Vol. 8, No. 2, April 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Optimizing Stroke Prediction Using Backward Elimination and SMOTE with C4.5 and K-Nearest Neighbors

Imam Bagus Pratama[1], Ahmad Zainul Fanani[2], M. Arief Soeleman[3], Via Indriani Kumalasari[4]

[123]Department of Computer Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia,

[4]Department of Physics Education, Faculty of Education, State University of Semarang, Indonesia

Received:
December 14, 2026
Revised:
March 10, 2026
Accepted:
April 11, 2026
Published:
April 26, 2026

Corresponding Author:
Author Name*:
Imam Bagus Pratama
Email*:
imambagus895@gmail.com

DOI:
10.63158/journalisi.v8i2.1521

Abstract. Early prediction of stroke risk is crucial for reducing mortality and the burden on the healthcare system, but class imbalance and irrelevant features often compromise model reliability. This study analyzes the impact of Backward Elimination and SMOTE on the performance of the C4.5 and K-NN algorithms in stroke prediction. The study used a fixed working subset of 1,239 data points and evaluated four modeling scenarios using Stratified 10-Fold Cross Validation. Model performance was measured using accuracy, precision, recall, F1-score, and AUC. The results showed that Backward Elimination improved model performance on the analyzed subsets. For C4.5, accuracy increased from 70.94% to 73.05%, stroke recall from 83.94% to 85.14%, and AUC from 0.776 to 0.806. For K-NN, accuracy increased from 72.31% to 74.82% and precision from 39.91% to 42.73%, while stroke recall remained relatively stable at 74.30%. These findings indicate that although the improvements are small numerically, the results remain practically relevant as they enhance the balance between sensitivity and class discrimination capability. In the context of stroke screening, reducing false negatives is more important because it helps minimize undetected high-risk cases, although false positives still need to be considered as a consequence of further testing. Overall, C4.5 with Backward Elimination demonstrates more balanced performance, although the results are still limited to the analyzed subset.

Keywords: stroke prediction, SMOTE, class imbalance, C4.5, K-nearest neighbors, backward elimination, feature selection

1. INTRODUCTION

Stroke is classified as a noncommunicable disease that has a significant impact on global public health. This condition occurs when blood flow to the brain is blocked, causing damage to nerve tissue and potentially leading to permanent disability or death.[1] The World Health Organization reports that stroke is one of the leading causes of death worldwide. In Indonesia, it is the second leading cause of death, indicating that the threat of stroke is not only a clinical concern but also a serious public health issue.

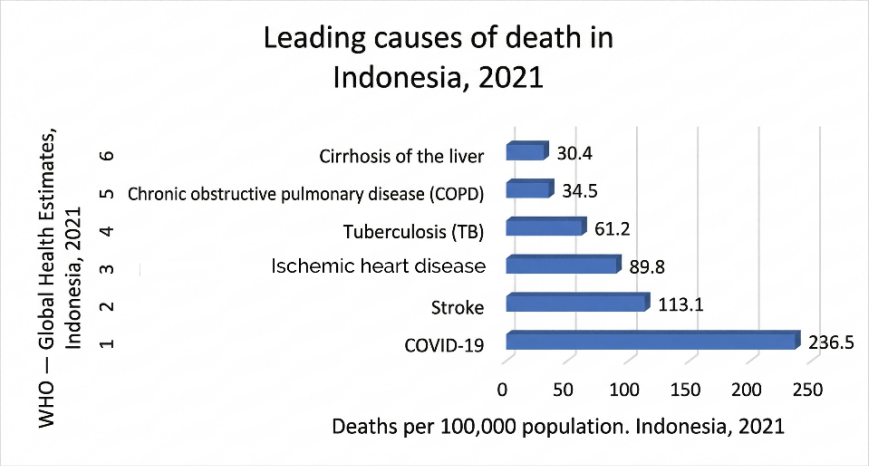

Figure 1 Distribution of Cases by Disease Type. WHO, 2021

Figure 1 shows the distribution of mortality rates by disease type. The data indicate that stroke is the second leading cause of death after COVID-19, with a rate of 113.1 deaths per 100,000 population. The disease with the highest mortality rate is COVID-19 (236.5), followed by stroke (113.1), ischemic heart disease (89.8), tuberculosis (TB) (61.2), Chronic Obstructive Pulmonary Disease (COPD) (34.5), and Cirrhosis of the liver (30.4). Figure 1 not only demonstrates that stroke is among the leading causes of death but also underscores that stroke is a public health issue requiring attention at the early detection stage. The high mortality burden from stroke indicates that delays in identifying risk factors can directly impact delays in prevention and initial management. Therefore, the development of a predictive model capable of identifying stroke risk more sensitively is crucial, particularly as an initial screening approach to support decision-making in the healthcare sector.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Advances in computing technology have driven the widespread adoption of machine learning in the healthcare sector, including for disease analysis and prediction. Various algorithms such as Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree (C4.5), and K-Nearest Neighbors (K-NN) have been widely used, with varying levels of accuracy.[2][3][4] These results indicate that the choice of algorithm, feature quality, and class balancing techniques significantly influence model performance. Healthcare datasets typically face challenges such as the presence of irrelevant features, correlation among attributes, and imbalanced class distributions.[5][6] This situation can reduce accuracy, precision, and recall, particularly for the minority class (stroke).

