**94.55**	**96.79**
SMOTE + RBF	96.50	99.13	93.83	96.41
SMOTE IPF + RBF	96.50	99.13	93.83	96.41
BORDERLINE SMOTE + Linear	96.30	97.87	94.65	96.23
SMOTEENN + RBF	96.24	99.18	93.01	96.00
SMOTE + Linear	96.14	98.59	93.62	96.04
SMOTE IPF + Linear	96.14	98.59	93.62	96.04
BORDERLINE SMOTE + Polynomial	95.73	96.54	94.86	95.69
SMOTEENN + Linear	95.55	97.41	94.54	95.59
SMOTE-ENN +Polynomial	95.15	95.83	95.41	95.62
RBF Kernel	95.11	100	0.00	0.00
Polynomial Kernel	95.11	100	0.00	0.00
SMOTE IPF +Polynomial	94.96	94.50	95.47	94.98
SMOTE +Polynomial	94.69	94.50	95.47	94.98
SMOTE + IG + RBF	94.96	95.14	94.75	94.95
BORDERLINE SMOTE + IG + RBF	94.76	93.67	95.99	94.82
SMOTEENN + IG + RBF	94.88	95.47	95.15	95.31
SMOTEENN + IG + Polynomial	94.48	96.61	93.12	94.83
SMOTE IPF + IG + RBF	94.40	94.90	93.83	94.36
BORDERLINE SMOTE + Chi2 + RBF	94.24	96.24	92.08	94.11
BORDERLINE SMOTE + IG + Polynomial	93.32	94.98	91.46	93.19
BORDERLINE SMOTE + Chi2 + Polynomial	93.32	94.98	91.46	93.19
SMOTEENN + Chi2 + RBF	92.91	95.60	91.16	93.33
SMOTE IPF + IG + Polynomial	92.24	95.36	88.79	91.96
SMOTE IPF + Chi2 + RBF	92.08	92.69	91.36	92.02
SMOTE + IG + Polynomial	91.98	94.74	88.89	91.72
SMOTEENN + Chi2 + Polynomial	91.93	95.96	88.89	92.29
SMOTE + Chi2 + RBF	91.57	96.87	85.91	91.06
SMOTE + Chi2 + Polynomial	90.18	95.67	84.16	89.55
SMOTE IPF + Chi2 + Polynomial	90.18	95.67	84.16	89.55
SMOTEENN + IG + Linear	85.95	85.24	89.92	87.52
BORDERLINE SMOTE + Chi2 + Linear	83.91	81.11	88.37	88.37
BORDERLINE SMOTE + IG + Linear	83.86	80.81	88.79	84.61
SMOTEENN SMOTE + Chi2 + Linear	83.43	83.79	86.17	84.96
Linear Kernel	81.80	17.62	74.00	28.46
SMOTE + IG + Linear	78.87	76.14	84.05	79.90
SMOTE IPF + IG + Linear	77.99	75.52	82.82	79.00
SMOTE + Chi2 + Linear	77.48	75.77	80.76	78.19
SMOTE IPF + Chi2 + Linear	77.48	75.77	80.76	78.19

C. COMPARISON OF THE BEST MODEL WITH THE MODEL WITHOUT OVERSAMPLING

The findings of this study demonstrate a significant improvement in performance with the application of oversampling techniques to an imbalanced dataset compared to models without oversampling. The analysis focuses on the best-performing model, Borderline-SMOTE with the RBF kernel, and the corresponding SVM models without oversampling for each kernel. The results of the best model with the model without oversampling are presented in TABLE 4.

TABLE 4
Comparison of the best model with the model without oversampling

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BORDERLINE SMOTE + RBF	**96.86**	**99.14**	**94.55**	**96.79**
No Oversampling + Linear	81.80	17.62	74.00	28.46
No Oversampling + RBF	95.11	100	00.00	00.00
No Oversampling + Polynomial	95.11	100	00.00	00.00

TABLE 4 presents a detailed comparison of models using oversampling techniques versus those without oversampling on the imbalanced stroke dataset. The Borderline-SMOTE with the RBF kernel demonstrated the highest performance among all models, achieving an accuracy of 96.86%, precision of 99.14%, recall of 94.55%, and F1-score of 96.79%. These results highlight the model's superior classification ability, effectively detecting true stroke cases (high recall) while maintaining a low false positive rate (high precision). The well-balanced F1-score further confirms that the model successfully balances sensitivity and specificity, making it highly suitable for real-world applications where both metrics are crucial in medical diagnostics.

The Linear kernel model, in contrast, achieved a lower accuracy of 81.80%, which, despite appearing reasonable, was accompanied by a very low precision of 17.62% and an F1-score of only 28.46%. While the model exhibited a moderately high recall of 74.00%, its inability to correctly classify positive cases (stroke patients) significantly weakened its overall predictive performance. This suggests that the Linear kernel struggles to handle imbalanced data effectively, resulting in many false positive predictions.

For the RBF and Polynomial kernels without oversampling, although their accuracy was seemingly high at 95.11%, their actual classification ability was severely compromised. Both models achieved a precision of 100% but had a recall and F1-score of 0.00%, indicating a complete failure in identifying any stroke cases. This stark contrast underscores the limitations of these models when dealing with class imbalance, as they tend to focus entirely on the majority class (non-stroke cases), completely neglecting the minority class. Such behavior renders these models unsuitable for medical applications, where accurately detecting stroke cases is critical.

The comparison clearly demonstrates the necessity of employing oversampling techniques such as Borderline-SMOTE in handling class imbalance. By generating synthetic samples for the minority class, Borderline-SMOTE allows the model to learn the characteristics of stroke cases more effectively, resulting in substantial improvements across all classification metrics. For instance, when comparing Borderline-SMOTE + RBF to the No Oversampling + Linear model, there was an 18.44% increase in accuracy (from 81.80% to 96.86%), a remarkable 81.52% improvement in precision (from 17.62% to 99.14%), a 20.55% boost in recall (from 74.00% to 94.55%), and an F1-score enhancement of 68.33% (from 28.46% to 96.79%).

