International journal of science, engineering, and information technology
Volume 06, Issue 02, July 2022
Journal homepage: https://journal.trunojoyo.ac.id/ijseit

THE IMPACT OF FEATURE SELECTION ON THE PROBABILISTIC MODEL ON ARRHYTHMIA DIAGNOSIS

Mula'ab [a], Husni [a], Mohammad Syarief [a], Bain Khusnul Khotimah [a], Dwi Kuswanto [a]

[a]*Universitas Trunojoyo Madura, Indonesia*

A B S T R A C T

Arrhythmia is a type of cardiac illness identified by an irregular heart rhythm that can be either too rapid or too slow. An electrocardiograph method is required to diagnose arrhythmia. Electrocardiogram (ECG) is the result of this Electrocardiograph process. The ECG is then utilized as a diagnostic tool for arrhythmia. Because the ECG data is so extensive, an adequate processing procedure is required. Understanding the ECG data can be done in various ways, one of which is classification. Naïve Bayes is a classification technique that can handle enormous amounts of data. ECG data has a lot of characteristics, which makes classification more difficult. Feature selection can be used to eliminate non-essential features from a dataset. This research aimed to determine the feature selection's impact on the Naïve Bayes classification. It was proven by increased accuracy, precision, recall, and f-measure by 4%, 0.13, 0.13, and 0.14, respectively. The computation time was 0.03 seconds faster. The highest performance was obtained by classification with 80 features. The accuracy was 93%, precision and recall were 0.45, f-measure was 0.42, and computation time was 0.10 seconds.

Keywords: Arrhythmia, electrocardiogram (ECG), classification, Naïve Bayes, feature selection.

Article History
Received 01 June 22 Received in revised form 01 July 22 Accepted 19 July 22

1. Introduction

Arrhythmia is heart disease characterized by an abnormal heart rhythm, becoming too fast or too slow. Diagnosing Arrhythmia cannot be done only by physical examination because some arrhythmia types do not have symptoms the patient feels. To diagnose an arrhythmia, an Electrocardiograph procedure is needed. An electrocardiograph records the heart's electrical activity by attaching leads (electrodes) to the patient's chest, arms, and legs. This Electrocardiograph will detect changes in each heartbeat's depolarization and repolarization pattern. The recording is in the form of an electrocardiogram or ECG. This ECG is then used as a reference for diagnosing an arrhythmia.

To diagnose an arrhythmia, an analysis of ECG data is needed to obtain a diagnosis by determining the actual patient's condition. Analyzing ECG data is not easy. The vast amount of ECG datasets becomes a hindrance in the analysis process.

Classification is one of the approaches that can be used to analyze ECG data. The analysis results can be used as a reference for diagnosing Arrhythmia based on the result of the ECG. One classification algorithm that is widely used is Naïve Bayes. This algorithm uses probability and statistics based on the Bayes theorem.

This study used the Naïve Bayes algorithm to classify arrhythmia data. The selection of Naïve Bayes is based on several studies that have been conducted before. Previously, there had been a lot of research using the Naïve Bayes algorithm. Some advantages of Naïve Bayes are: it has better performance rather than some other algorithms [1] [2], it does not require extensive data to conduct the training process [3], and Naïve Bayes has a high level of accuracy and speed when it is applied to large amounts of data [4]. Therefore, this study used Naïve Bayes as a classification algorithm due to its performance.

A large amount of data contains many features; either they are relevant, irrelevant, or redundant features. Ignoring the irrelevant and redundant features will confuse the data classification process. Thus, it will reduce the speed of classification, increase computational costs and memory usage, and significantly influence the classification result [4] [5]. The number of features can cause overfitting on the model, causing a decrease in the model performance. Therefore, preprocessing stages were needed to select the relevant parts.

Feature selection is a technique for selecting relevant features based on specific criteria. Thus, it can improve training performance, such as increasing classification accuracy, reducing computational costs, and making better model interpretation [6]. Generally, feature selection has three models: wrapper, filter, and embedded. The filter feature selection has some advantages. They have lower computation time than other types, are simple and fast, quickly measure high dimensional data, and are independent of the classification algorithm. One of the filter feature selection algorithms is Information Gain. This algorithm can handle the feature selection process quickly [7]. Also, it is more effective in removing

* *Corresponding author*.
E-mail address:syarief@trunojoyo.ac.id.

