Classification of Stroke Risk Based on Machine Learning: A Comparative Study of Naive Bayes and Decision Tree

Syahid Bahauddin
CV. Mitra Jaya Sejahtera, Madiun, Jawa Timur, Indonesia[1]
svidid2@gmail.com[1]

***Abstract*—Stroke is the leading cause of death in Indonesia, accounting for 19.42% of total mortality, with prevalence increasing by 56% from 7 per 1,000 population in 2013 to 10.9 per 1,000 population in 2018 based on Riskesdas data. Beyond its impact on morbidity and mortality, stroke is a catastrophic disease with the third-highest healthcare expenditure in Indonesia, reaching IDR 5.2 trillion in 2023. This condition underscores the urgency of early stroke risk detection to reduce mortality rates and the economic burden on the healthcare system. This study aims to implement and compare the performance of the Naive Bayes and Decision Tree algorithms in stroke risk classification. The dataset used contains 4,981 patient records with 10 stroke risk features, including age, gender, hypertension, heart disease, marital status, occupation type, residence type, average glucose level, body mass index, and smoking status. The research methodology encompasses data collection, preprocessing, splitting the data into training and testing sets at an 80:20 ratio (3,984 training records and 997 testing records), implementation of both algorithms, and model evaluation using accuracy, precision, recall, and F1-score metrics. The results indicate that the Decision Tree algorithm outperforms Naive Bayes, achieving an accuracy of 91.68%, precision of 91.91%, recall of 91.68%, and F1-score of 91.79%, while Naive Bayes achieved an accuracy of 86.06%, precision of 92.09%, recall of 86.06%, and F1-score of 88.72%. Confusion matrix analysis shows that Decision Tree has a lower misclassification rate, with an accuracy margin of 5.62% over Naive Bayes. This study concludes that decision trees are more effective for stroke risk classification. In practical terms, this model implies the provision of a decision-support tool for medical personnel to perform initial triage of high-risk patients prior to further clinical examination, thereby accelerating intervention and reducing the cost burden of stroke care. Theoretically, the findings enrich the literature on classification algorithm comparisons in the healthcare domain and serve as a baseline for the development of advanced models incorporating class imbalance handling and hyperparameter optimization in future research.**

Keywords—Stroke risk classification; Naive Bayes; Decision Tree; Comparative study; Disease prediction

I. INTRODUCTION

Stroke is one of the cerebrovascular diseases that has become a global health threat with significant impacts on morbidity and mortality. According to the World Stroke Organization report, stroke is the leading cause of disability and the second leading cause of death worldwide, with more than 12 million new cases annually [1]. In Indonesia, data from the Institute for Health Metrics and Evaluation (IHME) in 2019 identified stroke as the leading cause of death, accounting for 19.42% of total national mortality [2].

The high burden of stroke disease is further reflected in the results of the 2018 Basic Health Research (Riskesdas), which showed that stroke prevalence in Indonesia increased by 56%, from 7 per 1,000 population in 2013 to 10.9 per 1,000 population in 2018[3]. In terms of healthcare expenditure, stroke is one of the catastrophic diseases with the third-highest financing after heart disease and cancer, reaching IDR 5.2 trillion in 2023 [2]. Risk factors contributing to stroke incidence include hypertension, diabetes mellitus, dyslipidemia, obesity, and unhealthy lifestyles closely associated with dietary patterns and other comorbid conditions in the Indonesian adult population[3]. This situation underscores the need for more effective early detection and prevention efforts to reduce the public health and economic burden on the country.

Advances in information technology, particularly machine learning, have opened new opportunities in healthcare for disease prediction and classification. Machine learning is a branch of artificial intelligence that enables computer systems to learn from data and make predictions or decisions without being explicitly programmed[4]. In the medical domain, the application of machine learning can analyze patterns from complex clinical data to support healthcare decision-making, especially in conditions requiring early identification such as stroke [5], [6].

Several classification algorithms have been widely used in stroke disease prediction research, among which are Naive Bayes and Decision Tree. The Naive Bayes algorithm is a probability-based classification method that applies Bayes' theorem with the assumption of feature independence[7], while Decision Tree (including the C4.5/J48 variant) produces a decision tree structure that is easily interpretable by medical personnel[8]. Previous studies on stroke datasets have shown that Naive Bayes can achieve an accuracy of up to 92.48% [9], while other comparative studies indicate that Decision Tree outperforms Naive Bayes in stroke patient classification [10],[11]. This type of comparative approach has also been widely applied in other disease domains, such as heart disease and diabetes, with varying results depending on data characteristics[12], [13].

Although numerous studies on machine learning-based stroke classification have been conducted, several gaps still need to be addressed. First, most prior studies used foreign datasets whose population characteristics differ from those of Indonesia. Second, few studies have specifically compared the performance of Naive Bayes and Decision Tree using the same dataset with comprehensive evaluation metrics (accuracy, precision, recall, F1-score). Third, the handling of class imbalance, which is common in stroke datasets, has not been extensively discussed in similar studies, even though techniques such as SMOTE have been shown to significantly improve model performance [14], [15], [16].