Furthermore, even models with higher baseline accuracy, such as No Oversampling + RBF and No Oversampling + Polynomial, exhibited significant performance improvements after incorporating Borderline-SMOTE. The accuracy for these models increased by 1.75% (from 95.11% to 96.86%), and the recall and F1-score, initially at 0.00%, experienced dramatic enhancements. While the precision for the RBF kernel decreased slightly by 0.86%, this trade-off is insignificant compared to the overall improvement in model reliability and robustness. The increase in recall is particularly valuable in stroke prediction, as it ensures that fewer actual stroke cases are missed.

These findings highlight that the combination of oversampling and an appropriate kernel function, such as RBF, is essential for handling class imbalance effectively. The Borderline-SMOTE + RBF model consistently delivers superior results, making it the optimal approach for stroke classification. This study strongly supports the adoption of oversampling techniques in future research and real-world applications, particularly in medical datasets where class imbalance is prevalent and early detection is crucial for patient outcomes. Additionally, these results emphasize the need for careful selection of both the oversampling method and the SVM kernel, as their interplay significantly impacts overall classification performance.

D. MODEL EVALUATION

1) CONFUSION MATRIX RESULT

The evaluation in this study utilizes a confusion matrix to assess the model's performance in classifying stroke disease. The confusion matrix provides a comprehensive breakdown of correctly and incorrectly classified instances for each class, including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). By analyzing the confusion matrix, essential evaluation metrics such as accuracy, precision, recall, and F1-score can be derived to quantify the model's effectiveness, particularly in addressing class imbalance. This analysis is crucial to ensure that the model not only achieves high overall accuracy but also effectively identifies cases from the minority class. The results from the confusion matrix are presented in FIGURE 4.

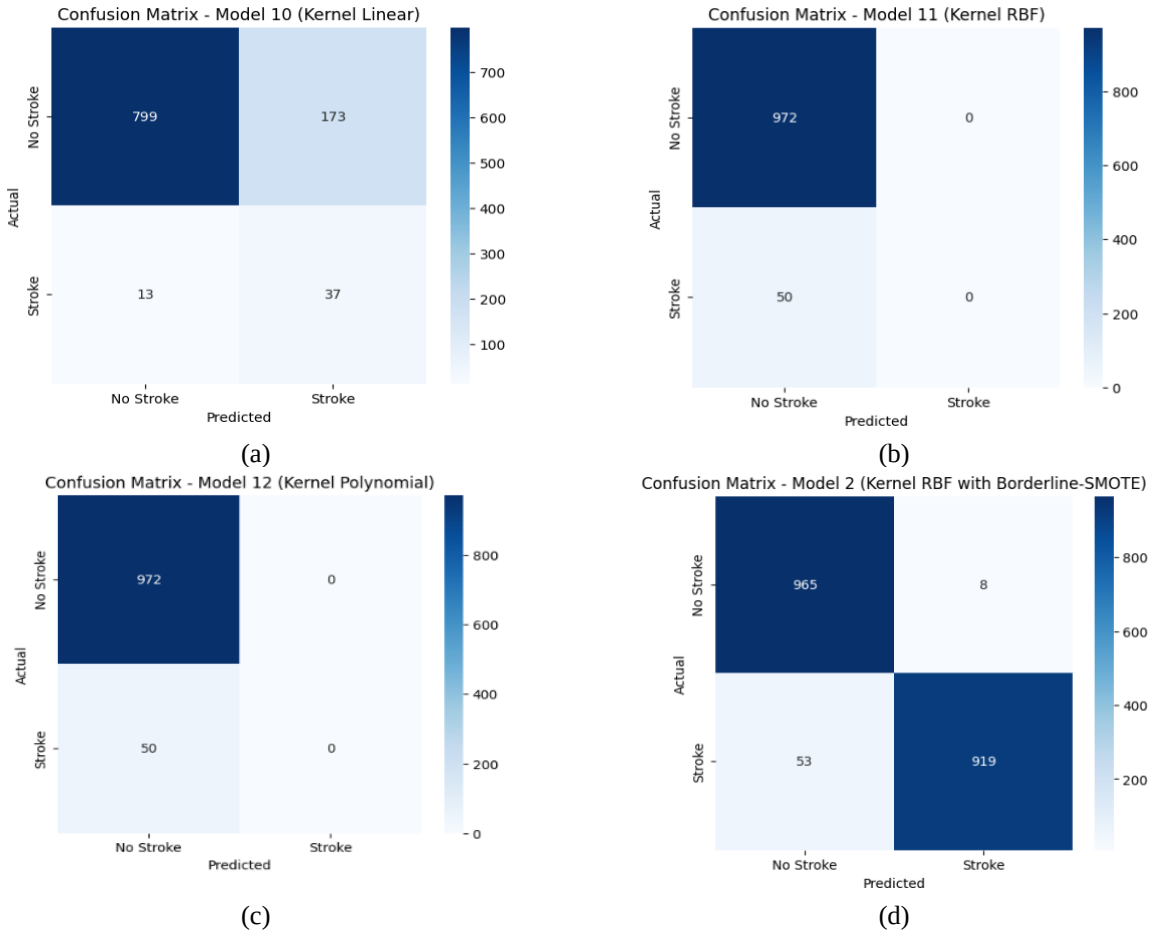

FIGURE 4. **Confusion Matrix Result: A detailed comparison between the baseline and best model. (a) Model no over sampling with linear kernel, (b) Model no over sampling with RBF kernel, (c) Model no over sampling with polynomial kernel, (d) Model Borderline-SMOTE with RBF kernel**

The results of the confusion matrix indicate that the Borderline-SMOTE method (Figure d) achieved the highest performance for Model 2 with the RBF kernel. This model successfully classified 919 stroke cases correctly (True Positive/TP) and 965 non-stroke cases correctly (True Negative/TN), yielding a high accuracy of 96.86% and the best F1-score of 96.79%. Moreover, the False Negative (FN) count was reduced to 53, compared to other methods, demonstrating an improved ability to detect stroke cases, while the False Positive (FP) count remained low at 8.