Several previous studies have applied various machine learning algorithms to predict stroke risk[7][3][4]. The implementation of SMOTE has been shown to significantly improve model performance.[8][9] The Random Forest algorithm performed best by improving accuracy, precision, recall, and F1-score [7]. Similar findings have also been reported by [9] which states that SMOTE is capable of improving the sensitivity of the K-NN model toward the minority class, even though precision remains limited. Another study by [10] assessed the effectiveness of the C4.5 algorithm in classifying stroke, whereas a study by [11] developing a neural network-based stroke risk prediction model. In addition, [12] introduces a feature selection approach to improve the performance of decision trees and neural networks on a stroke prediction dataset. The results show that feature quality and preprocessing techniques have a significant impact on the accuracy and sensitivity of the models.

Based on this review, there remains a research gap in stroke prediction, namely that few studies have directly tested the combination of Backward Elimination feature selection and SMOTE class balancing within a single comparative evaluation framework for the C4.5 and K-NN algorithms [3][13]. The novelty of this study lies in the controlled testing of these two preprocessing techniques to examine their impact on the performance of two algorithms with different characteristics: the rule-based C4.5 and the distance-based K-NN. This study aims to analyze the impact of Backward Elimination on model performance, compare the effectiveness of C4.5 and K-NN based on accuracy, precision, recall, F1-score, and AUC, and identify attributes that remain after feature selection. The contribution of this study is to provide a more comprehensive evaluation of imbalanced

stroke data,[14][15] with an emphasis on the sensitivity of stroke classification and the trade-off between false positives and false negatives in the context of initial screening.

2. METHODS

This study employs a quantitative approach to evaluate the performance of classification models in predicting stroke risk. The data used comes from a stroke dataset available on Kaggle and was collected in 2021. The original dataset consists of 5,110 records with an imbalanced class distribution, namely 250 stroke cases and 4,860 non-stroke cases. Before determining the final working subset, the data first underwent a data validity check. At this stage, 1 stroke record was excluded because it did not meet the validity criteria. Following this process, a fixed working subset of 1,239 records was established, consisting of 249 stroke cases and 990 non-stroke cases. This working subset was formed in a controlled manner for comparative experimental purposes, not through random sampling. As many stroke records as possible that met the criteria were retained, while non-stroke data were selected non-randomly based on record order until 990 records were reached. With this approach, the working subset is not intended to represent the true prevalence of the dataset, but rather to provide a fixed working dataset that allows for the consistent evaluation of the effects of preprocessing and feature selection across all testing scenarios. To provide an overview of the composition of the data used, Table 1 presents a comparison of class distributions between the original dataset and the working subset.

Table 1 Comparison of Class Distributions Between the Original Dataset and the Working Subset

Characteristics	Original Dataset	Working subset
Total Records	5110	1239
Stroke	250	249
Non-Stroke	4860	990
Proportions of Stroke	4.89%	20.10%
Non-Stroke Proportion	95.11%	79.90%

Table 1 shows that the working subset retains nearly all stroke cases but does not retain the original prevalence of the dataset. The proportion of strokes in the working subset

increases to 20.10%, higher than the 4.89% in the original dataset. This confirms that the subset is used as working data for comparative experiments, not as an inferential sample representing the full distribution of the original dataset. With this composition, results on the full dataset—which has more extreme class imbalance—may differ, particularly regarding stroke class precision, which could potentially decrease due to the increased proportion of false positives. Therefore, further validation on the full dataset or an external dataset remains necessary before drawing broader implications.

To ensure the validity of the evaluation, the same subset is consistently used across all experimental scenarios. Additionally, the class balancing process using SMOTE is applied only to the training data in each fold of the Stratified 10-Fold Cross-Validation, so that the test data is not involved in the oversampling process. This approach was taken to minimize evaluation bias and reduce the risk of data leakage during the modeling process. Modeling and evaluation were conducted by comparing four scenarios: C4.5 without feature selection, C4.5 with Backward Elimination, K-NN without feature selection, and K-NN with Backward Elimination. The entire data processing, modeling, and evaluation process was performed using Altair RapidMiner Studio version 2026.0.2 on a device with the following specifications: Intel Core i7-8700, 8 GB RAM, and Windows 11 Pro. This information is included to support transparency regarding the experimental environment and to facilitate interpretation of the computational times obtained.

2.1. Data Collection

The dataset contains 10 predictor features and 1 target label, namely stroke status, which indicates whether a patient has had a stroke or not. The features used in this study are as shown in Table 2. This dataset has common issues, including missing values in the BMI attribute and class imbalance, so preprocessing is required before modeling.

