Prediction of Post-Operative Survival Expectancy in Thoracic Lung Cancer Surgery Using Extreme Learning Machine and SMOTE

Ajwa Helisa, Triando Hamonangan Saragih, Irwan Budiman, Fatma Indriani, Dwi Kartini
Computer Science Lambung Mangkurat University, Jalan A.Yani Km 36, Banjarbaru 70714, Indonesia

ARTICLE INFO

Article history:

Received March 21, 2023
Revised April 17, 2023
Published April 19, 2023

Keywords:

Classification;
Extreme learning machine;
Smote;
Imbalance data;
Lung cancer;
Thoracic surgery

ABSTRACT

Lung cancer is the most common cause of cancer death globally. Thoracic surgery is a common treatment for patients with lung cancer. However, there are many risks and postoperative complications leading to death. In this study, we will predict life expectancy for lung cancer patients one year after thoracic surgery The data used is secondary data for lung cancer patients in 2007-2011. There are 470 data consisting of 70 death class data and 400 survival class data for one year after surgery. The algorithm used is Extreme learning machine (ELM) for classification, which tends to be fast in the learning process and has good generalization performance. Synthetic Minority Over-sampling (SMOTE) is used to solve the problem of imbalanced data. The proposed solution combines the benefits of using SMOTE for imbalanced data along with ELM. The results show ELM and SMOTE outperform other algorithms such as Naïve Bayes, Decision stump, J48, and Random Forest. The best results on ELM were obtained at 50 neurons with 89.1% accuracy, F-Measure 0.86, and ROC 0.794. In the combination of ELM and SMOTE, the accuracy is 85.22%, F-measure 0.864, and ROC 0.855 on neuron 45 using a data division proportion of 90:10. The test results show that the proposed method can significantly improve the performance of the ELM algorithm in overcoming class imbalance. The contribution of this study is to build a machine learning model with good performance so that it can be a support system for medical informatics experts and doctors in early detection to predict the life expectancy of lung cancer patients.

Corresponding Author:

Triando Hamonangan Saragih, Computer Science Lambung Mangkurat University, Banjarbaru 70714, Indonesia
Email: triando.saragih@ulm.ac.id

1. INTRODUCTION

Lung cancer is the most common cause of cancer death globally [1]–[3]. In 2013 there were 14.9 million cancer cases, with 8.2 million deaths, and 196.3 million people with disabilities. Cancers of the trachea, bronchi, and lungs are the leading cause of cancer deaths in men and women, with 1.6 million deaths [1]. World Health Organization (WHO) report states that 9.6 million people died from cancer in 2018, making it the second largest cause of death worldwide [4]. The most common cause of impairment in men, accounting for 24.9 million cases, is lung, tracheal, and bronchial cancer [1]. Thoracic surgery is typically used to treat lung cancer patients [5]. Postoperative care of thoracic surgical patients is a crucial aspect of patient recovery and a challenge in its own right [6]. Based on these problems, there is a need to classify patient deaths after thoracic surgery to assist physicians in early detection so that diagnosis can be more efficient and effective.

Over the past decades, various machine learning algorithms and classification methods for disease prediction and prognosis have been applied. Extensive research has been conducted on machine learning algorithms in cancer detection, recurrence, and survival prediction [7]. In research on the classification of life expectancy after lung cancer surgery with the same data, Zie̡ba *et al.* [8] used the Boosted SVM method by applying an oracle-based approach to extract decision rules in solving unbalanced data problems. Another by

Roshan *et al.* [9] used four algorithms: Naive Bayes, Random Forest, J48, and Decision Stump. The combination of J48 and Naive Bayes yielded the best results ROC of 0.738 and an accuracy of 88.73%. There are still few studies and clear guidelines in this domain using machine learning approaches to predict postoperative life expectancy in lung cancer patients [5].

Several Neural Network training methods provide deeper learning to make decisions on detection systems, one of which is the Extreme Learning Machine (ELM) [10]. ELM is a shallow network and many advantages as fast learning, easy convergence, and less randomness. Previous research conducted by Ghoneim *et al.* [11] Extreme Learning Machine for cervical cancer classification to get 99.7% accuracy in classifying two classes and 97.2% accuracy for seven classes. Other research used Extreme Learning Machine by Lahoura *et al.* [12] for breast cancer diagnosis with an accuracy of 0.9868.

The main difficulty in learning classification models is the character of the data. Usually, the collected raw data cannot be used directly in the training process due to various circumstances. One of them is the data imbalance problem [8]. Imbalances in the data set severely affect the performance of most classifiers [13], [14]. Unbalanced data between classes continues to be a common and challenging problem in supervised learning [15] because standard classification algorithms are designed to handle balanced class distributions [16]. In medical decision-making, especially postoperative risk assessment, there are many problems related to data imbalances. During the 1-year planning period, the number of surviving patients is often much higher than the number of deaths during the assumed interval [8]. It will affect the classification results because it tends to be biased toward the majority class [17].