In contrast, the evaluation of Model 10, Model 11, and Model 12 revealed suboptimal performance in stroke classification when no feature selection or SMOTE was applied. Model 10 (Figure a) with a linear kernel achieved an accuracy of 81.80%, yet exhibited a low precision of 17.62% due to a high number of false positives, indicating a substantial proportion of incorrect positive predictions. Despite having a relatively high recall of 74.00%, the imbalance between precision and recall resulted in a low F1-score of 28.46%. The confusion matrix for this model indicates that the number of true positives is 37, true negatives is 799, false positives is 173, and false negatives is 13, underscoring the model's limited reliability in making accurate positive predictions.

Furthermore, Model 11 and Model 12, utilizing RBF and polynomial kernels, respectively, achieved a high accuracy of 95.11% and a perfect precision score of 100%. However, both models exhibited 0% recall and F1-score, indicating a complete failure to detect any stroke cases within the dataset. The confusion matrices for these models (Figures b and c) show that the number of true positives is 0, true negatives is 972, false positives is 0, and false negatives is 50, signifying an inability to identify the minority class (stroke cases) effectively. Although these models may appear to perform well based on accuracy and precision, their overall effectiveness is inadequate due to their inability to classify positive cases.

The application of oversampling techniques such as Borderline-SMOTE has been demonstrated to significantly enhance stroke detection performance. Conversely, models trained on imbalanced data without appropriate balancing techniques tend to fail in recognizing minority class instances, leading to misleadingly high accuracy but poor recall and F1-score.

2) RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE AND AREA UNDER THE CURVE (AUC) ANALYSIS

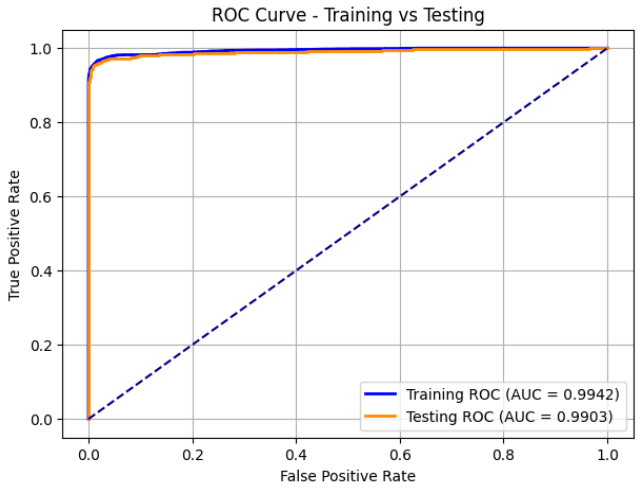

FIGURE 5. ROC Curve of the Optimal Model with Borderline SMOTE

FIGURE 5 illustrates the ROC curve of the Borderline SMOTE RBF Kernel model, which represents the optimal model. It demonstrates that the SVM model with an RBF kernel utilizing the Borderline SMOTE technique exhibits excellent performance in distinguishing between positive and negative classes. This is evidenced by the high AUC (Area Under the Curve) values, specifically 99.42% on the training data and 99.03% on the testing data. The minimal difference between the AUC scores for training and testing datasets indicates that the model possesses strong generalization capabilities and does not suffer from overfitting. If overfitting were present, a significant drop in the AUC score on the testing data would be expected compared to the training data. However, in this case, both values remain consistently high and stable.

Furthermore, the ROC curve, which closely approaches the top-left corner, suggests that the model achieves a high True Positive Rate (TPR) while maintaining a low False Positive Rate (FPR), signifying that it effectively identifies positive cases with minimal misclassification of negative cases.

Within the context of imbalanced data, the ROC curve also serves as a critical indicator that the model does not exhibit overfitting. Overfitted models trained on oversampled data often demonstrate excellent performance on training data but significantly poorer performance on testing data [27], which is typically reflected in a large discrepancy between the AUC scores. However, in this case, the consistently high AUC values above 99% for both datasets indicate that the application of Borderline SMOTE has not led to excessive adaptation to synthetic data. Instead, it has enhanced the model's ability to handle class imbalance while maintaining robust performance on unseen data.

Therefore, based on this ROC curve analysis, it can be concluded that the application of Borderline SMOTE in the SVM model with an RBF kernel successfully enhances classification performance without inducing overfitting, making it an optimal model for stroke prediction on an imbalanced dataset.

IV. DISCUSSION

Oversampling techniques play a crucial role in enhancing the model's sensitivity to the minority class, which contributes to improving accuracy in stroke disease prediction. The application of this technique helps the model recognize patterns in the minority class more effectively. Further analysis is provided in FIGURE 6.

