**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

# Diabetes Classification using Gain Ratio Feature Selection in Support Vector Machine Method

**Nabila Al Rasyid, Iis Afrianty*, Elvia Budianita, Siska Kurnia Gusti**

Faculty of Science and Technology, Informatics Engineering, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia

Email: [1]12150120329@students.uin-suska.ac.id, [2,*]iis.afrianty@uin-suska.ac.id, [3]elvia.budianita@uin-suska.ac.id, [4]siskakurniagusti@uin-suska.ac.id

Correspondence Author Email: iis.afrianty@uin-suska.ac.id

**Abstract**

Diabetes is a major cause of many chronic diseases such as visual impairment, stroke and kidney failure. Early detection especially in groups that have a high risk of developing diabetes needs to be done to prevent problems that have a wide impact. Indonesia is ranked seventh in the world with a prevalence of 10.7% of the total number of people with diabetes. This research aims to determine the attributes in the diabetes dataset that most affect the classification and apply the Support Vector Machine method for diabetes classification. For the determination process, Gain Ratio feature selection technique is applied. The dataset used consists of 768 data with 8 attributes. In this classification process, 3 SVM kernels (Linear, Polynomial, and RBF) are used with three possible data divisions using the ratio (70:30; 80:20; 90:10). Before applying feature selection, there were 8 attributes used and achieved the highest accuracy of 94.81% at a ratio of 80:20 using the RBF kernel with a combination of two parameters namely C = 100, Gamma = 3 and C = 100, Gamma = Scale. Feature selection parameters in the form of thresholds used include 0.02; 0.03; and 0.05. After applying feature selection, the attribute that produces the highest accuracy uses 6 attributes. The highest accuracy after applying feature selection reached 95.45% at a threshold of 0.02 with a ratio of 80:20 using the RBF kernel with parameters C = 100 and Gamma = Scale. The results showed that there was an increase in accuracy after applying feature selection.

**Keywords**: Data Mining; Diabetes; Feature Selection; Gain Ratio; Support Vector Machine

## 1. INTRODUCTION

Diabetes Mellitus is a chronic disease caused by elevated glucose levels in the blood due to the failure of the pancreas to produce adequate amounts of the hormone insulin [1], [2], [3]. Diabetes is the main cause of various chronic diseases such as visual impairment, heart disease, stroke, and kidney failure [2], [4]. According to [5] the United States with a percentage of 31%, India with a percentage of 77%, and China with a percentage of 116.4% are the three countries with the highest prevalence of diabetes in the world. Indonesia is ranked seventh with a prevalence of 10.7% in the number of people with diabetes [3], [5]. The high prevalence of diabetes in Indonesia, which is a developing country with a large population, makes it difficult for certain groups of people to consult with medical personnel for examination [6]. Early detection, especially in groups that have a high risk of developing diabetes, needs to be done to prevent problems that have a wide impact [2].

Data mining is a method of analyzing patterns and characteristics in large datasets to gather unexpected knowledge or information that is not yet owned [7], [8]. The results of data mining can be applied in the future to improve the quality of decision making [9]. In data mining there are various main functions, such as estimation, prediction, clustering, association and classification [8], [10]. Classification is a data analysis method to determine the class or category of data samples and find relationships or patterns between attributes contained in the data [11]. According to [12] the classification process has two steps including learning and classification. Learning (training phase) is the first stage, where the training data is analyzed by the classification algorithm that has been made until it can be applied to the form of classification rules. Next is the classification phase, where test data is used as an estimate of the accuracy of the classification rules. Applying classification to diseases based on medical history and symptoms can help speed up diagnosis to plan effective treatment [8].

Support Vector Machine (SVM) is one of the algorithms from machine learning techniques with a high level of accuracy in predicting the potential classification of data [1]. In research [13] with the Pima Indians Diabetes Dataset using the SVM method on the RBF kernel with a data ratio of 90: 10 resulted in an accuracy of 87%. Furthermore, research conducted by [14] on the Pima Indians Diabetes Dataset resulted in the highest accuracy in the benchmark model using a polynomial kernel with C = 100 and degree = 3 with 87% accuracy, while the highest accuracy in the scratch model using a polynomial kernel with C = 1, gamma = scale, and degree = 3 resulted in 78% accuracy. Another study conducted by [15] on skull bone data using SVM resulted in 91.3% at a ratio of 90:10 using the RBF kernel with C = 2, gamma = 'auto'. Another research on Pima Indians Diabetes Dataset with SVM method applying one of the feature selection techniques, namely forward selection by [1] resulted in a high increase in accuracy to reach 91% accuracy when using 20% test data and 80% training data.

Feature selection is a form of attribute reduction to improve data quality and enhance the performance of classification algorithms [10]. Feature selection can help algorithms process data faster because it helps select the most relevant attributes, so that irrelevant attributes will be reduced [8], [11]. According to [16], there is a need for a feature selection approach to select important features that are useful for the learning modeling process to improve its accuracy. One feature selection that has been proven to improve classification algorithms is the gain ratio [17], [18]. Gain Ratio is

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

one of the methods for selecting features that serves to determine the level of influence of an attribute on the target variable to be predicted [9], [19]. According to [20] the selected feature is determined by a value limit called threshold, which can be determined freely. According to [17], just like information gain, the gain ratio also requires determining the minimum limit to determine the features used by repeatedly testing the minimum limit. In research [21] used a threshold value of 0.01 to 0.1. Research conducted by [18] using feature selection gain ratio in the Naive Bayes method for heart disease resulted in an accuracy of 91.2% higher than the performance of Naive Bayes without feature selection which only resulted in 90.4%. From other research conducted by [9] on hypertension complications using feature selection gain ratio in the Naive Bayes method obtained an increase in accuracy of 20% which initially had an accuracy of 75% to 95%. In addition, research conducted by [22] on credit approval datasets obtained higher accuracy after applying feature selection gain ratio which initially only used C45 resulting in 94.12% to 95.29%.

