Recursive Journal of Informatics
Vol. 2, No. 2, September 2024

p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Optimizing Random Forest for Predicting Thoracic Surgery Success in Lung Cancer Using Recursive Feature Elimination and GridSearchCV

Deonisius Germandy Cahaya Putra[1], Anggyi Trisnawan Putra[2]
[1,2]Informatics Engineering Study Program, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Lung cancer is one of the deadliest forms of cancer, claiming numerous lives annually. Thoracic surgery is a strategy to manage lung cancer patients; however, it poses high risks, including potential nerve damage and fatal complications leading to mortality. Predicting the success rate of thoracic surgery for lung cancer patients can be accomplished using data mining techniques based on classification principles. Medical data mining involves employing mathematical, statistical, and computational methods. In this study, the prediction of thoracic surgery success employs the random forest algorithm with recursive feature elimination for feature selection. The feature selection process yields the top 8 features. The 8 best features include 'DGN', 'PRE4', 'PRE5', 'PRE6', 'PRE10', 'PRE14', 'PRE30', and 'AGE'. Hyperparameter using GridSearchCV is then applied to enhance classification accuracy. The results of this method implementation demonstrate a predictive accuracy of 91.41%.

Purpose: The study aims to develop and evaluate a Random Forest model with a Recursive Feature Elimination feature selection and applies hyperparameter GridSearchCV for predicting thoracic surgery success rate.

Methods: This study uses the thoracic surgery dataset and applies various data preprocessing techniques. The dataset is then used for classification using the Random Forest algorithm and applies the Recursive Feature Elimination feature selection to obtain the best features. GridSearchCV is used in this study for hyperparameter.

Result: The accuracy using the Random Forest algorithm and Recursive Feature Elimination feature selection with hyperparameters tuning GridSearchCV resulted in an accuracy of 91,41%. The accuracy was obtained from the following parameters values: bootstrap set to false, criterion set to gini, n_estimator equal to 100, max_depth set to none, min_samples_split equal to 4, min_samples_leaf equal to 1, max_features set to auto, n_jobs set to -1, and verbose set to 2 with 10-fold cross validation.

Novelty: This study comparison and analysis of various dataset preprocessing methods and different model configurations are conducted to find the best model for predicting the success rate of thoracic surgery. The study also employs feature selection to choose the best feature to be used in classification process, as well as hyperparameter tuning to achieve optimal accuracy and discover the optimal values for these hyperparameters.

Keywords: Random Forest, Recursive Feature Elimination, Lung Cancer, Thoracic Surgery, GridSearchCV
Received August 14, 2023 / **Revised** May 02, 2024 / **Accepted** September 17, 2024

INTRODUCTION

The lungs are a respiratory system organ that facilitates oxygen-carbon dioxide exchange in the blood. Unfortunately, many individuals remain indifferent to lung health, often neglecting it, exemplified by unhealthy habits such as smoking. Cigarettes contains over 4,000 chemical compounds and has been conclusively proven to be a leading cause of cancer. Individuals who smoke more than one pack of cigarettes per day face a 20-25 times higher risk of developing lung cancer compared to those who have never smoked [1]. According to the 2014 Sample Registration System (SRS) survey in Indonesia, there were identified 10 of the most lethal diseases, and among them was lung cancer [2]. Lung cancer is a medical condition characterized by the uncontrollable growth of cells within the lungs, attributed to exposure to various carcinogens [3]. This uncontrolled cellular growth is the result of the transformation of normal body tissues into malignant ones, leading to the aggressive spread of these cells to other parts of the body and potentially resulting in fatal outcomes [4].

The management of lung cancer patients can include thoracic surgical procedures. The thorax is the portion of the human body encompassing the chest and back, housing vital organs such as the heart, lungs, and

[1]*Corresponding author.
Email addresses: gerrydeonisius@students.unnes.ac.id (Gerry)
DOI: 10.15294/rji.v2i2.73154

respiratory tract. These vital structures are protected by the thorax. Patients requiring medical intervention, specifically surgical procedures, for lung diseases often undergo thoracic surgery. Thoracic surgery constitutes a medical discipline focused on diagnosing and performing surgical procedures for health disorders stemming from diseases or injuries affecting the esophagus, lungs, or other thoracic organs [5]. However, thoracic surgical procedures carry a significant risk, including nerve system damage, infections, and potentially fatal complications that could lead to death. Many complications occur in patients with cardiovascular diseases, such as heart disorders and vascular disruptions that could result in strokes. Consequently, the success rate of thoracic surgery is notably low. A crucial aspect in deciding upon thoracic surgery is selecting suitable patients, considering both the risks and benefits in the short and long term [6].

Selecting the appropriate patients remains a challenge in the decision-making process for thoracic surgery. Parameters must be taken into consideration to ensure that the risks and benefits to the patients are adequately weighed, both in the short term, such as post-operative complications or the first-month mortality rate, and in the long term, such as 1 to 5-year survival rates. The primary focus of this research is to predict the success rate of thoracic surgery for lung cancer patients using a Computer-Aided Diagnosis (CAD) System. Computer-Aided Diagnosis (CAD) is a technology that employs algorithms and mathematical models to analyze medical data, aiding doctors or medical experts in the disease diagnosis process [7]. The utilization of Computer-Aided Diagnosis (CAD) will assist in forecasting the success rate of thoracic surgery for lung cancer patients by conducting an analysis of the patient's condition.

Prediction of the success rate of thoracic surgery in lung cancer patients can be done by using the principles of data mining classification. Data mining is a set of mathematics, statistic, and computational methods and techniques [8]. Medical data mining is a highly significant field or research in the development of various applications within the growing healthcare domain [9]. In constructing data mining, a set of trial data is required as the basis for prediction, using data mining classification methods that are determined according to the characteristics of the trial data set [10]. The use of algorithms in the data mining process has become crucial in predicting the success rate of thoracic surgery operations in lung cancer patients. The Random Forest algorithm is one of the widely used and effective decision tree methods for data classification. Random Forest works by creating a set of structured classifications trees, each depending on independently sampled values and vectors with the same distribution for all trees [11]. Random Forest is designed to produce accuracy predictions and prevent overfitting on data. This algorithm possesses effective classification methods in terms of combining and selecting random features [12]. In the implementation of the Random Forest algorithm, feature selection can be applied. Feature selection is an essential step in the classification process, as the selected feature significantly influence the accuracy level of classification [13].

Recursive Feature Elimination is a feature selection method aimed at estimating which features are most helpful in distinguishing the desired class [14]. Recursive feature elimination avoids the repetition of model formation at each search step. Once the complete model is constructed, it calculates the importance measure of variables, ranking the attributes from the most important to the least important. Recursive feature elimination will undergo a computational process, which involves eliminating redundant or irrelevant features from the dataset of thoracic surgery. To enhance the accuracy of the classification model, hyperparameter tuning can be employed. Hyperparameter tuning works to find the parameter with optimal model estimation accuracy more efficiently [15]. There is no available method to precisely determine the combination of parameters for the model other than trying them one by one, as each best parameter will vary depending on the selected dataset. Therefore, to address the parameter combinations and achieve the best accuracy result in classification model, GridSearchCV can be employed. GridSearchCV is the process of selecting the best hyperparameters for a model by exploring combinations of hyperparameters and calculating the average cross-validation score each combination [16]. GridSearchCV will perform validation for multiple models and automatically provide their respective hyperparameters to optimize the accuracy value [17].

Based on above the explanation, this research focuses on and aims to improve the accuracy in the predicting the success rate of thoracic surgery operations in lung cancer patients using the random forest classification algorithm and recursive feature elimination for feature selection with the application of hyperparameter using GridSearchCV. Feature selection is used to choose the most influential features for the classification process, while hyperparameter using GridSearchCV is employed to obtain the best parameters and enhance the accuracy of the performed classification model. The dataset used in this research is a publicly available thoracic surgery dataset, and previous study result are used as a benchmark. The researcher decided to conduct research titled " Optimizing Random Forest for Predicting Thoracic Surgery Success in Lung Cancer Using Recursive Feature Elimination and GridSearchCV".

METHODS

This study focuses on the accuracy of the classification results using the random forest algorithm and recursive feature elimination for predicting the success rate of thoracic surgery operations in lung cancer patients, with the implementation of GridSearchCV. The process starts from inputting the dataset, data preprocessing, data splitting, feature selection, and classification. The process stage can be seen in Figure 1.

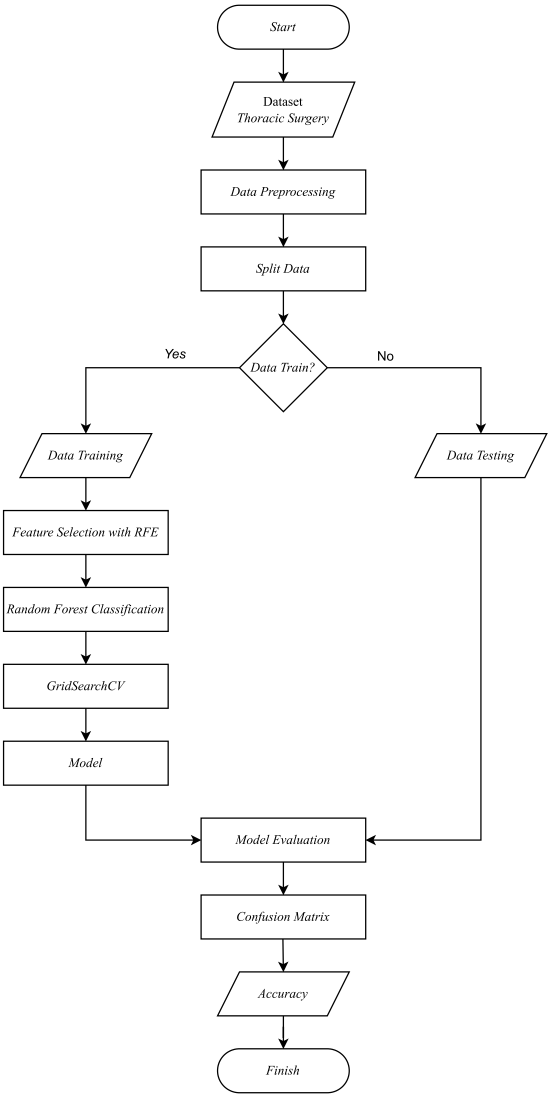

Figure 1. Flowchart of research processes

Data Collection

The dataset used in this study is the thoracic surgery data obtained from UCI Machine Learning Repository. The data utilized in this study is a collection sourced from the Wroclaw Thoracic Surgery Centre. This dataset comprises information on patients afflicted with lung cancer who underwent thoracic surgical procedures spanning from 2009 to 2014 [18]. This dataset contains information for each patient represented by 16 attributes, which encompass pre and post-operative thoracic surgery conditions. These attributes comprise nominal, numeric, and binary data. The post-thoracic surgery patient dataset encompasses two classes: 'die,' indicating mortality within one year, and 'survival,' signifying patients who survived. The dataset comprises 70 samples for the 'die' class and 400 samples for the 'survival' class.

Table 1. Thoracic Surgery Dataset

ID	Attributes	Descriptions	Data Type
1.	DGN	Diagnosis - a specific combination of ICD-10 codes for primary and secondary tumors, as well as multiple tumors if applicable.	Nominal
2.	PRE4	The forced vital capacity from the lungs after maximal inhalation (FVC)	Numeric
3.	PRE5	The volume of air exhaled in the first second of FVC (FEV1)	Numeric
4.	PRE6	General patient's ability in daily activities (Zubrod Scale)	Nominal
5.	PRE7	Pre-operative pain	Binary
6.	PRE8	Pre-operative hemoptysis	Binary
7.	PRE9	Pre-operative dyspnea	Binary
8.	PRE10	Pre-operative cough	Binary
9.	PRE11	Pre-operative weakness	Binary
10.	PRE14	Tumor size (TNM)	Nominal
11.	PRE17	Diabetes	Binary
12.	PRE19	Myocardial Infarction (MI) within 6 months	Binary
13.	PRE25	Peripheral Arterial Disease (PAD)	Binary
14.	PRE30	Smoking	Binary
15.	PRE32	Asthma	Binary
16.	AGE	Age at operation	Numerik
17.	RISK	Survival status after 1 year | valued as 'T' if deceased	Binary