Many oversampling techniques have proven effective in addressing the problem of data imbalances in the real-world domain. SMOTE is the most popular over-sampling method proposed to improve random oversampling [16]. Research related to the SMOTE method by Wei *et al.* [18] to predict the level of drug risk based on ADRs using four classification models: Random Forest, Gradient Boost, Logistic Regression, and AdaBoost, combined with SMOTE, the best results from Random Forest in combination with SMOTE were used was found with an accuracy score of 0.95. Another study by Ishaq *et al.* [19] used nine classification models to predict the survival of heart patients, based on research showing that the SMOTE technique significantly improved the performance of some classifier algorithms and ETC outperformed other models with an accuracy of 0.9262.

Based on some of the above problems, this study will use the Extreme Learning Machine to predict the survival of postoperative patients, and SMOTE to overcome the problem of unbalanced data can be expected to be able to provide optimal solutions so that it can predict the survival of postoperative thoracic lung cancer patients with good results. This study will be a support system for medical informatics and doctors to analyze retrospective data by utilizing large amounts of data that are often collected in daily operations to extract useful information, hoping to contribute to early detection to predict the life expectancy of lung cancer patients based on several indicators to determine the most appropriate treatment for patients and minimize the risk of death in postoperative.

2. RESEARCH METHOD

The main objective of this research is to predict survival in lung cancer patients after thoracic surgery. The data used is unbalanced, so this research wants to know the effect of SMOTE on classification results using ELM. The proposed system for predicting the survival of postoperative thoracic lung cancer patients can be seen in Fig. 1.’

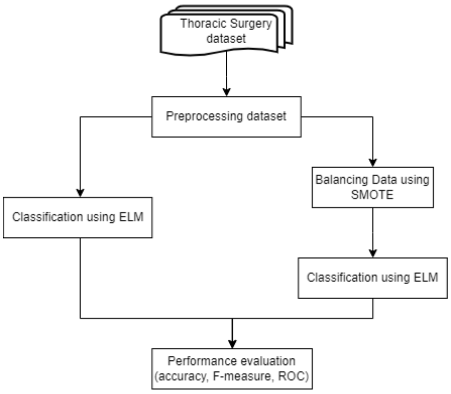

Fig. 1. The proposed system architecture

2.1. Data Collection

The data set is from the UCI Machine Learning Repository. According to the main repository site, the data was collected retrospectively at Wroclaw Thoracic Surgery Center for primary lung cancer patients in 2007-2011. The biomedical data used consisted of 470 samples and were based on classification issues related to postoperative life expectancy in lung cancer patients. It consists of two classes: death or survival within one year after surgery. There were 70 death class data and 400 survival class data for one year after surgery. In this work, the data is divided into training and testing.

Based on clinical studies was identified that there are many risks to patients after thoracic surgery, most common contributing factors are age, preoperative pulmonary function test, cardiovascular comorbidities, chronic obstructive pulmonary disease, and smoking status [17]. In this dataset, 16 features are indicative of lung cancer patients can be seen in Table 1.

Table 1. Descriptions of Thoracic Surgery dataset attributes

No	ID	Description	Category	Range Values
1	DGN	ICD-10 code diagnostic-specific combinations for primary and secondary tumors as well as multiple tumors, if present	Nominal	{DGN1, DGN2, DGN4, DGN5, DGN6, DGN8, DGN3}
2	PRE4	*FVC - Forced vital capacity*	Numeric	{1.44, 6.3}
3	PRE5	FEV1 - Volume is exhaled at the end of the first second of forced expiration.	Numeric	{0.96, 86.3}
4	PRE6	Performance status - Zubrod scale	Nominal	{PRZ0, PRZ1, PRZ2}
5	PRE7	Pain before surgery	Nominal	{T or F}
6	PRE8	Haemoptysis before surgery	Nominal	{T or F}
7	PRE9	Dyspnoea before surgery	Nominal	{T or F}
8	PRE10	Cough before surgery	Nominal	{T or F}
9	PRE11	Weakness before surgery	Nominal	{T or F}
10	PRE14	T in clinical TNM - size of the original tumor, from smallest to largest	Nominal	{OC11, OC12, OC13, OC14}
11	PRE17	Type 2 DM - diabetes mellitus	Nominal	{T or F}
12	PRE19	MI up to 6 months	Nominal	{T or F}
13	PRE25	PAD - peripheral arterial diseases	Nominal	{T or F}
14	PRE30	Smoking	Nominal	{T or F}
15	PRE32	Asthma	Nominal	{T or F}
16	AGE	Age at surgery	Numeric	{21, 87}
17	Risk1Y	1 year survival period - (T)rue value if died (T, F)	Nominal	{T or F}

2.2. Preprocessing Data

Preprocessing is a step used in data mining to transform raw data into a form that is easy to understand so that it can represent data effectively so that there is not much excessive information that is not related or noisy [20]. Data preprocessing in this research consists of transforming the data from nominal and binary to numeric so that the algorithm can process it. Outliers were removed after data transformation. Outliers are data that are significantly different and do not correspond to the normal behavior of the data [21]. There are 16 outliers in the data, and after removing these outliers the new data set contains 454 instances from the original 470. This data does not have such problems as missing values and duplicate data. The research uses split data with a proportion of 60:40, 70:30, 80:20, and 90:10.