This study aims to implement and compare the Naive Bayes and Decision Tree algorithms for stroke risk classification based on accuracy, precision, recall, and F1-score metrics in order to determine the algorithm with the best performance. Theoretically, this study is expected to contribute additional scientific references on the application of machine learning in healthcare and to advance the fields of data mining and public health[17]. Practically, the findings can provide recommendations for the best algorithm for an early stroke risk detection system, assist medical personnel in identifying high-risk patients, and support stroke prevention efforts through more accurate and efficient early detection. Several prior studies relevant to the topic of stroke classification using machine learning are described as follows.

Study 1: Conducted research on the application of data mining for stroke disease classification using the Naive Bayes algorithm. The dataset consisted of 10 stroke-related attributes including gender, age, hypertension history, heart disease history, marital status, occupation type, residence type, glucose level, body mass index, and smoking status. The study reported an accuracy of 92.48% [9].

Study 2: Compared the Naive Bayes and Decision Tree (J48) methods on stroke patients at Abdul Wahab Sjahranie Regional Hospital in Samarinda. The study utilized 7 risk factors: age, gender, blood pressure, diabetes mellitus, dyslipidemia, uric acid levels, and heart disease. The results showed that Decision Tree (J48) achieved a higher accuracy than Naive Bayes [10].

Study 3: Performed stroke prediction analysis by comparing three classification methods: Decision Tree, Naive Bayes, and Random Forest. The dataset consisted of 5,110 respondents with 12 attributes. The results indicated that Decision Tree achieved the highest accuracy of 95.13%, followed by Random Forest and Naive Bayes [11].

Study 4: Published in JATI (Journal of Informatics Engineering Students, ITN Malang), this study compared Naive Bayes and K-Nearest Neighbors algorithms for stroke disease classification using the same 10 attributes as this study. The results showed that Naive Bayes delivered competitive classification performance for stroke datasets with similar attribute schemas[18].

Study 5: Published in JOINTECS, this study developed a stroke prediction model based on the Random Forest algorithm augmented with SHAP interpretability. The study underscored the importance of model interpretability in the medical domain and confirmed that age, blood glucose level, and BMI are the primary contributors to stroke risk findings consistent with the features used in this stroke risk classification study [1], [5].

Study 6: Published in Jurnal Riset Informatika, this study evaluated the effect of data balancing techniques (SMOTE, oversampling, undersampling) on the performance of stroke prediction models. The study highlighted that class imbalance is a major challenge in stroke datasets, and that appropriate class imbalance handling can significantly improve the recall and F1-score of classification models [14].

Study 7: Published in Jurnal INSOLOGI, this study performed early stroke detection using various machine learning algorithms including Decision Tree, Naive Bayes, Random Forest, and Gradient Boosting. Results showed that Gradient Boosting achieved the highest True Positive and True Negative rates; however, Decision Tree remained a competitive baseline due to its interpretability for medical personnel [4].

Study 8: Compared the C4.5 (Decision Tree) and Naive Bayes algorithms in predicting cerebrovascular disease. The results showed that the C4.5 algorithm achieved an accuracy of 95.3%, while Naive Bayes achieved 91.3% [19].

Study 9: Published in the internationally indexed journal Stroke Research and Treatment, this study performed stroke risk prediction using a data-driven approach on a dataset of 5,110 patient records identical to those used in this study. Descriptive statistical analysis identified age, glucose level, BMI, hypertension, and heart disease as the most influential risk factors. Naive Bayes was noted to have the highest recall (0.404) for detecting stroke cases, albeit with a higher false positive rate compared to other algorithms such as XGBoost [6].

Study 10: Published in Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK, SINTA 2), this study optimized the Naive Bayes algorithm through K-Means discretization for heart disease diagnosis. The study demonstrated that preprocessing techniques such as discretization can significantly improve the accuracy of Naive Bayes, reinforcing the argument that appropriate preprocessing is key to the performance of probability-based classification models [12].

Study 11: Published in JTIIK (SINTA 2), this study analyzed the performance of Decision Tree, Naive Bayes, and K-Nearest Neighbor algorithms for classifying COVID-19 risk zones in Indonesia. Although the domain differs, this study is relevant as it provides a comparative overview of the performance of these three popular classification algorithms in the context of health risk in Indonesia, with Decision Tree and Naive Bayes standing out in terms of accuracy and computational efficiency [17].

Study 12: Published in Jurnal CoSciTech, this study compared classification algorithms (Logistic Regression, Random Forest, SVM, KNN) with the SMOTE technique for stroke disease classification using a dataset of 4,891 patient records. The results showed that SMOTE significantly improved model performance on imbalanced datasets such as stroke datasets [15].

Study 13: Published in the Journal of Cardiovascular Development and Disease (Q2 Scopus), this study applied five supervised algorithms (LR, RF, SVM, XGBoost, LightGBM) for stroke risk prediction on 7,389 participants from the Suita

Study. Random Forest emerged as the best-performing algorithm; however, the study also emphasized the importance of identifying risk clusters and key clinical variables that align with the features used in this study's dataset[20].

Study 14: Published in JUITA (Jurnal Informatika), this study compared data mining classification algorithms (C4.5, Naive Bayes, KNN) for stroke prediction using SMOTE Upsampling. The results showed that C4.5 with SMOTE was the best-performing algorithm, predicting stroke risk more accurately than models without SMOTE [16].

