Predicting life expectancy of lung cancer patients after thoracic surgery using SMOTE and machine learning approaches

SELLY ANASTASSIA AMELLIA KHARIS1*, ARMAN HAQQI ANNA ZILI2

1Faculty of Science and Technology, Universitas Terbuka, Tangerang Selatan, Indonesia  
2Faculty of Mathematics and Science, Universitas Indonesia, Depok, Indonesia

Abstract. Lung cancer is a life-threatening condition characterized by the uncontrolled growth and spread of abnormal cells in the lungs. Thoracic surgery is a commonly employed diagnostic and treatment procedure for lung cancer. The objective of this study is to utilize machine learning techniques to predict the life expectancy of lung cancer patients one year after thoracic surgery. The study utilizes the Thoracic Surgery Data Set, consisting of 454 data, with 385 data representing surviving patients and 69 data representing patients who passed away. Due to an imbalance in the data, the Synthetic Minority Oversampling Technique (SMOTE) process is applied to balance the dataset. Multiple machine learning algorithms, including Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), are employed for prediction. Validation is performed using 5-fold cross validation, repeated three times. The results indicate that the KNN model achieves the highest mean accuracy of 84.80% before the SMOTE process, although all models exhibit a low mean F1-score. Following the SMOTE process, the RF model attains the highest mean accuracy of 79.52%, while the KNN model demonstrates the highest mean F1-score of 26.54%. This research contributes valuable insights to clinicians in making informed decisions and improving patient outcomes.

Keywords: lung cancer, machine learning, SMOTE, thoracic surgery

INTRODUCTION

Cancer is a spectrum of illnesses characterized by uncontrolled growth and spread of abnormal cells [1]. Death may result if the spread of these cells is not stopped. The Asian continent is the largest contributor to both new cancer cases and cancer deaths. According to estimates, 21.7 million new cases of cancer will be diagnosed by 2030, 13 million of which will result in death [2]. Cancer cells are malignant and have the ability to spread through the blood or lymph nodes. One type of cancer is lung cancer. The abnormal proliferation of the epithelial cells that line the lungs’ airways is the cause of this malignancy. Active smokers have a 10-fold higher chance of acquiring lung cancer compared to non-smokers. According to the Global Cancer Observatory, there were 2,206,771 new lung cancer cases in 2020 [3]. Lung cancer alone has caused the deaths of 1,796,144 people in the world.

One of the diagnostic and treatment procedures for lung cancer is thoracic surgery. Thoracic surgery is performed if the cancer is still at an early stage. This surgery is done to inhibit the spread of cancer cells by removing the tumor and some other healthy tissue. The health condition and lung function of patients after thoracic surgery are the main focus in predicting their life expectancy. Lung cancer remains a significant global health challenge, with high mortality rates even after thoracic surgery. Recognizing the importance of early treatment interventions in reducing mortality rates, this study aims to explore the potential of utilizing a combination of the Synthetic Minority Oversampling Technique (SMOTE) and machine learning algorithms for predicting the life expectancy of lung cancer patients post thoracic surgery.

This research uses a two-stage approach. First, This research apply SMOTE, a resampling technique, to address the imbalance data in lung cancer dataset. Imbalanced data is a situation where one class has fewer observations...
Predicting life expectancy of lung cancer patients after thoracic surgery

(Selly Anastasia Amelia Kharis, Arman Haqqi Anna Zili)

This study uses K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) for classification data. K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) have been widely used to make a classification or prediction using health or disease data. Data on diseases have long been classified using this KNN Method. In 2012, pulmonary disease was classified using this approach [9], and it is still utilized to categorize or early detection of COVID data in 2022 [10]. Random forest already used for classification of Alzheimer disease data since 2013 [11] and still used by researcher in 2022 to make a classification of cancer data [12]. While SVM method has long been used to classify disease data. SVM method was used to classify Alzheimer disease data in 2009 [13] and is still relevantly used to classify skin tumour disease in 2022 [14]. The KNN, Random Forest, and Support Vector Machine methods are part of guided learning that studies the relationship between the independent variable and the target variable [15].