2.3. Extreme Learning Machine

Extreme Learning Machine is one of the learning methods in a feedforward neural network with a single hidden layer. ELM is a least-square-based learning algorithm that can be applied as the estimator in regression or classification problems [22]. The algorithm consists of three layers: input, hidden, and output layer [23]–[25]. The input weight in the ELM is determined randomly, while the output weight is determined analytically so that the output on the results of this algorithm is only one [10]. ELM has two parameters set by the user, the number of neurons of the hidden layer (L) and the variance of the input weight of the hidden layer (w). Improper initialization of (L) or (w) may interfere with the performance of ELM models [26]. The ELM model can be seen in Fig. 2.

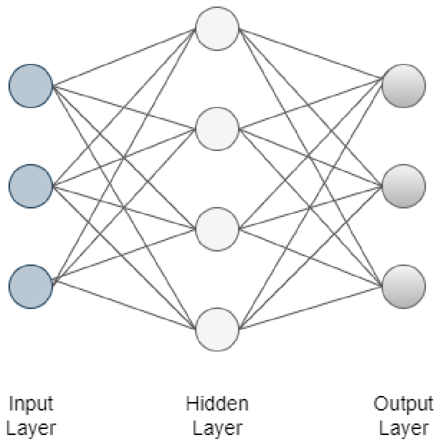

Fig. 2. ELM architecture model

In ELM, during the training process, the weight of the hidden layer is randomly assigned but never updated, and only the weight of the output layer is changed. The training dataset is (x_j, t_j), where $x_j = [x_{j1}, x_{j2}, \ldots, x_{jN}]^T$is the input vector and t_j is the output vector. The output of the i^{th} hidden layer neuron is given by $g(w_i, b_i, x_j)$, where w_i is the weight vector connecting the input neuron to the i^{th} hidden neuron, b_i is the bias of the i^{th} hidden neuron, and g is the activation function. Each hidden layer neuron in ELM is also connected to each output layer neuron with some associated weights, denoting the weights connecting the i^{th} hidden layer neuron to the output neuron with b_i. This architecture can be mathematically described as in (1).

$$\sum_{i=1}^{L} \beta_i g(w_i, b_i, x_j) = t_j \tag{1}$$

where, L is the number of hidden neurons, and j is the input/output instances of total N training instances.

2.4. SMOTE

Synthetic Minority Over-sampling (SMOTE) proposed by Chawla is an improved scheme based on a random oversampling algorithm [27], [28]. SMOTE is a popular and effective method to handle the class imbalance problem in many domains. The goal of SMOTE is to overcome the overfitting rendered by simply oversampling with replication and help classifiers to improve generalization on testing data [29]. The basic idea of the SMOTE algorithm is to analyze the minority class samples and add new artificially synthesized samples corresponding to the minority class samples to the data set [30], [31].

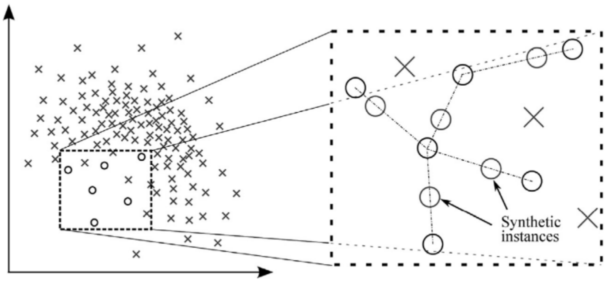

Fig. 3. Generation of Synthetic Instances using SMOTE [32]

The steps in SMOTE are as follows [33]:

1. Sample x_i is randomly selected from S_{min}. The Euclidean distance is used as a standard to calculate its distance to other samples in S_{min}, and its k nearest neighbors are determined.
2. Sample x_j is randomly selected from the k neighbors, and a new sample x_{new} is generated using equation (2).

$$x_{new} = x_i + rand(0,1) \times |x_i - x_j| \tag{2}$$

3. Repeat the first two steps until the number of samples for the category in the training set equals the largest number of samples.

2.5. Assessment Indices

In this work, evaluation for classification uses accuracy, F-Measure, and ROC to predict the survival of lung cancer patients following thoracic surgery. The existence of an assessment index aims to measure the model's success rate and facilitates comparison with other algorithms.

Accuracy is used to measure the closeness between predicted and actual values. The formulas are shown in (3).

$$Accuracy = \frac{correctprediction}{totalprediction} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

F-measure is an assessment index using an average that combines precision and recall values [34]. The F-measure calculation is shown in (4)

$$F - Measure = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

ROC (Receiver operating characteristic) is a graph or curve used to measure classifiers and visualize their performance. The goal is to compare diagnostic tests. ROC is widely used in machine learning, decision-making, and data mining [35]. The closer the ROC plot is to the upper left corner, approaching the values of 100% true positive rate (TPR) and 0% false positive rate (FPR) indicates high resulting accuracy [36]. ROC is obtained by calculating the area under the ROC curve that plots FPR as (5) on the x axis and TPR as (6) for the y axis [37].

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

True positives (TP) are the rate of positive tuples correctly labeled as positive by the classifier. False positives (FP) are the rate of negative tuples falsely flagged as positive. True Negative (TN) is the rate of negative tuples correctly labeled as negative. False negatives (FN) are the fraction of positive tuples incorrectly labeled as negative [35].