Study 15: Published in JNTETI (Jurnal Nasional Teknik Elektro dan Teknologi Informasi, UGM SINTA 2), this study performed early detection of diabetes using machine learning with the Logistic Regression algorithm. This study serves as an important methodological reference due to its application of a similar framework (preprocessing, SMOTE, confusion matrix evaluation) for early detection of chronic diseases that are risk factors for stroke [13].

From the review of prior studies, it can be concluded that both Naive Bayes and Decision Tree are effective algorithms for stroke classification, achieving high accuracy levels. However, the performance of each algorithm varies depending on dataset characteristics, preprocessing strategies, class imbalance handling, and the parameters used.

Based on the prior studies reviewed, the following research gaps can be identified: (1) Performance variation there is variation in performance between Naive Bayes and Decision Tree across different studies, indicating the need for further investigation with consistent datasets and preprocessing strategies; (2) Class imbalance handling most prior studies have not explicitly addressed class imbalance, even though it can significantly affect model performance [14], [15], [16]; (3) Comprehensive evaluation metrics some studies report only accuracy as an evaluation metric without considering precision, recall, and F1-score, which are critical in class imbalance scenarios; (4) Hyperparameter optimization few studies document the hyperparameter optimization process for achieving best performance; (5) Model validation not all studies apply cross-validation to ensure the stability and generalizability of the model on new data.

II. METHOD

A. Research Framework

This study employs a quantitative approach with an experimental method to compare the performance of the Naive Bayes and Decision Tree algorithms in stroke risk classification. The experimental approach was chosen because it enables objective measurement of both algorithms' performance through the same evaluation metrics on an identical dataset, ensuring that comparisons are fair and reproducible. The research framework is systematically structured through six main stages, as illustrated in Figure 1, ranging from data collection to analysis and result comparison. Each stage is designed to minimize bias and ensure consistency in the process between the Naive Bayes and Decision Tree models.

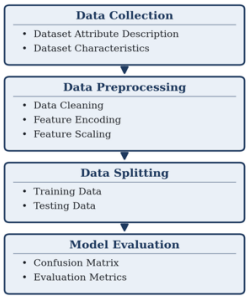

Fig. 1. Research Framework

B. Data Collection

This study uses a stroke dataset obtained from the Kaggle platform [13], containing patient health information relevant to stroke risk identification. The dataset was selected because it includes clinical features consistent with the stroke risk factors identified in the Introduction, namely age, hypertension, heart disease, glucose level, BMI, and smoking status. The dataset had already undergone a cleaning process with no missing values and consists of 4,981 records with 11 columns (10 input features and 1 target). This sample size is sufficiently adequate for an 80:20 training-testing split while still providing representative data for the classification model training process.

1) Dataset Attribute Description

The dataset consists of 10 input attributes and 1 target attribute with a diverse composition of data types: three continuous numerical attributes (age, avg_glucose_level, bmi), two binary attributes (hypertension, heart_disease), and several nominal categorical attributes (gender, ever_married, work_type, Residence_type, smoking_status). This diversity in data types requires different preprocessing strategies, namely encoding for categorical attributes and normalization for numerical attributes. A complete description of the attributes is presented in Table I.

TABLE I. STROKE DATASET ATTRIBUTE DESCRIPTION

No	Attribute	Data Type	Description	Value/Range
1	gender	Categorical	Patient's gender	Male, Female
2	age	Numerical	Patient's age	0.08 – 82 years
3	hypertension	Binary	History of hypertension	0 = No, 1 = Yes
4	heart_disease	Binary	History of heart disease	0 = No, 1 = Yes
5	ever_married	Categorical	Marital status	Yes, No
6	work_type	Categorical	Occupation type	children, Govt_job, Never_worked, Private, Self-employed
7	Residence_type	Categorical	Residence type	Urban, Rural
8	avg_glucose_level	Numerical	Average glucose level	mg/dL
9	bmi	Numerical	Body Mass Index	kg/m²

No	Attribute	Data Type	Description	Value/Range
10	smoking_status	Categorical	Smoking status	formerly smoked, never smoked, smokes, Unknown
11	stroke (target)	Binary	Stroke status	0 = No, 1 = Yes

2) Dataset Characteristics

The dataset characteristics identified after initial exploration are as follows: a total of 4,981 records consisting of 10 input attributes and 1 target attribute; no missing values after the cleaning process; and a significant class imbalance, with 4,733 patients (95.02%) in the non-stroke class and 248 patients (4.98%) in the stroke class.

This class imbalance with a 19:1 ratio is a critical finding that has a major impact on the modeling strategy and result interpretation. Classification models risk being biased toward the majority class (non-stroke), such that even when overall accuracy is high, the model's ability to detect the minority class (stroke) may be very low a phenomenon known as the accuracy paradox. This characteristic is consistent with other medical datasets measuring low-prevalence diseases and represents one of the research gaps identified in the Literature Review.

C. Data Preprocessing

Data preprocessing is a critical stage that determines the quality of the input to the classification model. Since the dataset was already in a clean condition with no missing values or duplicates, the preprocessing stage in this study focused on data type transformation and normalization of the numerical value ranges to ensure compatibility with both algorithms being compared.