Research on SVM methods with feature selection gain ratio has been done before by [23] for skull bone classification with a threshold of 0.01 resulting in an accuracy of 92.01%, while without feature selection only produces 91.39%. In research [24] for sentiment analysis using feature selection gain ratio in SVM method can increase accuracy compared to before using feature selection. The results showed the use of 1732 attributes with a threshold weight of less than 0.0001 increased the accuracy of 61.63% to 71.51%. While the use of 518 attributes with a threshold weight of less than 0.002 increases the accuracy of 61.63% to 62.79%.

Based on previous research, the gain ratio feature selection technique and SVM method have proven effective in various studies. Therefore, this study applies a combination of gain ratio to Support Vector Machine in diabetes classification. The purpose of this research is to improve the performance of the prediction model to better predict the risk of diabetes.

## 2. RESEARCH METHODOLOGY

The research method consists of several processes including problem identification, literature study, diabetes data collection, data preprocessing, data transformation using min max normalization, feature selection using gain ratio, classification using SVM, evaluation using confusion matrix, and conclusion. The research flow is shown in Figure 1.

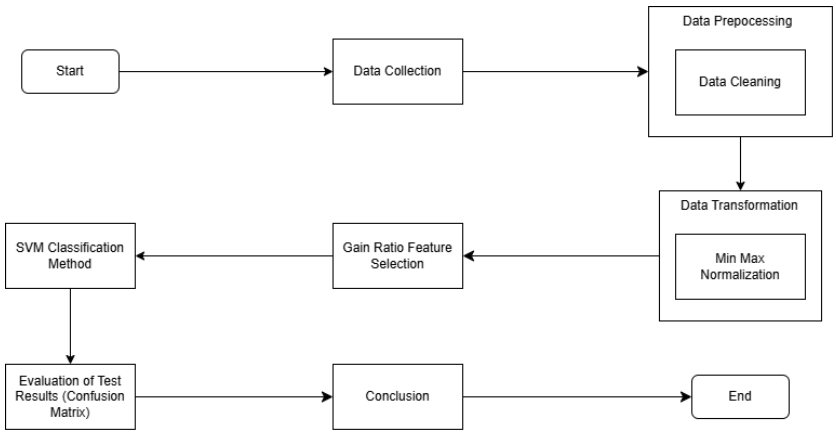


**Figure 1.** Research method

### 2.1 Data Collection

The data in this study are secondary data in the form of datasets taken from the kaggle platform. The license listed is CC0: Public Domain and can be accessed via https://www.kaggle.com/datasets/jamaltariqcheema/pima-indians-diabetes-dataset/data. Pima Indians Diabetes Dataset totals 768 with 8 attributes. The class of data consists of diabetes totaling 268 data and non-diabetes totaling 500 data. The diabetes dataset attributes can be seen in Table 1.

**Table 1.** Dataset Attributes

| No | Atribut |
|---|---|
| 1 | Pregnancies |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin Thickness |
| 5 | Insulin |
| 6 | BMI/Body Mass Index |
| 7 | Diabetes Pedigree Function |
| 8 | Age |
| 9 | Outcome |

The diabetes dataset can be seen in Table 2.

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

**Table 2.** Diabetes dataset

| Patient | Pregnancies | Glucose | Blood Presure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 148 | 72 | 35 | 169.5 | 33.6 | 0.627 | 50 | 1 |
| 2 | 1 | 85 | 66 | 29 | 102.5 | 26.6 | 0.351 | 31 | 0 |
| 3 | 8 | 183 | 64 | 32 | 169.5 | 23.3 | 0.672 | 32 | 1 |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 768 | 1 | 93 | 70 | 31 | 102.5 | 30.4 | 0.315 | 23 | 0 |

### 2.2 Data Prepocessing

This stage performs data cleaning from duplication, missing values, and not filled with inappropriate data, thus rearranging the data to fit the modeling to be done [25]. Cleaning data removes errors in the data, such as handling missing values and removing duplicate data. Aims to avoid bias that can be caused by missing diabetes data used, so as to improve the performance of the prediction model. Data balancing was not applied because the data difference between the two classes was only 232 data.

### 2.3 Data Transformation

This stage converts the data type to be in accordance with the provisions [26]. Data is changed to be simpler without changing the basic content [25]. This stage performs cleaning, changing and rearranging diabetes data to suit the modeling to be carried out. Data normalization measures the feature value of the dataset within the specified value range. This research uses the Min Max Normalization method. The Min Max Normalization method is a normalization method that changes the range of data values to be in the range of 0 to 1 [27].

$$X' = \frac{X_i - min_{(x)}}{max_{(x)} - min_{(x)}} \quad (1)$$

Description:
$X'$ = normalized value
$X_i$ = the specific value to be normalized
$min_{(x)}$ = minimum value of an attribute
$max_{(x)}$ = maximum value of an attribute

### 2.4 Gain Ratio Feature Selection

This stage performs feature selection using the gain ratio. Feature selection is an important process that aims to identify and select the most influential set of attributes [19]. Gain ratio is the development of information gain and is the best feature selection model and is widely used by researchers [10]. Feature selection can help find the ranking results of each attribute in diabetes data, so that it can help the learning modeling process and improve its accuracy. The threshold gain ratio used in this study includes 0.02, 0.03, and 0.05. According to [23] the gain ratio stage includes:

1. Calculating the entropy value of each attribute

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * log2pi \quad (2)$$

2. Calculating the information gain value of each attribute

$$Information\ Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|}\ x\ Entropy(Si) \quad (3)$$

3. Calculating the split information value

$$Split\ Info\ (D) = -\sum_{j=1}^{v} \frac{Dj}{D}\ x\ log2\ \frac{Dj}{D} \quad (4)$$

