Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link

Hairani Hairani a,*, Anthony Anggrawan a, Dadang Priyanto a

a Department of Computer Science, Universitas Bumigora, Mataram, 83127, Indonesia
Corresponding author: *hairani@universitasbumigora.ac.id

Abstract—Most of the health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to classify the majority data more and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. Early diagnosis of diabetes is needed to minimize the occurrence of more severe complications. In the diabetes dataset used, there is an imbalanced data between positive and negative diabetes classes. Diabetes negative class data (500 data) is more than diabetes positive class (268), so it can affect the performance of the classification method. Therefore, this study aims to apply the Smote-Tomeklink and Random Forest methods in the classification of diabetes. The research methodology used is the collection of diabetes data obtained from Kaggle, as many as 768 data with eight input attributes and 1 output attribute as a class, pre-processing data is used to balance the dataset with Smote-Tomeklink, classification using the random forest method, and performance evaluation based on accuracy, sensitivity, precision, and F1-score. Based on the tests conducted by dividing data using 10-fold cross-validation, the Random Forest algorithm with Smote-Tomeklink gets the highest accuracy, sensitivity, precision, and F1-score compared to Random Forest with Smote. The Random Forest algorithm with Smote-Tomeklink has 86.4% accuracy, 88.2% sensitivity, 82.3% precision, and 85.1% F1-score. Thus, using Smote-Tomeklink can improve the performance of the random forest method based on accuracy, sensitivity, precision, and F1-score.

Keywords—Class Imbalance; smote-tomeklink; random forest method; diabetes disease.

I. INTRODUCTION
Most of the Health data contained unbalanced data that affected the performance of the classification method. Unbalanced data causes the classification method to classify the majority data more and ignore the minority class. One of the health data that has unbalanced data is Pima Indian Diabetes. Diabetes is a deadly disease caused by increased blood sugar in the body. Diabetes is caused by the body's inability to produce enough insulin. Complications of diabetes can cause heart attacks and strokes. One way to improve the performance of the classification method is to manage balanced data by adding minority data so that the number is equal to the majority class. The diabetes dataset has 768 instances of data. However, the problem is an imbalance of data in the dataset, namely the negative diabetes class with 500 data (majority class), while the positive diabetes class with 268 data (minority class). Data imbalance is the amount of data in one class more than in the other class. The problem of data imbalance causes the classification method to be more dominant in classifying the majority class than the minority class, or in other words, the classification method ignores the minority class. The problem of unbalanced data can be handled with a data sampling approach.

Several data sampling methods that can be used to solve the problem of data imbalance are oversampling [1]-[4], undersampling [5], [6], and Hybrid Sampling [6],[7]. Oversampling works by adding the minority class, while Undersampling works by removing the majority class to produce balanced data. However, both methods have their respective weaknesses. The weakness of the oversampling method is that there are too many repetitions of samples which can cause overfitting of the classification method, while the weakness of undersampling is that it will lose information from most of the samples in the dataset and cannot take full advantage of the available information[9].

The Smote method was developed to overcome these weaknesses to avoid overfitting the oversampling method. Smote is an oversampling method to generate new synthesis
training data by linear interpolation on minority classes[10]. However, the Smote method has a weakness, namely overgeneralization, and the addition of a minority class randomly can generate noise data because it does not differentiate between classes[11]. Therefore, the undersampling method is used to improve the performance of the oversampling method by cleaning the noise data in the majority class. The noise data is the majority class instance which is closest to the minority class instance. Usually, noise data reduces the level of accuracy for classification methods[5]. One method to remove noise data in the majority class is Tomeklink[12]. Tomeklink is an undersampling method that cleans noise data from the majority class, which has similar characteristics and overlapping. However, Tomeklink only deletes instances defined as “Tomek Links” so that the analyzed data cannot be balanced, and in its implementation, the method is combined with other methods. Combining Tomeklink and Smote oversampling can improve accuracy better than individual performance[13].

Data mining research in Health plays an important role, especially in predicting various types of diseases using different techniques or methods[14]. Research [15] uses a statistical approach to analyze the success rate of students following subjects using online or face-to-face learning. The results show that online students have significantly higher average grades than face-to-face classes.