Data Preprocessing

Data Preprocessing is the initial stage in data analysis aimed at cleansing and preparing data before further processing. The preprocessing stage involves preliminary data manipulation processes that are essential to obtaining well-prepared data prior to the data analysis phase. Data preprocessing consists of the following stages that can be explained below.

1. Checking for missing values, is the first stage in data preprocessing, aiming to inspect the values in the data to be used in the classification process, as missing values can lead to errors in data analysis and result in inaccurate models.
2. Oversampling is a second stage in data processing. Oversampling aims to address issues of data imbalance within a dataset. Data imbalance occurs when one class significantly outnumbers the other classes, leading to reduced classification performance for the minority class.
3. Data transformation is the final stage in data preprocessing. Data transformation aims to convert data into the required format, typically into a numeric form. Data transformation is necessary to enable subsequent data computation and analysis processes.

Data Splitting

Data splitting is the next stage after performing data preprocessing. The data splitting process involves dividing the data into two parts, there are training data and testing data. In this study, the dataset is split using an 8:2 ratio, which means 80%:20%. 80% portion of the data will be used for training model, while the remaining 20% will serve as the testing data.

Feature Selection Recursive Feature Elimination

Feature selection is a stage to choose and produce more optimal accuracy results. In this stage, we need a feature selection technique that can select features with the highest scores and eliminate features with the lowest scores, using the recursive feature elimination technique. recursive feature elimination feature selection employe a classification algorithm as the estimator to rank and eliminate features that have less significant impact on the classification process. Recursive feature elimination can combine attributes that contribute to predicting the target variable or class to explore and identify which features are the most prominent and dominant [19].

Random Forest Algorithm

Random Forest is a machine learning algorithm that builds multiple decision tree classifiers on several subsamples of the dataset and uses averaging to improve prediction accuracy and control overfitting [20]. Specifically, Random Forest employs random sampling with replacement to extract training samples from the original the dataset and constructs decision tree models for each iteration of the training samples [13]. The training results from multiple decision tree models using the Random Forest algorithm are combined to determine the prediction categories [21].

GridSearchCV

Hyperparameters have a critical role in optimizing the performance of machine learning algorithms. The values of hyperparameters cannot be determined from data and are taken as given during model definition, meaning hyperparameters must be set before a model undergoes the learning process [22]. In this study, GridSearchCV will be applied to the Random Forest classification model to find the best combination of parameters to optimize the classification results [23]. This method can be used on a set of parameters with specified upper and lower bounds for each independent variable.

RESULT AND DISCUSSION

The section is divided into two parts, results, and discussion. The results are a description of the data and findings obtained using the methods and procedures described in the data collection method. The discussion is an explanation of the results that answer research questions more comprehensively.

Result

The result of this study uses the Random Forest algorithm and Recursive Feature Elimination with the application of hyperparameter tuning using GridSearchCV to predict thoracic surgery success in lung cancer. Recursive feature elimination is used for feature selection in the thoracic surgery dataset, the Random Forest algorithm is used for the classification process in thoracic surgery, and hyperparameter tuning GridSearchCV is used to obtain the best parameters to improve the accuracy of the classification performed. This study was conducted in several stages, there are data preprocessing, feature selection, and classification. The following is a more detailed explanation of the research results.

Preprocessing Results

In data preprocessing, several steps are implemented in this study there are checking missing values, checking imbalanced data, and handling data outliers.

1. Checking Missing Values

   In checking for missing values, df.isna().sum() is a method used to calculate the number of ‘NaN’ or ‘null’ values in each column of dataframe ‘df’. In the dataset used for checking, there were no missing values in any of its features.
2. Oversampling

   The oversampling method used is random oversampling (ROS). Random oversampling is one technique for addressing class imbalance in training data by randomly adding instances of the minority class. The oversampling process is performed using 'ros = RandomOverSampler()' to create a 'RandomOverSampler' object and store it in the variable 'ros'. Then, 'X_resampled, y_resampled = ros.fit_resample(df.drop('Risk1Yr', axis=1), df['Risk1Yr'])' is employed to carry out the oversampling process. The dataset used is the 'df' dataframe, where the target column 'Risk1Yr' requires balancing. 'df.drop('Risk1Yr', axis=1)' separates the target column 'Risk1Yr' from the 'df' dataframe, resulting in a new dataframe 'X'. 'df['Risk1Yr']' selects only the 'Risk1Yr' target column, creating a 'y' series containing the data labels. Figure 2 illustrates the data graph prior to oversampling, while Figure 3 depicts the data graph after the oversampling process.

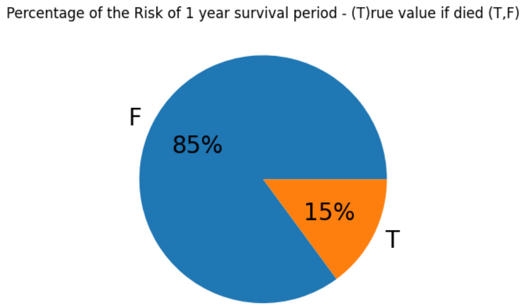

Figure 2. Imbalance Data

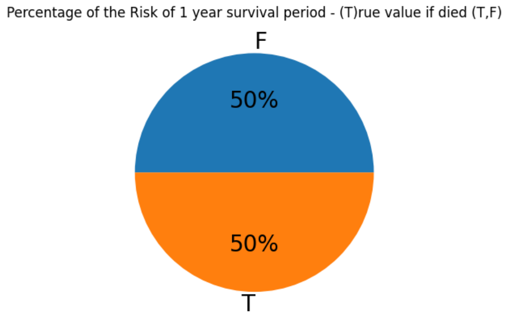

Figure 3. After Oversampling

The figure depicts the graph before the implementation of oversampling, displaying data imbalance with a proportion of 85%:15%. However, after applying oversampling, the data becomes balanced.

3. Data Transformation
   This data transformation will replace specific values in the invoked columns with 1s and 0s, where 1 represents 'T' and 0 represents values other than 'F'. The data transformation applies a lambda function to the selected columns. This lambda function is employed to change values in the columns to 1s and 0s based on specific conditions. In this scenario, if the value in the column is 'T', it will be changed to 1, and if the value is anything other than 'T', it will be changed to 0. Subsequently, data transformation continues for nominal data types, replacing values in the designated columns with numerical values according to the provided mapping. Table 2 illustrates the cleaned data that has undergone the data preprocessing process.

Table 2. Cleaned Data

	DGN	PRE4	PRE5	PRE6	…	PRE25	PRE30	PRE32	AGE
0	2	2,88	2,16	1	…	0	1	0	60
1	3	3,40	1,88	0	…	0	1	0	51
2	3	2,76	2,08	1	…	0	1	0	59
3	3	3,68	3,04	0	…	0	0	0	54
4	3	2,44	0,96	2	…	0	1	0	73
…	…	…	…	…	…	…	…	…	…
795	3	2,08	1,52	1	…	0	1	0	49
796	3	2,64	2,12	1	…	0	1	0	72
797	3	3,88	2,92	0	…	0	1	0	67
798	5	2,38	1,72	1	…	1	1	0	87
799	3	3,04	3,60	1	…	0	1	0	62

Feature Selection Results

After the dataset has been divided into training data and testing data, with each proportion of data training 80% or 640 data and for testing data is 20% or 160 data. The next step involves using the recursive feature elimination method as the feature selection process or attribute selection. This process aims to select features with the highest scores or rankings, using the Random Forest algorithm as the estimator. The selected features will have a ranking value of 1, while features that are not selected or eliminated from the classification process will have a ranking value other than 1. The results of the feature selection process are shown in Table 3.

Table 3. Selected attributes of the Recursive Feature Elimination

ID	Attributes	Ranking
1.	DGN	1
2.	PRE4	1
3.	PRE5	1
4.	PRE6	1
5.	PRE7	4
6.	PRE8	2
7.	PRE9	5
8.	PRE10	1
9.	PRE11	3
10.	PRE14	1
11.	PRE17	6
12.	PRE19	9
13.	PRE25	7
14.	PRE30	1
15.	PRE32	8
16.	AGE	1

Classification

At this stage, the researcher conducted data processing to obtain accuracy results in the classification. First, the classification process using the Random Forest algorithm on the thoracic surgery dataset. Second, the classification process was carried out using the Random Forest algorithm on the thoracic surgery dataset that had implemented feature selection using the recursive feature elimination method. Third, the classification process using the Random Forest on the thoracic surgery dataset that had implemented feature selection using the Recursive Feature Elimination method and applied hyperparameter tuning using GridSearchCV

Default Random Forest Parameters

The determination of Random Forest algorithms parameters in the classification was implemented using the default parameters provided by the scikit-learn library. The following are the default parameters of the Random Forest classifier algorithm, which can be seen in Table 4.

Table 4. Default parameters of Random Forest

Parameters	Value
n_estimator	100
criterion	gini
min_samples_split	2
min_samples_leaf	1
max_features	sqrt
n_jobs	None
verbose	0

Random Forest Algorithm Classification Results

At this stage, the thoracic surgery dataset it classified using the Random Forest algorithm without using feature selection. Then, the training data is processed using the Random Forest algorithm for model testing with default Random Forest parameters in Table 4. The results of classification model can be seen in Table 5.

Table 5. Accuracy classification with classification algorithms

Algorithm	Accuracy
Random Forest	84,38%

Classification Results of Random Forest Algorithm by Applying Recursive Feature Elimination

The thoracic surgery dataset was classified using the Random Forest algorithm with the application of Recursive Feature Elimination feature selection. Next, the training data was processed using the Random Forest algorithm for model testing. The model testing used 8 selected features from the Recursive Feature Elimination feature selection process, including 'DGN', 'PRE4', 'PRE5', 'PRE6', 'PRE10', 'PRE14', 'PRE30', and 'AGE'. Additionally, the classification was carried out using the default parameters of the Random Forest algorithm, which can be seen in Table 4. The accuracy results can be seen in Table 6.

Table 6. Accuracy classification with adding Recursive Feature Elimination

Algorithm	Accuracy
Random Forest + Recursive Feature Elimination	88,75%

GridSearchCV Parameters

The selection of hyperparameters can be determined manually (trial and error), but it is much more efficient and effective to perform parameter tuning in combination with model testing. Table 7 shows the parameter combinations used in the Random Forest algorithm model with application of GridSearchCV.

Table 7. GridSearchCV Parameters Determination

Parameters	Value
Bootstrap	True, False
N_estimator	100, 200, 300
Max_depth	None, 5, 10, 15
Min_samples_split	1, 2, 4, 6
Min_samples_leaf	1, 2, 4, 6
Max_features	Auto, Sqrt, Log2
N_jobs	-1
Verbose	2

Classification Results of Random Forest Algorithm by Applying Recursive Feature Elimination and GridSearchCV

The thoracic surgery dataset is classified using the Random Forest algorithm with the addition of Recursive Feature Elimination feature selection and applying hyperparameter tuning using GridSearchCV with 10-fold cross validation. GridSearchCV is used to optimize the accuracy of the classification model in this study. The best parameter results obtained from the GridSearchCV process and the accuracy results after applying hyperparameter tuning with GridSearchCV can be seen in Table 8 and Table 9.