The KNN method uses a measure of similarity between agencies using the Euclidean distance formula [16]. One of the advantages of the KNN method is that it is relatively simple and nonparametric [17]. The limitation is KNN encounters the challenge known as the Curse of Dimensionality, resulting in irrelevant features becoming problematic [18]. KNN remains a relevant and useful algorithm for classification in health or disease data despite its drawbacks, including sensitivity to the choice of k and the curse of dimensionality for high-dimensional data. This is because of KNN's simplicity, adaptability, interpretability, and capacity for handling various data types.

In the field of bioinformatics, Random Forest is a popular choice due to its nonparametric nature, applicability to various datasets, efficiency, and high predictive accuracy across different types of data. Due to its high accuracy, capability to handle high-dimensional and non-linear data, robustness against overfitting, calculation of the relevance of features, management of imbalanced datasets, and potential interpretability, Random Forest is also relevant for classification in health or disease data. These capabilities make Random Forest an invaluable tool for delving into complicated health datasets and assisting in the diagnosis and prognosis of diseases, as well as the identification of pertinent traits and variables linked to certain disorders. In addition, Random Forest has the advantage of being able to manage small data sizes, high-dimensional feature spaces, and complex data structures.
[19]. However, it may struggle with small sample sizes, resulting in overfitting or reduced predictive accuracy [20].

Another method used in this research is the support vector machine. A decision boundary known as a hyperplane will be created using the SVM approach. Finding a hyperplane that optimizes margins poses the difficulty in SVM [21]. SVM has a simple and flexible method to overcome various classification problems [22]. SVM are useful for analyzing complex health datasets and assisting in disease diagnosis or prognosis due to their capacity to handle non-linearity, high-dimensional data, robustness against overfitting, capacity to handle imbalanced datasets, and potential interpretability. However, determining the optimal hyperplane to maximize margins becomes computationally intensive, particularly for large datasets [23].

This study aims to predict thoracic surgery outcomes using machine learning techniques, specifically focusing in the impact of preprocessing methods and addressing the class imbalance challenge. While previous research has investigated the prediction of surgical outcomes, this study offers several novel contributions that differentiate it from previous studies. Firstly, this study incorporates preprocessing techniques, including data cleansing, normalization, and the utilization of the SMOTE. These preprocessing steps are crucial in handling real world healthcare datasets characterized by class imbalance and noise. Secondly, the study emphasized the comparison of the multiple supervised learning algorithms, namely Random Forest, KNN, and SVM. This comparative analysis enables a comprehensive assessment of the model’s performance and provides insights into their strengths and weaknesses in predicting thoracic surgery outcomes. These factors distinguish it from predicting thoracic surgery outcomes using machine learning methods.

One limitation of using SMOTE is the potential introduction of synthetic samples that may not entirely represent the true distribution of the minority class. This study took steps to carefully evaluate the impact of synthetic samples on the predictive model’s performance. We conducted extensive validation and testing procedures to assess the generalizability of our predictions and mitigate the risk of bias introduced by the synthetic samples. By doing so, we aimed to ensure that our model’s predictions accurately reflect the life expectancy of lung cancer patients.

By acknowledging and addressing these limitations, we aimed to establish a clear connection between the specific case of utilizing a lung cancer dataset to predict life expectancy and the challenges associated with SMOTE and machine learning techniques. By elucidating the significance of employing SMOTE and machine learning, this study aims to contribute to the field of predictive medicine and emphasize the necessity of implementing accurate and robust methods for predicting the life expectancy of lung cancer patients post thoracic surgery.

**METHODODOLOGY**

**Material**

The data used in this study is the Thoracic Surgery Data Set from the UCI Machine Learning Repository [24]. The data consists of 454 instances, which are divided into 2 classes, namely, 69 samples who passed away within one year (died) and 385 samples who were able to