Table 2 Research Dataset

No	Attributes	Description
1	Gender	Gender (Male/Female)
2	Age	Age of respondents
3	Heart Disease	History of heart disease (yes/no)
4	Hypertension	Hypertension (yes/no)

No	Attributes	Description
5	Ever Married	Marital status (married/never married)
6	Residence Type	Type of residence (urban/rural)
7	Work Type	Type of work
8	Smoking Status	Smoking status
9	Average Glucose Level	Average blood glucose level
10	Body Mass Index (BMI)	Body Mass Index
11	Stroke	A diagnostic label indicating whether or not the patient has had a stroke

2.2. Experimental Framework

This study proposes an experimental framework that combines preprocessing, class balancing, and feature selection to compare the performance of the C4.5 and K-NN algorithms in stroke prediction. The experimental workflow is conducted in stages as follows. First, the stroke dataset is collected and organized into a fixed working subset used consistently across all scenarios. Second, data preprocessing is performed, which includes handling missing values and preparing predictor attributes. Third, model evaluation is conducted using Stratified 10-Fold Cross Validation to ensure the proportions of stroke and non-stroke classes remain consistent in each fold. Fourth, in each fold, the data was first split into training and test data. Fifth, SMOTE was applied only to the training data to reduce bias toward the majority class and prevent data leakage. Sixth, in the feature selection scenario, Backward Elimination was applied to the training data as a wrapper method to select the most relevant attributes before model training. Seventh, the C4.5 and K-NN models are trained and tested across four scenarios: C4.5 without feature selection, C4.5 with Backward Elimination, K-NN without feature selection, and K-NN with Backward Elimination. Finally, model performance is compared using accuracy, precision, recall, F1-score, and AUC.[15],[16].

2.3. Data processing

The preprocessing stage is conducted to ensure that the dataset is in optimal condition before the modeling process begins. Some of the steps carried out in this study include:

1) Handling Missing Values

The dataset contains missing values in the Body Mass Index (BMI) attribute, with a total of 52 missing values. To address this issue, imputation was performed using the Replace Missing Value operator. The imputation process was carried out prior to the modeling stage to ensure that all data were complete and could be processed consistently across all experimental scenarios. [17]

2) Addressing Class Imbalances

The dataset has an imbalanced class distribution between the stroke and non-stroke classes. This imbalance can cause the model to be biased toward the majority class, thereby neglecting the minority class. To address this issue, a sampling process was performed using the SMOTE (Synthetic Minority Oversampling Technique) method. SMOTE was applied only to the training data in each evaluation fold to prevent data leakage. The SMOTE technique works by synthetically generating new samples based on the similarity between data points in the minority class. This process helps improve class distribution so that the model can learn more evenly and improve classification performance, particularly on the recall and F1-score metrics.[18]

2.4. Feature Selection

The feature selection stage was conducted using Backward Elimination with a wrapper approach to identify the attributes most relevant to stroke risk prediction. In this method, the selection process begins with all available predictor attributes; attributes are then gradually eliminated based on their contribution to model performance until a more optimal subset of attributes is obtained. In this study, Backward Elimination was applied as part of the experimental scenario and evaluated consistently alongside a scenario without feature selection. In each fold, the feature selection process was performed only on the training data after the class balancing stage, so that information from the test data was not involved in attribute selection. Thus, the feature selection process remained within the training workflow and did not cause data leakage [19].

2.5. Model Parameters

Model parameters were set consistently across all evaluation folds. For the K-NN algorithm, the parameters used included k = 3, a distance metric using mixed measures with MixedEuclideanDistance, and a voting scheme using weighted voting, such that

neighbors closer to the test data exert a greater influence on class determination. In the C4.5 algorithm, the model parameters used include criterion = gain ratio, maximal depth = 10, apply pruning = true, confidence = 0.1, apply prepruning = true, minimal gain = 0.01, minimal leaf size = 2, and minimal size for split = 4. All of these parameters are used consistently in every experimental scenario to ensure that the results of the comparison between models remain consistent.

2.6 Flowchart

The flowchart in Figure 2 illustrates the sequence of research steps conducted systematically, from data collection to the drawing of final conclusions. The first stage begins with the data collection process, namely obtaining the stroke dataset that is the subject of the research. The data obtained then undergoes cleaning and preprocessing, which includes handling missing values and class balancing using SMOTE to meet the needs of the modeling process. Once the data is prepared, the research proceeds to the initial evaluation stage using the cross-validation technique.

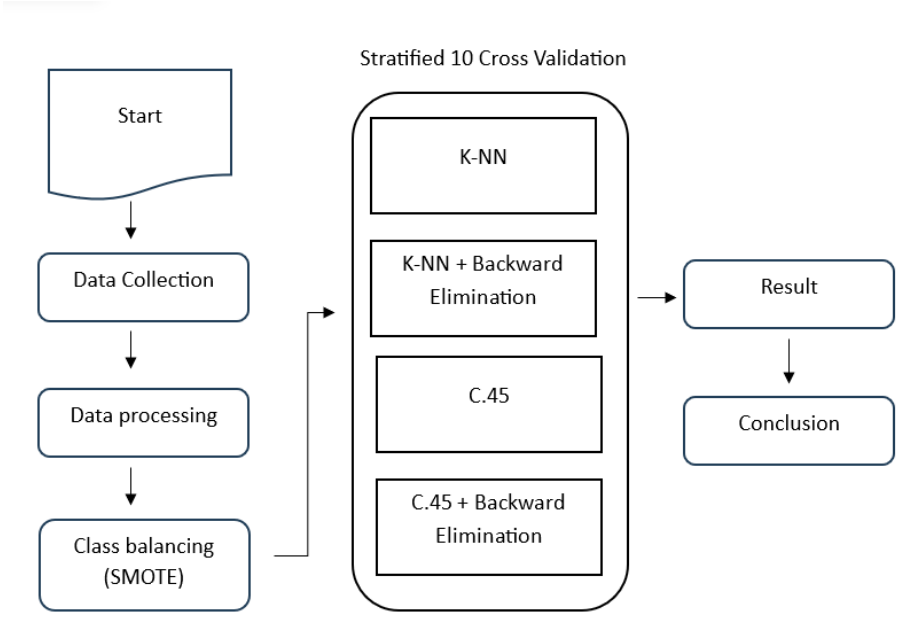

Figure 2. Research Flow

This evaluation was conducted to ensure data readiness and to measure baseline performance prior to optimization through feature selection. The next stage is the modeling process, which is carried out through four scenarios, namely: (1) K-Nearest