Several previous studies have focused on the classification of diabetes; namely, Kaur [16] predicts diabetes using the k-NN method with an accuracy of 83%. The weakness of the research is that it does not address the problem of data imbalance. Azrar et al. [17] classify diabetes using the C4.5 method with an accuracy of 75.65%. The weakness of the research is that the accuracy obtained is low, so it can be improved, and it does not address the data imbalance problem. Barik et al. [18] Using XGBoost to predict diabetes with 74% accuracy. The weakness of the research is that the accuracy obtained is low, so it can be improved, and it does not address the data imbalance problem.

Hairani, Inuddin, and Rahardi [19] used the Correlated Naïve Bayes method with correlation-based feature selection to predict diabetes with an accuracy of 69.51%. The weakness of the research is that the accuracy obtained is low, so it can be improved, and it does not address the data imbalance problem. Fiarani, Sipayung, and Maemunah [20] used the C4.5 method for diabetes detection with an accuracy of 68%.

Erlin et al. [21] used logistic regression and smote methods to detect diabetes with 82% accuracy, 81% precision, 79% recall, and 80% F1-score. The weakness of the research is that the accuracy is good but can be improved using Tomeklink to clean noise data in the majority class. Azaz et al. [22] used the C4.5 and Smote methods to predict diabetes with 82% accuracy, 80% precision, and 86% sensitivity. Shi et al. [23] used logistic and Smote-ENN methods to predict kidney disease with 75.2% accuracy, 70.6% recall, 4.9% precision, and 30% F1-score. The weakness of the research is the low accuracy, so it can be improved using Tomeklink to clean noise data in the majority class. Wang et al. [24] studied SME-XGBoost with Smote-ENN for heart disease prediction with 80% AUC.

Several previous studies have applied various approaches to improve diabetes classification methods, such as the oversampling approach with SMOTE. However, there are weaknesses in previous studies; namely, the accuracy of the proposed method still ranges from 82% to 83%, so there is a gap in improving its accuracy. So, this study proposes the Smote-Tomeklink hybrid sampling method to overcome the imbalance in diabetes data and improve the accuracy of the classification method.

Smote-Tomeklink is a good way to avoid the drawbacks of SMOTE and Tomeklink techniques [9]. The classification method used in this research is Random Forest. The Random Forest method was chosen because it has several advantages: high accuracy [25], the ability to handle noise data, fast performance in training data, overfitting control, and ease of implementation [26].

This study aims to apply the Smote-Tomeklink hybrid sampling method to balance the data on diabetes data so as to improve the performance of the Random Forest classification method. Measurement of the performance of the random forest method based on accuracy, sensitivity (recall), precision, and F1-score.

II. MATERIALS AND METHOD

This research consists of several stages, as shown in Figure 1.

| Data Collection : 
Pima Indian Diabetes | 
Data Preprocessing : 
1. SMOTE 
2. Smote-Tomeklink | 
Classification Method : 
Random Forest | 
Evaluation Performance : 
Accuracy, Sensitivity, Precision, F1-Score |

Fig. 1 Research Stages

A. Data Collection

The dataset used in this study is a diabetes dataset obtained from Kaggle, which consists of 768 instances and 9 attributes. The description of the attributes and the sample data used are shown respectively in Table I and Table II.

<table>
<thead>
<tr>
<th>No</th>
<th>Attribute</th>
<th>Description</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pregnancies</td>
<td>Number of Pregnancy</td>
<td>X1</td>
</tr>
<tr>
<td>2</td>
<td>Glucose</td>
<td>Glucose level 2 hours after eating</td>
<td>X2</td>
</tr>
<tr>
<td>3</td>
<td>Blood Pressure</td>
<td>Blood Pressure</td>
<td>X3</td>
</tr>
<tr>
<td>4</td>
<td>Skin Thickness</td>
<td>Skin Thickness</td>
<td>X4</td>
</tr>
<tr>
<td>5</td>
<td>Insulin</td>
<td>Insulin</td>
<td>X5</td>
</tr>
<tr>
<td>6</td>
<td>BMI</td>
<td>Body Massa Index</td>
<td>X6</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes Pedigree Function</td>
<td>Diabetes Pedigree Function</td>
<td>X7</td>
</tr>
<tr>
<td>8</td>
<td>Age</td>
<td>Age</td>
<td>X8</td>
</tr>
<tr>
<td>9</td>
<td>Outcome</td>
<td>Diabetes Status (1 = Positive Diabetes, 2 = Negative Diabetes)</td>
<td>Y</td>
</tr>
</tbody>
</table>