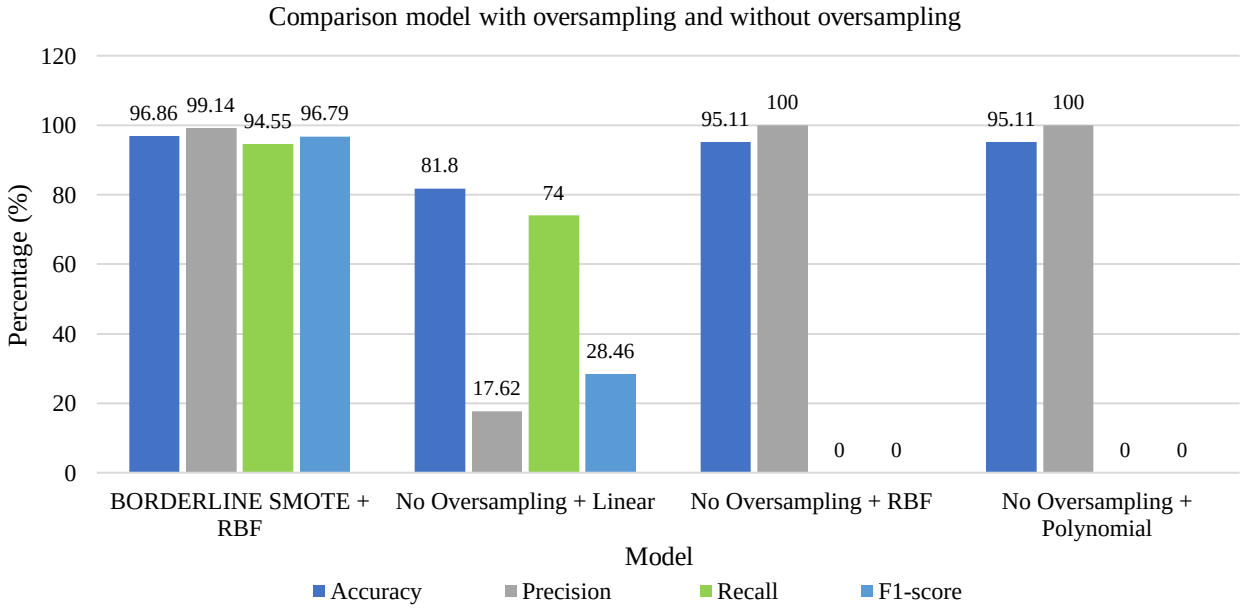

FIGURE 6. Comparison Model with Oversampling and Without Oversampling

The bar chart (FIGURE 6) presents a comparative analysis of models with and without oversampling techniques, specifically evaluating accuracy, precision, recall, and F1-score across different configurations. The Borderline-SMOTE with the RBF kernel demonstrates the highest overall performance, achieving an accuracy of 96.86%, a precision of 99.14%, a recall of 94.55%, and an F1-score of 96.79%. These results indicate that the combination of Borderline-SMOTE and the RBF kernel is highly effective in handling class imbalance while maintaining a balanced trade-off between precision and recall.

In contrast, models without oversampling techniques exhibit a substantial decline in performance, particularly in recall and F1-score. The Linear SVM without oversampling shows a significant drop, with a recall of 74% and an F1-score of only 28.46%, indicating a severe inability to correctly identify positive instances. More notably, both the RBF and Polynomial SVMs without oversampling fail to detect the minority class entirely, as reflected in their recall and F1-score values of zero. This highlights the critical impact of class imbalance, where the model becomes biased toward the majority class, rendering it ineffective for stroke classification.

The results strongly emphasize the necessity of oversampling techniques in addressing class imbalance. The Borderline-SMOTE method effectively enhances recall, ensuring that the model correctly identifies a greater number of stroke cases without significantly compromising precision. Additionally, the strong performance of the RBF kernel suggests that stroke classification exhibits non-linear characteristics, making it more suitable than linear or polynomial alternatives. Overall, the findings demonstrate that applying Borderline-SMOTE significantly improves classification performance, particularly with the RBF kernel, ensuring both robust generalization and reduced bias toward the majority class.

A. COMPARISON OF BORDERLINE-SMOTE AND COST-SENSITIVE LEARNING IN SVM MODELS

The Support Vector Machine (SVM) model with cost-sensitive learning was implemented to assess its effectiveness in handling class imbalance in stroke classification. This approach modifies the misclassification penalty by assigning higher costs to the minority class, aiming to enhance the model's ability to detect positive cases without relying on oversampling techniques. This study specifically investigates whether Borderline-SMOTE remains the superior technique compared to cost-sensitive learning in addressing class imbalance. By incorporating cost-sensitive learning into the SVM model, this experiment evaluates its impact on accuracy, precision, recall, and F1-score, particularly in detecting instances of the minority class.

The comparison between Borderline-SMOTE and cost-sensitive learning provides valuable insights into the most effective method for improving model performance on an imbalanced dataset. While oversampling techniques generate synthetic minority instances to balance class distribution, cost-sensitive learning directly influences the decision boundary by imposing penalties on misclassifications. The results of this evaluation will determine whether cost-sensitive learning alone is sufficient or if oversampling remains the more robust approach for stroke classification. The results of the SVM model with cost-sensitive learning are presented in Table 5, while the comparison of the performance between the cost-sensitive learning model and Borderline-SMOTE is illustrated in FIGURE 7.

TABLE 5
Evaluation results of the model using cost-sensitive learning

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Training	99.93	98.51	100	99.25
Testing	91.10	02.33	02.00	02.15

The evaluation results of the SVM model with cost-sensitive learning reveal a significant disparity between the training and testing phases. During training, the model achieved an accuracy of 99.93%, a precision of 98.51%, a recall of 100%, and an F1-score of 99.25%, indicating that it performed exceptionally well in classifying the data within a controlled environment. However, when tested on unseen data, there was a drastic decline in precision, recall, and F1-score, with values dropping to 2.33%, 2.00%, and 2.15%, respectively, despite maintaining a relatively high accuracy of 91.10%.

This sharp contrast suggests that the model suffers from overfitting, meaning it has learned patterns from the training data too well but fails to generalize effectively to new data. The low recall and F1-score indicate that the model struggles to detect minority class instances (stroke cases), increasing the risk of misclassification for positive cases.

Compared to Borderline-SMOTE, which achieved a precision of 99.14%, a recall of 94.55%, and an F1-score of 96.79%, the cost-sensitive learning approach showed significantly lower performance, particularly in identifying the minority class. While cost-sensitive learning can address class imbalance by adjusting misclassification penalties, it may not be sufficient for handling highly imbalanced data distributions. In contrast, Borderline-SMOTE enhances the representation of the minority class by generating synthetic samples, enabling the model to better capture the patterns of the underrepresented class. This technique has proven to be more effective in improving model performance on imbalanced datasets, as it allows for better generalization and pattern detection in the minority class. Furthermore, Borderline-SMOTE provides a more robust solution by focusing on the boundary areas where the minority and majority classes are most likely to overlap, which can lead to improved decision-making by the model. As a result, the integration of Borderline-SMOTE helps mitigate the risks of misclassification in the minority class.