4. Calculating the gain ratio value

$$Gain\ Ratio\ (A) = \frac{Gain(A)}{Split\ Info(A)} \quad (5)$$

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

Description:
S = Sample
n = Number of values in the classification class
pi = Number of samples in class i
A = Attributes
|Si| = Number of samples of value i
|S| = Total data samples
D = Total number of samples in the dataset or data subset being processed
Dj = Number of samples in the Jth category or subset after separation based on a particular attribute
v = Number of categories or subsets resulting from attribute separation
Gain (A) = Information gain value of attribute A
Split Info (A) = Split info value of attribute A

### 2.5 SVM Classification Method

This stage conducts training and testing to create SVM modeling. Previously, diabetes data was divided into training data and testing data using the ratio of training data and testing data 90:10; 80:20; and 70:30. Next, classify using three kernels, namely linear, RBF, and polynomial. The value of parameter C to be used for all kernels is 1, 10, and 100. For the polynomial kernel, it includes the degree value. The degree values to be used are 1, 2, and 3. For the RBF kernel, the values used are gamma and scale. The gamma values to be used are 1, 2, and 3. According to [28] the kernel equation includes the equation:

1. Linear

$$K(Xi, Xj) = Xi^T \,.\, Xj \tag{6}$$

2. Polynomial

$$K(Xi, Xj) = (\Upsilon(Xi^T \,.\, Xj) + r)^d \tag{7}$$

3. RBF (Radial Basis Function)

$$K(Xi, Xj) = e^{-(\Upsilon \,||\, Xi - Xj \,||^2)} \tag{8}$$

Description:
$d$ = degree of polynomial
r = constant
$\Upsilon$ = kernel parameters

### 2.6 Evaluation of Test Results

This stage is the final stage in the classification algorithm calculation process [26]. This stage checks the suitability of patterns or information with previously existing facts or hypotheses [25]. This stage conducts accuracy testing by comparing the performance results of using various SVM algorithm kernels in diabetes classification. Previously, a single data split was performed using a ratio followed by the process of generating accuracy, recall, and precision measurements using a confusion matrix. According to [1] the representation of the results of the classification process on the confusion matrix has four terms including TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative).

1. Accuracy, indicating the level of accuracy of the model applied in classification

$$accuracy = \left(\frac{(TP + TN)}{(TP + FP + TN + FN)}\right) \times 100\% \tag{9}$$

2. Precision, indicates the accuracy between the requested data and the prediction results provided by the model.

$$precision = \left(\frac{TP}{(TP + FP)}\right) \times 100\% \tag{10}$$

3. Recall, indicating the success of the model in retrieving information

$$recall = \left(\frac{TP}{(TP + FN)}\right) \times 100\% \tag{11}$$

4. F-1 score, shows the weighted average comparison of precision and recall

$$F - 1\ score = \left(\frac{(2 \times \text{precision} \times \text{recall})}{(precision + recall)}\right) \times 100\% \tag{12}$$

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

# 3. RESULT AND DISCUSSION

The results and discussion include a discussion of the results of the research that has been done in measuring the effectiveness of applying feature selection gain ratio to the SVM method to improve the accuracy of diabetes classification.

## 3.1 Data Prepocessing

In this step, the data is checked first before the data is cleaned. After checking, the data has no missing values as shown in Figure 2. Since there are no missing values, the process of deleting or changing missing values is not necessary so we can proceed to the data transformation process.

| | 0 |
|---|---|
| Outcome | 0 |
| Pregnancies | 0 |
| Glucose | 0 |
| BloodPressure | 0 |
| SkinThickness | 0 |
| Insulin | 0 |
| BMI | 0 |
| DiabetesPedigreeFunction | 0 |
| Age | 0 |

**Figure 2.** Check for Missing Values

## 3.2 Data Transformation

In this step using Min Max Normalization. Data that has been normalized can be seen in Table 3.

**Table 3.** Data Normalization

| Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age |
|---|---|---|---|---|---|---|---|
| 0,353 | 0,671 | 0,490 | 0,304 | 0,187 | 0,315 | 0,234 | 0,483 |
| 0,059 | 0,265 | 0,429 | 0,239 | 0,106 | 0,172 | 0,117 | 0,167 |
| 0,471 | 0,897 | 0,408 | 0,272 | 0,187 | 0,104 | 0,254 | 0,183 |
| 0,059 | 0,290 | 0,429 | 0,174 | 0,096 | 0,202 | 0,038 | 0,000 |
| … | … | … | … | … | … | … | … |
| … | … | … | … | … | … | … | … |
| 0,059 | 0,316 | 0,469 | 0,261 | 0,106 | 0,249 | 0,101 | 0,033 |

## 3.3 Feature Selection Gain Ratio

The output generated by the selection of gain ratio features is the ranking order of all attributes from highest to lowest, as well as the selection results of ranking all attributes according to the threshold used. Figure 3 shows the results of the gain ratio calculation.

```
Fitur: Insulin, Gain Ratio: 0.1438
Fitur: DiabetesPedigreeFunction, Gain Ratio: 0.0737
Fitur: SkinThickness, Gain Ratio: 0.0622
Fitur: BMI, Gain Ratio: 0.0459
Fitur: Glucose, Gain Ratio: 0.0456
Fitur: Age, Gain Ratio: 0.0280
Fitur: BloodPressure, Gain Ratio: 0.0188
Fitur: Pregnancies, Gain Ratio: 0.0178
```

**Figure 3.** Gain Ratio Calculation Results

The representation of the Gain Ratio calculation is shown in Figure 4.