Neighbor (K-NN) without feature selection, (2) K-NN with Backward Elimination feature selection, (3) C4.5 without feature selection, and (4) C4.5 with Backward Elimination. Each scenario is run to analyze the impact of feature selection on the classification algorithm's performance. All modeling results are then incorporated into the Results section, which includes a comparison of accuracy, precision, recall, F1-Score, and AUC values. The final stage is the Conclusion, which presents findings based on the model performance analysis and their implications for future research development. Thus, the methodological contribution of this study lies not only in the comparison of two classification algorithms, but also in the experimental framework that consistently combines class imbalance handling using SMOTE and feature selection using Backward Elimination in the comparative evaluation of stroke prediction.

3. RESULTS AND DISCUSSION

3.1 Decision Tree Model

The modeling process using the Decision Tree (C4.5) algorithm without feature selection took 1 minute to complete. The confusion matrix results show the number of correct and incorrect predictions for each class, namely the stroke class (positive) and the non-stroke class (negative). The classification as shown in Table 3 shows that the algorithm produced. Based on these results, the Decision Tree model demonstrated a high recall rate of 83.94% for the stroke class, indicating the model's ability to detect the majority of stroke cases with a relatively low false-negative rate. This is a critical factor in a medical context, as failure to detect stroke patients has the potential to cause serious clinical consequences.

Table 3 Confusion Matrix for the DT Algorithm, Including Precision and Recall Values

Accuracy: 70.94%

	True Stroke	**True non-stroke**	**Class Precision (%)**
Pred. Stroke	209	320	39.51
Pred. non-stroke	40	670	94.37
Class Recall (%)	83.94	67.68	

However, the precision value for the stroke class remains relatively low at 39.51%, indicating that some stroke predictions are still false positives. This suggests that while

the model is capable of detecting stroke cases effectively, the accuracy of positive predictions still needs to be improved. The low precision may be influenced by the complexity of the relationships between features as well as the tendency of the Decision Tree algorithm to form rules that are too specific, thereby increasing the number of inaccurate positive predictions.

3.2 Decision Tree Model Using Backward Elimination

The implementation of the Decision Tree (C4.5) algorithm combined with feature selection took 3 minutes of computation time. Based on the evaluation process, the model produced 212 data points classified as True Positive (TP), 297 as False Positive (FP), 37 as False Negative (FN), and 693 as True Negative (TN). These values indicate that the model has good classification ability in distinguishing between stroke and non-stroke classes. These values show that the model was able to correctly identify 212 stroke patients, but still misclassified 37 stroke patients as non-stroke (false negative). Meanwhile, there were 297 non-stroke patients who were incorrectly predicted as having a stroke, indicating a tendency for the model to produce an overabundance of positive predictions.

Table 4 Confusion Matrix for the DT+BE Algorithm, Including Precision and Recall Values

Accuracy: 73.05%

	True Stroke	True non-stroke	Class Precision (%)
Pred. Stroke	212	297	41.65
Pred. non-stroke	37	693	94.93
Class Recall (%)	85.14	70.00	

Based on Table 4, an average accuracy of 73.05% ± 3.68% was obtained using the 10-fold stratified cross-validation scheme, indicating that the model successfully classified the majority of the data. The recall for the stroke class reached 85.14%, suggesting that the model performs well in detecting stroke cases and effectively minimizes false negatives. This is a critical aspect in the healthcare context, as failure to detect stroke patients could potentially lead to delays in medical treatment. However, the precision value for the stroke class remains relatively low at 41.65%, indicating a tendency for the model to generate excessive positive predictions (false positives). Conversely, the non-stroke class has a high precision value of 94.93%, so non-stroke predictions can be considered more

reliable. The balance between precision and recall in the stroke class is reflected in the F1-score of 56.05%, indicating that the model is more focused on improving sensitivity rather than the accuracy of positive predictions. Table 5 shows the results of backward elimination:

Table 5 Results of attribute elimination from DT+BE

Attribute	weight
Gender	1
Age	1
Hypertension	1
Heart Disease	1
Ever Married	1
Residence Type	1
Work Type	0
Smoking Status	1
Average Glucose Level	1
Body Mass Index (BMI)	1

Table 5 shows the results of feature selection using the Backward Elimination method in the Decision Tree (C4.5) algorithm. Based on these results, out of the total of 10 predictor attributes used in the initial stage, one attribute was eliminated—namely, Work Type—while the other attributes were retained with a weight of 1. The weight values in the table indicate the status of the attributes after the feature selection process, where a value of 1 indicates that the attribute was retained because it contributes to the model's performance, while a value of 0 indicates that the attribute was eliminated because its contribution to classification was deemed insignificant. The elimination of the Work Type attribute indicates that the information contained in that attribute does not provide a significant improvement in model performance when combined with other attributes. These results show that attributes other than "Work Type" play a more dominant role in the stroke classification process. By removing less relevant attributes, the model becomes simpler, more stable, and capable of improving performance, particularly in terms of recall and accuracy metrics.