3. RESULTS AND DISCUSSION

This study is limited to using only ELM as a classification algorithm and SMOTE to overcome the challenge of unbalanced data. In addition, the dataset used focuses only on a single source retrospectively collected at the Wroclaw Center for Thoracic Surgery: postoperative survival data for lung cancer patients. There are many factors associated with the survival of postoperative lung cancer patients. There are many factors associated with the survival of post-surgical lung cancer patients. There are many factors associated with the survival of post-surgical lung cancer patients. In the data used, many other indicators were not included, and prognostic variables consisting of neoadjuvant therapy, biomarkers, anatomopathological findings, and tumor genome analysis are limitations of this study [38].

Several tests were conducted to determine the proportion of training and testing data that can provide good results based on accuracy, F-Measure, and ROC. The composition of training data and test data used in this work is 60:40, 70:30, 80:20, and 90:10. Additionally, neuron testing was performed to ascertain the number of neurons required in the ELM process to get the best results. This test uses 5 to 50 neurons in multiples of 5. Each neuron is tested ten times, and the resulting accuracy, F-measure, and ROC results are averaged. The first test results for the proportion of training data and test data 60:40 are shown in Table 2.

In Table 2, it can be seen that the composition of 60% training data and 40% test data provides quite good performance results with an average accuracy value of 87.334%, 0.828 for F-Measure, and ROC 0.657 for ELM. In the classification using additional SMOTE, accuracy, and F-Measure decreased with an average accuracy of 70.543 and F-Measure of 0.735, while ROC increased to 0.696. Furthermore, the second test results for the composition of training data and test data 70:30 are shown in Table 3.

Table 2. Experimental results for 60% training data and 40% test data

Neuron	Accuracy		F-Measure		ROC	
	ELM	ELM + SMOTE	ELM	ELM + SMOTE	ELM	ELM + SMOTE
5	86.978	77.582	0.81	0.75	0.551	0.6126
10	86.318	65.879	0.81	0.70	0.598	0.677
15	86.813	66.538	0.81	0.70	0.615	0.674
20	87.966	70.879	0.83	0.73	0.644	0.695
25	87.802	71.263	0.83	0.74	0.667	0.698
30	87.802	70.714	0.83	0.73	0.686	0.701
35	87.582	70.769	0.84	0.75	0.693	0.712
40	87.417	69.065	0.84	0.74	0.698	0.715
45	87.582	71.263	0.84	0.75	0.71	0.74
50	87.088	71.483	0.84	0.76	0.717	0.742
Average	87.334	70.543	0.828	0.735	0.657	0.696

Table 3. Experimental results for 70% training data and 30% test data

Neuron	Accuracy (%)		F-Measure		ROC	
	ELM	ELM + SMOTE	ELM	ELM + SMOTE	ELM	ELM + SMOTE
5	86.593	79.56	0.81	0.78	0.556	0.630
10	85.244	70.069	0.80	0.74	0.625	0.664
15	86.373	65.911	0.81	0.71	0.638	0.717
20	85.057	68.491	0.79	0.73	0.680	0.732
25	84.807	68.175	0..79	0.73	0.690	0.727
30	85.046	70.584	0.80	0.75	0.699	0.735
35	85.569	71.605	0.80	0.75	0.713	0.722
40	86.728	71.751	0.82	0.75	0.715	0.726
45	86.788	72.116	0.82	0.75	0.725	0.726
50	87.618	72.992	0.84	0.76	0.723	0.733
Average	85.982	71.125	0.81	0.745	0.676	0.711

In Table 3, it can be seen that classification using ELM provides lower accuracy and F-Measure compared to the previous test results and there is a slight increase in the ROC value of 0.676, accuracy of 85.982%, and F-Measure of 0.81. While using ELM and SMOTE has improved with 71.125% accuracy, F-Measure 0.745, and ROC 0.711. The highest result in the 70:30 proportion is at the number of neurons as much as 50, accuracy 87.618%, F-Measure 0.84, and ROC 0.723 for ELM and 71.992% accuracy, F-Measure 0.76, ROC 0.733 for ELM and SMOTE. Furthermore, the third test results for the composition of training data and test data 80:20 are shown in Table 4.

Table 4. Experimental results for 80% training data and 20% test data

Neuron	Accuracy		F-Measure		ROC	
	ELM	ELM + SMOTE	ELM	ELM + SMOTE	ELM	ELM + SMOTE
5	84.066	79.560	0.77	0.78	0.579	0.630
10	83.956	78.131	0.78	0.79	0.589	0.661
15	84.615	65.911	0.79	0.71	0.564	0.717
20	84.835	66.373	0.79	0.71	0.633	0.730
25	83.736	65.824	0.78	0.71	0.695	0.749
30	85.384	67.254	0.80	0.72	0.689	0.735
35	87.472	68.792	0.83	0.72	0.670	0.746
40	87.253	71.538	0.83	0.75	0.689	0.752
45	87.142	73.736	0.83	0.76	0.698	0.725
50	87.143	73.626	0.83	0.76	0.679	0.695
Average	85.56	71.074	0.803	0.741	0.648	0.714

Table 4 shows that the composition of 80% training data and 20% testing data gives worse results than the test results of the previous two scenarios, the accuracy of 85.56%, F-Measure 0.803, and ROC 0.648 for ELM. Similarly, for the ELM and SMOTE experiments, there was only a slight improvement in the average ROC value of 0.714, accuracy of 71.074, and F-Measure of 0.741. Finally, the results of testing the 90:10 training and test data composition are shown in Table 5.

