Proceeding of The International Conference of Inovation, Science, Technology, Education, Children, and Health

e-ISSN : 2776-9062

Research Article

Stroke Disease Prediction Using Support Vector Machine Method

Gayatri Dwi Santika[1*], Valiant Shabri Rabbani[2]

[1] Universitas Jember, Indonesia. E-mail: gayatri@unej.ac.id
[2] Universitas Jember, Indonesia.

***** Corresponding Author : Valiant Shabri Rabbani

Abstract: Stroke is one of the leading causes of death globally and is particularly prevalent in Indonesia. Early prediction of stroke is critical to reducing the risk of long-term disability and mortality. This study aims to build a stroke prediction model using the Support Vector Machine (SVM) classification method. The dataset used is sourced from Kaggle, containing 5,110 records with class imbalance. To address the imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during preprocessing. The study evaluates model performance across multiple data splits (70:30, 80:20, 90:10) and k-fold cross-validation values (k=5, 7, 10). The SVM was tested with various kernel types—linear, polynomial, and radial basis function (RBF)—along with parameter tuning for C, gamma, and degree. The results show that the polynomial kernel yielded the highest prediction accuracy of 92%. The model performance was evaluated using accuracy, precision, recall, and F1-score metrics..

Keywords: Stroke, Prediction, Support Vector Machine, SMOTE.

1. Introduction

Stroke is an acute clinical condition resulting from neurological dysfunction caused by disturbances in the blood vessels of the brain, spinal cord, or retina. It can lead to partial or complete paralysis and even death. Stroke ranks as the third leading cause of death in Indonesia, following cancer and heart disease [1] Approximately 25% of stroke survivors experience a recurrent stroke within 1 to 5 years, with a mortality risk twice as high as those experiencing their first stroke [2]. The high mortality rate associated with stroke highlights the urgent need for early detection to prevent its occurrence and reduce fatality. Early detection and assessment of stroke can minimize the severity of the condition and significantly lower the risk of death [3]. In medical practice, stroke detection often relies on reviewing patients' medical records and analyzing health check data, which can be vast and complex. This makes it difficult for medical professionals to manually extract meaningful patterns. To overcome this challenge, data mining techniques can be applied to uncover hidden patterns and rules within large datasets, offering useful insights for stroke prediction [4]. Support Vector Machine (SVM) is one of the most effective and widely used classification algorithms in medical diagnostics. Several studies have shown that SVM can achieve high accuracy in clinical predictions, making it a reliable method for disease classification[5] Moreover, incorporating class-balancing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) further improves the robustness of predictions, especially when dealing with imbalanced medical datasets.

Previous research [6] proposed a hybrid method combining Random Forest Regression to estimate missing values and a Deep Neural Network (DNN) for classification. This approach, applied to an imbalanced medical dataset from HealthData.gov, achieved an accuracy of 71.6%. [7] Another research used data from the Korean National Health Insurance Service (KNHIS) and employed Logistic Regression to select 48 significant features associated with ischemic stroke. They further applied several machine learning models including DNN, Extreme Gradient Boosting (XGBoost), and Random Forest, with the DNN producing the highest AUROC value of 0.72. Another research has been done by

Received: 07 May, 2025
Revised: 31 May, 2025
Accepted: 05 July, 2025
Published: 08 July, 2025
Curr. Ver.: 08 July, 2025

utilizing the National Health Insurance Research Database (NHIRD) and comparing multiple classifiers such as Deep Neural Network, Gradient Boosting Decision Trees (GBDT), Logistic Regression, and Support Vector Machine[8]. The DNN achieved the best performance, with an accuracy of 86%. In contrast,[5] employed the Taiwan Stroke Registry (TSR) to evaluate several machine learning models—Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN). Their findings showed that SVM outperformed other methods with an accuracy of 94%. While these studies demonstrate the effectiveness of various algorithms, they also highlight the need for robust preprocessing, especially in handling class imbalance. However, none of the reviewed works focused specifically on optimizing SVM using a combination of kernel variations, hyperparameter tuning, and the SMOTE technique for stroke prediction. This research aims to fill that gap by developing an SVM-based model with improved generalization and accuracy for stroke classification. This study aims to develop a predictive model for stroke disease using the SVM algorithm, combined with the SMOTE technique and optimized parameters. The performance of the model is evaluated using accuracy, precision, recall, and F1-score metrics under different data-splitting and cross-validation scenarios. Numerous studies have explored the application of machine learning in stroke prediction. These studies vary in terms of datasets used, algorithms implemented, and performance metrics achieved.