---

**Table 1. Attributes of the dataset.**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGN</td>
<td>Diagnosis-specific combination of ICD-10 codes for primary and secondary tumors and more than one tumor, if any</td>
<td>Nominal</td>
</tr>
<tr>
<td>PRE4</td>
<td>The amount of air forcibly exhaled from the lungs after taking as deep a breath as possible is known as the Forced Vital Capacity (FVC)</td>
<td>Numerical</td>
</tr>
<tr>
<td>PRE5</td>
<td>The amount of air that has been exhaled at the end of the first second of FVC (FEV1)</td>
<td>Numerical</td>
</tr>
<tr>
<td>PRE6</td>
<td>A measure of the general ability of cancer patients in daily activities</td>
<td>Nominal</td>
</tr>
<tr>
<td>PRE7</td>
<td>Pain before surgery</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE8</td>
<td>Hemoptysis before surgery</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE9</td>
<td>Dyspnea before surgery</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE10</td>
<td>Cough before surgery</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE11</td>
<td>Weak condition before surgery</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE14</td>
<td>Tumor size (TNM)</td>
<td>Nominal</td>
</tr>
<tr>
<td>PRE17</td>
<td>Diabetes</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE19</td>
<td>Myocardial infarction (MI) up to 6 months</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE25</td>
<td>Diseases that attack the arteries/blood flow (PAD)</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE30</td>
<td>Smoke</td>
<td>Binary</td>
</tr>
<tr>
<td>PRE32</td>
<td>Asthma</td>
<td>Binary</td>
</tr>
<tr>
<td>AGE</td>
<td>Age at surgery</td>
<td>Numerical</td>
</tr>
</tbody>
</table>
Figure 1. Research Stages

Data collection → Data cleaning → Divide data into training and testing data → Data normalization → Train models with training data → Testing the model with testing data → Evaluation

Research Method

The stages of this research are shown in Figure 1. The first step in this research is data collection. After that, data is cleaned first to detect errors and inconsistencies in the data. It is means that the primary purpose of data cleansing is to identify any errors or inconsistencies that may exist within the dataset. This involves examining the dataset for various types of issues, such as missing values, outliers, duplicate records, incorrect data formats, and conflicting information [25]. Furthermore, because it has different data types, nominal and binary data are transformed into numeric data so that all data has the same data type, namely numerical data type. Furthermore, the data is normalized, namely by changing the size or scale of a variable in the data so that each has the same size. Normalization will scale back to the new range while maintaining the difference between the values of the variable.

Data normalization is a preprocessing technique used to ensure that each feature in the dataset contributes equally. The quality of the data plays a crucial role in the success of machine learning algorithms for creating accurate and generalized predictive models. Numerous studies have highlighted the significance of data normalization in enhancing data quality and subsequently improving the performance of machine learning algorithms [26]. One of the normalization techniques is Min-Max scaling. The Min-max scaling technique is performed using the formula (1).

\[ a'_i = \frac{a_i - \text{min}(a)}{\text{max}(a) - \text{min}(a)} \times (A - B) + B \]  

(1)

where, \( a'_i \) represents the value of the i-th observation on variable \( a \) after normalization. \( a_i \) represents the value of the i-th observation on variable \( a \) before normalization, \( \text{min}(a) \) represents the minimum value on variable \( a \), \( \text{max}(a) \) represents the maximum value for variables \( a \), \( A \) and \( B \) is the range, with \( A > B \) [27]. In this study, a range from 0 to 1 will be used so that formula (1) becomes formula (2).

\[ a'_i = \frac{a_i - \text{min}(a)}{\text{max}(a) - \text{min}(a)} \]  

(2)

After the data is normalized, the data is divided by a cross-validation technique. Cross validation is a method used to assess a model's performance. The dataset is split into two pieces by cross validation: training data and test data. K-fold cross validation is one of the most popular cross validation methods [28]. This study will employ five-fold cross validation, with each step being repeated three times.

Data generation on SMOTE using KNN is to find the nearest neighbor based on distance. The distance used in this study uses the Euclidean distance formula in (3).

\[ d(p,q) = \sqrt{\sum_{n=1}^{N}(q_n - p_n)^2} \]  

(3)

The scalar distance between observations \( p \) and \( q \) is \( d(p,q) \), the magnitude of the observation value on the \( n \)-th variable is \( p_n \), and the magnitude of the observation value of \( q \) on the \( n \)-th variable is \( q_n \), and \( N \) is the number of variables in the data.

Since the SMOTE process makes use of KNN, it is required to ascertain the quantity \( k \) or the number of nearest neighbors. In this study, the value of \( k = 5 \) is used because it is the \( k \) value commonly used for the SMOTE method [6].