1) Data Cleaning

An initial inspection of the dataset confirmed no missing values and no duplicate records. Therefore, the data cleaning stage required no imputation or record deletion operations, allowing all 4,981 records to be used directly in the encoding and normalization stages. This also avoids any potential bias that could arise from the imputation of missing values.

a) Feature Encoding

Categorical attributes were converted to numerical values so that they could be processed by the classification algorithms. Label Encoding was applied with the following mappings:

- gender: Male = 1, Female = 0
- ever_married: Yes = 1, No = 0
- Residence_type: Urban = 1, Rural = 0
- work_type and smoking_status: converted using multi-class label encoding with numerical ordering (children=0, Govt_job=1, Never_worked=2, Private=3, Self-employed=4 for work_type; formerly smoked=0, never smoked=1, smokes=2, Unknown=3 for smoking_status).

Label Encoding (rather than One-Hot Encoding) was chosen to keep the feature dimensionality low, given that the relatively small dataset (4,981 records) is more susceptible to the curse of dimensionality if features are expanded using One-Hot Encoding.

b) Feature Scaling

Normalization of numerical data was performed to standardize the value ranges so that all features contribute equally to the classification process. The numerical features normalized include age (range 0.08–82 years), avg_glucose_level (mg/dL scale, typically 55–272 mg/dL), and bmi (kg/m² scale, typically 10–98 kg/m²). Without normalization, features with larger value ranges (such as glucose) can dominate the classification process and reduce sensitivity to smaller-scale features. The Min-Max Scaling technique was applied so that all values fall within the range [0, 1].

D. Data Splitting

The dataset was divided into two subsets using a random split approach with an 80:20 ratio:

1) Training Data

80% of the total data (3,984 records) used to train the classification models.

2) Testing Data

20% of the total data (997 records) used to evaluate the generalization performance of the models on new data.

The 80:20 ratio was chosen as it is a common practice in machine learning research for datasets of several thousand records, providing sufficient training data for model learning while retaining a representative testing set. The random_state was set to a fixed value (seed=42) to ensure reproducibility of results.

E. Model Evaluation

Model performance evaluation was conducted on the testing data using a combination of confusion matrix analysis and four primary metrics. The selection of four metrics (Accuracy, Precision, Recall, F1-Score) was made to address the limitation of relying solely on accuracy in class imbalance scenarios, as identified in the research gaps in the Literature Review.

1) Confusion Matrix

The Confusion Matrix describes classification performance by comparing predictions against actual values, as shown in Table II.

TABLE II. CONFUSION MATRIX STRUCTURE

		Predicted Negative	**Predicted Positive**
Actual	Negative	TN (True Negative)	FP (False Positive)
	Positive	FN (False Negative)	TP (True Positive)

In the context of stroke classification, each cell has the following clinical interpretation: TN (True Negative) means a patient is correctly identified as not at risk of stroke; TP (True Positive) means a patient at stroke risk is successfully identified and can be intervened early; FN (False Negative) is the most

dangerous case because a patient at stroke risk is not detected and loses the opportunity for early intervention; while FP (False Positive) leads to unnecessary follow-up examinations but does not endanger the patient. In a medical context, minimizing FN is the primary priority even if it means accepting a slight increase in FP.

2) Evaluation Metrics

Four primary metrics were used to measure classification performance:

a) Accuracy

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

Measures the proportion of correct predictions out of all predictions. For balanced datasets, accuracy is a good indicator; however, for class-imbalanced datasets such as stroke datasets, accuracy can be misleading.

b) Precision

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

Measures how many positive predictions are actually correct. In the stroke context, high precision means that when the model predicts stroke, the prediction is generally correct.

c) Recall (Sensitivity)

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

Measures how many actual positive cases are successfully detected. Recall is the most critical metric in early stroke detection because false negatives can be fatal.

d) F1-Score

$$F1 - Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (4)$$

The harmonic mean of precision and recall, providing a balance between the two. It is suitable as a single metric in class imbalance scenarios.

III. Result And Discussion

A. Data Exploration

Prior to the classification process, an exploration of the dataset was conducted to understand the characteristics and distribution of the data. The stroke dataset used contains a total of 4,981 records with 10 predictive features and 1 target variable (stroke). This data exploration aims to identify patterns, anomalies, and statistical characteristics that may influence the choice of modeling strategy.

1) Target Class Distribution

Analysis of the target class distribution reveals a significant class imbalance, as shown in Table 3.

TABLE III. Target Class Distribution

Class	Count	Percentage
No Stroke (0)	4,733	95.02%
Stroke (1)	248	4.98%
Total	4,981	100.00%

This class imbalance with an approximately 19:1 ratio is a common characteristic in medical datasets, particularly for diseases with low prevalence such as stroke. This pattern is also consistent with the national stroke prevalence reported by Riskesdas 2018 of 10.9 per 1,000 population (approximately 1.09%), although the proportion of stroke patients in the Kaggle dataset is higher because the dataset is clinical in nature (not a general population sample). This class imbalance condition requires careful consideration when interpreting model results, as models tend to be biased toward the majority class. Consequently, overall accuracy alone is insufficient to evaluate model performance per-class metrics (especially recall for the minority class) become critically important.