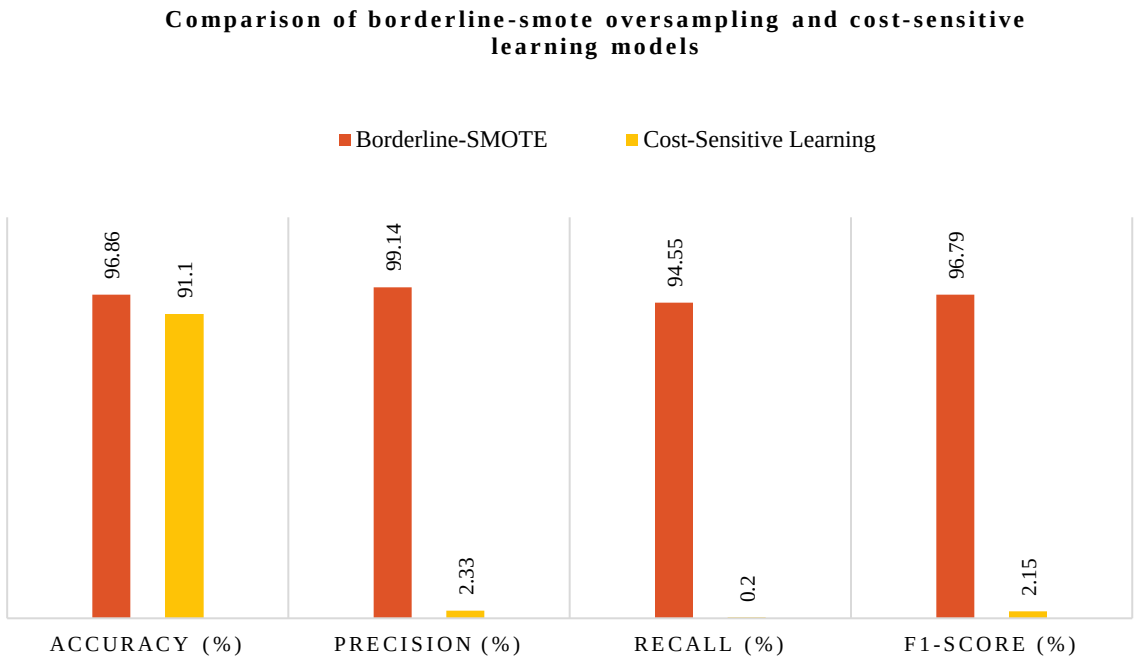

FIGURE 7. **Comparison of Borderline-SMOTE Oversampling and Cost-Sensitive Learning Models**

B. ANALYSIS OF THE BORDERLINE-SMOTE MODEL WITH RBF KERNEL

Overfitting occurs when a model performs exceptionally well on training data but fails to generalize effectively to unseen data, resulting in a significant decline in test performance. In this study, the Borderline-SMOTE RBF kernel model demonstrates consistent performance across both training and testing phases, indicating the absence of overfitting. The evaluation metrics, including accuracy, precision, recall, and F1-score for both phases, are presented in FIGURE 8.

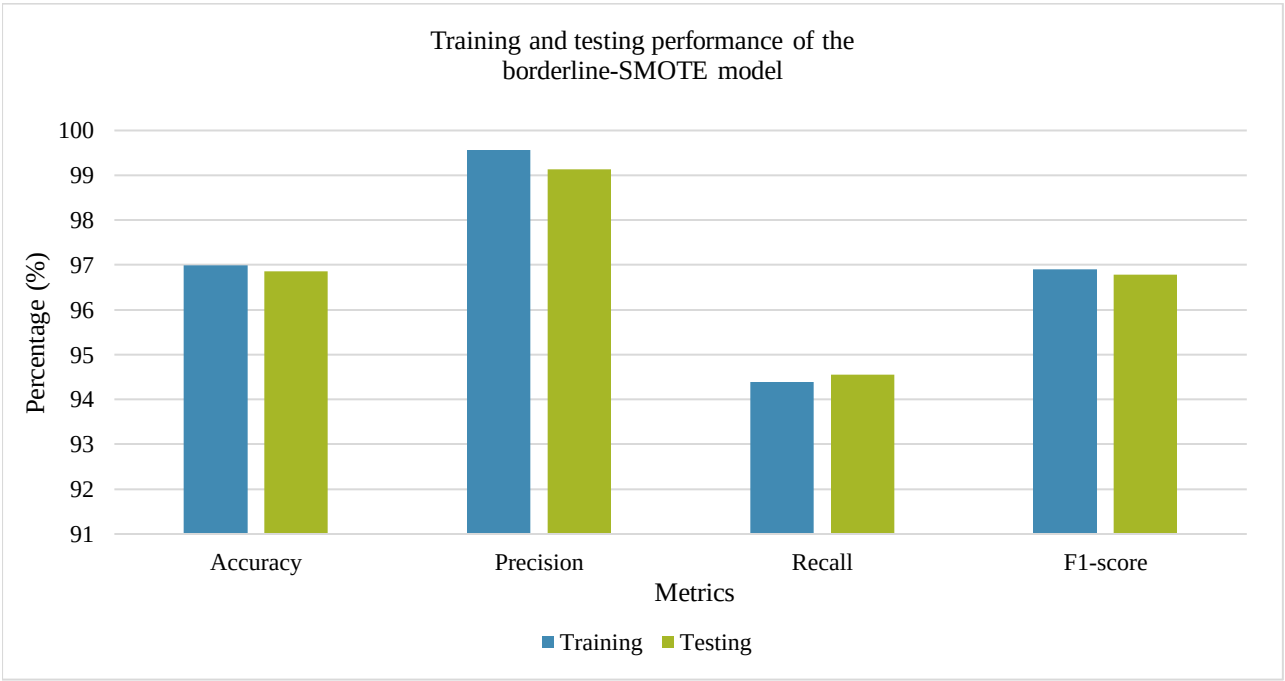

FIGURE 8. **Training and Testing Performance of the Borderline-SMOTE Model**

The evaluation results of the Borderline-SMOTE model with the RBF kernel indicate that the model exhibits excellent performance in stroke classification, with minimal differences between evaluation metrics on the training and testing datasets. On the training dataset, the model achieved an accuracy of 96.99%, precision of 99.57%, recall of 94.39%, and F1-score of 96.91%. These results remain consistent in the testing phase, with an accuracy of 96.86%, precision of 99.14%, recall of 94.55%, and F1-score of 96.79%.

The minimal discrepancy between training and testing performance suggests that the model possesses strong generalization capability and does not suffer from overfitting. In the case of overfitting, a significant gap would be expected between the training and testing metrics, where the model performs exceptionally well on training data but experiences a drastic performance drop on unseen data. However, the obtained results demonstrate that the model successfully identifies relevant patterns rather than memorizing training data, allowing it to maintain optimal performance on new data.

The visualization in FIGURE 8 further supports this claim, as the evaluation metrics exhibit a balanced distribution across both the training and testing phases. Consequently, the Borderline-SMOTE model with the RBF kernel can be considered an effective solution for addressing data imbalance while simultaneously mitigating the risk of overfitting and underfitting.