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

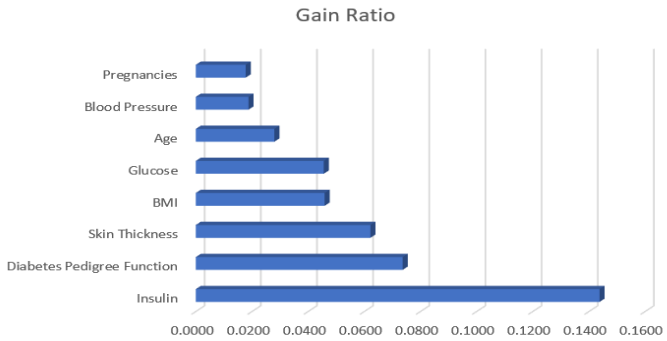


**Figure 4.** Gain Ratio Calculation

Table 4 shows the attributes used at various thresholds. Threshold 0.02 takes attributes that are above the 0.02 threshold, as well as the thresholds 0.03 and 0.05. Threshold 0.02 uses more attributes than other thresholds. The smaller the threshold, the more attributes are used and the larger the threshold, the fewer attributes are used.

**Table 4.** Attributes of Various Thresholds

| Threshold | Total Attribute | Attributes | | | | | |
|---|---|---|---|---|---|---|---|
| 0,02 | 6 Attribute | Insulin | Diabetes Pedigree Function | Skin Thickness | BMI | Glucose | Age |
| 0,03 | 5 Attribute | Insulin | Diabetes Pedigree Function | Skin Thickness | BMI | Glucose | |
| 0,05 | 3 Attribute | Insulin | Diabetes Pedigree Function | Skin Thickness | | | |

### 3.4 SVM Classification Method

This stage conducts training and testing to create SVM modeling. Previously, diabetes data was divided into training data and test data using a ratio of training data and test data of 90:10; 80:20; and 70:30. Next, classify using three kernels, namely linear, RBF, and polynomial. The values of parameter C to be used for all kernels are 1, 10, and 100. For the polynomial kernel, it includes the degree value. The degree values to be used are 1, 2, and 3. For the RBF kernel, the values used are gamma and scale. The gamma values to be used are 1, 2, and 3. The parameters of each kernel can be seen in Table 5.

**Table 5.** Kernel Parameters

| | C | Gamma | Degree |
|---|---|---|---|
| Linear | 1 | - | - |
| | 10 | - | - |
| | 100 | - | - |
| RBF | 1 | 1 | - |
| | | 2 | - |
| | | 3 | - |
| | | Scale | - |
| | 10 | 1 | - |
| | | 2 | - |
| | | 3 | - |
| | | Scale | - |
| | 100 | 1 | - |
| | | 2 | - |
| | | 3 | - |
| | | Scale | - |
| Polynomial | 1 | - | 1 |
| | | - | 2 |
| | | - | 3 |
| | 10 | - | 1 |
| | | - | 2 |
| | | - | 3 |
| | 100 | - | 1 |
| | | - | 2 |
| | | - | 3 |

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

### 3.5 Evaluation of Test Results

This study shows that the application of feature selection gain ratio in SVM can increase the accuracy of diabetes classification by 0.64% from 94.81% to 95.45% after applying a threshold of 0.02 at a ratio of 80:20 with RBF kernel at parameters C = 100, Gamma = Scale. There was an increase in accuracy from research [13] and [14] which only produced the highest accuracy of 87% using the SVM method on the same dataset. Table 6 shows the results of SVM testing on all kernels before applying the selection feature getting the highest accuracy of 94.81% at a ratio of 80:20 with the RBF kernel at 2 parameter combinations, namely at parameter C = 100, Gamma = 3 and at parameter C = 100, Gamma = Scale.

**Table 6.** Test Results Without Selection Features

| Ratio | Kernel | Parameter | Accuracy |
|---|---|---|---|
| 70:30 | Linear | C = 1 | 76,62% |
| | RBF | C = 100, Gamma = Scale | 93,51% |
| | Polynomial | C = 100, Degree = 3 | 90,48% |
| 80:20 | Linear | C = 100 | 79,22% |
| | RBF | C = 100, Gamma = 3 | 94,81% |
| | | C = 100, Gamma = Scale | |
| | Polynomial | C = 100, Degree = 3 | 91,56% |
| 90:10 | Linear | C = 100 | 77,92% |
| | RBF | C = 100, Gamma = 3 | 92,21% |
| | | C = 100, Gamma = Scale | |
| | Polynomial | C = 100, Degree = 3 | 89,61% |

The highest accuracy produced is 94.81% at a ratio of 80:20 with the RBF kernel with a combination of 2 parameters including C = 100, Gamma = 3 and C = 100, Gamma = Scale. All kernels show that the 80:20 ratio produces optimal performance.. The highest accuracy results before applying feature selection to each ratio are represented in Figure 5.

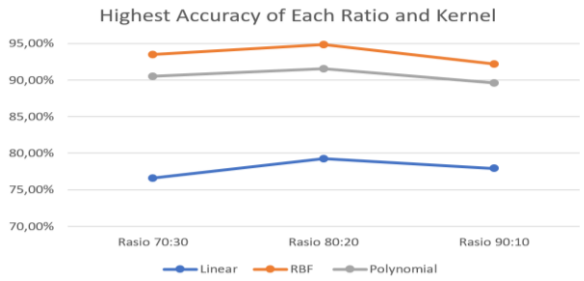


**Figure 5.** Testing Results Before Applying Feature Selection

Table 7 shows the results of SVM testing without gain ratio and with gain ratio at each threshold and each ratio represented by the highest accuracy on the linear kernel.