2. Literature Review

Table 1, it can be observed that Support Vector Machine (SVM) is a suitable method for disease detection due to its ability to produce high accuracy scores. In medical diagnostics, accuracy is a critical metric, as errors in detecting medical conditions can lead to fatal consequences for patients. However, there is currently a lack of studies that specifically evaluate the performance of SVM as a standalone method by incorporating parameter tuning and the Synthetic Minority Over-sampling Technique (SMOTE) during the preprocessing stage for stroke prediction.

Table 1. Previous Studies on Stroke Prediction Using Machine Learning.

Reference	Dataset	Methodologies	Results
Hung et al., 2017	NHRID (900,000 records)	DNN, SVM, LR, GBDT	Accuracy = 86% AUC = 0.92
Lin et al., 2020	TSR (records)	SVM, RF, ANN, HANN	AUC = 0.97
Arslan et al., 2016	Turgut Ozal Medical Centre (192 records)	SVM, SGB, PLR	Accuracy = 97% AUC = 0.97
Akter et al., 2022	Kaggle (5,110 records)	RF, DT, SVM	Accuracy = 95.3%

Various machine learning algorithms have been applied for stroke prediction with differing levels of effectiveness. Deep Neural Networks (DNN) consistently showed strong performance across multiple studies, particularly in large-scale datasets, as demonstrated [8][7]. However, while DNN achieved high accuracy, it often required extensive computational resources and preprocessing techniques such as dimensionality reduction and feature engineering. Gradient Boosting and Random Forest also yielded competitive results but were slightly outperformed by DNN in most cases. Notably, Support Vector Machine (SVM) proved to be a strong contender, especially in the study [5] where it outperformed both Random Forest and Artificial Neural Network with an AUC of 0.97. Despite its proven accuracy, there is limited research evaluating SVM as a standalone classifier using kernel optimization and SMOTE for handling class imbalance. This indicates a research gap, which this study aims to address by analyzing the impact of different kernel functions and parameter tuning on SVM performance in stroke prediction

3. Proposed Method

This study proposes a stroke prediction model using the Support Vector Machine (SVM) algorithm with the integration of the Synthetic Minority Over-sampling Technique (SMOTE) during the preprocessing stage to address class imbalance in the dataset. The method is implemented through the following steps:

3.1. Data Collection

The dataset used in this study is sourced from Kaggle and consists of 5,110 records with 11 features including age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and stroke diagnosis (target variable). The dataset is imbalanced, with significantly fewer positive stroke cases compared to non-stroke cases.

3.2. Data Preprocessing

Preprocessing is a crucial step to prepare the data for classification. The following techniques are applied:

- **Handling Missing Values**: Missing values in the BMI column are imputed using the mean method.
- **Encoding Categorical Features**: Categorical variables are converted into numerical values using one-hot encoding or label encoding where appropriate.
- **Normalization**: Numerical features such as age, average glucose level, and BMI are normalized using Min-Max scaling.
- **Balancing the Dataset**: SMOTE is applied to synthesize new samples from the minority class to balance the dataset and improve model performance.

3.3. Model Construction Using SVM

Support Vector Machine is used as the main classification method. SVM works by finding the optimal hyperplane that separates data into different classes. The objective is to maximize the margin between support vectors of each class. The decision function is defined as:

$$\boldsymbol{f(x) = sign(\Sigma_{i=1}^{n} \alpha i \gamma i K(xi, x) + b)} \quad (1)$$

Where:

$\boldsymbol{\alpha i}$ are the Lagrange multipliers
$\boldsymbol{\gamma i}$ are the class labels
$\boldsymbol{K(xi, x)}$ is the kernel function
$\boldsymbol{b}$ is the bias term

4. Results and Discussion

4.1 Dataset Collection Results

This study utilizes a secondary dataset titled **"Stroke Prediction Dataset"**, obtained from the Kaggle platform. The dataset consists of **5,110 records** and includes **10 independent variables**, namely: **Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Average Glucose Level, BMI, and Smoking Status**, along with **one target variable** that indicates whether or not a patient has experienced a stroke. An overview of the dataset features and their statistical descriptions is presented in Table 2.

Table 2. Previous Studies on Stroke Prediction Using Machine Learning.