Overfitting typically occurs when a model is excessively complex, capturing noise in the training data, leading to high training performance but poor generalization on unseen data [27]. However, in this model, the stability of precision, recall, and F1-score across both phases indicates that it does not merely memorize the training data but instead effectively extracts meaningful patterns. Furthermore, the integration of Borderline-SMOTE mitigates data imbalance by generating synthetic samples near the decision boundary, allowing the model to learn in a more representative manner without introducing excessive bias toward the majority class. Thus, the combination of oversampling techniques and the RBF kernel, which is well-suited for capturing non-linear patterns, contributes to enhanced generalization, ensuring that the model maintains optimal performance on new data.

In the healthcare-dataset-stroke-data, which consists of 5110 samples, there is a significant class imbalance, where the number of minority class samples (stroke) is considerably lower than that of the majority class (non-stroke). This imbalance often causes machine learning models to be biased toward the majority class, leading to low recall and F1-score in stroke classification. Various oversampling techniques have been explored to address this issue, with Borderline-SMOTE proving to be the most effective method in this study.

In contrast to standard SMOTE, which generates synthetic data randomly across the feature space, Borderline-SMOTE specifically targets minority class samples located in the decision boundary or borderline region, where misclassification is most likely to occur. By generating synthetic data only around borderline samples, this technique ensures that the SVM model becomes more sensitive to minority class patterns without disrupting the overall data distribution.

The primary advantage of Borderline-SMOTE over other oversampling methods lies in its ability to identify minority class samples situated near the decision boundary between the two classes. Machine learning models often struggle to classify samples in this region, and Borderline-SMOTE helps establish a clearer decision boundary. Consequently, SVM models can significantly improve recall, F1-score, and overall accuracy, particularly in detecting stroke cases that were previously difficult to classify correctly.

In this study, Borderline-SMOTE utilizes the k-Nearest Neighbors (k-NN) algorithm with $k = 5$ to determine whether a minority class sample is within the borderline region. Suppose more than half (≥ 3) of its five nearest neighbors belong to the majority class. In that case, the sample is categorized as a borderline sample, indicating that it is in a decision region prone to misclassification. By generating synthetic data exclusively around these samples, Borderline-SMOTE maintains a balanced data distribution while avoiding the creation of irrelevant synthetic samples.

The combination of Borderline-SMOTE with the Radial Basis Function (RBF) kernel in SVM has proven to be the most effective approach in this study. The RBF kernel is capable of capturing complex relationships in data that are not linearly separable. With the presence of more representative synthetic data in the borderline region, SVM with the RBF kernel can establish a more flexible and accurate decision boundary compared to other kernels. This results in a significant improvement in recall and F1-score, making this combination the most effective method for handling class imbalance in stroke classification.

Based on the studies conducted by Han *et al.* [28], Anna Glazkova [29], Brandt and Lanzen [30], and Elreedy and Atiya [31], the application of oversampling techniques such as SMOTE, Borderline-SMOTE, and other resampling methods has been shown to significantly improve the performance of classification models in imbalanced datasets. These studies indicate that implementing oversampling techniques enhances the model's sensitivity to the minority class, resulting in notable improvements in accuracy, precision, recall, and F1-score. Furthermore, models utilizing oversampling consistently outperform those trained on imbalanced data without resampling, particularly in medical classification tasks such as stroke prediction, where accurate detection of the minority class is critical.

C. COMPARISON OF RESEARCH USING THE SVM METHOD WITH PREVIOUS STUDIES

This section presents a comparative analysis between the proposed study and previous research that employed the Support Vector Machine (SVM) method for stroke classification on healthcare datasets. The objective of this comparison is to evaluate the effectiveness of the proposed approach in relation to prior studies that utilized the same classification technique. The comparison of previous research with the proposed method is presented in Table 6.

TABLE 6
Comparison with previous studies

Previous Studies	Metode	Oversampling	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Priyanka Bathla and Rajneesh Kumar [2]	SVM	SMOTE	74.72	78.15	73.13	75.56
Saad Sahriar *et al.* [11]	SVM	SMOTE	90.36	85.18	95.02	89.83
Hanqing Zhang [32]	SVM	-	79.20	71.20	91.2	80.0
Yifan Feng [33]	SVM	-	77.00	74.00	73.0	78.0
Aakanshi Gupta *et al.* [34]	SVM	-	95.04	90.33	95.04	92.63
Windy Junita Sari *et al.* [35]	SVM	-	94.11	88.57	99.41	-
Arya Syifa Hermiati *et al.* [36]	SVM	SMOTE	87.75	-	-	-
Proposed Method	**SVM**	**Borderline-SMOTE (RBF)**	**96.86**	**99.14**	**94.55**	**96.79**

TABLE 6 shows a comparison between the proposed method and previous studies that used healthcare stroke datasets for stroke disease classification using the Support Vector Machine (SVM) method. The studies listed utilized SVM for classification, with some applying oversampling techniques such as SMOTE, SMOTENC, and Borderline-SMOTE to address class imbalance in the datasets. The accuracies recorded in these studies vary, with studies using SMOTE showing results between 74.72% and 95.04%. The proposed method using Borderline-SMOTE with the RBF kernel achieved the highest accuracy at 96.86%, with a precision of 99.14%, recall of 94.55%, and F1-score of 96.79%. These results indicate that Borderline-SMOTE with the RBF kernel is a more effective approach for handling class imbalance, significantly contributing to the improvement of the SVM classification model's performance in stroke disease classification. The limitations of this study lie in the dataset used, which consists of 5,110 samples and 12 features that may not fully represent a broader population. Future research could utilize a larger dataset to improve the generalizability of the findings. Additionally, this study can be further developed by implementing deep learning models to enhance classification accuracy and performance, particularly in handling imbalanced datasets.