3.3 K-NN Model

The modeling process using the K-NN algorithm without feature selection takes 1 minute of computation time. Table 6 is Confusion Matrix for the K-NN Algorithm, Including Precision and Recall Values. The model evaluation results showed 186 data points classified as True Positive (TP), 280 as False Positive (FP), 63 as False Negative (FN), and 710 as True Negative (TN). These findings indicate that the model successfully identified 186 patients with stroke accurately, although there were still 63 stroke cases predicted as non-stroke (false negatives). Additionally, 280 patients who did not actually have stroke were classified as having stroke by the model.

Table 6 Confusion Matrix for the K-NN Algorithm, Including Precision and Recall Values

Accuracy: 72.31%

	True Stroke	True non-stroke	Class Precision (%)
Pred. Stroke	186	280	39.91
Pred. non-stroke	63	710	91.85
Class Recall (%)	74.70	71.72	

The average accuracy obtained was 72.31% ± 4.38% using the 10-fold stratified cross-validation scheme, indicating that the model was able to classify the majority of the data fairly well. The recall value for the stroke class was 74.70%, indicating that the model was able to detect about three-quarters of stroke cases, although there was still a clinical risk due to the presence of false negatives. Meanwhile, the precision for the stroke class was recorded at 39.91%, indicating that positive predictions are still dominated by false positives. Conversely, the non-stroke class demonstrated more stable performance with a recall of 71.72% and a precision of 91.85%, making non-stroke predictions relatively more reliable. The balance between precision and recall in the stroke class is reflected in an F1-score of 52.00%, indicating that the model's performance falls into the moderate category. These results suggest that the K-NN model without feature selection is more focused on case detection than on the accuracy of positive predictions, thus requiring further optimization to improve stroke prediction accuracy.

3.4 K-NN Model Using Backward Elimination

Table 7 present Confusion Matrix for the K-NN + BE Algorithm, Including Precision and Recall. The modeling process using the K-NN algorithm with Backward Elimination and

feature selection required 3 minutes of computation time. Based on the test results, the model achieved 185 True Positives (TP), 248 False Positives (FP), 64 False Negatives (FN), and 742 True Negatives (TN). These results indicate that the model was able to correctly identify 185 stroke patients, but there were still 64 stroke patients who were incorrectly classified as non-stroke (false negatives). Additionally, 248 non-stroke patients were predicted to have stroke, indicating that the model still tends to produce an excessive number of positive predictions.

Table 8 Confusion Matrix for the K-NN + BE Algorithm, Including Precision and Recall

Accuracy: 74.82%

	True Stroke	**True non-stroke**	**Class Precision (%)**
Pred. Stroke	185	248	42.73
Pred. non-stroke	64	742	92.06
Class Recall (%)	74.30	74.95	

The modeling process using the K-NN algorithm with Backward Elimination and feature selection required 3 minutes of computation time. Based on the test results, the model achieved 185 True Positives (TP), 248 False Positives (FP), 64 False Negatives (FN), and 742 True Negatives (TN). These results indicate that the model was able to correctly identify 185 stroke patients, but there were still 64 stroke patients who were incorrectly classified as non-stroke (false negatives). Additionally, 248 non-stroke patients were predicted to have stroke, indicating that the model still tends to produce an excessive number of positive predictions.

Based on the confusion matrix, the model achieved an accuracy of 74.82% ± 4.64%, with a micro-average of 74.82%. These results indicate that the model was able to correctly classify the majority of the data and performed relatively better than previous results. The recall or sensitivity for the stroke class is 74.30%, indicating that the model can detect approximately 185 out of 249 patients who actually experienced a stroke. This value reflects fairly good detection capability, although there remains a clinical risk due to unidentified stroke cases. Precision for the stroke class was recorded at 42.73%, meaning that less than half of the stroke predictions were actual stroke cases. In the non-stroke class, the model demonstrated stable performance with a recall of 74.95% and a precision of 92.06%. This indicates that predictions for non-stroke patients have a

high level of reliability, meaning the model is quite effective at identifying healthy patients. The balance between precision and recall in the stroke class is reflected in an F1-score of approximately 54%, indicating the model's performance falls into the moderate category. This value suggests that the model prioritizes sensitivity over the accuracy of positive predictions.

Table 9 Results of attribute elimination from K-NN+BE

Attribute	weight
Gender	1
Age	1
Hypertension	1
Heart Disease	1
Ever Married	1
Residence Type	1
Work Type	1
Smoking Status	1
Average Glucose Level	1
Body Mass Index (BMI)	0

Table 8 shows the results of feature selection using the Backward Elimination method in the K-Nearest Neighbors (K-NN) algorithm. Based on these results, out of the total predictor attributes used, one attribute was eliminated—Body Mass Index (BMI)—while the other attributes were retained with a weight of 1.

