Combination of Smote and Random Forest Methods for Lung Cancer Classification

Christopher Michael Lauw¹, Hairani Hairani², Ilham Saifudin², Juvalna Ximenes Guterres³, Muhammad Maariful Huda⁴, Mayadi Mayadi⁵

¹Universitas Bumigora, Mataram, Indonesia
²Universitas Muhammadiyah Jember, Jember, Indonesia
³Universidade Oriental Timor Lorosa’e, Unital Becora Dili, Timor Leste
⁴Politeknik Angkatan Darat, Malang, Indonesia
⁵University Teknologi Mara, Selangor, Malaysia

ABSTRACT

Lung cancer is a network of cells that grow abnormally in the lungs. Lung cancer has four severity levels, namely stages 1 to 4. If lung cancer is not treated quickly, it is at risk of causing death. This research aimed to combine Synthetic Minority Over-sampling (Smote) and Random Forest methods for lung cancer classification. The method used was a combination of Smote and Random Forest. Smote was used to balance the data, while Random Forest was used to classify lung cancer data. The results showed that the combination of Smote and Random Forest methods obtained an accuracy of 94.1%, sensitivity of 94.5, and specificity of 93.7%. Meanwhile, without Smote, the accuracy is 89.1%, sensitivity is 55%, and specificity is 94.5%. The use of Smote can improve the performance of the Random Forest classification method based on accuracy and sensitivity. There was an increase of 5% in accuracy and a 39% increase in sensitivity.

Keywords: Data Mining, Lung Cancer, Prediction Method, Random Forest

Copyright ©2023 International Journal of Engineering and Computer Science Applications (IJECSA). This is an open access article under the CC BY-SA license.
1 INTRODUCTION

Lung cancer is an abnormal cell tissue that grows in the lungs. The problem found is that many patients are not saved due to lung cancer because it is not detected early [1]. The cause of the slow detection of lung cancer is the risk of causing death [2]. Early detection of lung cancer must be done to get fast and proper treatment to reduce the risk of more severe [3]. Many lung cancer screening applications can be utilized in today's technology era. Early detection of lung cancer independently can be done by entering the symptoms experienced in the application installed on the smartphone. The problem is that the level of accuracy of the resulting diagnosis does not have high accuracy. Therefore, this research not only focuses on developing an early screening application for lung cancer but also considers a very high level of accuracy. To get high-accuracy results, looking at the complexity of the lung disease dataset used is necessary. This research uses a lung cancer dataset obtained from Kaggle with a positive number of cancer of 238 instances and negative cancer of 38 instances. The lung cancer dataset has an unbalanced data problem, with more positive than negative cancer classes. This will cause the prediction method to recognize more cancer-positive classes than cancer-negative classes so that the performance of the prediction method is low.


Based on several previous studies that have used various approaches for predicting lung cancer, several gaps can be concluded from previous research, such as the level of accuracy is not optimal and solving the problem of unbalanced data on lung disease prediction has not been done. Therefore, it is necessary to improve the performance of classification methods in lung cancer prediction by solving the problem of unbalanced data on lung cancer disease data. An approach that can be used to solve unbalanced data on lung cancer data is Smote [12], [13]. The SMOTE method is used to overcome data imbalance by generating artificial data in the minority class so that the number is the same as the majority class [14]. At the same time, the classification method used is Random Forest because it performs better than several other classification methods in lung cancer prediction [8]. Therefore, this research aims to combine Smote and Random Forest methods for lung disease prediction to obtain high accuracy.

2 RESEARCH METHOD

The research framework is shown in Figure 1. In the initial stage, the data collection process obtains lung cancer data from the Kaggle site. The lung cancer dataset obtained from Kaggle has 276 data and 15 attributes. Lung cancer attributes can be seen in Table 1.
Table 1. Attributes of the Lung Cancer Dataset.