V. CONCLUSION

This study evaluated the impact of oversampling techniques on the performance of stroke classification models using Support Vector Machine (SVM) in addressing data imbalance issues. The dataset consisted of 5,110 samples with 12 features, and 39 models were tested under four primary scenarios: no feature selection with oversampling, feature selection using Information Gain with oversampling, feature selection using Chi-Square with oversampling, and no feature selection and no oversampling. The results demonstrated that oversampling techniques, including SMOTE, Borderline-SMOTE, SMOTE-ENN, and SMOTE-IPF, significantly enhanced the model's ability to recognize patterns in the minority class. Additionally, feature selection methods, such as Information Gain and Chi-Square, improved model performance by selecting relevant features and mitigating the risk of overfitting. Among the tested SVM kernels, the Radial Basis Function (RBF) kernel exhibited the best performance, with the combination of Borderline-SMOTE and the RBF kernel achieving the highest accuracy (96.86%), precision (99.14%), recall (94.55%), and F1-score (96.79%).

Class imbalance in the stroke dataset causes machine learning models to be biased toward the majority class, leading to low recall and F1-score in stroke classification. Therefore, the application of oversampling techniques is crucial in enhancing the model's sensitivity to the minority class. Borderline-SMOTE proved to be the most effective method in this study, as it not only generates synthetic data randomly like standard SMOTE but also specifically targets minority class samples near the decision boundary, where misclassification is most likely to occur. This approach enables the model to better recognize minority class patterns without disrupting the overall data distribution.

The evaluation results indicate that the Borderline-SMOTE model with the RBF kernel exhibits excellent classification performance with minimal differences between evaluation metrics on the training and testing datasets. During training, the model achieved an accuracy of 96.99%, precision of 99.57%, recall of 94.39%, and F1-score of 96.91%. Similarly, in the testing phase, the model maintained a high accuracy of 96.86%, with a precision of 99.14%, recall of 94.55%, and F1-score of 96.79%. The minimal discrepancy between these metrics suggests strong generalization capability, indicating that the model does not suffer from overfitting. If overfitting were present, a significant drop in performance on unseen data would be expected. However, the stable performance across both datasets confirms that the model effectively learns meaningful patterns rather than memorizing training data.

The ROC curve analysis further supports this finding, as the Borderline-SMOTE RBF model achieved high AUC

values of 99.42% for training and 99.03% for testing, with minimal deviation between the two. This demonstrates the model's ability to distinguish between positive and negative cases effectively without overfitting. The ROC curve, which closely approaches the top-left corner, also indicates a high True Positive Rate (TPR) while maintaining a low False Positive Rate (FPR), confirming the model's reliability in stroke classification. In the context of imbalanced data, the consistently high AUC values suggest that the application of Borderline-SMOTE has enhanced model robustness without excessive adaptation to synthetic data.

The findings of this study suggest that the combination of Borderline-SMOTE and the RBF kernel is the most effective approach for improving stroke classification in imbalanced datasets. This study highlights the importance of oversampling techniques in addressing class imbalance while ensuring strong generalization capabilities. Future research could explore deep learning approaches, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to further enhance classification accuracy.

REFERENCES

[1] 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.

[2] P. Bathla and R. Kumar, "A Hybrid System To Predict Brain Stroke Using A Combined Feature Selection And Classifier," *Intell. Med.*, vol. 4, no. April 2023, pp. 75–82, 2024, doi: 10.1016/j.imed.2023.06.002.

[3] M. M. Chowdhury, R. S. Ayon, and M. S. Hossain, "An Investigation Of Machine Learning Algorithms And Data Augmentation Techniques For Diabetes Diagnosis Using Class Imbalanced BRFSS Dataset," *Healthc. Anal.*, vol. 5, no. December 2023, p. 100297, 2024, doi: 10.1016/j.health.2023.100297.

[4] T. G.S., Y. Hariprasad, S. S. Iyengar, N. R. Sunitha, P. Badrinath, and S. Chennupati, "An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets," *Machine Learning with Applications*, vol. 8. p. 100267, 2022, doi: 10.1016/j.mlwa.2022.100267.

[5] K. Iscra *et al.*, "Optimizing machine learning models for classification of stroke patients with epileptiform EEG pattern : the impact of dataset balancing techniques," vol. 00, 2024, doi: 10.1016/j.procs.2024.09.324.

[6] S. Paliwal, S. Parveen, M. A. Alam, and J. Ahmed, "Improving Brain Stroke Prediction through Oversampling Techniques: A Comparative Evaluation of Machine Learning Algorithms," *Preprints*, vol. 44, no. 6, pp. 1484–1502, 2023, [Online]. Available: www.preprints.org.

[7] S. Saha and K. Nur, "Prediction of Stroke Disease Using Deep CNN Based Approach," no. January 2022, 2023, doi: 10.12720/jait.13.6.604-613.

[8] T. Swathi Priyadarshini and M. A. Hameed, "Collaboration Of Clustering And Classification Techniques For Better Prediction Of Severity Of Heart Stroke Using Deep Learning," Meas. Sensors, vol. 37, no. September 2024, p. 101405, 2025, doi: 10.1016/j.measen.2024.101405.

[9] Y. He *et al.*, "Construction of a machine learning-based prediction model for unfavorable discharge outcomes in patients with ischemic stroke," *Heliyon*, vol. 10, no. 17, p. e37179, 2024, doi: 10.1016/j.heliyon.2024.e37179.

[10] V. P. Prasetyo, M. F. A. Ulin Nuha, M. H. Hakiki, R. A. Vinarti, and A. Djunaidy, "Comparison of Data Mining Techniques on Stroke Clinical Dataset," *Procedia Comput. Sci.*, vol. 234, pp. 502–511, 2024, doi: 10.1016/j.procs.2024.03.033.

[11] S. Sahriar *et al.*, "Unlocking Stroke Prediction: Harnessing Projection-Based Statistical Feature Extraction With ML Algorithms," *Heliyon*, vol. 10, no. 5, p. e27411, 2024, doi: 10.1016/j.heliyon.2024.e27411.

[12] F. Fachruddin, E. Rasywir, and Y. Pratama, "Increasing the Accuracy of Brain Stroke Classification using Random Forest Algorithm with Mutual Information Feature Selection," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 4, pp. 555–562, 2024.

[13] Z. Rustam, Arfiani, and J. Pandelaki, "Cerebral Infarction Classification Using Multiple Support Vector Machine With Information Gain Feature Selection," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1578–1584, 2020, doi: 10.11591/eei.v9i4.1997.