TABLE I

DESCRIPTION ATTRIBUTE DATASET
TABLE III
SAMPLE DATASET

<table>
<thead>
<tr>
<th>No</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>148</td>
<td>72</td>
<td>35</td>
<td>0</td>
<td>33.6</td>
<td>0.627</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>85</td>
<td>66</td>
<td>29</td>
<td>0</td>
<td>26.6</td>
<td>0.351</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>183</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>23.3</td>
<td>0.672</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>89</td>
<td>66</td>
<td>23</td>
<td>94</td>
<td>28.1</td>
<td>0.167</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>137</td>
<td>40</td>
<td>35</td>
<td>168</td>
<td>43.1</td>
<td>2.288</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>116</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>25.6</td>
<td>0.201</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>78</td>
<td>50</td>
<td>32</td>
<td>88</td>
<td>31</td>
<td>0.248</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>115</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35.3</td>
<td>0.134</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>197</td>
<td>70</td>
<td>45</td>
<td>543</td>
<td>30.5</td>
<td>0.158</td>
<td>53</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>125</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>76.6</td>
<td>0.351</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>110</td>
<td>92</td>
<td>0</td>
<td>0</td>
<td>37.6</td>
<td>0.191</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>168</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>0.537</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>139</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>27.1</td>
<td>1.441</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>189</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0.672</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>166</td>
<td>72</td>
<td>19</td>
<td>175</td>
<td>25.8</td>
<td>0.587</td>
<td>51</td>
<td>1</td>
</tr>
</tbody>
</table>

B. Data Pre-processing

Data Pre-processing is one of the important stages in data mining to improve the quality of datasets. This study focuses on dealing with unbalanced data contained in the diabetes dataset. The dataset used has 268 instances of negative diabetes and 500 instances of Positive Diabetes. The algorithms used to handle unbalanced data in the dataset are SMOTE (Synthetic Minority Oversampling Technique) and Smote-Tomeklink.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes randomly by creating synthetic data for oversampling purposes [10]. Creating new data on the minority class using the equation (1).

$$Y' = Y' + (Y - Y') \times \gamma$$

(1)

$Y'$ is the representation of the addition of the minority class. $Y'$ is the representation of minority class, $Y'$ is a value chosen at random from the k-nearest neighbors of the minority class on $Y$ and $\gamma$ is a value in a randomly selected vector with a range of 0 to 1 [2]. SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class, as shown in Figure 2 below.

Tomeklink is an undersampling method that cleans noise data from the majority class with similar characteristics and overlapping[12]. Tomeklink eliminates the majority class instances closer to the minority class by applying the nearest neighbor rule to select instances. The combination of Tomeklink and Smote oversampling can improve accuracy better than individual performance [13].

C. Random Forest Method

Random Forest is a decision tree-based ensemble learning method [28]. The Random Forest method has the advantages of high accuracy, handling noise data, fast performance in training data, overfitting control, and easy implementation [26]. The working process of the Random Forest method in classifying data is shown in Figure 3.

---

Fig. 2 Process of Synthetic Minority Oversampling Technique (SMOTE) Algorithm [27]

---

260
Figure 3 shows how the Random Forest algorithm works by creating a set of decision trees from a randomly selected subset, getting predictions from each decision tree, voting for each predicted outcome, and choosing the best prediction result based on the most votes assigned as the final prediction.

D. Evaluation Performance

Performance testing uses a confusion matrix table. The confusion matrix is a table that is used to describe the performance of the classification method on a dataset whose true value is known. The confusion matrix can visualize the amount of data that is classified as true and false, as shown in Table III.[29]

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>Negative</td>
<td>TN</td>
</tr>
<tr>
<td>Positive</td>
<td>FN</td>
</tr>
</tbody>
</table>

The formula used to calculate accuracy (6), Sensitivity (7), Precision (8) [30] [31] [32], and F1-score (5) [33].

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{8}
\]

\[
F1 - score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{9}
\]

True Positive (TP) is a class of positive diabetes that is predicted correctly. False Positive (FP) is a diabetes-negative class but is predicted to be diabetes positive. True Negative (TN) is a diabetes negative class that is predicted correctly. False Negative (FN) is a positive diabetes class but is predicted to be diabetes negative.

III. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is diabetes data obtained from Kaggle. The pre-processing of this study used the Smote and Smote-Tomeklink algorithms to deal with class imbalances in diabetes data. The classification method of this research is Random Forest. The performance test is based on accuracy, sensitivity, precision, and F1-score. The results of the comparison of the original data with the data from Smote and the results of Smote-Tomeklink are shown in Figure 4.