<table>
<thead>
<tr>
<th>No</th>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
<td>Gender is an attribute of the patient’s sex</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>Age is the age attribute of the patient</td>
</tr>
<tr>
<td>3</td>
<td>Smoking</td>
<td>An attribute that describes whether the patient is a smoker</td>
</tr>
<tr>
<td>4</td>
<td>Yellow_Fingers</td>
<td>Attributes in the form of a question whether the patient has yellow fingers</td>
</tr>
<tr>
<td>5</td>
<td>Anxiety</td>
<td>Excessive panic when breathing out of breath rhythm</td>
</tr>
<tr>
<td>6</td>
<td>Peer_Pressure</td>
<td>Psychological stress or feeling pressured by the environment (shortness of breath in crowds)</td>
</tr>
<tr>
<td>7</td>
<td>Chronic Disease</td>
<td>Having a chronic disease</td>
</tr>
<tr>
<td>8</td>
<td>Fatigue</td>
<td>Tired quickly during daily activities</td>
</tr>
<tr>
<td>9</td>
<td>Allergy</td>
<td>Allergic disease</td>
</tr>
<tr>
<td>10</td>
<td>Wheezing</td>
<td>Breathing sounds</td>
</tr>
<tr>
<td>11</td>
<td>Alcohol Consuming</td>
<td>History of alcohol consumption</td>
</tr>
<tr>
<td>12</td>
<td>Coughing</td>
<td>Coughing is an attribute of coughs</td>
</tr>
<tr>
<td>13</td>
<td>Shortness Of Breath</td>
<td>Shortness of Breath is an attribute of shortness of breath</td>
</tr>
<tr>
<td>14</td>
<td>Swallowing Difficulty</td>
<td>Swallowing difficulty is an attribute of shallowing difficulty</td>
</tr>
<tr>
<td>15</td>
<td>Lung Cancer</td>
<td>Prediction Class</td>
</tr>
</tbody>
</table>

The next stage is data processing, which is needed to transform raw data into better-quality data so that the performance of the classification method is better. The processing part uses sampling techniques to balance the data on the lung cancer dataset because the majority class has 238 instances, and the minority class has 38 instances. The Smote method balances the minority class so that the number equals the majority class. The Smote method generates new synthetic data based on the number of nearest neighbors between minority data. The third stage is implementing the Random Forest classification method for predicting lung cancer data by dividing training and testing data based on 10-fold cross-validation. The last stage is a performance evaluation that is needed to determine the level of accuracy of the Random Forest method in predicting lung cancer. Performance testing based on accuracy, recall, and specificity using formula (1), (2), and (3).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{3}
\]

3 RESULTS AND ANALYSIS

This section will present the research results obtained at each stage. The research stages start from collecting lung cancer data, preprocessing data using the Smote sampling approach, implementing the Random Forest method, and testing performance based on accuracy, sensitivity, and specificity. Lung cancer data is obtained from the Kaggle site, which has 276 data instances. Before class balancing using SMOTE, there were 38 negative and 238 positive cancer. Lung cancer data that has not been balanced, then performed data balancing using Smote so that the distribution of each class becomes 238 instances—the distribution of each class before and after Smote can be seen in Table 2. The next process is to classify lung cancer data using Random Forests. The classification results of the Random Forest method without Smote can be seen in Figure 2, while the results of the combination of Smote and Random Forest are shown in Figure 3.

Table 2. Class Ratio of Lung Cancer Dataset

<table>
<thead>
<tr>
<th>Sampling Methods</th>
<th>Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>Original</td>
<td>38</td>
</tr>
<tr>
<td>SMOTE</td>
<td>238</td>
</tr>
</tbody>
</table>
In Figure 2, the Random Forest method without Smote correctly predicts the cancer class in as many as 21 instances out of 38 data. In comparison, the non-cancer class is correctly predicted in as many as 225 instances out of 238 data. The Random Forest method with Smote correctly predicts the cancer class in as many as 225 instances out of 238 data, while the non-cancer class correctly predicted as many as 223 instances out of 238 data. The Random Forest method without Smote obtained an accuracy of 89.1%, a sensitivity of 55%, and a specificity of 94.5%. At the same time, the Smote and Random Forest methods combination obtained an accuracy of 94.1%, a sensitivity of 94.5, and a specificity of 93.7%. The use of Smote can improve
the performance of the Random Forest classification method based on accuracy and sensitivity. In accuracy, there was an increase of 5%, and the sensitivity of the increase was 39%; this was reinforced by research [15], [16].

4 CONCLUSION

Based on the test results, the combination of the Smote method with Random Forest gets an accuracy of 94.1%, sensitivity of 94.5%, and specificity of 93.7%. In contrast, without Smote, it gets an accuracy of 89.1%, sensitivity of 55%, and specificity of 94.5%. The Smote and Random Forest combination increased accuracy by 5% and sensitivity by 39% compared to without Smote. This result indicates that using Smote can increase the accuracy and sensitivity of the Random Forest method in lung cancer prediction. Future research suggestions are hybrid samplings such as Smote-Tomek Link or Smote-ENN to solve the problem of unbalanced data in lung cancer to improve the performance of the classification method used better than Smote individually.

REFERENCES


[16] M. M. Hassan and D. Kadir, “Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE