 ISSN:2302-8149
e-ISSN:2540-9719

Implementation of the Random Forest Algorithm with Optuna Optimization in Lung Cancer Classification

[1]Ahmad Ainul Yaqin*, [2]Mula Agung Barata, [3]Nur Mahmudah
[1,2]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri
[3]Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Ulama Sunan Giri
[1,2,3]Jl. Ahmad Yani No.10, Jambean, Sukorejo, Kec. Bojonegoro, Kabupaten Bojonegoro, Jawa Timur 62115, Indonesia
*e-mail: ahmadainulyaqin131@gmail.com

(***received*:** 4 December 2024, ***revised*:** 30 January 2025, ***accepted*:** 30 January 2025)

Abstract

Lung cancer remains one of the leading causes of death worldwide, with many sufferers unaware of their condition until it is too late for treatment. Therefore, high-accuracy prediction methods are urgently needed for early detection of lung cancer. This research uses the Random Forest algorithm, known for its excellent performance in medical data classification. In this study, modeling was optimized by implementing hyperparameter optimization using Optuna. The results of the generated model show an accuracy rate of 98.6%, which is highly significant in the context of early lung cancer detection. Additionally, this algorithm demonstrated 100% recall for the positive class and 97% for the negative class, indicating that the model is highly effective in identifying patients who truly have lung cancer. Another advantage of this model is seen in the AUC (Area Under the Curve) value reaching 1, indicating 100% accurate predictions. With these results, this research affirms the importance of using the Random Forest algorithm in developing early detection systems for lung cancer. This not only can improve treatment success rates but also significantly reduce mortality rates from lung cancer.

Keywords: lung cancer, random forest, optuna, hyperparameter optimization, classification

1 Introduction

Lung cancer is one of the most common and deadly cancers in the world, with the incidence increasing every year. As per the 2020 Global Cancer Burden statistics, roughly 2.21 million fresh lung cancer diagnoses occurred, resulting in 1.80 million deaths. In Indonesia, lung cancer ranks third after breast and cervical cancer, with the number of cases reaching 34,783 or 8.8% of the total 396,914 diagnosed cancer cases. This high mortality rate is largely due to late diagnosis, with around 70% of cases only detected at an advanced stage when treatment is already ineffective [1].

The development of machine learning technology in recent years has opened up new opportunities to improve the accuracy and speed of lung cancer diagnosis. Random Forest algorithm, as one of the powerful machine learning methods, has shown promising potential in various medical classification cases. It has the advantage of handling complex datasets and can provide accurate classification results. However, the performance of Random Forest is highly dependent on proper hyperparameter settings. Manual optimization of hyperparameters often takes a long time and does not always result in an optimal configuration [2].

Optuna is a relatively new hyperparameter optimization framework with three main advantages in model selection or hyperparameter determination. First, Optuna offers a define-by-run style API, which allows users to dynamically define the hyperparameter search space, providing flexibility in experiment setup. Second, the efficient pruning and sampling mechanisms, including efficient search and performance estimation, utilize cost-effective optimization methods such as Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Tree-structured Parzen Estimator (TPE) and enable customizable sampling procedures. In addition, Optuna's pruning mechanism involves periodically monitoring the interim objective value and stopping the experiment when certain conditions are not met, thus keeping the focus on the relevant target. Finally, ease of setup is one of Optuna's significant advantages. It can be easily configured for both lightweight experiments and heavy distributed

 ISSN:2302-8149
e-ISSN:2540-9719

computing in a versatile architecture, making it suitable for various types of experiments, ranging from small research to large-scale applications [3] [4].

Based on the background that has been presented, the Random Forest algorithms show great potential for improving the accuracy and efficiency of lung cancer diagnosis, particularly when optimized with appropriate hyperparameter settings. However, manual optimization presents a significant challenge due to its time and resource requirements. The Optuna framework serves as an ideal solution for hyperparameter optimization. With its define-by-run capabilities, efficient pruning mechanisms, and flexible configuration options, Optuna can significantly enhance Random Forest model performance. The implementation of this approach is expected to make substantial contributions to medical practice, particularly in improving early lung cancer detection accuracy and speed, thereby helping reduce mortality rates associated with late diagnosis.

2 Literature Review

Previous research Sinaga et al. [5] used a lung cancer dataset taken from the Kaggle website. This study compared the performance of the Random Forest algorithm with a combination of the Adaboost and Random Forest algorithms in lung cancer classification. The accuracy results from combining Adaboost with Random Forest reached 95.4%, superior to the Random Forest algorithm which only achieved 93.2% accuracy.