The classification method of this research is Random Forest. Performance testing is based on accuracy, sensitivity, precision, and F1-score using a confusion matrix table. Based on testing the Random Forest method using 10-fold cross-validation, the results obtained in the form of a confusion matrix table as shown in Table IV for the Random Forest method on the original data, Table V for the results of the Random Forest method with Smote, and Table VI for the results of the Random Forest method with Smote-Tomeklink. The results of the comparison of the performance of the Random Forest method as a whole are shown in Figure 5.
In Table IV, the Random Forest method succeeded in correctly classifying the negative class (TN) as many as 429 instances and the negative class classified incorrectly (FP) as many as 17 instances. While the correctly classified positive class (TP) is 155 instances, and the incorrectly classified positive class is 113 instances.

In Table V, the Random Forest method with Smote succeeded in correctly classifying the negative class (TN) in as many as 390 instances and the negative class classified incorrectly (FP) as many as 110 instances. While the positive class that is classified correctly (TP) is 429 instances, and the positive class that is classified incorrectly is 71 instances.

In Table VI, the Random Forest method with Smote-Tomeklink succeeded in correctly classifying the negative class (TN) in as many as 385 instances and the negative class classified incorrectly (FP) as 90 instances. While the positive class that is classified correctly (TP) is 419 instances and the positive class that is classified incorrectly is 56 instances.

Based on Figure 4, there was an increase in the performance of the Random Forest method with Smote-Tomeklink based on accuracy, sensitivity, precision, and F1-score. In the original dataset, the Random Forest method has 76% accuracy, 57.8% sensitivity, 68.6% precision, and 62.7% F1-score. The Random Forest method with Smote has an accuracy of 81.9%, the sensitivity of 85.8%, precision of 79.6%, and F1-score of 82.6%. Meanwhile, using the Random Forest method with Smote-Tomeklink resulted in an accuracy of 86.4%, a sensitivity of 88.2%, a precision of 83.3%, and an F1-score of 85.1%.

Sensitivity has a very important role in improving the Accuracy and F1-score performance of the Random Forest method with Smote-Tomeklink. The Random Forest method with Smote-Tomeklink gives higher accuracy, sensitivity, precision, and F1-score results than smote and without sampling.

Random Forest method with Smote an increase in performance indicators Accuracy, sensitivity, precision, and F1-score. The increase in accuracy scores is 5.9%, Sensitivity is 28%, precision is 11%, and F1-score is 19.9%. The Random Forest method with Smote-Tomeklink showed an increase in the indicators of accuracy by 10.4%, Sensitivity by 30.4%, precision by 13.7%, and F1-score by 22.4%. Therefore, the
Table VII. COMPARISON OF THE PROPOSED MODEL PERFORMANCE WITH PREVIOUS STUDIES

<table>
<thead>
<tr>
<th>No</th>
<th>Author (Year)</th>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[16]</td>
<td>Pima Indian Diabetes</td>
<td>KNN</td>
<td>83%</td>
</tr>
<tr>
<td>2</td>
<td>[17]</td>
<td>Pima Indian Diabetes</td>
<td>Decision Tree C.45</td>
<td>75.65%</td>
</tr>
<tr>
<td>4</td>
<td>[21]</td>
<td>Pima Indian Diabetes</td>
<td>Logistic regression + Smote</td>
<td>82%</td>
</tr>
<tr>
<td>5</td>
<td>[22]</td>
<td>Pima Indian Diabetes</td>
<td>C4.5 Method + Smote</td>
<td>82%</td>
</tr>
<tr>
<td>6</td>
<td>The Proposed Method</td>
<td>Pima Indian Diabetes</td>
<td>Random Forest + SMOTE Tomek links</td>
<td>86%</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

This study applies the Smote-Tomeklink algorithm to the Random Forest method for the classification of diabetes. The implementation of Smote-Tomeklink can improve the performance of accuracy, sensitivity, precision, and F1-score in the Random Forest method. The combination of Random Forest and Smote-Tomeklink got the best accuracy, sensitivity, and precision compared to Smote and without sampling for the classification of diabetes. There was an increase in performance indicators of 10.4% accuracy, 30.4% sensitivity, 13.7% precision, and 22.4 F1-score. Further research can apply Smote-Tomeklink to deal with the problem of data imbalance in multiclass data.

REFERENCES