Table 5. Experimental results for 90% training data and 10% test data

Neuron	Accuracy		F-Measure		ROC	
	ELM	ELM + SMOTE	ELM	ELM + SMOTE	ELM	ELM + SMOTE
5	84.268	72.79	0.79	0.75	0.595	0.74
10	84.52	76.3	0.79	0.78	0.639	0.79
15	83.86	75.21	0.77	0.78	0.603	0.805
20	83.35	70.2	0.77	0.75	0.673	0.857
25	83.18	76.5	0.77	0.79	0.661	0.801
30	85.17	79.6	0.80	0.82	0.704	0.887
35	84.27	78.3	0.79	0.80	0.728	0.85
40	85.38	74.2	0.80	0.77	0.72	0.81
45	86.46	85.22	0.81	0.86	0.744	0.855
50	89.1	82.3	0.86	0.83	0.794	0.819
Average	84.955	77.062	0.795	0.793	0.686	0.821

The test results in Table 5 show that the accuracy, F-Measure, and ROC values generated using ELM and SMOTE are the best compared to other test scenarios, with an average accuracy value of 77.062%, F-Measure 0.793, and ROC 0.821. In ELM, the average accuracy is 84.955%, F-Measure 0.795, and ROC 0.686. As for the highest results from all experiments with a scenario of 90% training data and 10% test data, there are experiments with the number of neurons 50 using ELM getting the highest results with accuracy reaching 89.1%, F-Measure 0.86, and ROC 0.794. ELM and SMOTE gave the best results in 45 neurons with 85.22% accuracy, F-measure 0.86, and ROC 0.855. Based on these results, it is shown that the composition of the training and test data and the number of neurons influence the classification results given.

To clarify the understanding of all test scenario results, we have provided a graphical comparison of the test results to all scenarios based on accuracy, F-measures, and ROC shown in Fig. 4-Fig. 6.

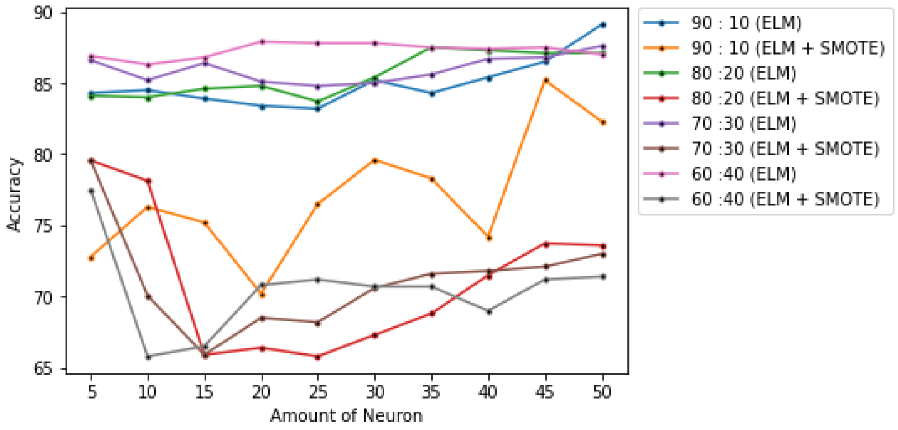

Fig. 4. Graph of accuracy for all test scenarios

The graph shows the results based on accuracy in all experimental scenarios, with the ELM model getting a higher average accuracy than using ELM and SMOTE. The highest accuracy in ELM is 89.1% in the proportion 90:10 with neurons 50. Similarly, with ELM and SMOTE, the best results are in scenarios 90:10 with neurons 45 with an accuracy of 85.22%.

The graph shows the results based on F-measure for all test scenarios. The ELM model shows very small differences in F- measure values. Based on the average F-Measure value, the 60:40 scenario has a higher value than the average F-Measure value with other scenarios because the F-Measure value on each neuron tends to be stable. Meanwhile, if we look at the highest result based on each neuron experiment, the 90:10 scenario is the best at 50 neurons with an F-Measure of 0.86. For ELM and SMOTE models, the best F-Measure value is also in the 90:10 scenario with an F-Measure of 0.864.

The graph based on ROC for all test scenarios shows that the 90:10 data split using ELM and SMOTE has the highest results compared to other scenarios. The highest result with 30 neurons has a ROC value of 0.887. Similarly, using ELM, the highest result is in the 90:10 used 50 neurons having a ROC of 0.794.

Comparison of accuracy, F-Measure, and ROC with other studies was also conducted to measure the performance of Random Forest to predict the life expectancy of lung cancer patients after thoracic surgery. In this test, the composition of training data and test data used is 90:10 to fit the comparative study. Table 6 shows the comparison of accuracy, F-Measure, and ROC values of several algorithms that have been used previously for the same case, as well as the Extreme Learning Machine and SMOTE used in this study.