Previous research Sitanggang and Sitompul [6] discussed the application of the Random Forest algorithm to the classification of heart failure. This research uses the Hyperparameter Tuning approach with the Grid Search method to optimize model performance. The results showed that the Random Forest algorithm optimized with Grid Search was able to achieve an accuracy of 85%. The use of Hyperparameter Tuning is proven to improve model performance, making it more effective in detecting heart failure in patients early.

Previous research Juliani and Soleh [7] addressed the topic of lung cancer classification using the Naïve Bayes algorithm combined with a chatbot system based on lung cancer datasets. This study proved that the Naïve Bayes algorithm can record a peak accuracy rate of 81%. The combination of the algorithm with the chatbot system is expected to help in providing early information related to lung cancer detection effectively.

Previous research Sari et al. [8] discussed lung cancer prediction analysis using the Random Forest algorithm optimized with K-fold cross-validation. This study uses a lung cancer dataset as the object of study. The results showed that the Random Forest algorithm optimized with K-fold cross-validation was able to produce an accuracy of 98.4%, higher than the Naïve Bayes algorithm. This finding confirms the superiority of Random Forest in handling lung cancer datasets for prediction and classification purposes.

Previous research Hanifi et al. [9] analyzed the effectiveness of various hyperparameter optimization techniques on machine learning models. This research compares the performance of models with default parameters and models that have been optimized. The results showed a significant increase in accuracy in models using hyperparameter optimization compared to models using default parameters. The hyperparameter optimization process proved to be an influential factor in improving the performance of the classification model.

Previous research Sipper [10] used 250 datasets covering regression and classification cases. This research compares the Optuna optimization method with Random Search and Grid Search in hyperparameter tuning. The results showed that Optuna, which uses Bayesian optimization, is superior to Random Search in improving model performance. Although this research succeeded in proving the superiority of Optuna, the focus of the experiment is still limited to general datasets and has not covered specific cases such as specific disease classification.

Based on the literature review described, there is a significant research gap in the optimization of machine-learning models for lung cancer classification. Although previous studies have shown the potential for performance improvement through hyperparameter optimization, there is no research that specifically uses Optuna to optimize Random Forest in the context of lung cancer classification. Therefore, this research will focus on the implementation of Random Forest with Optuna optimization to improve the accuracy and efficiency of lung cancer classification models.

 ISSN:2302-8149
e-ISSN:2540-9719

3 Research Method

The research stages conducted in this study are presented in Figure 1. The process consists of several steps, which start from data collection to model evaluation. Each stage is further explained to provide a comprehensive view of the research process.

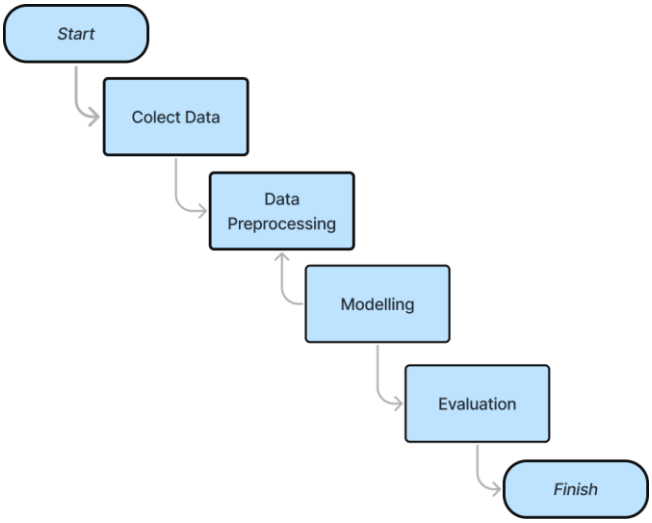

Figure 1. Stages of research

3.1 Collect Data

The data used is a dataset on lung cancer obtained from www.kaggle.com. This data set has a CSV format consisting of 16 features and 309 rows. Where columns describe variables and rows indicate the number of respondents. For more details, an example of the data is shown in Table 1, while an explanation of each column can be found in Table 2.

Tabel 1. Lung cancer dataset

No	GENDER	AGE	SMOKING	…	SWALLOWING DIFFICULTY	LUNG_CANCER
1	M	69	1	…	2	YES
2	M	74	2	…	2	YES
3	M	59	1	…	1	YES
…	…	…	…	…	…	…
309	M	60	1	…	2	YES

Tabel 2. Feature details of lung cancer dataset

No	Feature	Data Type	Description
1	Gender	Object	Gender (M for male, F for female)
2	Age	Integer	Age in year
3	Smoking	Integer	Smoking status (1: No Smoking, 2: Smoking)
4	Yellow_fingers	Integer	Finger yellowing (1: No, 2: Yes)
5	Anxiety	Integer	Anxiety (1: No, 2: Yes)
6	Peer_pressure	Interger	Peer pressure (1: No, 2: Yes)
7	Chronic disease	Integer	Chronic disease (1: No, 2: Yes)
8	Fatigue	Integer	Fatigue (1: No, 2: Yes)
9	Allergy	Integer	Allergy (1: No, 2: Yes)