E-ISSN 2548-4214

features with excellent accuracy [8]. Due to the advantages described above, the feature selection used in this study was Information Gain.

Therefore, this study will discuss the effect of feature selection on the performance of the Naïve Bayes classification model in diagnosing an arrhythmia. The feature selection use would be expected to improve the performance of the Naïve Bayes classification model.

2. Literature Review

2.1. Electrocardiogram

Electrocardiography is a term in use in the cardiovascular field. Electrocardiography is used to examine and diagnose abnormalities in the heart by recording the electrical activity of the heart using leads placed on the chest, arms, and legs to detect changes in the depolarization pattern and repolarization of each heartbeat. The recording result is in the form of an electrocardiogram or commonly known as an ECG. ECG is used to diagnose specific heart disease types, such as arrhythmias.

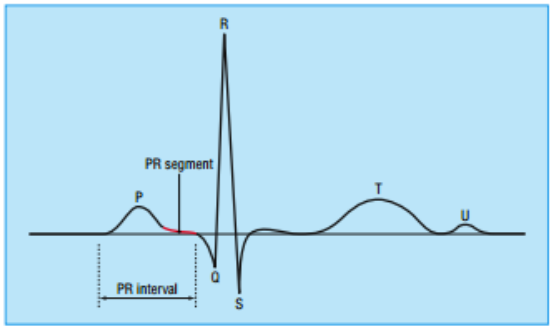

Figure 1. Illustration of ECG Waves

There are several types of ECG waves, as shown in Figure 1. The P-wave depicted atrial depolarization. The Q-wave is the beginning of ventricular depolarization. PR is the interval between the P wave's start and the RQS complex's start. The PR segment is the end of the P-wave until the beginning of the QRS complex. The QRS complex is the ventricular depolarization time interval. ST described the period of ventricular depolarization. T-wave describes ventricular repolarization [1].

2.2. Arrhythmia Dataset

The dataset used in this study was the arrhythmia dataset. Arrhythmia is a disorder of abnormal heart rhythm which causes the heart rate to be faster or slower than the usual rhythm. This dataset was downloaded from the Large Dataset, UCI Machine Learning Repository. This dataset had 279 features with 452 labeled data.

The arrhythmia dataset had 279 features. They were the result of ECG interpretation and recorded data from 452 patients. There were 73 features with nominal data types, and 206 other features with numeric data types. The dataset had 16 classes, namely normal class and several classes that refer to the arrhythmia types. The first Class had 245 data, and 185 other data were divided into different classes. There were three classes, including those which did not appear in the dataset. They were the 11th, 12th, and 13th classes. The class distributions in the dataset are shown in Table 1.

Table 1. Class Distribution in the Dataset

Class Code	Class	Amount of Data
1	*Normal*	245
2	*Ischemic changes (Coronary Artery Disease)*	44
3	*Old Anterior Myocardial Infarction*	15
4	*Old Inferior Myocardial Infarction*	15
5	*Sinus tachycardia*	13
6	*Sinus bradycardia*	25
7	*Ventricular Premature Contraction (PVC)*	3
8	*Supraventricular Premature Contraction*	2
9	*Left bundle branch block*	9
10	*Right bundle branch block*	50
11	*degree AtrioVentricular block*	0
12	*degree AV block*	0
13	*degree AV block*	0
14	*Left ventricular hypertrophy*	4
15	*Atrial Fibrillation or Flutter*	5
16	*Other classes*	22

2.3. Feature Selection

Feature selection is a technique for selecting relevant features based on specific criteria with the slightest possible elimination of information; thus, it could improve training performance such as increasing classification accuracy, decreasing computation costs and memory usage [2], and better model interpretation [3]. The number of features could cause overfitting in the model, which causes a decrease in the model's performance.

Feature selection can reduce the dimension of the feature, cut the required storage space, eliminate irrelevant and excessive data and noise in the data, speed up running time in the learning algorithm, enhance data quality and improve the accuracy of the resulting model [4].

Feature selection is divided into three techniques: filter, wrapper, and embedded [3] [4] [5].

2.3.1. Filter

Filter is distinct from the classification process; thus the feature selection process is not affected by the bias of the learning algorithm. Filter technique will sort features based on specific criteria; then the top features will be used in the classification process. It is a simple and fast technique to easily measure high-dimensional data. Some algorithms included in the filter technique are Relief, Fisher Score, and Information Gain.