B. Model Evaluation Results

Model evaluation was conducted on the testing data consisting of 997 records (20% of the total dataset), with a distribution of 947 non-stroke cases and 50 stroke cases. The following are the performance evaluation results for both algorithms.

1) Naive Bayes Performance

The Naive Bayes model (Gaussian Naive Bayes variant) yielded the overall performance shown in Table 4.

TABLE IV. Naive Bayes Evaluation Results

Metric	Value	Percentage
Accuracy	0.8606	86.06%
Precision	0.9209	92.09%
Recall	0.8606	86.06%
F1-Score	0.8872	88.72%

At the overall level, Naive Bayes achieved an accuracy of 86.06% with an F1-Score of 88.72%. This figure is lower than the result reported by Riany & Testiana (2023) [9], who achieved 92.48% on a similar dataset.

The evaluation of the Naive Bayes classifier yielded an accuracy of 86.06%, demonstrating that the model correctly classified the majority of instances in the dataset. The precision value of 92.09% indicates that the model exhibited a high positive predictive value, suggesting a low rate of false positives in the classification output. Meanwhile, the recall value of 86.06% reflects the model's capability to identify true positive instances, which is consistent with the overall accuracy obtained. The F1-Score of 88.72% confirms a well-balanced trade-off between precision and recall, indicating that the Naive Bayes algorithm performs robustly in this classification task. The relatively higher precision compared to recall suggests that the model demonstrates a conservative prediction tendency, which may be attributed to the probabilistic nature of the Naive Bayes approach in handling the feature distribution of the dataset.This difference is likely attributable to differences in class distribution, the number of records (4,981 vs. the smaller original dataset), and the preprocessing strategy employed. For a deeper analysis, the per-class classification report is presented in Table V.

TABLE V. NAIVE BAYES CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
No Stroke (0)	0.96	0.89	0.92	947
Stroke (1)	0.14	0.34	0.20	50
Accuracy			0.86	997
Macro Avg	0.55	0.61	0.56	997
Weighted Avg	0.92	0.86	0.89	997

From the Naive Bayes evaluation results, several important findings can be analyzed. The overall accuracy of 86.06% appears high on the surface; however, this figure is misleading as it is dominated by predictions of the majority class. Precision for the non-stroke class is very high (96%), but for the stroke class it is very low (14%) meaning that out of every 100 stroke predictions made by the model, only 14 are correct. Recall for the stroke class reaches 34%, indicating the model successfully detects approximately one-third of actual stroke cases. Although higher than Decision Tree, this figure is still far from ideal in a medical context. There is a clear trade-off between precision and recall for the minority class: to increase recall, Naive Bayes sacrifices precision, resulting in a high number of false positives. The Macro Average of 0.55–0.61 (far lower than the Weighted Average of 0.86–0.92) confirms that Naive Bayes has not been able to equally elevate the performance of the minority class.

2) Decision Tree Performance

The Decision Tree model (CART variant with Gini Index criterion) yielded the overall performance shown in Table 6.

TABLE VI. DECISION TREE EVALUATION RESULTS

Metric	Value	Percentage
Accuracy	0.9168	91.68%
Precision	0.9191	91.91%
Recall	0.9168	91.68%
F1-Score	0.9179	91.79%

Decision Tree achieved an accuracy of 91.68%, which is 5.62% higher than Naive Bayes. All four primary metrics fall within the range of 91.68%–91.91%, demonstrating better overall performance consistency.

The Decision Tree classifier demonstrated strong and consistent performance across all evaluation metrics. The model achieved an accuracy of 91.68%, indicating that it correctly classified over nine out of ten instances in the test dataset. The precision value of 91.91% and recall value of 91.68% were nearly identical, reflecting a highly balanced classification behavior with minimal divergence between false positive and false negative rates. Consequently, the F1-Score of 91.79% further confirms the model's robust and well-calibrated performance. Notably, the marginal gap between precision and recall (0.23%) suggests that the Decision Tree algorithm maintains a stable decision boundary, which is characteristic of tree-based models when applied to structured, feature-rich datasets. Compared to the Naive Bayes classifier, the Decision Tree model demonstrated superior performance in accuracy (+5.62%), recall (+5.62%), and F1-Score (+3.07%), while achieving a comparable precision value, thereby establishing Decision Tree as the more effective algorithm for this classification task. The per-class classification report for Decision Tree is presented in Table 7.

TABLE VII. DECISION TREE CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
No Stroke (0)	0.96	0.95	0.96	947
Stroke (1)	0.19	0.20	0.19	50
Accuracy			0.92	997
Macro Avg	0.57	0.58	0.58	997
Weighted Avg	0.92	0.92	0.92	997

Analysis of the Decision Tree results reveals the following findings. Overall accuracy reached 91.68%, which is 5.62% higher than Naive Bayes this is the primary finding of this study. Precision for the non-stroke class is very high (96%), while for the stroke class it increases to 19% (vs. 14% in Naive Bayes), meaning Decision Tree is more precise when predicting stroke. However, recall for the stroke class decreases to 20% (vs. 34% in Naive Bayes), revealing an opposing trade-off: Decision Tree is more precise but misses more actual stroke cases. Performance on the majority class (non-stroke) is excellent, with a recall of 95% and an F1-score of 96%, far more consistent than Naive Bayes.