 ISSN:2302-8149
e-ISSN:2540-9719

No	Feature	Data Type	Description
10	Wheezing	Integer	Wheezing (1: No, 2: Yes)
11	Alcohol consuming	Integer	Alcohol consumption (1: No, 2: Yes)
12	Coughing	Integer	Coughing (1: No, 2: Yes)
13	Shortness of breath	Integer	Shortness of breath (1: No, 2: Yes)
14	Swallowing difficulty	Integer	Swallowing difficulty (1: No, 2: Yes)
15	Chest pain	Integer	Chest pain (1: No, 2: Yes)
16	Lung_cancer	Object	Lung cancer diagnosis (YES: Positive, NO: Negative)

3.2 Data Preprocessing

The data preprocessing stage in this research is to clean the data, preprocess the data encoding label, and balance the data.

a. Data Cleaning

Data preprocessing includes a crucial stage involving the elimination of duplicate data and missing values. This stage encompasses the identification and removal of similar or redundant entries within the dataset. Data quality is enhanced through an iterative elimination process, which subsequently contributes to improving the developed model's accuracy. This step is essential as duplicate data can introduce bias into the analysis results and adversely affect the model's overall performance [11].

b. Label Encoding Preprocessing

The Label Encoding process is a technique used to convert categorical variables into numerical values, which is necessary because most machine learning algorithms can only work with numerical data [12]. In the lung cancer dataset we are discussing, two categorical features need to be encoded so that the machine learning model can process them: Gender and lung cancer.

c. Data Balancing

ADASYN (Adaptive Synthetic Sampling) can be implemented as an effective oversampling technique to address data imbalance issues. This approach aims to balance class distributions in datasets by increasing the number of samples in the minority class. ADASYN generates synthetic instances based on the local density distribution of individual data points within the minority class. The technique adaptively synthesizes more samples for minority instances that are more challenging to learn, thereby enabling the model to better comprehend minority class characteristics and enhance overall classification accuracy. Empirical research demonstrates that ADASYN implementation can significantly improve model performance across diverse scenarios, making it a robust solution for machine learning tasks involving imbalanced datasets [13].

3.3 Modelling

This research implements the Random Forest machine learning model for lung cancer classification. This model was chosen due to its ability to handle the complexity of medical data. After model selection, the Optuna hyperparameter optimization technique is used to improve the performance of the Random Forest model. The following is a detailed explanation of the two components that are the focus of this research.

a. Random Forest

Random Forest is a highly effective statistical learning algorithm for prediction tasks, introduced by Breiman in 2001. It works by building many decision trees that are each trained on a random subset of the training data, using a technique known as bootstrap aggregating or bagging. Each tree in the random forest segments the data based on certain criteria, such as entropy for classification or mean squared error for regression. The advantage of Random Forest lies in its ability to reduce overfitting that often occurs with a single decision tree, resulting in better prediction accuracy. In addition, Random Forest can also handle datasets with many independent variables, even when the number of variables exceeds the number of observations [14].

The construction of decision trees within the Random Forest algorithm adheres to the principles of the Classification and Regression Tree (CART) methodology, with a notable distinction being the

 ISSN:2302-8149
e-ISSN:2540-9719

omission of the pruning phase in Random Forest. For feature selection at each internal node of the decision tree, this algorithm employs the Gini Index calculation as its selection criterion. The Gini Index value can be computed using equation (1).

$$Gini(S_i) = 1 - \Sigma_{i=0}^{c-1} p_i^2 \quad (1)$$

pi represents the relative frequency for class Ci in the data set. The class Ci is set for each i ranging from 1 to c - 1, where c denotes the total number of predefined classes.

The quality is split into subsets Si based on k features. This is the total samples belonging to class Ci, which is then calculated as the sum of the Gini indication considerations of the formed subsets. Information can be calculated using equation (2).

$$Gini_{split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right) Gini(S_i) \quad (2)$$

$Gini$ is the number of samples in the subset Si. After splitting n represents the number of samples in the given node.

b. Optuna

Optuna is implemented as an optimization framework that applies the Bayesian approach with the Tree-structured Parzen Estimator (TPE) algorithm for the hyperparameter optimization process. The framework works by analyzing previous evaluation results to determine more potential search directions in hyperparameter space. This method allows Optuna to identify the optimal hyperparameter configuration more effectively and efficiently compared to the use of conventional grid search methods [15].