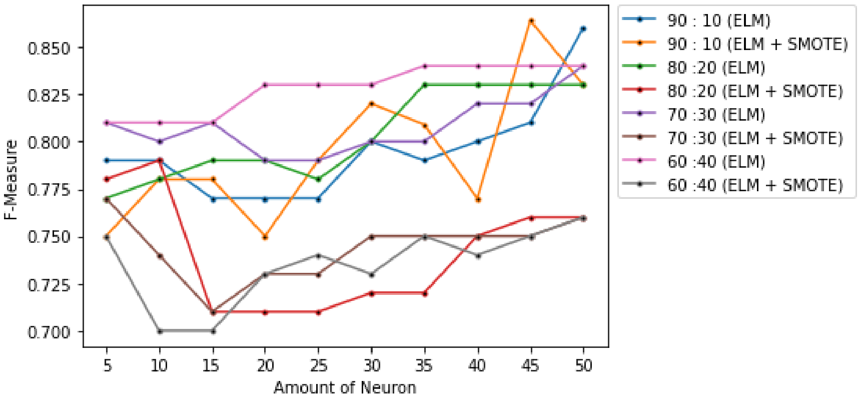

Fig. 5. Graph of F-Measure for all test scenarios

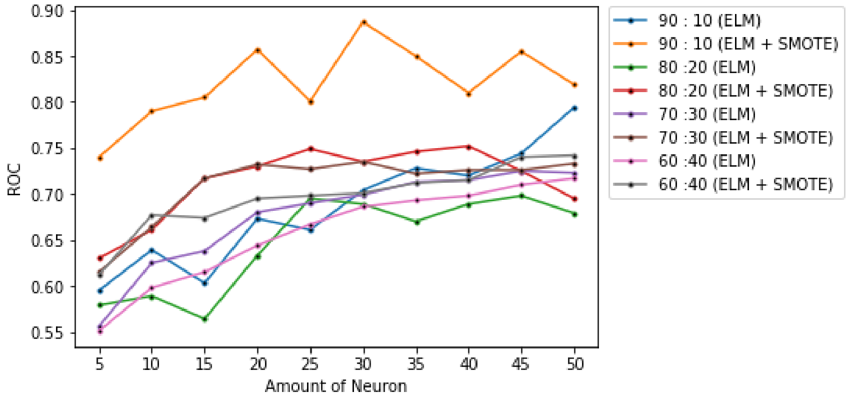

Fig. 6. Graph of ROC for all test scenarios

Table 6. Comparison of accuracy, F-Measure, and ROC with other algorithms

Algorithms	Accuracy (%)	F-Measure	ROC
Naïve Bayes [9]	84.51	0.829	0.738
Decision stump [9]	88.73	0.834	0.493
J48 [9]	88.73	0.834	0.5
Random Forest [9]	88.73	0.834	0.681
J48+ Naïve Bayes [9]	88.73	0.834	0.738
J48+Random Forest [9]	88.73	0.834	0.681
Extreme Learning Machine	89.1	0.86	0.794
Extreme Learning Machine + SMOTE	85.22	0.864	0.855

Based on the comparison results with previous studies using the same data shown in Table 6, it can be seen that ELM and SMOTE provide better results for solving cases related to the life expectancy of lung cancer patients after thoracic surgery. The case in this study shows that the ELM algorithm can work well for classification models by combining SMOTE as a technique for data balancing. SMOTE has a positive impact when the data is not balanced because the balancing process minimizes the possibility of biased learning toward the majority class [39]. The results show that SMOTE can improve the performance of the model, as evidenced by the increased F and ROC values compared with ELM.

However, it should be noted that data is fundamental in intelligent learning and plays a very important role we can improve the results of the proposed system by training new data from sources in various other health centers. Based on this, it shows that the method and dataset used greatly affected the accuracy of the results.

4. CONCLUSION

Based on the test results, it can be concluded that the composition of the training data and test data and the number of neurons influence the classification or prediction results characterized by changes in accuracy, F-Measure, and ROC. Test using four split data scenarios shows different results. Moreover, the more neurons used, the better the ability to classifier is shown by the F-Measure value tends to increase along with the increase in neurons. The test results show that ELM can produce good accuracy and F-Measure and can outperform other algorithms, but the resulting ROC value is not so good, only in a few trials ROC with a good value. It is because unbalanced data causes the classification results to be biased toward the majority class. In addition, the ELM and SMOTE models can increase the ROC results, while the accuracy results in decrease because the data has been balanced with SMOTE so that the classification results are not biased towards certain classes.

This study is limited to using ELM and SMOTE as methods to overcome unbalanced data. In addition, the dataset used focuses only on a single source retrospectively collected at the Wroclaw Center for Thoracic Surgery: postoperative survival data for lung cancer patients. It should be taken into account as the method and dataset used greatly affect the accuracy of the results. In future studies, optimization techniques such as Genetic Algorithms can be considered to determine the optimum values of the parameters and the number of hidden neurons. In addition, using Particle Swarm Optimization (PSO) can help to obtain the optimal number of hidden neurons and selection of input feature subsets to improve ELM performance [40].