These findings confirm that the choice of the best algorithm depends on priorities: if the priority is overall accuracy, Decision Tree is superior; if the priority is detecting as many stroke cases as possible (minority class recall), Naive Bayes is actually more appropriate.

C. Model Performance Comparison

A systematic comparison of the performance of both algorithms is presented in Table 8.

TABLE VIII. COMPARATIVE PERFORMANCE OF NAIVE BAYES AND DECISION TREE

Metric	Naive Bayes	Decision Tree	Difference	Best Model
Accuracy	86.06%	91.68%	+5.62%	Decision Tree
Precision	92.09%	91.91%	−0.18%	Naive Bayes
Recall	86.06%	91.68%	+5.62%	Decision Tree
F1-Score	88.72%	91.79%	+3.07%	Decision Tree

Based on Table 8, Decision Tree outperforms in three out of four evaluation metrics (Accuracy, overall Recall, and F1-

Score), while Naive Bayes only marginally leads in Precision with a difference of 0.18%. The accuracy difference of 5.62% is statistically significant for a testing dataset of 997 records and demonstrates that Decision Tree is generally more effective in classifying stroke risk on this dataset. This finding is consistent with the results of Aulia et al. (2024) [11] and Lishania et al. (2021) [10], which also found Decision Tree to be superior to Naive Bayes on stroke datasets.

However, it is important to note that the very small Precision margin (0.18%) indicates that both algorithms actually have comparable ability to distinguish true positive cases. The key differentiator lies in how each algorithm manages false negatives and false positives, as will be analyzed in the Confusion Matrix section below.

D. Confusion Matrix Analysis

1) Naive Bayes Confusion Matrix

The Naive Bayes confusion matrix on the testing data (997 records) is presented in Table 9.

TABLE IX. NAIVE BAYES CONFUSION MATRIX

		Predicted: No Stroke	**Predicted: Stroke**	
Actual	No Stroke	843 (TN)	104 (FP)	**Actual**
	Stroke	33 (FN)	17 (TP)	

Analysis of the Naive Bayes confusion matrix reveals four important findings:

a) True Negative (TN)

843 out of 947 non-stroke cases (89% recall for the majority class) the model performs reasonably well in identifying non-at-risk patients.

b) True Positive (TP)

17 out of 50 stroke cases (34% recall for the minority class) the model successfully detects approximately one-third of actual stroke cases.

c) False Positive (FP)

104 cases 104 patients who are actually non-stroke are predicted as stroke; this clinically leads to unnecessary follow-up examinations but does not endanger patients.

d) False Negative (FN)

33 cases 33 patients who are actually at risk of stroke are not detected; this is a serious concern in a medical context as these patients lose the opportunity for early intervention.

Clinically, the combination of FN=33 and FP=104 indicates that Naive Bayes tends to be aggressive in predicting stroke (many false alarms), which may actually be beneficial as an initial screening tool over-prediction is preferable to missing actual stroke cases.

2) Decision Tree Confusion Matrix

The Decision Tree confusion matrix is presented in Table X.

TABLE X. DECISION TREE CONFUSION MATRIX

		Predicted: No Stroke	**Predicted: Stroke**	
Actual	No Stroke	900 (TN)	47 (FP)	**Actual**
	Stroke	40 (FN)	10 (TP)	

Analysis of the Decision Tree confusion matrix reveals:

a) True Negative (TN)

900 out of 947 non-stroke cases (95% recall for the majority class) significantly better than Naive Bayes (89%).

b) True Positive (TP)

10 out of 50 stroke cases (20% recall for the minority class) the model detects only one-fifth of actual stroke cases.

c) False Positive (FP)

47 cases fewer than Naive Bayes (104), meaning fewer unnecessary follow-up examinations.

d) False Negative (FN)

40 cases higher than Naive Bayes (33), meaning more stroke patients are missed by the detection system.

The differing strategies of the two models become clear from this confusion matrix: Decision Tree is more conservative and accurate overall, while Naive Bayes is more sensitive to the minority class. In the context of early stroke detection, this trade-off must be considered in accordance with the clinical priorities of the healthcare facility: hospitals with limited resources may prefer Decision Tree (fewer false alarms), while primary care clinics performing initial screening may prefer Naive Bayes (higher stroke recall).

E. Discussion

1) Interpretation of Results

The results of this study indicate that Decision Tree demonstrates superior performance over Naive Bayes in stroke risk classification, with an accuracy margin of 5.62%. Several theoretical factors that explain these results are as follows:

a) Handling Non-Linear Relationships.

Decision Tree is capable of capturing non-linear relationships and complex interactions between features without requiring data distribution assumptions. In the case of stroke, risk factors such as age, hypertension, and glucose level exhibit complex interactions (e.g., the effect of hypertension on stroke risk differs between younger and elderly age groups). Decision Tree captures such interactions naturally through its branching structure, whereas Naive Bayes disregards inter-feature interactions. This finding is consistent with Aulia et al. (2024) [11] and Vu et al. (2024) [20], who found that tree-based algorithms outperform on stroke datasets.

b) Frequently Violated Independence Assumption.