**Table 7.** Test Results After Applying Feature Selection on Linear Kernel

| Threshold | Total Attribute | Ratio | Parameter | Accuracy |
|---|---|---|---|---|
| No Gain Ratio | 8 Attribute | 70:30 | C = 1 | 76,62% |
| | | 80:20 | C = 100 | 79,22% |
| | | 90:10 | C = 100 | 77,92% |
| Threshold 0,02 | 6 Attribute | 70:30 | C = 1 | 76,19% |
| | | 80:20 | C = 1 | 77,92% |
| | | 90:10 | C = 10 | 76,62% |
| Threshold 0,03 | 5 Attribute | 70:30 | C = 10 | 77,06% |
| | | 80:20 | C = 1 | 81,17% |
| | | | C = 10 | |
| | | 90:10 | C = 1 | 85,71% |
| | | | C = 10 | |
| | | | C = 100 | |
| Threshold 0,05 | 3 Attribute | 70:30 | C = 100 | 82,25% |
| | | 80:20 | C = 100 | 84,42% |
| | | 90:10 | C = 10 | 85,71% |

Table 8 shows the results of SVM testing without gain ratio and with gain ratio at each threshold and each ratio represented by the highest accuracy on the RBF kernel.

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

**Table 8.** Test Results After Applying Feature Selection to RBF Kernel

| Threshold | Total Attribute | Ratio | Parameter | Accuracy |
|---|---|---|---|---|
| No Gain Ratio | 8 Attribute | 70:30 | C = 100, Gamma = Scale | 93,51% |
| | | 80:20 | C = 100, Gamma = 3 | 94,81% |
| | | | C = 100, Gamma = Scale | |
| | | 90:10 | C = 100, Gamma = 3 | 92,21% |
| | | | C = 100, Gamma = Scale | |
| Threshold 0,02 | 6 Attribute | 70:30 | C = 100, Gamma = Scale | 94,37% |
| | | 80:20 | C = 100, Gamma = Scale | 95,45% |
| | | 90:10 | C = 100, Gamma = Scale | 94,81% |
| Threshold 0,03 | 5 Attribute | 70:30 | C = 100, Gamma = Scale | 90,48% |
| | | 80:20 | C = 10, Gamma = Scale | 92,21% |
| | | | C = 100, Gamma = 2 | |
| | | | C = 100, Gamma = Scale | |
| | | 90:10 | C = 100, Gamma = 2 | 93,51% |
| Threshold 0,05 | 3 Attribute | 70:30 | C = 100, Gamma = Scale | 88,31% |
| | | 80:20 | C = 100, Gamma = Scale | 88,96% |
| | | 90:10 | C = 100, Gamma = Scale | 89,61% |

Table 9 shows the results of SVM testing without gain ratio and with gain ratio at each threshold and each ratio represented by the highest accuracy on the polynomial kernel.

**Table 9.** Test Results After Applying Feature Selection to the Polynomial Kernel

| Threshold | Total Attribute | Ratio | Parameter | Accuracy |
|---|---|---|---|---|
| No Gain Ratio | 8 Attribute | 70:30 | C = 100, Degree = 3 | 90,48% |
| | | 80:20 | C = 100, Degree = 3 | 91,56% |
| | | 90:10 | C = 100, Degree = 3 | 89,61% |
| Threshold 0,02 | 6 Attribute | 70:30 | C = 100, Degree = 3 | 87,01% |
| | | 80:20 | C = 100, Degree = 3 | 88,96% |
| | | 90:10 | C = 10, Degree = 3 | 88,31% |
| | | | C = 100, Degree = 3 | |
| Threshold 0,03 | 5 Attribute | 70:30 | C = 100, Degree = 3 | 87,88% |
| | | 80:20 | C = 100, Degree = 3 | 90,91% |
| | | 90:10 | C = 1, Degree = 2 | 88,31% |
| | | | C = 1, Degree = 3 | |
| | | | C = 100, Degree = 3 | |
| Threshold 0,05 | 3 Attribute | 70:30 | C = 10, Degree = 1 | 82,25% |
| | | | C = 100, Degree = 1 | |
| | | 80:20 | C = 1, Degree = 1 | 85,71% |
| | | | C = 10, Degree = 1 | |
| | | | C = 100, Degree = 1 | |
| | | 90:10 | C = 1, Degree = 1 | 87,01% |
| | | | C = 1, Degree = 2 | |
| | | | C = 10, Degree = 1 | |
| | | | C = 10, Degree = 2 | |
| | | | C = 100, Degree = 1 | |
| | | | C = 100, Degree = 2 | |

The results of the highest accuracy of the linear kernel before applying feature selection and after applying feature selection at each threshold and each ratio are represented in Figure 6.

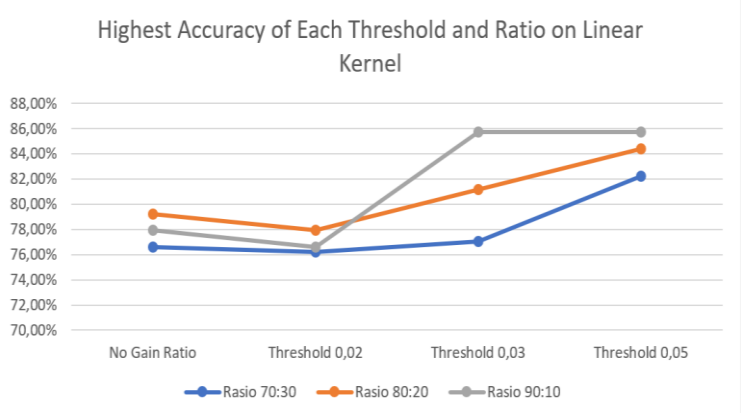


**Figure 6.** Testing Results After Applying Feature Selection on Linear Kernel

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

The highest accuracy results of the RBF kernel before applying feature selection and after applying feature selection at each threshold and each ratio are represented in Figure 7.