REFERENCES

[1] G. B. of Disease Cancer Collaboration, “The Global Burden of Cancer 2013,” *JAMA Oncol.*, vol. 1, no. 4, pp. 505–527, 2015, https://doi.org/10.1001/jamaoncol.2015.0735.

[2] D. D. E. M. A. Detillon, E. J. M. Driessen, M. J. Aarts, M. L. G. Janssen-Heijnen, C. H. J. van Eijck, and E. J. Veen, “Changes in treatment patterns and survival in elderly patients with stage I non–small-cell lung cancer with the introduction of stereotactic body radiotherapy and video-assisted thoracic surgery,” *Eur. J. Cancer*, vol. 101, pp. 30–37, 2018, https://doi.org/10.1016/j.ejca.2018.06.016.

[3] H. Yang and Y.-P. P. Chen, “Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information,” *Expert Syst. Appl.*, vol. 42, no. 15, pp. 6168–6176, 2015, https://doi.org/10.1016/j.eswa.2015.03.019

[4] S. Ronoud and S. Asadi, “An evolutionary deep belief network extreme learning-based for breast cancer diagnosis,” *Soft Comput.*, vol. 23, no. 24, pp. 13139–13159, 2019, https://ijassa.ipu.ru/index.php/ijassa/article/view/351.

[5] A. S. Desuky and L. M. El Bakrawy, “Improved prediction of post-operative life expectancy after thoracic surgery,” *Adv. Syst. Sci. Appl.*, vol. 16, no. 2, pp. 70–80, 2016, https://doi.org/10.5772/55351.

[6] A. Iyer and S. Yadav, “Postoperative Care and Complications After Thoracic Surgery,” in *Principles and Practice of Cardiothoracic Surgery*, *IntechOpen*, pp. 57-84, 2013, https://books.google.co.id/books?id=E3afDwAAQBAJ.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, https://doi.org/10.1016/j.csbj.2014.11.005.

[8] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients,” *Appl. Soft Comput.*, vol. 14, pp. 99–108, 2014, https://doi.org/10.1016/j.asoc.2013.07.016.

[9] S. Roshan and V. Rohini, “Prediction of Post-Surgical Survival of Lung Cancer Patients after Thoracic Surgery using Data Mining Techniques,” *Int. J. Soft Comput.*, vol. 16, no. 3, pp. 34–38, 2021, http://dx.doi.org/10.21474/IJAR01/3852.

[10] T. H. Saragih, D. M. N. Fajri, W. F. Mahmudy, A. L. Abadi, and Y. P. Anggodo, “Jatropha Curcas Disease Identification with Extreme Learning Machine,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 2, pp. 883–888, 2018, https://doi.org/10.11591/ijeecs.v12.i2.pp883-888.

[11] A. Ghoneim, G. Muhammad, and M. S. Hossain, “Cervical cancer classification using convolutional neural networks and extreme learning machines,” *Futur. Gener. Comput. Syst.*, vol. 102, pp. 643–649, 2020, https://doi.org/10.1016/j.future.2019.09.015.

[12] V. Lahoura *et al.*, “Cloud Computing-Based Framework for Breast Cancer Diagnosis Using Extreme Learning Machine,” *Diagnostics*, vol. 11, no. 2, 2021, https://doi.org/10.3390/diagnostics11020241.

[13] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, 2020, https://doi.org/10.1016/j.neucom.2019.10.118.

[14] S. Cateni, V. Colla, and M. Vannucci, “A method for resampling imbalanced datasets in binary classification tasks for real-world problems,” *Neurocomputing*, vol. 135, pp. 32–41, 2014, https://doi.org/10.1016/j.neucom.2013.05.059.

[15] M. Koziarski, “Radial-Based Undersampling for imbalanced data classification,” *Pattern Recognit.*, vol. 102, p. 107262, 2020, https://doi.org/10.1016/j.patcog.2020.107262.

[16] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, https://doi.org/10.1016/j.ins.2018.06.056.

[17] G. Kovács, “Smote-variants: A python implementation of 85 minority oversampling techniques,” *Neurocomputing*, vol. 366, pp. 352–354, 2019, https://doi.org/10.1016/j.neucom.2019.06.100.

[18] J. Wei, Z. Lu, K. Qiu, P. Li, and H. Sun, “Predicting Drug Risk Level from Adverse Drug Reactions Using SMOTE and Machine Learning Approaches,” *IEEE Access*, vol. 8, pp. 185761–185775, 2020, https://doi.org/10.1109/ACCESS.2020.3029446.

[19] A. Ishaq *et al.*, “Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques,” *IEEE Access*, vol. 9, pp. 39707–39716, 2021, https://doi.org/10.1109/ACCESS.2021.3064084.

[20] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, “Data preprocessing in predictive data mining,” *Knowl. Eng. Rev.*, vol. 34, p. e1, 2019, https://doi.org/10.1017/S026988891800036X.

[21] H. Wang, M. J. Bah, and M. Hammad, “Progress in Outlier Detection Techniques: A Survey,” *IEEE Access*, vol. 7, pp. 107964–108000, 2019, https://doi.org/10.1109/ACCESS.2019.2932769.