The weight values in the table indicate the status of the attributes after the feature selection process, where a value of 1 indicates that the attribute was retained because it contributes to the model's performance, while a value of 0 indicates that the attribute was eliminated because its contribution was deemed insignificant in the classification process. The elimination of the BMI attribute indicates that the information contained in that attribute does not significantly improve the performance of the K-NN model when combined with other attributes. In the distance-based K-NN algorithm, the presence of certain numerical features can affect the model's sensitivity to data distribution. In this context, the BMI attribute is correlated with other attributes such as age and blood glucose levels, so its presence has the potential to add redundancy and noise to the

distance calculation process. By removing this attribute, the model becomes more stable and is able to improve classification performance, particularly in terms of accuracy and F1-score metrics. Figure 3 shown the comparison of accuracy for all classification models.

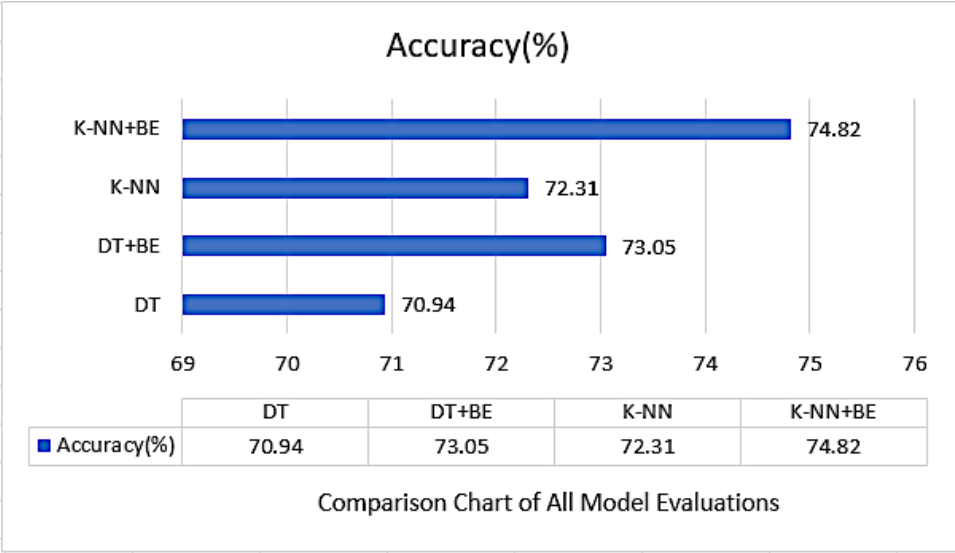

Figure 3 Bar Chart Comparing the Accuracy of All Model Evaluations

Figure 3 shows a comparison of the accuracy scores of all the classification models tested, namely Decision Tree (DT), Decision Tree with Backward Elimination (DT+BE), K-Nearest Neighbors (K-NN), and K-Nearest Neighbors with Backward Elimination (K-NN+BE). The evaluation was conducted using a Stratified 10-Fold Cross-Validation scheme on a dataset that had undergone preprocessing and class balancing using SMOTE. Based on the graph, it is evident that the application of Backward Elimination feature selection improves accuracy for both algorithms. For the Decision Tree algorithm, accuracy increased from 70.94% to 73.05%, while for the K-NN algorithm, accuracy increased from 72.31% to 74.82%. This improvement indicates that removing less relevant attributes can enhance the effectiveness of the classification process. When comparing the algorithms, K-NN demonstrates a higher accuracy value than the Decision Tree, both in scenarios without feature selection and with feature selection. The K-NN+BE model achieved the highest accuracy of 74.82%. However, in the context of stroke prediction, model performance evaluation cannot be based solely on accuracy; it is also necessary to consider recall, F1-score, and AUC to assess the ability to detect stroke cases more comprehensively.

Table 10 A comparison of the performance of four classification models based on accuracy, precision, recall, and F1-score in percentage (%), as well as the AUC value for stroke risk prediction

Model	Accuracy	Precision	Recall	F1 score	AUC
C4.5	70.94	39.51	83.94	53.86	0.776
K-NN	72.31	39.91	74.70	52.10	0.766
C4.5+BE	73.05	41.65	85.14	56.05	0.806
K-NN+BE	74.82	42.73	74.30	54.39	0.759

Based on Table 9, the comparison table of the four classification algorithms above, although the K-NN+BE model produced the highest accuracy score, the C4.5+BE model actually delivered the best overall performance. This is evident from its higher recall, F1-score, and AUC values compared to the other models. In the context of stroke prediction, where the number of cases is far fewer than non-stroke cases, the model's ability to not miss patients who are truly at risk is of utmost importance. The C4.5+BE model achieves a recall of 85.14%, indicating that it is more sensitive in identifying stroke cases. Additionally, an AUC of 0.806 demonstrates that this model is better at distinguishing between stroke and non-stroke patients. With this combination of performance metrics, the C4.5+BE model exhibits a more balanced performance on the analyzed data subset.

3.5 Discussion

The results of the study show that the application of Backward Elimination has a positive impact on both tested algorithms, although the patterns of improvement differ. In the C4.5 algorithm, feature selection increased accuracy from 70.94% to 73.05%, stroke class recall from 83.94% to 85.14%, and AUC from 0.776 to 0.806. Based on the confusion matrix, the number of false negatives decreased from 40 to 37 and false positives decreased from 320 to 297. These findings indicate that removing less relevant attributes helps rule-based models form a classification structure that is more focused on patterns important to the stroke class [5], [20]. In the context of medical prediction, improving recall is particularly important because it indicates that the model is able to reduce the likelihood of undetected stroke cases.