2.3.2. Wrapper

Unlike filter, wrapper technique performs feature selection along with the classification process using accuracy estimation of the classification model.

This type of feature selection is not recommended to handle data with vast number of features. Compared to some filter techniques, the wrappers have higher computational costs and increase risk of overfitting.

2.3.3. Embedded

Embedded is a feature selection technique embedded in classification construction. This feature selection utilizes all features to train the classification model and removes less influential features with a coefficient close to 0. This technique has better computational costs compared to the wrapper technique. Decision Trees is an algorithm of the embedded method.

2.4. Information Gain

Information Gain (IG) is one of feature selection algorithms in the filter model. IG calculates the gain value of each feature and gives a score based on the gain value, then ranks the scores. The gain value shows how much influence a feature has on data classification. The higher gain value of a feature indicates a more relevant feature.

Information Gain uses the entropy concept to measure uncertainty of dataset features. The following is the equation for calculating entropy [6]:

$$Info(D) = \sum_{i=1}^{m} p_i log_2(p_i) \quad (1)$$

D : Set of cases
m : Number of classification class
p_i : Probability of feature i

Below is an equation to calculate the entropy of each feature:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info\ (D_j) \quad (2)$$

A : Attribute
|D| : Number of all data samples
|Dj| : Number of samples for j value
v : A possible value for attribute A

$$Gain\ (A) = |Info\ (D) - Info_A(D)| \quad (3)$$

2.5. Classification

Classification is a data analysis method to group data into appropriate classes in order to understand data more efficiently. Classification can be applied to various fields, such as marketing targets, manufacturing, diagnosis in medical field, etc. Classification has four fundamental components [7]:

1. Class is dependent categorical variable form that represents label as the classification result, such as customer loyalty, earthquake types, etc.
2. Predictors represent data characteristics or attributes such as blood pressure, season, marital status, wind speed and direction, etc.
3. Training dataset, a data set with class and predictors. This data is used for training the model to recognize the appropriate class based on the available predictors.
4. Testing dataset, new data that will be classified by the model that has been constructed.

The stages in data classification process involve a process of constructing a model (learning step) and application of the model (classification step) [6]. In the learning phase, the model is built based on data that has complete information, such as features or class labels. The training data is analyzed by the classification algorithm that has been constructed. While at the classification stage, the model will be used to determine class of the testing data. This step will calculate the accuracy rate of the classification algorithm based on the percentage of testing data that the model appropriately classifies.

Several things that are used as considerations in choosing a method in classification model are accuracy rate of the model in classifying data, speed in processing data, the reliability when it faces noises in the data, easy-to-understand model interpretation, and the simplicity of the model [7]. Some classification algorithms frequently used are *Decision Tree, Naïve Bayes, Neural Network, K-nearest Neighbor*, etc.

2.6. Naïve Bayes

One of the classification algorithms widely used is Naïve Bayes. It predicts future opportunities based on previous experience using probability and statistical methods as the Bayes theorem concept. Compared to other algorithms, some advantages of Naïve Bayes are easy to use, lower error rate, high accuracy rate, and fast if applied to extensive data because it does not require complex repetition scheme parameters [6]. Naïve Bayes with the naïve assumption can reduce computation time by multiplying the probabilities. Because of its simplicity, Naïve Bayes can handle datasets with many features.

Bayes theorem states that the occurrence probability of specific characteristic samples in C class (posterior probability) is the probability of C class (prior) multiplied by the probability of sample characteristics in C class (likelihood), divided by the probability of global sample characteristics (evidence). Below is the Bayes theorem equation:

$$P\ (C|X) = \frac{P(X|C)\ P(C)}{P(X)} \quad (4)$$

$$posterior = \frac{prior \times likelihood}{evidence} \quad (5)$$

X : Data with an unknown class
C : Data hypothesis
P(C|X) : Probability of hypothesis C based on the condition of X (posterior probability)
P(C) : Probability of hypothesis C (prior probability)
P(X|C) : Probability of X based on the condition of hypothesis C
P(X) : Probability of X (evidence)

For the classification with continuous data, the Gauss Density (Gaussian distribution) formula is used as follows:

$$P = (X_i = x_i \mid Y_j = y_j) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma^2 ij}} \quad (6)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (7)$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5} \quad (8)$$

P	:	Probability
Xi	:	I attribute
x_i	:	Value of attribute i
Yj	:	Class j
y_j	:	Sub Class j
μ	:	Mean, an average of all attributes
σ	:	Standard deviation

2.7. K-fold Cross-Validation

Cross-Validation is a technique used to assess the performance of a model or algorithm by partitioning data into training data and testing data. K-fold Cross Validation is one of the Cross Validation methods. It will divide data into K partitions. (K-1) partition is used as testing data, and the remaining is used as training data. Then the Cross-Validation process is repeated for K times with different test data [6].