[22] W. Zong, G.-B. Huang, and Y. Chen, “Weighted extreme learning machine for imbalance learning,” *Neurocomputing*, vol. 101, pp. 229–242, 2013, https://doi.org/10.1016/j.neucom.2012.08.010.

[23] Y. Ma, L. Wu, Y. Guan, and Z. Peng, “The capacity estimation and cycle life prediction of lithium-ion batteries using a new broad extreme learning machine approach,” *J. Power Sources*, vol. 476, p. 228581, 2020, https://doi.org/10.1016/j.jpowsour.2020.228581.

[24] S. Yahia, S. Said, and M. Zaied, “Wavelet extreme learning machine and deep learning for data classification,” *Neurocomputing*, vol. 470, pp. 280–289, 2022, https://doi.org/10.1016/j.neucom.2020.04.158.

[25] W. Cai, J. Yang, Y. Yu, Y. Song, T. Zhou, and J. Qin, “PSO-ELM: A Hybrid Learning Model for Short-Term Traffic Flow Forecasting,” *IEEE Access*, vol. 8, pp. 6505–6514, 2020, https://doi.org/10.1109/ACCESS.2019.2963784.

[26] M. A. Shehab and N. Kahraman, “A weighted voting ensemble of efficient regularized extreme learning machine,” *Comput. Electr. Eng.*, vol. 85, p. 106639, 2020, https://doi.org/10.1016/j.compeleceng.2020.106639.

[27] X. Xiaolong, C. Wen, and S. Yanfei, “Over-sampling algorithm for imbalanced data classification,” *J. Syst. Eng. Electron.*, vol. 30, no. 6, pp. 1182–1191, 2019, https://doi.org/10.21629/JSEE.2019.06.12.

[28] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, “LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM,” *Knowledge-Based Syst.*, vol. 196, p. 105845, 2020, https://doi.org/10.1613/jair.1.11192..

[29] A. Fernández, S. García, F. Herrera, and N. V. Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary,” *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, https://doi.org/10.1613/jair.1.11192.

[30] F. Shen, X. Zhao, G. Kou, and F. E. Alsaadi, “A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique,” *Appl. Soft Comput.*, vol. 98, p. 106852, 2021, https://doi.org/10.1016/j.asoc.2020.106852.

[31] C.-R. Wang and X.-H. Shao, “An Improving Majority Weighted Minority Oversampling Technique for Imbalanced Classification Problem,” *IEEE Access*, vol. 9, pp. 5069–5082, 2021, https://doi.org/10.1109/ACCESS.2020.3047923.

[32] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, “RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, Part A, pp. 5059–5074, 2022, https://doi.org/10.1016/j.jksuci.2022.06.005.

[33] H. Dong, D. He, and F. Wang, “SMOTE-XGBoost using Tree Parzen Estimator optimization for copper flotation method classification,” *Powder Technol.*, vol. 375, pp. 174–181, 2020, https://doi.org/10.1016/j.powtec.2020.07.065.

[34] A. Berger and S. Guda, “Threshold optimization for F measure of macro-averaged precision and recall,” *Pattern Recognit.*, vol. 102, p. 107250, 2020, https://doi.org/10.1016/j.patcog.2020.107250.

[35] A. I. Al-issa, M. Al-Akhras, M. S. ALsahli, and M. Alawairdhi, “Using Machine Learning to Detect DoS Attacks in Wireless Sensor Networks,” in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp. 107–112, 2019, https://doi.org/10.1109/JEEIT.2019.8717400.

[36] C. Scotto, A. Ippolito, and D. Sabbagh, “A method for automatic detection of equatorial spread-F in ionograms,” *Adv. Sp. Res.*, vol. 63, no. 1, pp. 337–342, 2019, https://doi.org/10.1016/j.asr.2018.09.019.

[37] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, “Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data,” in *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 14–18, 2018, https://doi.org/10.1109/IC3INA48034.2019.8949568.

[38] M. S. Iraji, “Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing,” *J. Appl. Biomed.*, vol. 15, no. 2, pp. 151–159, 2017, https://doi.org/10.1016/j.jab.2016.12.001.

[39] F. M. Palechor and A. de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” *Data Br.*, vol. 25, p. 104344, 2019, https://doi.org/10.1016/j.dib.2019.104344.

[40] R. Ahila, V. Sadasivam, and K. Manimala, “An integrated PSO for parameter determination and feature selection of ELM and its application in classification of power system disturbances,” *Appl. Soft Comput.*, vol. 32, pp. 23–37, 2015, https://doi.org/10.1016/j.asoc.2015.03.036.

BIOGRAPHY OF AUTHORS

Ajwa Helisa is an undergraduate student in the Department of Computer Science, Lambung Mangkurat University. Her research interest is centered on Data Mining. Email: ajwahlsa@gmail.com

Triando Hamonangan Saragih is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science. Email: triando.saragih@ulm.ac.id

Irwan Budiman is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science. Email: irwan.budiman@ulm.ac.id

Fatma Indriani is a lecturer in Department of Computer Science, Lambung Mangkurat University. Her research interest is centered on Data Science. Email: f.indriani@ulm.ac.id

Dwi Kartini is a lecturer in Department of Computer Science, Lambung Mangkurat University. Her research interest is centered on Data Science. Email: dwikartini@ulm.ac.id