Variable	count	Data Type	Mean	Std Dev	Min	Max
Gender	5110	Object	-	-	-	-
Age	5110	Float64	43.2	22.6	0.08	82
Hypertension	5110	Int64	0.09	0.29	0	1
Heart Disease	5110	Int64	0.05	0.22	0	1
Ever Married	5110	Object	-	-	-	-
Work Type	5110	Object	-	-	-	-

Residence Type	5110	Object	-	-	-	-
Avg Glucose Level	5110	Float64	106.14	45.28	55.12	271.74
BMI	4909	Float64	28.89	7.85	10.3	97.6
Smoking Status	5110	Object	-	-	-	-
Stroke (Target)	5110	Int64	0.04	0.21	0	1

4.2 Model Development Results

To facilitate the data processing workflow, several libraries were utilized in this study such as NumPy that was used for efficient numerical operations and to work with arrays. Pandas that were employed for data analysis, data manipulation, and data cleaning tasks. Matplotlib was applied to support data visualization, including the creation of various types of graphs such as line plots, histograms, and bar charts. These libraries were instrumental in streamlining preprocessing, model training, and performance visualization throughout the development of the stroke prediction model. This section also presents the Python implementation used in developing the most optimal machine learning model on algorithm 1. During the entire model-building process, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied prior to separating the dataset into features and target labels. SMOTE is specifically designed to address **class imbalanced** classification problems, where the minority class is significantly underrepresented. Ignoring class imbalance in the dataset may lead to negative effects such as **model bias**, where the machine learning model tends to favor the majority class, achieving high accuracy while failing to correctly classify minority class instances. Such biased models may not be suitable for real-world applications, particularly in critical cases like stroke prediction, where correct identification of the minority class (stroke cases) is essential.

Algorithm 1.Stroke Prediction Model with SMOTE and SVM

```
X = df.drop(columns="stroke")
y = df.stroke
ct = ColumnTransformer(transformer= [('encoder ',
OneHotEncoder(), [0,4,5,6,9])], remainder='passthrough')
X = ct.fit_transform(X)
sm = SMOTE(random_state=42)
X_train_res, y_train_res, = sm.fit_transform(X, y)
X_train, X_test, y_train, y_test =
train_test_split(X_train_res, y_train_res,
stratify=y_train_res, test_size=0.1, randoms_state=42)
parameters = {
'SVM__kernel' : ['linear', ' poly','rbf'],
'SVM__C' : [0.1,1,10],
'SVM__gamma' : [0.1,1,10],
'SVM__degree' : [1,2,3]
}
pipeline = Pipeline ([
("scaler", StandardScaler()),
("SVM", SVC())
model = GridSearchCV(pipeline, parameters, cv=5,
n_jobs=-1, verbose=1)
model.fit(X_train, y_train)
```

4.3 Model Performance Testing Results

The performance evaluation of the Support Vector Machine (SVM) algorithm for stroke prediction was conducted using four key metrics: accuracy, precision, recall, and F1-score. These metrics were calculated by comparing the model's predicted labels with the actual true labels on the test dataset. The calculation of these metrics is based on the confusion matrix, which is automatically generated during model evaluation. This matrix illustrates the model's ability to correctly classify both positive and negative cases, serving as the basis for computing evaluation scores. To compute and visualize these metrics, several libraries such as scikit-learn and matplotlib were utilized. The Python code used to generate the confusion matrix and retrieve the performance scores is presented in Algorithm 2.

Algorithm 2. .Confusion Matrix and Evaluation Metrics in Python

```
1. print(confusion_matrix(y_test, y_pred))
2. print(classification_report(y_test, y_pred))
3. print(accuracy_score(y_test, y_pred))
4.  print(precision_score(y_test, y_pred))
5.  print(recall_score(y_test, y_pred))
6. print(f1_score(y_test, y_pred))
```

Based on the classification report, the performance of the Support Vector Machine algorithm—measured using accuracy, precision, recall, and F1-score—can be visualized using a heatmap as seen on Figure 2. This visual representation allows for an easier comparison of the model's performance across different kernel functions and evaluation metrics, highlighting the effectiveness of each configuration in stroke prediction.