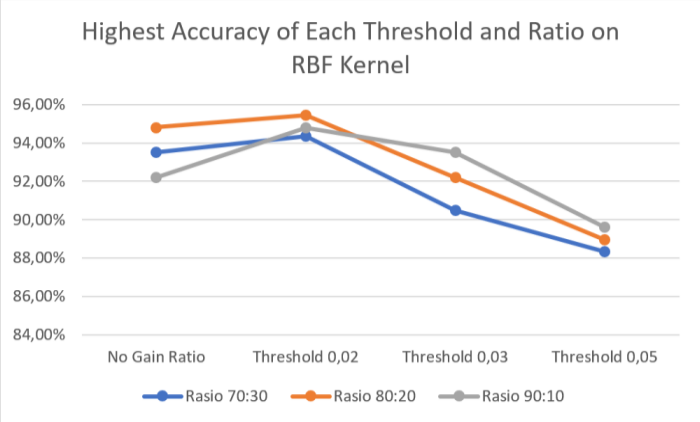


**Figure 7.** Test Results After Applying Feature Selection to the RBF Kernel

The highest accuracy results of the Polynomial kernel before applying feature selection and after applying feature selection at each threshold and each ratio are represented in Figure 8.

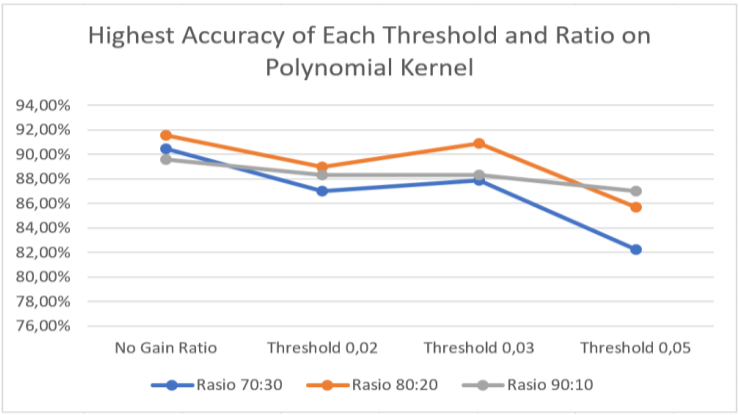


**Figure 8.** Test Results After Applying Feature Selection to the Polynomial Kernel

The highest accuracy test with SVM without feature selection resulted in the highest accuracy of 94.81% using a ratio of 80:20 on the RBF kernel with a combination of parameters Cost = 100, Gamma = 3 and Cost = 100, Gamma = Scale resulting in a confusion matrix shown in Figure 9. Using 154 test data, data classification successfully predicted data that was actually diabetic correctly as diabetes (True Positive) as much as 50 data, but there were 5 diabetic data that were incorrectly predicted as not diabetic (False Negative). Data classification also successfully predicts data that is actually not diabetic correctly as not diabetic (True Negative) as much as 96 data, but there are 3 non-diabetic data that are wrongly predicted as diabetes (False Positive). False Positive is negative data detected as positive data, while False Negative is positive data detected as negative data. The results show that the model is quite good at identifying cases of no diabetes and diabetes, but there are still errors in detecting diabetes as much as 5 data and no diabetes as much as 3 data.

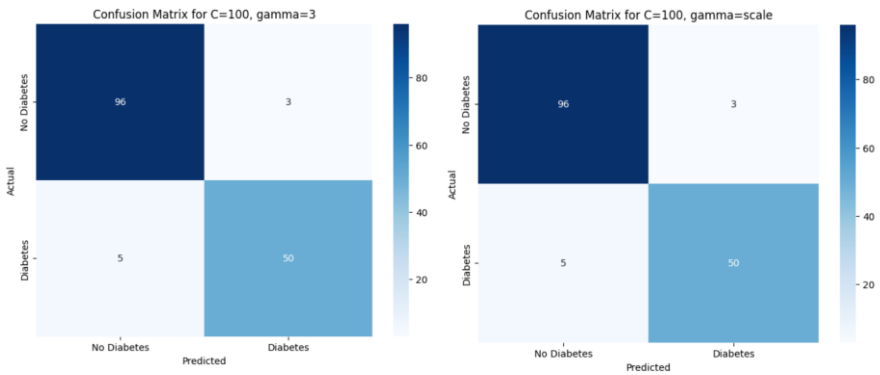


**Figure 9.** Confusion Matrix of Best Test Results Without Feature Selection

**Bulletin of Informatics and Data Science**
Vol. 4 No. 1, May 2025, Page 10−21
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.114
https://ejurnal.pdsi.or.id/index.php/bids/index

Testing the highest accuracy with SVM after applying feature selection resulted in the highest accuracy of 95,45% using a ratio of 80:20 on the RBF kernel with parameters Cost = 100 and Gamma = Scale produces a confusion matrix shown in Figure 10. Using 154 test data, data classification successfully predicts data that is actually diabetic correctly as diabetes (True Positive) as much as 52 data, but there are 3 diabetic data that are incorrectly predicted as not diabetic (False Negative). Data classification also successfully predicts data that is actually not diabetic correctly as not diabetic (True Negative) as much as 95 data, but there are 4 non-diabetic data that are wrongly predicted as diabetes (False Positive). False Positive is negative data detected as positive data, while False Negative is positive data detected as negative data. The results show that the model is quite good at identifying cases of no diabetes and diabetes, but the error in detection is reduced to 7 data including 3 diabetes data and 4 no diabetes data.