Data generation using SMOTE using formula (4).

\[ x^*_n = x^0_n + (\text{rand}(0,1) \times (x^m_n - x^0_n)) \]  

(4)

Where \( x^*_n \) is the synthesis data from the SMOTE process on the \( n \)-th variable, \( x^0_n \) is the object of observation in the minority class on the \( n \)-th variable, \( x^m_n \) is the minority class observation on the \( n \)-th variable which is chosen randomly based on the closest neighbors. with the object of observation \( x^0_n \), \( n \) is a variable in
the data \( n \in 1, 2, \ldots, N \) and \( \text{rand}(0,1) \) is a random number with an interval of 0 to 1.

The step by step procedure to generating additional synthesis data using SMOTE method are as follows [29]:

1. Determine the object of observation in the minority class
2. Calculate the distance between the minority class observation objects and all other minority classes
3. Determine the closest \( k \) observations to the object of observation
4. Choose one observation at random based on the closest \( k \) observations
5. Generating synthesis data using equation (5)

The SMOTE method algorithm also described by the flowchart in the Figure 2.

The next stage is machine learning using training data using the KNN, Random Forest, and Support Vector Machine methods to predict the life expectancy of lung cancer patients after 1-year thoracic surgery.

Accuracy, recall, precision, and F1-score in the context of machine learning are commonly used metrics to evaluate the performance of classification models. Accuracy is a measure of how correctly the model predicts the labels or classes of the samples in the dataset. It is calculated as the ratio of the number of correctly classified samples to the total number of samples [30]. While accuracy provides an overall assessment of the model’s performance. After the normal data, the SMOTE process is carried out.

Recall is a metric that measures the ability of a model to correctly identify all positive instances from a dataset. It is defined as the ratio of the number of true positive predictions to the sum of true positive and false negative predictions. In other words, recall measures how well a model detects positive cases when they are present [31]. Precision is a metric that measures the ability of a model to correctly identify only the positive instances from a dataset. It is defined as the ratio of the number of true positive predictions to the sum of true positive and false positive predictions [32]. In other words, precision measures how often a model’s positive predictions are correct. F1-score is a metric that combines precision and recall into a single score that provides a balanced evaluation of a model’s performance. It is the harmonic mean of precision and recall [33], and is defined in (5).

\[
\frac{F_1 \text{score}}{2} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} (5)
\]

**RESULTS AND DISCUSSION**

**Preprocessing data**

Preprocessing carried out in this study consisted of data cleansing and data normalization. After checking the data, no empty data and duplicate data were found. Furthermore, data normalization is performed using Min-max scaling. After the normal data, the SMOTE process is carried out. The data after preprocessing the data amounted to 454 with the ratio of patients surviving and dying around 40: 7 or 85%: 15%. The class distribution of the thoracic surgery data is visible in Figure 3.

---

**Figure 2. The Flowchart of SMOTE Algorithm**
Data balancing with SMOTE

The amount of data that is not balanced causes the need to balance the data. The data balancing method used in this study uses SMOTE. SMOTE balances the data by generating data on the minority class. The additional synthesis data that has been generated is then combined with the entire original data. In this research, the generation of synthesis data was carried out in the dead class because it is a minority class. The generation of minority class data was carried out until the number of it is equal to the majority class data which is illustrated in Figure 4.

Results before SMOTE

The preprocessed data is then fitted to the Random Forest, K-Nearest Neighbor and Support Vector Machine models. The model generates predictive data that is compared to the testing dataset, resulting in a confusion matrix to draw conclusions. From the confusion matrix, four measurements are calculated and commonly used in predictions process using machine learning, they are accuracy, recall, precision and F1-score. Since the process repeated three times, the value of accuracy, recall, precision and F1-score presented here is the mean value. The mean of the four measurements is presented in Table 2.

Results after SMOTE

The imbalance data is then going through the SMOTE process so that it is become a balance data. The balance data once again fitted into the three models to be concluded into a confusion matrix. From the confusion matrix, the new mean accuracy, mean recall, mean precision and mean F1-score presented in Table 3.

Before SMOTE vs After SMOTE

According to result presented in the Table 2 and 3, a comparison between the three models is possible before and after applying SMOTE. The comparison of Random Forest, K-Nearest Neighbor and Support Vector Machine presented in Figure 5, 6 and 7 respectively.
Based on Figure 5, 6, and 7, the model accuracy decreased after the SMOTE process, but it became more reliable as indicated by the increasing values of Recall, Precision and F1-score. So, the model performance after SMOTE is better than before SMOTE although the accuracy value decreasing but it become more reliable.