[14] N. Nasution, F. Nasution, E. Erlin, and M. Hasan, "Evaluation Study of the Chi-Square Method for Feature Selection in Stroke Prediction with Random Forest Regression," 2024, doi: 10.4108/eai.30-10-2023.2343096.

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

[16] S. Ray, "Chi-Squared Based Feature Selection for Stroke Prediction using AzureML," 2020.

[17] J. Gao and G. Zhang, "Tennis action recognition and evaluation with inertial measurement unit and SVM," *Systems and Soft Computing*, vol. 6. 2024, doi: 10.1016/j.sasc.2024.200154.

[18] R. Gholami and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," 1st ed. Elsevier Inc., 2017. doi: 10.1016/B978-0-12-811318-9.00027-2.

[19] D. Fitria, T. H. Saragih, D. Kartini, and F. Indriani, "Classification of Appendicitis in Children Using SVM with KNN Imputation and SMOTE Approach to Improve Prediction Quality," vol. 6, no. 3, pp. 302–311, 2024.

[20] T. O. Omotehinwa, D. O. Oyewola, and E. G. Moung, "Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease," *Informatics Heal.*, vol. 1, no. 2, pp. 70–81, 2024, doi: 10.1016/j.infoh.2024.06.001.

[21] A. Tharwat, T. Gabel, and A. E. Hassanien, "Classification Of Toxicity Effects Of Biotransformed Hepatic Drugs Using Optimized Support Vector Machine," Adv. Intell. Syst. Comput., vol. 639, pp. 161–170, 2018, doi: 10.1007/978-3-319-64861-3_15.

[22] Y. Han and I. Joe, "Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging," *Appl. Sci.*, vol. 14, no. 21, 2024, doi: 10.3390/app14219772.

[23] G. A. Pradipta and Putu Desiana Wulaning Ayu, "Kombinasi Inisial Filtering Oversampling dengan Metode Ensemble Classifier pada Klasifikasi Data Imbalanced," *J. Sist. dan Inform.*, vol. 17, no. 2, pp. 137–145, 2023, doi: 10.30864/jsi.v17i2.591.

[24] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[25] M. Mahmud, I. Budiman, F. Indriani, D. Kartini, and M. R. Faisal, "Implementation of C5.0 Algorithm using Chi-Square Feature Selection for Early Detection of Hepatitis C Disease," vol. 6, no. 2, pp. 116–124, 2024.

[26] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing Two SVM Models Through Different Metrics Based On The Confusion Matrix," *Comput. Oper. Res.*, vol. 152, no. December 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.

[27] A. K. P. Anil and U. K. Singh, "An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models," *J. Syst. Eng. Inf. Technol.*, vol. 2, no. 2, pp. 77–84, 2023, doi: 10.29207/joseit.v2i2.5460.

[28] W. Sun, Z. Cai, and X. Chen, "Region-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Commun. Comput. Inf. Sci.*, vol. 1944 CCIS, pp. 151–160, 2024, doi: 10.1007/978-981-99-7743-7_9.

[29] A. Glazkova, "A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification," no. 18, pp. 1–12, 2020, [Online]. Available: http://arxiv.org/abs/2008.04636.

[30] J. Brandt and E. Lanzen, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," p. 42, 2020.

[31] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.

[32] H. Zhang, "Stroke Prediction Based on Support Vector Machine," *Highlights Sci. Eng. Technol.*, vol. 31, pp. 53–59, 2023, doi: 10.54097/hset.v31i.4812.

[33] Y. Feng, "Support Vector Machine for Stroke Risk Prediction," *Highlights Sci. Eng. Technol.*, vol. 38, pp. 917–923, 2023, doi: 10.54097/hset.v38i.5977.
[34] A. Gupta *et al.*, "Predicting stroke risk: An Effective Stroke Prediction Model Based On Neural Networks," *J. Neurorestoratology*, vol. 13, no. 1, p. 100156, 2024, doi: 10.1016/j.jnrt.2024.100156.
[35] W. J. Sari *et al.*, "Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 34–43, 2024, doi: 10.57152/predatecs.v2i1.1119.
[36] A. S. Hermiati, R. Herteno, F. Indriani, T. H. Saragih, Muliadi, and Triwiyanto, "A Comparative Study: Application of Principal Component Analysis and Recursive Feature Elimination in Machine Learning for Stroke Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 3, pp. 231–242, 2024, doi: 10.35882/jeeemi.v6i3.446.

AUTHOR BIOGRAPHY

LUH AYU MARTINI began her undergraduate studies in Computer Systems at the Institut Teknologi dan Bisnis STIKOM Bali in 2018 and graduated in the 2021/2022 odd semester. During this time, she developed expertise in system analysis and network engineering.
Currently, she is pursuing a master's degree (S2) in Computer Science at the same institution, starting in 2023. Her research focuses on analyzing the impact of oversampling techniques for imbalanced datasets, particularly in stroke disease classification. By leveraging machine learning algorithms, she aims to enhance the accuracy and reliability of predictive models in healthcare analytics. Her final project aims to explore innovative oversampling strategies to improve the performance of machine learning algorithms for handling imbalanced data, addressing key challenges in medical datasets. She can be contacted at email: 232011013@stikom-bali.ac.id.

GEDE ANGGA PRADIPTA holds a Doctor of Computer Science from Department of Computer Science and Electronics, Faculty of Natural Sciences, Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia., in 2021. He also received a bachelor's degree in computer informatics from Universitas Atma Jaya (UAJY), Yogyakarta, Indonesia, in 2012 and a master's degree in information technology from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2014. His research interests include Machine Learning, Pattern Recognition, And Image Processing. He is currently lecturing with Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. He can be contacted at email: angga_pradipta@stikom-bali.ac.id.

ROY RUDOLF HUIZEN graduated with Doctor of Computer Science (2018) from Universitas Gadjah Mada (UGM) Yogyakarta, Indonesia. Lecturer and researcher at the Department of Magister Information System at the Institut Teknologi dan Bisnis STIKOM Bali, with research interests in the fields of Object Identification, Signal Processing, Cyber Security Forensics and Artificial Intelligence. Email: roy@stikom-bali.ac.id. ORCID: 0000-0002-3671-6030.