Naive Bayes assumes feature independence; however, in medical reality, many stroke risk factors are mutually correlated. For instance, age correlates positively with hypertension and heart disease; BMI correlates with glucose level and hypertension. This violation of the independence assumption reduces the accuracy of Naive Bayes because the calculated

probabilities become inaccurate. This is consistent with Wickramasinghe & Kalutarage's (2021) [7] critique regarding Naive Bayes's vulnerability on data with inter-feature correlations.

c) Handling Class Imbalance.

Although both algorithms face the challenge of class imbalance (95.02% vs. 4.98%), Decision Tree with its decision tree structure is more flexible in adjusting the decision boundary for the minority class, particularly when using split criteria based on Gini or Information Gain that are sensitive to class distribution. However, the still-low minority class recall (20%) indicates that both algorithms still require explicit class imbalance handling techniques such as SMOTE, as demonstrated by Pertiwi et al. (2023) [16] and Handayani & Taufiq (2024) [15].

d) Interpretability for Medical Personnel.

Decision Tree produces a decision tree structure that can be visualized and directly interpreted by medical personnel, enabling clinical domain knowledge validation against the decision rules learned by the model. Naive Bayes, while fast and lightweight, generates probabilistic output without an explicit structure that can be validated clinically. This interpretability becomes an important added value for implementation in healthcare facilities, as emphasized by Jatiprasetya & Utami (2024) [1].

2) Minority Class (Stroke) Performance Analysis

Both algorithms faced significant challenges in classifying the minority class (stroke), as reflected in the per-class metrics:

Naive Bayes on Stroke class: Precision 14%, Recall 34%, F1-Score 20%.

Decision Tree on Stroke class: Precision 19%, Recall 20%, F1-Score 19%.

The low performance on the stroke class is attributable to three main factors:

a) Severe Class Imbalance.

The 95:5 (or 19:1) ratio causes the model to be biased toward the majority class. In the training set, the model encounters far more examples of the non-stroke class, causing the decision boundary to shift in favor of the majority class. Indra et al. (2024) [14] affirm that at such class imbalance ratios, balancing techniques such as SMOTE, oversampling, or cost-sensitive learning become mandatory for more equitable outcomes.

b) Limited Positive Samples.

Only 248 stroke cases exist in the full dataset (approximately 198 in the training set) this number is too small for the model to capture the variation in characteristics of stroke patients. Stroke patients span a wide spectrum (ischemic vs. hemorrhagic, mild vs. severe, with varying comorbidities), and 198 samples are insufficient to cover this spectrum representatively.

c) Feature Overlap.

The characteristics of stroke and non-stroke patients likely overlap across several features for example, many hypertensive patients do not experience stroke, and some patients without hypertension do have strokes. Without strongly discriminative features (such as CT-scan results, cholesterol levels, or family history), the model struggles to differentiate between the two classes. Sutcu et al. (2025) [6] demonstrated that clinical feature enrichment (feature engineering) can significantly improve minority class recall.

3) Comparison with Prior Studies

The results of this study contribute to the literature on stroke classification algorithm comparisons in Indonesia. The Decision Tree accuracy of 91.68% falls within a range consistent with prior studies Lishania et al. (2021) [10] found Decision Tree (J48) to outperform Naive Bayes on the RSUD Samarinda dataset; Aulia et al. (2024) [11] obtained a Decision Tree accuracy of 95.13% on a dataset of 5,110 respondents; and Kohsasiah et al. (2022) [19] recorded C4.5 achieving 95.3% on cerebrovascular disease. This variation in results confirms that algorithm performance is sensitive to dataset characteristics, preprocessing strategy, and the number of features.

The unique contribution of this study is the confusion matrix analysis revealing the trade-off between recall and minority class precision a dimension rarely analyzed in prior studies, which typically report only overall accuracy.

IV. CONCLUSION

Based on the results of stroke risk classification using the Naive Bayes and Decision Tree algorithms, the following conclusions are drawn:

The Decision Tree algorithm demonstrates better performance than Naive Bayes in stroke risk classification, achieving an accuracy of 91.68%, which is 5.62% higher than Naive Bayes (86.06%). This advantage is further supported by a higher F1-Score (91.79% vs. 88.72%), indicating that Decision Tree achieves a better balance between precision and recall overall.

Decision Tree outperforms in three out of four primary evaluation metrics: Accuracy (91.68%), overall Recall (91.68%), and F1-Score (91.79%), while Naive Bayes only marginally leads in Precision (92.09% vs. 91.91%) with a difference of 0.18%, which is not practically significant.

Both algorithms face significant challenges in classifying the minority class (stroke), with low recall values Naive Bayes at 34% and Decision Tree at 20%. The primary causes are severe class imbalance (95.02% vs. 4.98%), limited stroke samples (248 out of 4,981), and feature overlap between classes. These findings address the research gap identified in the Literature Review regarding the importance of explicit class imbalance handling.

Decision Tree is more effective at capturing non-linear relationships and interactions between stroke risk features (age, hypertension, glucose, BMI), while Naive Bayes is constrained by violations of the feature independence assumption that commonly occur in medical data. This is consistent with both theory and prior research.

Confusion matrix analysis reveals an important clinical trade-off: Naive Bayes has a higher stroke recall (34% vs. 20%), making it more suitable for initial screening that prioritizes detecting as many cases as possible, whereas Decision Tree is more accurate overall, making it more appropriate for resource-

limited environments seeking to minimize unnecessary follow-up examinations.