**Model Performance Comparison**

Four measurements of the model performance before the SMOTE process in Table 2 are represented in a graph shown in Figure 8.

The highest accuracy refers to the highest percentage of correct predictions made by a machine learning model. It measures the overall correctness of the model’s predictions by comparing the to the actual values. A higher accuracy indicates that the model is making more correct predictions [34]. On the other hand, the F1-score is a measure of a model’s accuracy, specifically for imbalanced datasets or situations where one class is more prevalent than the other [35]. It considers both precision and recall to provide a single score that balances the trade-off between them.

If the performance is only analyzed from the accuracy value, KNN is the best model. But accuracy value is not the right or dependable measurement if the data used is imbalance. When the prior probability of the classes is extremely different, accuracy in classification problems might be deceiving [36]. A more appropriate measure to use in this case is recall or precision value. This study uses the F1-score which includes the values of recall and precision as a measure of model performance. When it is viewed from the F1-score, it is evident that the three models that have been applied to the original data prior to the SMOTE process did not perform well enough because all of them had small F1-scores. However, SVM model is the best in three terms: recall, precision and F1-score.

The four measurements of the model performance after the SMOTE process in Table 3 are depicted in a graph shown in Figure 9.
Predicting life expectancy of lung cancer patients after thoracic surgery...
(Selly Anastassia Amellia Kharis, Arman Haqqi Anna Zili)

The accuracy of RF model slightly decreased while accuracy of SVM and KNN model heavily decreased. However, the recall, precision and F1-score of all models now are greater than before. If the performance is now analyzed from the recall, precision and F1-score value, KNN is still the best model even the accuracy is just around 60 percent. Another perspective leads to the conclusion that Random Forest is the best model since the accuracy value still around 80 percent while the F1-score greatly increased after SMOTE. However, the F1-score of RF model is the smallest among those three.

Anis and Ali in their research revealed that the imbalance ratio of a dataset significant impacts the results and introduces bias. Balanced dataset showed limited performance improvements even after generating synthetic instance. On the other hand, extremely imbalanced datasets performed poorly even with a 100% increase in synthetic samples, proving unsatisfactory in addressing the issue [37]. Therefore, it is essential not only generate minority class samples but also to remove certain majority class samples and bias features without disturbing the original data distribution.

CONCLUSION

Based on the simulations that have been carried out regarding the implementation of machine learning to predict life expectancy of lung cancer patients after thoracic surgery, several conclusions have been obtained. Predictions made without using SMOTE have lower performance according to the value of recall, precision and recall compared to the predictions after SMOTE. Before SMOTE, The KNN model is the best model in terms of accuracy while the SVM model is the best in terms of recall, precision and F1-score. The use of SMOTE in balancing data has an impact to the accuracy value of the models. The accuracy value of all models decreased but the recall, precision and F1-score value increased. After SMOTE, The RF model is the best model in term of accuracy but with the lowest F1-score while the KNN model is the best model in terms of recall, precision and F1-score.

Based on the assessment findings, there are several potential areas for future research and improvement in the prediction of thoracic surgery outcomes using machine learning techniques. First, enhanced handling of imbalanced data. Although SMOTE was effective in improving the performance of the models, further exploration of advanced techniques for handling class imbalance should be considered. Second, feature engineering and selection. Investigating additional relevant features or performing feature engineering could potentially enhance the performance of the models. Third, advanced ensemble techniques. Ensemble methods, such as boosting or stacking, could be explored to further enhance the predictive capabilities of the models. Combining the strengths of multiple classifiers leads to improved generalization and better handling of class imbalance, potentially yielding more accurate predictions for thoracic surgery outcomes. Addressing these suggestions in future research would advance the field of predicting life expectancy for lung cancer patients after thoracic surgery, resulting in improved clinical decision-making and patient care. For future research, enhancing the study by engaging with specialist and incorporating primary data holds great promise for its application in the medical field. This approach will not only elucidate the study’s advantages and contributions but also render it more pertinent and impactful for clinical utilization.

ACKNOWLEDGMENT

This research is funded by Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Terbuka (LPPM UT). We would like to express our gratitude to LPPM UT, without whose support this research would not have been possible.

REFERENCES


cic+Sur+gery+Data