2.8. Classification Evaluation

To find out the performance of a classification model, it is necessary to conduct a classification evaluation process. The evaluation method used in this study was Confusion Matrix as a measure of accuracy [7]. Table 2 describes the Confusion Matrix model.

Evaluation using Confusion Matrix will produce accuracy rate, precision, recall, and f-measure values. Confusion Matrix contains several cases that are correctly classified and incorrect ones.

Table 2. Confusion Matrix Model

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

Accuracy is the accurateness percentage of classification prediction results. The following describes how to calculate accuracy with a *confusion matrix* table:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (9)$$

Precision is case proportion predicted as positive, which is also true positive on the actual data. In other words, precision is the exactness level by dividing the number of relevant items selected with all selected items. The following is a precision calculation with a confusion matrix table:

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

Recall is the proportion of actual positive cases that are correctly predicted as positive. In other word, *recall* is success level (completeness) in finding relevant items by dividing the number of relevant items selected with the total number of relevant items available. The following is a *recall* calculation with a *confusion matrix* table:

$$Recall = \frac{TP}{(TP+FN)} \quad (11)$$

F-measure is used to evaluate classification performance which is a combination of precision and *recall*. Below is the equation to calculate *f-measure*:

$$F - measure = \frac{2 \times presisi \times recall}{presisi + recall} \quad (12)$$

TP (*True Positive*) : Positive prediction detected by the system that matches the actual state

TN (*True Negative*) : Negative prediction detected by the system that matches the actual state

FP (*False Positive*) : Positive prediction detected by the system but not in accordance with the actual state

FN (*False Negative*) : Negative prediction detected by the system but not in accordance with the actual state

3. Research Methodology

The system constructed in this study was a classification system with the *Naïve Bayes* method. The dataset used was a diagnosis of *Arrhythmia* with no *missing value* and is ready to use. In this system, a feature selection process was used to determine the effect of applying the feature selection on the performance of the *Naïve Bayes* classification. Several experimental scenarios will be carried out to determine the impact of feature selection on Naïve Bayes classification. They were the classification with and without feature selection with several different conditions. An evaluation process

would be carried out using the *Confusion Matrix*. The system design is shown in Figure 2 below.

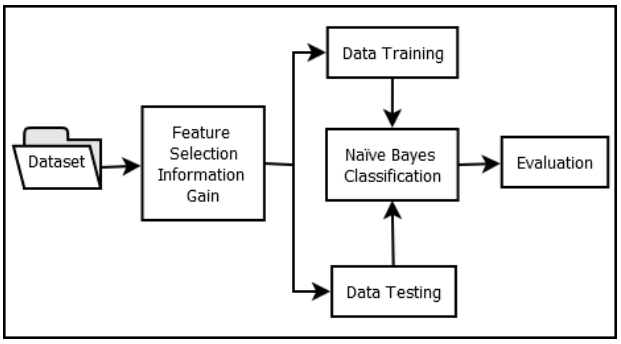

Figure 2. System design

3.1. Input Dataset

It was *the Arrhythmia* dataset, the diagnostic data based on patient's heart activities. The dataset was downloaded from the *Large Dataset, UCI Machine Learning Repository*. It consisted of 452 data, including 279 features such as age, gender, weight and height of the patient, heart rate and other patient data.

In the dataset, there were five features with *missing values*: 375 *missing values* for J-feature, 22 for P-feature, and 8 for T-feature, while QRST-feature and Heart, each had one *missing value*. In addition, there were 17 features with a single data value, i.e SS1, SSAVL, 'Exst Ragged R Wave AVL', 'Exst Ragged P Wave AVF', 'Exst Ragged P WaveV4', 'Exst disphasic P WaveV4', SSV5, 'Exst Ragged R WaveV5', 'Exst Ragged P WaveV5', 'Exst Ragged T WaveV5', SSV6, 'Exst disphasic P WaveV6', 'Exst Ragged T WaveV6', 'SS wave 1', 'SS wave AVL', 'SS wave V5', 'SS wave V6'.

Data cleaning was performed by filling the *missing value* with the average value of each feature: P, T, QRST, and Heart. Meanwhile, the feature J was omitted because there were many *missing values*. Data cleaning was undertaken by eliminating 17 features with a single value because they did not have variation. Therefore, the numbers of data and features used in this study were 452 data with 261 features.

3.2. Feature Selection

The input data will go through the feature selection process using the *Information Gain* (IG) method. In this step, the *gain* value of each feature will be calculated. It will be ranked; the greater the gain value of a feature shows how relevant the feature is to the classification process. The result of this feature selection process is relevant features that will be used in the classification process.

3.3. Classification

The next step is constructing a classification model using the *Naïve Bayes* method with *Gaussian* distribution. The construction of the model starts by calculating the prior of each class, the mean, and the standard deviation of each feature in each class. The mean and the standard deviation will be used to calculate the *likelihood* of each feature. Based on the *prior* and *likelihood* values. The value of *posterior* would be used as a standard classification.

3.4. Testing and Evaluation

To measure the method performance, then the test was carried out with several scenarios as follows:

1. 1st scenario, *Naïve Bayes* classification test was conducted with no feature selection.
2. 2nd scenario was the *Naïve Bayes* classification test with *Information Gain* feature selection. The *Information Gain* would be applied by using the ranking limit of certain features (n = 40, 80, 120, 160, 200, 240).

Each test above would produce accuracy, precision, *recall, f-measure* value, and computation time to assess the performance of the *Naïve Bayes* classification model.

The evaluation phase was performed by using *K-fold Cross-Validation* with k = 5 in each testing scenario. To find out the performance of the classification model, calculations of accuracy, precision, *recall, and f-measure* were performed using the *Confusion Matrix* and ROC (*Receiver Operating Characteristic*) curve. The performance of the two testing scenarios would be compared to find out the best scenario.

4. Finding and Discussion

As explained earlier that the test was carried out with two scenarios: classification without feature selection and with feature selection. The goal was to get the best procedure out of the two scenarios.

Table 3. Experiment Results of the 1st Testing Scenario

Scenario 1	Experiment Results				
	Accuracy	**Precision**	***Recall***	***F-measure***	**Time**
	86%	0.20	0.22	0.16	0.14 s

Tables 3 and 4 resulted from implementing the first and the second testing scenarios.

Table 4. Experiment Results of the 2nd Testing Scenario

Number of Features	Experiment Results				
	Accu-racy	**Preci-sion**	***Recall***	***F-measure***	**Time (second)**
n=40	91%	0.33	0.34	0.30	0.09
n=80	93%	0.45	0.45	0.42	0.10
n=120	93%	0.43	0.40	0.40	0.11
n=160	92%	0.33	0.39	0.34	0.12
n=200	87%	0.23	0.27	0.19	0.12
n=240	86%	0.21	0.25	0.17	0.13
Average	**90%**	**0.33**	**0.35**	**0.30**	**0.11**

In the second testing scenario, the experiments were conducted six times by applying the *Naïve Bayes* classification model and feature selection using different amounts of features: 40, 80, 120, 160, 200, and 240 features. From the series of experiments above, we got the average accuracy rate of 90%, while the average value of precision, recall, and f-measure obtained were 0.33, 0.35, and 0.30, respectively. The experiment was carried out by spending an average of 0.11 seconds.

A series of experiments in the second testing scenario produced a comparison graph of the accuracy rate, as shown in Figure 3.