In the K-NN algorithm, the application of Backward Elimination improved accuracy from 72.31% to 74.82% and precision from 39.91% to 42.73%, while stroke-class recall remained

relatively stable, changing slightly from 74.70% to 74.30%. Based on the confusion matrix, false positives decreased from 280 to 248, but false negatives slightly increased from 63 to 64. These results indicate that feature selection in K-NN contributed more to improving the correctness of positive predictions and reducing false alarms than to increasing sensitivity for stroke detection. Importantly, although accuracy improved, the AUC decreased from 0.766 to 0.759. This pattern suggests that the benefit of Backward Elimination in K-NN was more apparent at the specific classification threshold used in this study, but it did not improve the model's overall ability to discriminate between stroke and non-stroke cases across decision boundaries. In other words, the reduced feature set helped the model make more accurate final classifications in some instances, yet it did not strengthen class separation in a broader sense. Therefore, the improvement in K-NN after feature selection should be interpreted cautiously: higher accuracy does not necessarily indicate better overall discrimination performance, especially in an imbalanced medical prediction task where AUC remains an important indicator of model robustness [5], [13].

The results of this study also indicate a clear trade-off between precision and recall across all tested models. In the context of medical screening, recall is a particularly important metric because it is directly related to the model's ability to minimize false negatives—that is, cases of stroke patients who go undetected [21][22]. False negatives are more critical than false positives because patients who are actually at risk of stroke may go unidentified and potentially experience delays in clinical evaluation and early intervention. Therefore, models with higher recall are more relevant for the initial screening stage. However, the precision for the stroke class across all models remains relatively low, ranging from 39.51% to 42.73%, indicating that more than half of the positive predictions are still false positives.

In real-world healthcare settings, a relatively high false-positive rate can still be tolerated if the model is used as an initial screening tool rather than a definitive diagnostic tool, and every positive result is followed up with confirmatory testing. However, this situation still requires careful consideration, as false positives can place an additional burden on the healthcare system [4], [18], [23], [24]. Therefore, the most appropriate model for this study is not merely the one with the highest accuracy, but the one that offers a better balance between detection sensitivity and class discrimination capability. Based on these

considerations, C4.5+BE is more relevant than the other models because it yields the highest recall, the highest F1-score, and the best AUC on the analyzed subset.

Compared to previous studies, the results of this study are consistent with the finding that SMOTE can improve the model's sensitivity on minority classes, although precision is often still limited [4], [18], [25], [26], Meanwhile, the improvement in performance of the C4.5 model following feature selection supports the study [12] which indicates that reducing less relevant attributes can improve model stability. The main difference between this study and previous ones lies in the comparative evaluation of two algorithms within the same experimental framework, as well as the use of a more comprehensive set of metrics, namely accuracy, precision, recall, F1-score, and AUC. Quantitatively, the most significant improvement in this study occurred in C4.5, with an increase in accuracy of 2.11 points, stroke recall of 1.20 points, and AUC of 0.030 after applying Backward Elimination. Meanwhile, for K-NN, the main improvements were in accuracy (2.51 points) and precision (2.82 points), but these were not accompanied by an increase in recall. This indicates that the impact of feature selection is not uniform across all algorithms but is influenced by the characteristics of the models used [19].

4. CONCLUSION

This study shows that the application of Backward Elimination improves the performance of the C4.5 and K-NN algorithms on the analyzed data subset. The improvement is most evident in the C4.5 model, with accuracy increasing from 70.94% to 73.05%, stroke class recall increasing from 83.94% to 85.14%, and AUC increasing from 0.776 to 0.806. Meanwhile, for K-NN, Backward Elimination improved accuracy from 72.31% to 74.82% and precision from 39.91% to 42.73%, although stroke class recall remained relatively stable. Overall, C4.5+BE demonstrated more balanced performance based on recall, F1-score, and AUC. However, these results are still limited to the working subset used, with class conditions remaining imbalanced and the stroke class precision still relatively low. Therefore, these findings need to be further validated on the full dataset or an external dataset before being applied to broader scenarios.

ACKNOWLEDGMENTS

The author would like to thank Dian Nuswantoro University for the academic support provided throughout the conduct and preparation of this research. The author also extends gratitude to the dataset providers for making the data openly available, thereby enabling the successful completion of this research. Furthermore, the author appreciates the contributions of all those who provided constructive feedback and suggestions in refining this article.

REFERENCES

[1] V. L. Feigin *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *International Journal of Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022, doi: 10.1177/17474930211065917.

[2] W. Heseltine-Carp *et al.*, "Machine learning to predict stroke risk from routine hospital data: A systematic review," *Int. J. Med. Inform.*, vol. 196, no. January, p. 105811, 2025, doi: 10.1016/j.ijmedinf.2025.105811.

[3] F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The most efficient machine learning algorithms in stroke prediction: A systematic review," *Health Sci. Rep.*, vol. 7, no. 10, 2024, doi: 10.1002/hsr2.70062.

[4] T. Vu *et al.*, "Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 7, 2024, doi: 10.3390/jcdd11070207.

[5] P. Chakraborty *et al.*, "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–23, 2024, doi: 10.1186/s12859-024-05866-8.