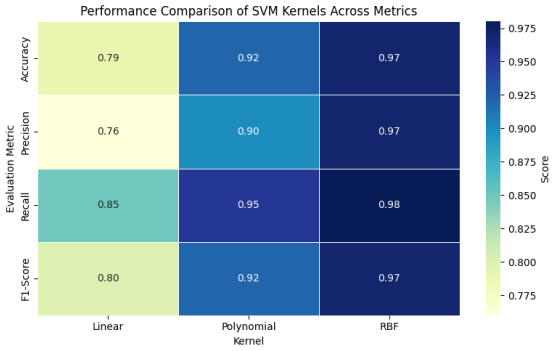

Figure 2. Comparison of SVM kernels

Based on the classification performance results of stroke prediction using the Support Vector Machine algorithm and the Stroke Prediction Dataset, the highest accuracy was achieved using the **RBF kernel** with parameters **C = 10** and **gamma = 0.1**, across various dataset split ratios (80:20 and 90:10) and k-fold values (k = 5, 7, and 10). The highest **precision** was also obtained using the RBF kernel with the same parameters and train-test split of 80:20, validated consistently across all k-fold values. The best **recall** score was achieved with the RBF kernel under the same configuration (C = 10, gamma = 0.1) across both 80:20 and 90:10 split ratios and all k values. Similarly, the highest **F1-score** was recorded under this same configuration, highlighting the effectiveness of the RBF kernel in stroke classification. These findings are consistent with previous studies. Arslan et al. (2016) stated that Support Vector Machine is among the most effective machine learning algorithms when compared to other models. Additionally, emphasized that SVM is regarded as one of the most robust and accurate algorithms in clinical prediction tasks. Both studies support the conclusion that SVM is a reliable and effective algorithm for stroke prediction.

5. Conclusions

This study aimed to build a machine learning model to predict stroke disease using the Support Vector Machine (SVM) algorithm. The model was trained and evaluated using the Stroke Prediction Dataset from Kaggle, with the integration of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Based on the experimental results, the **RBF kernel** consistently outperformed other kernel types, achieving the highest accuracy, precision, recall, and F1-score. The best performance was obtained using the configuration of **C = 10** and **gamma = 0.1**, with dataset split ratios of **80:20 and 90:10**, and k-fold values of **5, 7, and 10**. These results demonstrate the effectiveness of combining SVM with proper kernel tuning and data balancing techniques such as SMOTE. Furthermore, the study confirms previous findings that Support Vector Machine is among the most reliable and accurate machine learning algorithms for clinical prediction tasks. The use of confusion matrices and heatmaps also facilitated a comprehensive evaluation of model performance across multiple configurations. In conclusion, the SVM model developed in this research can serve as a useful tool for early stroke prediction, supporting healthcare professionals in decision-making and risk assessment. Future work may involve comparing the model with deep learning approaches or testing it with real-time clinical datasets for further validation.

References

[1] Agustiyawan and E. Prabowo, "Pembekalan kemampuan deteksi dini dan asesmen stroke," *J. Pengabdian Masyarakat Multidisiplin*, vol. 4, no. 1, pp. 1–5, 2020.

[2] H. Lin et al., "Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry," *Comput. Methods Programs Biomed.*, vol. 190, p. 105338, 2020. doi: 10.1016/j.cmpb.2019.105338.

[3] Y. Hung et al., "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jeju, Korea, 2017, pp. 3110–3113. doi: 10.1109/EMBC.2017.8037515 .

[4] 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. Inform.)*, vol. 8, no. 4, pp. 555–562, 2024 .

[5] 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 .

[6] K. M. Park et al., "Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke," *Front. Neurol.*, vol. 14, p. 1234046, 2023. doi: 10.3389/fneur.2023.1234046 .

[7] L. A. Martini, G. A. Pradipta, and R. R. Huizen, "Analysis of the impact of data oversampling on the support vector machine method for stroke disease classification," *J. Electr. Electron. Eng. Med. Inform.*, vol. 4, no. 2, pp. 96–105, 2022. doi: 10.35882/jeeemi.v4i2.698 .

[8] L. Despitasari, "Hubungan hipertensi dengan kejadian stroke berulang pada penderita pasca stroke," *MIDWINERSLION: J. Kesehatan STIKes Buleleng*, vol. 5, no. 1, pp. 124–131, 2020.

[9] M. Huda, *Algoritma Data Mining: Analisis Data Dengan Komputer*. Yogyakarta: Bisakimia, 2019.

[10] S. Jung et al., "Predicting ischemic stroke in patients with atrial fibrillation using machine learning," *Front. Biosci.-Landmark*, vol. 27, no. 3, p. 80, 2022. doi: 10.31083/j.fbl2703080.

[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] S. Susilawati and S. K. Nurhayati, "Faktor resiko kejadian stroke," *J. Ilm. Keperawatan Sei Betik*, vol. 14, no. 1, pp. 41–48, 2018.

[13] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on an imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, 2019. doi: 10.1016/j.artmed.2019.101723.

[14] Y. He et al., "An exploration on the machine-learning-based stroke prediction model," *Front. Neurol.*, vol. 15, p. 1372431, 2024. doi: 10.3389/fneur.2024.1372431 .

[15] Z. Rustam, Arfiani, and J. Pandelaki, "Cerebral infarction classification using multiple support vector machine with information gain feature selection," *Bull. Electr. Eng. Inform.*, vol. 9, no. 4, pp. 1578–1584, 2020. doi: 10.11591/eei.v9i4.1997.