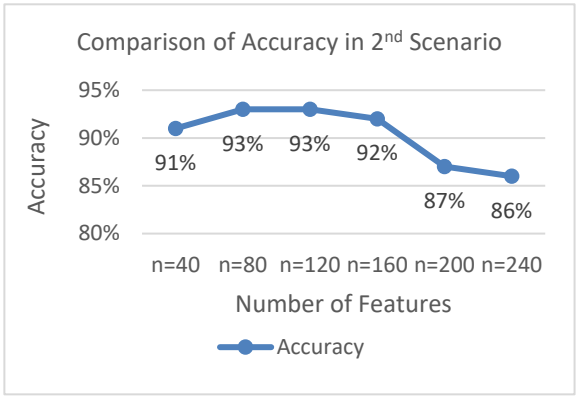

Figure 3. Comparison of Accuracy of 2nd Testing Scenario

As shown in Figure 3, the accuracy rates of the 2nd testing scenario series with 40, 80, 120, 160, 200, and 240 features were 91%, 93%, 93%, 92 %, 87%, and 86%, respectively. The highest accuracy rate (93%) was obtained in experiments with 80 and 120 features, while the lowest accuracy rate (86%) was gained in an experiment with 240 features.

Figure 4 describes a comparison graph of the precision, *recall*, and *f-measure* values of the 2nd testing scenario series.

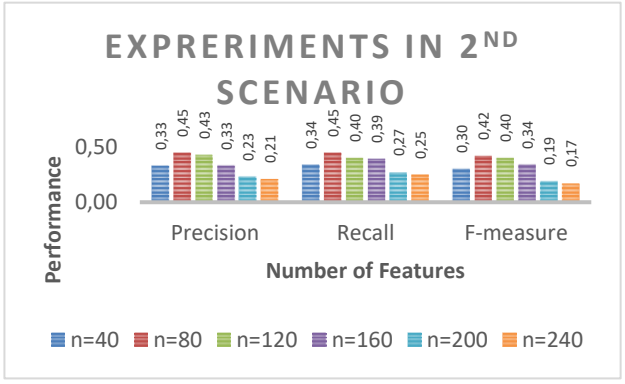

Figure 4. Comparison of Precision, Recall, and F-measure of 2nd Scenario experiments

The graph in Figure 4 clearly showed that the experiment with 80 features had the highest value of precision, *recall*, and *f-measure* compared to other experiments. In an experiment with 80 features, the precision and *recall* values were 0.45, and the *f-measure* value was 0.42.

In contradiction, the experiment with 240 features had the lowest value of precision, *recall*, and *f-measure* compared to other experiments. The precision, *recall*, and *f-measure* value gained in the experiment were 0.21, 0.25, and 0.17, respectively.

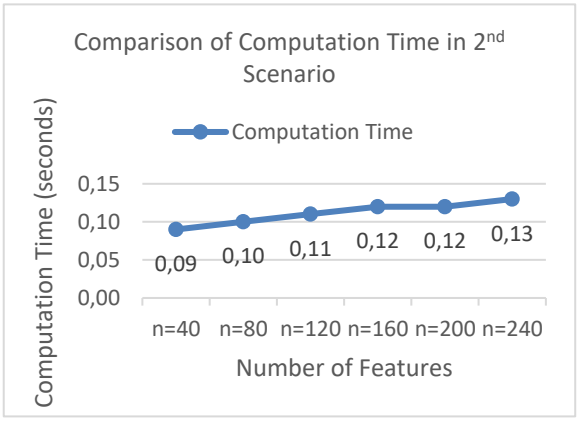

Figure 5. Comparison of Computation Time of 2nd Testing Scenario

Figure 5 describes a comparison graph of the computation time for each experiment of the 2nd Testing Scenario. In this scenario, each experiment required different execution time. Experiments with 40, 80, and 120 features took 0.09 seconds, 0.10 seconds, and 0.11 seconds, respectively, while experiments with 160 and 240 features needed 0.12 seconds.

Primarily, the computation time increased linearly. The more features used, the longer it took to complete the classification process.

Table 4. 3 Experiment Results Comparison

Scenario	Comparison of Experiment Results				
	Accuracy	Precision	*Recall*	*F-measure*	Time (seconds)
1	86 %	0.20	0.22	0.16	0.14
2	90 %	0.33	0.35	0.30	0.11

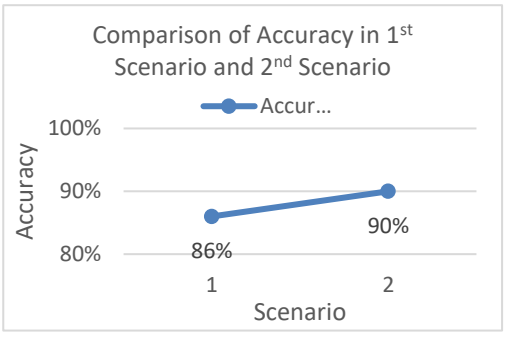

Figure 6. Comparison of Accuracy of 1st Scenario and 2nd Scenario

Table 4.3 compares the experiment results between 1st scenario and 2nd scenario. The 2nd scenario's result was the average value of six experiments using feature selection of several different features.