Decision Tree is recommended as the foundation for developing an early stroke risk detection system in healthcare facilities due to its superior overall performance and high interpretability for medical personnel. However, for practical implementation, the model requires further refinement through class imbalance handling techniques (such as SMOTE) and hyperparameter optimization to improve stroke class recall.

REFERENCES

[1] Harumawan Jatiprasetya and E. Utami, "Pembelajaran Mesin Untuk Prediksi Stroke Berdasarkan Random Forest Dan SHAP," *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 9, no. 3, pp. 135–148, Aug. 2025, doi: 10.31328/jointecs.v9i3.7158.

[2] Kementerian Kesehatan RI, "World Stroke Day 2023: Greater than stroke — Kenali dan kendalikan stroke. ," *Direktorat Jenderal Pelayanan Kesehatan*, Jakarta, Oct. 29, 2023.

[3] A. Syauqy, L. R. Wiragapa, M. Y. E. Soekatri, F. Ernawati, C. Nissa, and F. F. Dieny, "Hubungan Antara Pola Makan Dan Kondisi Penyerta Dengan Prevalensi Strok Pada Usia Dewasa Di Indonesia: Analisis Data Riskesdas 2018," *Gizi Indonesia*, vol. 46, no. 1, pp. 121–132, Mar. 2023, doi: 10.36457/gizindo.v46i1.785.

[4] Kevinda Sari, Muhammad Fadli, M. F. Fudholi, and E. R. Susanto, "Deteksi Dini Stroke Menggunakan Machine Learning," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 4, no. 4, pp. 706–720, Aug. 2025, doi: 10.55123/insologi.v4i4.5590.

[5] M. Mirafuddin and S. Supriyatna, "Penerapan Machine Learning pada Aplikasi Kesehatan HEALTIME," *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 7, no. 6, pp. 1601–1606, Dec. 2024, doi: 10.32672/jnkti.v7i6.8233.

[6] M. Sutcu, D. Jouda, B. Yildiz, J. Katrib, and K. M. Almustafa, "Predicting Stroke Risk Using Machine Learning: A Data-Driven Approach to Early Detection and Prevention," *Stroke Res. Treat.*, vol. 2025, no. 1, Jan. 2025, doi: 10.1155/srat/2892726.

[7] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft comput.*, vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[8] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[9] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *Jurnal SAINTEKOM*, vol. 13, no. 1, pp. 42–54, Mar. 2023, doi: 10.33020/saintekom.v13i1.352.

[10] A. Wahab, S. Samarinda, I. Lishania, R. Goejantoro, and Y. N. Nasution, "Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Comparison of the Classification for Naive Bayes Method and the Decision Tree Algorithm (J48) for Stroke Patients in Abdul Wahab Sjahranie Samarinda Hospital," *Jurnal EKSPONENSIAL*, vol. 10, no. 2, 2021.

[11] Y. Aulia, A. Andriyansyah, S. Suharjito, and S. W. Nensi, "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest," *Jurnal Ilmu Komputer dan Informatika*, vol. 3, no. 2, pp. 89–98, Jan. 2024, doi: 10.54082/jiki.90.

[12] N. Fajriati and B. Prasetiyo, "Optimasi Algoritma Naive Bayes dengan Diskritisasi K-Means pada Diagnosis Penyakit Jantung," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 3, pp. 503–512, Jul. 2023, doi: 10.25126/jtiik.2023106510.

[13] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, May 2022, doi: 10.22146/jnteti.v11i2.3586.

[14] Muhamad Indra, Siti Ernawati, and Ilham Maulana, "Machine Learning for Stroke Prediction: Evaluating the Effectiveness of Data Balancing Approaches," *Jurnal Riset Informatika*, vol. 6, no. 4, pp. 211–222, Sep. 2024, doi: 10.34288/jri.v6i4.344.

[15] Fitri Handayani and Reny Medikawati Taufiq, "Komparasi Algoritma Menggunakan Teknik Smote Dalam Melakukan Klasifikasi Penyakit Stroke Otak," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 5, no. 2, pp. 367–372, Aug. 2024, doi: 10.37859/coscitech.v5i2.7439.

[16] R. Sebastian and C. Juliane, "Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method," *JUITA : Jurnal Informatika*, vol. 11, no. 2, p. 311, Nov. 2023, doi: 10.30595/juita.v11i2.17348.

[17] A. Ainurrohmah and D. T. Wiyanti, "Analisis Performa Algoritma Decision Tree, Naive Bayes, K-Nearest Neighbor untuk Klasifikasi Zona Daerah Risiko Covid-19 di Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 1, pp. 115–122, Feb. 2023, doi: 10.25126/jtiik.2023105935.

[18] K. Akmal, A. Faqih, and F. Dikananda, "Perbandingan Metode Algoritma Naïve Bayes Dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 470–477, Mar. 2023, doi: 10.36040/jati.v7i1.6367.

[19] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *Jurnal Informatika*, vol. 9, no. 1, pp. 13–17, Apr. 2022, doi: 10.31294/inf.v9i1.11931.

[20] T. Vu *et al.*, "Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 7, p. 207, Jul. 2024, doi: 10.3390/jcdd11070207.