Table 8. Best parameters results using GridSearchCV

Parameters	Value
Bootstrap	False
N_estimator	100
Max_depth	None
Min_samples_split	4
Min_samples_leaf	1
Max_features	Auto
N_jobs	-1
Verbose	2

Table 9. Accuracy classification with adding Recursive Feature Elimination and hyperparameter tuning GridSearchCV

Algorithm	Accuracy
Random Forest + Recursive Feature Elimination + GridSearchCV	91,41%

Discussion

Based on the accuracy obtained from the conducted classification, this classification yields novelty when compared to previous studies. The combination of the Random Forest classification algorithm with Recursive Feature Elimination and GridSearchCV produces higher accuracy compared to previous studies with the same topic and object using the purposed method in this research. The comparison of this research with previous studies can be seen in Table 10.

Table 10. Comparison with previous study

Writer	Algorithm	Accuracy
Ravichandran and Gds (2021)	Random Forest	83%
Roshan and Rohini (2017)	Random Forest, Decision Stump and J48	88,73%
Abdulhadi dan Talabani (2021)	Random Forest and Gain Ratio	81,70%
Setyadi et al (2020)	Naïve Bayes Classifier and Genetic Algorithm	85,31%
Prasetio and Susanti (2019)	Boosted K-Nearest Neighbor	85,11%
Purposed method (2023)	Random Forest + Recursive Feature Elimination + GridSearchCV	91,41%

The study's strength lies in its capacity to enhance accuracy through the application of Recursive Feature Elimination for feature selection and hyperparameter tuning via GridSearchCV. The integration of Recursive Feature Elimination and GridSearchCV has been empirically validated to bolster the accuracy of the Random Forest algorithm. This advancement has been effectively showcased within the classification phase. Additionally, the incorporation of these methodologies within this research introduces innovation, resulting in heightened accuracy compared to prior investigative efforts. However, certain limitations exist within this study, notably the employment of hyperparameter tuning through GridSearchCV, which necessitates a substantial time investment to identify the optimal parameter combinations for optimizing accuracy outcomes in the classification process.

CONCLUSION

Classification to predict thoracic surgery success rate in lung cancer patients from a publicly accessible thoracic surgery dataset through UCI Machine Learning Repository. The classification in this study utilizes Random Forest and Recursive Feature Elimination feature selection, along with applying hyperaparmeter using GridSearchCV, which has proven to achieve good accuracy. With the dataset used in this study, combination of Random Forest and Recursive Feature Elimination feature selection with 8 selected features, there are 'DGN', 'PRE4', 'PRE5', 'PRE6', 'PRE10', 'PRE14', 'PRE30', and 'AGE', along with hyperparameter using GridSearchCV with the following parameter values are bootstrap = false, n_estimator = 100, max_depth = none, min_samples_split = 4, min_samples_leaf = 1, max_features = auto, n_jobs = -1, and verbose = 2 with a cross validation of 10, results in an accuracy of 91,41%.

REFERENCES

[1] Roshan and Rohini, “Prediction of Post-Surgical Survival of Lung Cancer Patients After Thoracic Surgery Using Data Mining Techniques.,” *Int. J. Adv. Res.*, vol. 5, no. 4, pp. 596–600, 2017, doi: 10.21474/ijar01/3852.

[2] R. T. Prasetio and S. Susanti, “Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted k-Nearest Neighbor,” vol. 1, no. 1, pp. 64–69, 2019.

[3] I. F. Anshori and D. Riana, “Prediksi Harapan Hidup Pasien Kanker Paru-Paru Pasca Operasi Bedah Thoraks Menggunakan Boosted Neural Network Dan Smote,” vol. 6, no. 1, pp. 9–15, 2021.

[4] H. Kenang, C. Alivian, W. Suharso, and A. Qurrota, “Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes , Multinomial Naïve Bayes , Dan Bernoulli Naïve Bayes Classification Of Breast Cancer And Lung Cancer Using The Gaussian Naïve Bayes Multinomial Nave Bayes And Berno,” vol. 3, no. 4, pp. 350–355, 2022.

[5] R. Sanjaya and F. Fitriyani, “Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 316, 2019, doi: 10.26418/jp.v5i3.35324.