The average accuracy rate in the second scenario increased by 4% more than in the first scenario, as shown in Figure 6.

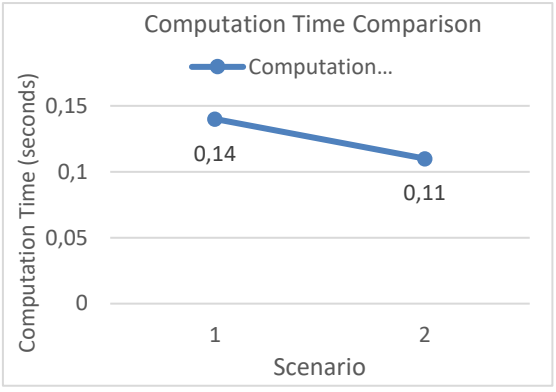

Figure 7. Comparison of Computation of Scenario 1 and Scenario 2

Figure 7 shows that the average time needed to classify data in the second scenario is 0.03 seconds shorter than in the first scenario.

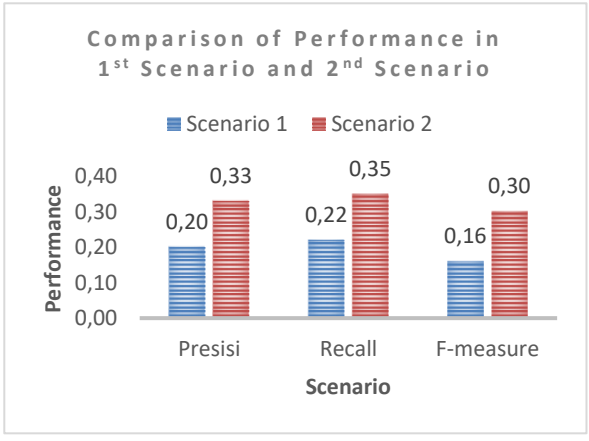

Figure 8. Comparison of Precision, Recall, and F-measure of Scenario 1 and Scenario 2

It can be seen clearly in Figure 8 that in the second scenario, precision, *recall*, and *f-measure* had increased significantly compared with the first scenario. Precision increased by 0.13, *recall* increased by 0.13, and *f-measure* increased by 0.14. Overall, Table 4.3 and Figure 8 show that 2nd scenario's result was better than 1st scenario's

According to the experiments' results above, feature selection implementation could increase the accuracy rate, precision, *recall*, and *f-measure*. The computation time needed for the classification process is also getting faster. Overall, it would improve the performance of the *Naïve Bayes* classification model.

5. Conclusion

From the results of the two scenarios carried out in this study, it can be concluded that feature selection influenced the performance of the *Naïve Bayes* classification model on the *Arrhythmia* diagnosis. The implementation of feature selection could increase accuracy rate by 4%, precision by 0.13, *recall* by 0.13, and *f-measure* by 0.14 while the computation time was 0.03 seconds faster. The highest performance is obtained by classification with 80 features. The accuracy was 93%, precision and recall were 0.45, f-measure was 0.42, and the computation time was 0.10 seconds.

REFERENCES

[1] F. Morris, J. Edhouse, W. J. Brady, and J. Camm, *ABC of Clinical Electrocardiography*. London: BMJ Books, 2003.
[2] W. N. Gansterer and G. F. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy," *Proc. Mach. Learn. Res.*, vol. 4, pp. 90–105, 2008.
[3] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification : A Review," *Semant. Sch.*, 2013.
[4] L. Ladha and T. Deepa, "Feature Selection Methods and Algorithms," *Int. J. Comput. Eng.*, vol. 3, no. 5, pp. 1787–1797, 2011.
[5] P. Larranaga and Y. Saeys, "Gene expression A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[6] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. Morgan Kaufman, Elsevier, 2012.
[7] F. Gorunescu, *Data Mining Concepts, Models and Techniques*, 12th ed. Springer-Verlag Berlin Heidelberg, 2011.