[6] J. Zhu *et al.*, "Processing imbalanced medical data at the data level with assisted-reproduction data as an example," *BioData Min.*, vol. 17, no. 1, 2024, doi: 10.1186/s13040-024-00384-y.

[7] F. Fadmadika, H. H. Handayani, T. Al Mudzakir, and J. Indra, "Pengaruh Smote Terhadap Performa Algoritma Random Forest Dan Algoritma Gradient Boosting Dalam Memprediksi Penyakit Stroke," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 2, p. 837, Dec. 2024, doi: 10.37600/tekinkom.v7i2.1575.

[8] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[9] Z. Khairi, R. Yanti, T. A. Fitri, and E. Fatdha, "Optimasi Algoritma Knn Menggunakan Smote Untuk Prediksi Stroke," *Jurnal Algoritma*, vol. 22, no. 2, pp. 164–175, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2474.

[10] F. Nabila, I. Afrianty, S. Sanjaya, and F. Syafria, "Implementasi Algoritma C4.5 dalam Melakukan Klasifikasi Penyakit Stroke Otak," *Jurnal Informatika Universitas Pamulang*, vol. 8, no. 2, pp. 229–235, 2023, doi: 10.32493/informatika.v8i2.31361.

[11] A. Gupta *et al.*, "Predicting stroke risk: An effective stroke prediction model based on neural networks," *Journal of Neurorestoratology*, vol. 13, no. 1, p. 100156, 2025, doi: 10.1016/j.jnrt.2024.100156.

[12] Indah Werdiningsih *et al.*, "Analisis Prediksi Stroke Menggunakan Pendekatan Decision Tree dengan Seleksi Fitur dan Neural Network," *Jurnal Sistem Cerdas*, vol. 6, no. 3, pp. 213–221, Dec. 2023, doi: 10.37396/jsc.v6i3.310.

[13] K. Moulaei, L. Afshari, R. Moulaei, B. Sabet, S. M. Mousavi, and M. R. Afrash, "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models.," *Sci. Rep.*, vol. 14, no. 1, p. 31392, Dec. 2024, doi: 10.1038/s41598-024-82931-5.

[14] P. Eini, M. Rezayee, M. Kassulke, and J. Tremblay, "Efficacy and comparative performance of machine learning models for stroke risk prediction in hypertensive patients: A systematic review and meta-analysis," *International Journal of Cardiology Cardiovascular Risk and Prevention*, vol. 28, no. October 2025, p. 200564, 2026, doi: 10.1016/j.ijcrp.2025.200564.

[15] B. Van Calster *et al.*, "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance," *Lancet Digit. Health*, vol. 7, no. 12, p. 100916, 2025, doi: 10.1016/j.landig.2025.100916.

[16] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01313-4.

[17] M. Liu *et al.*, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artif. Intell. Med.*, vol. 142, Aug. 2023, doi: 10.1016/j.artmed.2023.102587.

[18] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. S. Philip Kegelmeyer, "synthetic minority over-sampling Technique," *J Artif Intell Res*, vol. 16, p. 16, 2018.

[19] D. Patel, A. Saxena, and J. Wang, "A Machine Learning-Based Wrapper Method for Feature Selection," *International Journal of Data Warehousing and Mining*, vol. 20, no. 1, pp. 1–33, 2024, doi: 10.4018/IJDWM.352041.

[20] D. Zhang, N. Yu, X. Yang, Y. De Marinis, Z. P. Liu, and R. Gao, "SRPNet: stroke risk prediction based on two-level feature selection and deep fusion network," *Front. Physiol.*, vol. 15, no. November, pp. 1–13, 2024, doi: 10.3389/fphys.2024.1357123.

[21] M. E. Klontzas *et al.*, "ESR Essentials: common performance metrics in AI—practice recommendations by the European Society of Medical Imaging Informatics," *Eur. Radiol.*, pp. 1528–1540, 2025, doi: 10.1007/s00330-025-11890-w.

[22] I. Aiyer, L. Shaik, A. Sheta, and S. Surani, "Review of Application of Machine Learning as a Screening Tool for Diagnosis of Obstructive Sleep Apnea," *Medicina (Lithuania)*, vol. 58, no. 11, 2022, doi: 10.3390/medicina58111574.

[23] M. Goyal *et al.*, "A bayesian framework to optimize performance of pre-hospital stroke triage scales," *J. Stroke*, vol. 23, no. 3, pp. 443–448, 2021, doi: 10.5853/jos.2021.01312.

[24] S. Patil, R. Rossi, D. Jabrah, and K. Doyle, "Detection, Diagnosis and Treatment of Acute Ischemic Stroke: Current and Future Perspectives," *Front. Med. Technol.*, vol. 4, no. June, 2022, doi: 10.3389/fmedt.2022.748949.

[25] M. Jacobs, N. Hammarlund, E. Evans, and C. Ellis, "Identifying predictors of stroke in young adults: a machine learning analysis of sex-specific risk factors," *Frontiers in Stroke*, vol. 3, no. MI, 2024, doi: 10.3389/fstro.2024.1488313.

[26] A. A. Soladoye, N. Aderinto, M. R. Popoola, I. A. Adeyanju, A. Osonuga, and D. B. Olawade, "Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps," *Int. J. Med. Inform.*, vol. 203, no. June, p. 106041, 2025, doi: 10.1016/j.ijmedinf.2025.106041.