[6] M. Koklu, H. Kahramanli, and N. Allahverdi, “Applications of Rule Based Classification Techniques for Thoracic Surgery,” *Jt. Int. Conf. 2015*, no. November, pp. 1991–1998, 2015.

[7] M. L. Giger and K. Suzuki, “Computer-Aided Diagnosis,” *Biomed. Inf. Technol.*, pp. 359–XXII, Jan. 2008, doi: 10.1016/B978-012373583-6.50020-7.

[8] D. Jollyta, W. Ramdhan, and M. Zarlis, *Konsep Data Mining Dan Penerapan*. Deepublish, 2020.

[9] D. Derisma, “Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining,” *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 84–88, 2020, doi: 10.30871/jaic.v4i1.2152.

[10] S. Maesaroh and Kusrini, “Sistem Prediksi Produktifitas Pertanian Padi Menggunakan Data Mining,” *J. Energi*, vol. 7, no. 2, pp. 25–30, 2017.

[11] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[12] L. Breiman, “Randomizing outputs to increase prediction accuracy,” *Mach. Learn.*, vol. 40, no. 3, pp. 229–242, 2000, doi: 10.1023/A:1007682208299.

[13] J. Saputra, Wahyu S., A. R. Sujatmika, and A. Z. Arifin, “Seleksi Fitur Menggunakan Random Forest Dan Neural Network,” *13 Th Ind. Electron. Semin. 2011 (IES 2011)*, vol. 1, no. Ies, pp. 93–97, 2011.

[14] L. Demarchi, A. Kania, W. Ciezkowski, H. Piórkowski, Z. Oświecimska-Piasko, and J. Chormański, “Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of poland based on airborne hyperspectral and LiDAR data fusion,” *Remote Sens.*, vol. 12, no. 11, 2020, doi: 10.3390/rs12111842.

[15] J. Lu *et al.*, “Estimation of monthly 1 km resolution PM2.5 concentrations using a random forest model over ‘2 + 26’ cities, China,” *Urban Clim.*, vol. 35, no. November 2020, p. 100734, 2021, doi: 10.1016/j.uclim.2020.100734.

[16] Andriana *et al.*, “Prediksi Gelombang Corona Dengan Metode Neural Network,” *JIKOMSI (Jurnal Ilmu Komput. dan Sist. Inf.*, vol. 3, no. 2, pp. 102–107, 2020.

[17] Z. Maisat, E. Darmawan, and A. Fauzan, “Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM,” vol. 13, no. 1, pp. 8–15, 2023.

[18] I. W. Septiani, A. C. Fauzan, and M. M. Huda, “Implementasi Algoritma K-Medoids Dengan Evaluasi Davies-Bouldin- Index Untuk Klasterisasi Harapan Hidup Pasca Operasi Pada Pasien Penderita Kanker Paru-Paru,” vol. 3, pp. 556–566, 2022, doi: 10.30865/json.v3i4.4055.

[19] S. Asadi, S. E. Roshan, and M. W. Kattan, “Random forest swarm optimization-based for heart diseases diagnosis,” *J. Biomed. Inform.*, vol. 115, no. August 2020, p. 103690, 2021, doi: 10.1016/j.jbi.2021.103690.

[20] “Heart Disease Dataset | Kaggle.” .

[21] A. R. I. Pratama, S. A. Latipah, and B. N. Sari, “Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (Svm) Dan Recursive Feature Elimination (Rfe),” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 2, pp. 314–324, 2022, doi: 10.29100/jipi.v7i2.2675.

[22] A. A. Mohammed, R. Basa, A. K. Kuchuru, S. P. Nandigama, and M. Gangolla, “Random Forest Machine Learning technique to predict Heart disease,” vol. 7, no. 4, p. 2020, 2020.

[23] G. A. Lujan-Moreno, P. R. Howard, O. G. Rojas, and D. C. Montgomery, “Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study,” *Expert Syst. Appl.*, vol. 109, pp. 195–205, 2018, doi: 10.1016/j.eswa.2018.05.024.