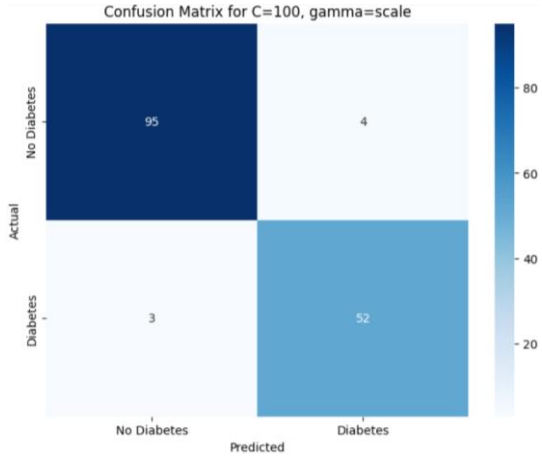


**Figure 10.** Confusion Matrix of Best Test Results with Feature Selection

This research shows that the application of feature selection gain ratio in SVM is able to increase the accuracy of diabetes disease classification by 0.64% from 94.81% using a ratio of 80:20 on the RBF kernel with parameters C = 100, Gamma = 3 and C = 100, Gamma = Scale to 95.45% at a threshold of 0.02 using a ratio of 80:20 on the kernel with parameters C = 100, Gamma = Scale. A ratio of 80:20 can produce optimal performance as in research [1]. RBF kernel often leads to high accuracy compared to other kernels as in research [1] and [13], both before applying feature selection and after applying feature selection. The polynomial kernel is proven to produce higher accuracy than the linear kernel. However, the accuracy of the polynomial kernel decreases after applying feature selection. The scale parameter in the RBF kernel also constantly produces the highest accuracy at each threshold variation and data ratio.

## 4. CONCLUSION

Tests by applying the Gain Ratio selection feature at various thresholds show that a threshold of 0.02 produces the highest accuracy. Tests on various ratios and kernels with various parameters show that the RBF kernel still provides optimal results. The application of feature selection, data ratio, and kernel parameters also affect the performance of the model. At a threshold of 0.02, the 80:20 data ratio produces higher accuracy compared to other data ratios on the RBF kernel. Threshold 0.02 produces the highest accuracy at a ratio of 80:20 across all kernels and produces the lowest accuracy at a ratio of 70:30. At a threshold of 0.03 there is a constant increase in accuracy from a ratio of 70:30 to 90:10 in the linear and RBF kernels, while the polynomial kernel produces the highest accuracy at a ratio of 80:20 and produces the lowest accuracy at a ratio of 70:30. At a threshold of 0.05 there is a constant increase in accuracy from a ratio of 70:30 to 90:10 across all kernels. Overall, the test results show that the application of selection features using gain ratio can improve model performance on all three kernels. After applying the selection feature, the highest accuracy increase was 95.45% at a threshold of 0.02 in a ratio of 80:20 using the RBF kernel with parameters Cost = 100 and Gamma = Scale. This research shows that the application of feature selection gain ratio in SVM is able to increase the accuracy of diabetes disease classification by 0.64% from 94.81%. A ratio of 80:20 can produce optimal performance on RBF and polynomial kernels. RBF kernel often leads to high accuracy compared to other kernels, both before applying feature selection and after applying feature selection. The polynomial kernel is proven to produce higher accuracy than the linear kernel. However, the accuracy of the polynomial kernel decreases after applying feature selection. The scale parameter in the RBF kernel also constantly produces the highest accuracy at each threshold variation and data ratio. The right combination of threshold, data ratio, and parameters of each kernel can produce a more reliable model in predicting the risk of diabetes. Future research is suggested to develop other combinations such as using data division other than ratios, such as cross validation. Due to the difference in the amount of data between classes 0 and 1, the use of data balancing techniques is recommended. In addition, it can also apply the gain ratio selection feature to other algorithms to improve accuracy or apply other selection features to the SVM method.

# REFERENCES

[1] H. Sohibul Wafa, A. I. Hadiana, and F. Rakhmat Umbara, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," vol. 4, no. 1, pp. 40–45, 2022
[2] Erika, "Meningkatkan Pemahaman Masyarakat Pentingnya Deteksi Dini Diabetes Melitus Melalui Penyuluhan Dan Pengukuran Gula Dan Tekanan Darah," *Jurnal Pengabdian Masyarakat*, vol. 1, no. 7, pp. 685–697, 2023.
[3] R. P. Febrinasari, T. A. Sholikah, D. N. Pakha, and S. E. Putra, *Buku Saku Diabetes Melitus Untuk Awam*, 1st ed. Surakarta, 2020. [Online]. Available: https://www.researchgate.net/publication/346495581
[4] A. Fanani and L. Sulaiman, "Faktor Obesitas dan Faktor Keturunan dengan Kejadian Kasus Diabetes Mellitus," *Riset Informasi Kesehatan*, vol. 10, no. 1, 2021, doi: 10.30644/rik.v8i2.464.
[5] H. Esti Ardiani, T. Astika Endah Permatasari, and Sugiatmi, "Obesitas, Pola Diet, dan Aktifitas Fisik dalam Penanganan Diabetes Melitus pada Masa Pandemi Covid-19," *Muhammadiyah Journal of Nutrition and Food Science (MJNF)*, vol. 2, no. 1, pp. 1–12, 2021, doi: 10.24853/mjnf.2.1.1-12.
[6] Ardiansyah. Aswin, E. C. O. Telaumbanua, A. S. Gultom, and A. A. S. M. Limbong, "Klasifikasi Penyakit Diabetes Menggunakan Metode SVM Dan KNN," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 3, no. 1, pp. 77–83, 2024, doi: 10.55606/juprit.v3i1.3151.
[7] S. Setyaningtyas, B. Indarmawan Nugroho, and Z. Arif, "Tinjauan Pustaka Sistematis Pada Data Mining: Studi Kasus Algoritma K-Means Clustering," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 10, no. 2, pp. 52–61, 2022, doi: 10.21063/jtif.2022.v10.2.52-61.
[8] P. W. Rahayu *et al.*, *BUKU AJAR DATA MINING*. Jambi: PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: https://www.researchgate.net/publication/377415198
[9] I. Made Arya Adinata Dwija Putra, I. Made Gede Sunarya, and I. Gede Aris Gunadi, "Perbandingan Algoritma Naive Bayes Berbasis Feature Selection Gain Ratio dengan Naive Bayes Kovensional dalam Prediksi Komplikasi Hipertensi," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 6, no. 1, pp. 37–49, 2024, doi: 10.35746/jtim.v6i1.488.
[10] Ivandari, M. Adib Al Karomi, and M. Rifqi Maulana, "Improved C45 Performance With Gain Ratio For Credit Approval Dataset," *JAICT Journal of Applied Communication and Information Technologies*, vol. 7, no. 2, pp. 135–139, 2022.
[11] S. Z. HR, A. Aziz, and W. Harianto, "Optimasi Algoritma K-Nearest Neighbor (Knn) Dengan Normalisasi Dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver," *Jurnal Mahasiswa Teknik Informatika*, vol. 6, no. 2, pp. 439–445, 2022
[12] N. Wijaya, M. Endah, and M. Feliati, "Penerapan Algoritma Decision Tree C.45 Untuk Klasifikasi Data Status Huni Rumah Rehabilitasi Pasca Erupsi Merapi Application Of C.45 Decision Tree Algorithm For Rehabilitation Household Data Classification Post Eruption Of Merapi," *Seminar Nasional UNRIYO*, pp. 424–430, 2020.
[13] M. Hilmy Haidar Aly, "Klasifikasi Diabetes Menggunakan Algoritma Support Vector Machine Radial Basis Function," *Jurnal Teknik Informatika dan Teknologi Informasi*, vol. 4, no. 1, pp. 28–38, 2024, doi: 10.55606/jutiti.v4i1.3420.
[14] A. Wildan Mucholladin, F. Abdurrachman Bachtiar, and M. Tanzil Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 2, pp. 622–633, 2021, [Online]. Available: http://j-ptiik.ub.ac.id
[15] S. Sri Rahayu, I. Afrianty, E. Budianita, and F. Syafria, "Klasifikasi Tulang Tengkorak Berdasarkan Jenis Kelamin Dalam Antropologi Forensik Menggunakan Metode Support Vector Machine," *Jurnal Inovtek Polbeng*, vol. 9, no. 1, pp. 243–256, 2024.
[16] N. Wijaya, M. Endah, and M. Feliati, "Penerapan Algoritma Decision Tree C.45 Untuk Klasifikasi Data Status Huni Rumah Rehabilitasi Pasca Erupsi Merapi," *Seminar Nasional UNRIYO*, pp. 424–430, 2020.
[17] Kurniabudi, A. Harris, and A. Edward Mintaria, "Komparasi Information Gain, Gain Ratio, CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali," *Jurnal Media Informatika Budidarma*, vol. 5, no. 1, pp. 332–343, 2021, doi: 10.30865/mib.v5i1.2258.
[18] A. Lutfia, Gunawan, R. Saepul Rohman, and A. Gunawan, "Penerapan Seleksi Fitur Gain Ratio Pada Prediksi Penyakit Jantung Berbasis Naïve Bayes," *Jurnal Responsif*, vol. 6, no. 1, pp. 1–10, 2024
[19] M. Yamin Amzah, Kusnadi, and L. Bayuaji, "Optimasi Algoritma Support Vector Machine Dengan Menggunakan Feature Selection Gain Ratio Untuk Analisis Sentimen," *Inovtek Polbeng*, vol. 9, no. 1, pp. 326–340, 2024.
[20] M. I. Prasetiyowati, N. Ulfa Maulidevi, and K. Surendro, "Determining Threshold Value On Information Gain Feature Selection To Increase Speed And Prediction Accuracy Of Random Forest," *J Big Data*, vol. 8, pp. 1–22, 2021, doi: 10.1186/s40537-021-00472-4.
[21] A. Wildan Attabi', L. Muflikhah, and M. A. Fauzi, "Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain," JPTIIK, vol. 2, no. 11, pp. 4548–4554, 2018
[22] Ivandari, M. Adib Al Karomi, and M. Rifqi Maulana, "Improved C45 performance with gain ratio for credit approval dataset," *Journal of Applied Communication and Information Technologies*, vol. 7, no. 2, p. 2022, 2022.
[23] Y. Harni, I. Afrianty, S. Sanjaya, R. Abdillah, F. Yanto, and F. Syafria, "Performance Analysis of LVQ 1 Using Feature Selection Gain Ratio for Sex Classification in Forensic Anthropology," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, 2023, doi: 10.47065/bits.v5i1.3625.
[24] M. Yamin Amzah, Kusnadi, and L. Bayuaji, "Optimasi Algoritma Support Vector Machine Dengan Menggunakan Feature Selection Gain Ratio Untuk Analisis Sentimen," Jurnal INOVTEK Polbeng, vol. 9, no. 1, pp. 326–340, 2024.
[25] A. Srirahayu and L. Setya Pribadie, "Review Paper Data Mining Klasifikasi Data Mining," *Jurnal Ilmiah Informatika Global*, vol. 14, no. 1, pp. 7–12, 2023, doi: 10.36982/jiig.v13i2.2307.
[26] L. Budhy Adzy, Asriyanik, and A. Pambudi, "Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Iuran Jaminan Kesehatan Pemerintah Daerah Kabupaten Sukabumi," *Jurnal MNEMONIC*, vol. 6, no. 1, 2023.
[27] I. Permana and F. N. Salisah, "Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *Indonesian Journal of Informatic Research and Software Engineering*, vol. 2, no. 1, pp. 67–72, 2022, Accessed: Apr. 16, 2025. [Online]. Available: https://journal.irpi.or.id/index.php/ijirse

[28] I. Afrianty, D. Nasien, and H. Haron, "Performance Analysis of Support Vector Machine in Sex Classification of The Sacrum Bone in Forensic Anthropology," *JURNAL TEKNIK INFORMATIKA*, vol. 15, no. 1, pp. 63–72, 2022, doi: 10.15408/jti.v15i1.25254.