3.4 Confusion Matrix

Confusion Matrix is a tabular evaluation tool that serves to evaluate the accuracy and performance of classification algorithms, both in the context of classification and prediction of attributes from test data. This evaluation technique is specifically designed to measure the effectiveness of machine learning models in solving classification problems. This matrix consists of four main assessment components, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), each of which represents the model's prediction results against actual data. The assumptions in the Confusion Matrix are shown in Table 3.

Tabel 3. Assumptions in confusion matrix

Prediction Class	Actual Class	
	Positi	Negatif
Positif	TP	FP
Negatif	FN	TN

True Positive (TP) represents instances where both the predicted and actual values are positive. False Negative (FN) indicates cases where the model predicts a negative outcome despite a positive actual value. False Positive (FP) occurs when the model predicts a positive outcome while the actual value is negative. Several key metrics are employed in model performance evaluation: accuracy measures the overall prediction correctness, recall quantifies the model's ability to identify positive cases, and precision assesses the accuracy of positive predictions. The formulae for calculating recall, precision, accuracy, and F1-Score are presented in Table 4.

Tabel 4. Model performance evaluation formula

Performance Matrix	Formula
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + TN}$
F1-Score	$\frac{TP}{TP + FN}$

 ISSN:2302-8149
e-ISSN:2540-9719

Accuracy represents the ratio of correct predictions to total observations in the dataset. Recall quantifies the proportion of correctly identified positive cases among all actual positive instances. Precision measures the ratio of correct positive predictions to all positive predictions made by the model. The Area Under the Curve (AUC) serves as an additional performance metric, indicating the model's discriminative ability between categories. A higher AUC value demonstrates the model's enhanced effectiveness in distinguishing between different types of conditions, and the higher the number of correct predictions, the higher the number of correct predictions.

4 Results and Analysis

In the initial step following data acquisition, preprocessing procedures are implemented. The graph presented in Figure 2 illustrates the changes in class distribution within the dataset across three main data processing phases: before processing, after missing value removal, and after duplicate data cleanup. The class distribution remains unchanged following the removal of missing values, indicating that all entries in the dataset contain complete target variables. After duplicate data cleanup, numerical changes occur, with the negative class decreasing from 270 to 238, while the positive class reduces from 39 to 38. This indicates the successful removal of similar entries, which is essential to ensure the model is not trained using redundant or unrepresentative data.

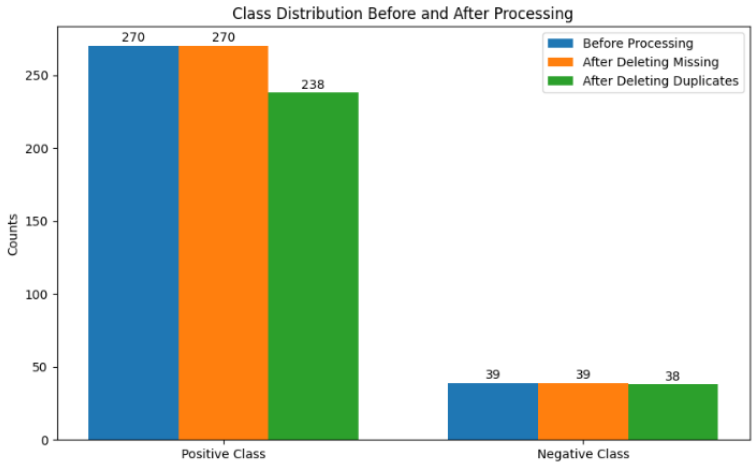

Figure 2. Plot of dataset changes before preprocessing, after missing value removal, and after duplicate data removal.

The next step is label encoding and oversampling using ADASYN. The results of the ADASYN process are presented in Figure 3.

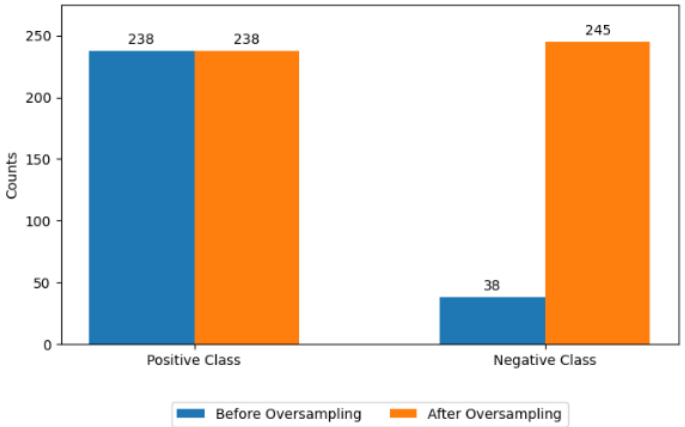

Figure 3. Plot of dataset changes before and after ADASYN.

The next step is to construct the data. In this step, the data is divided into two parts, namely train data and test data. The train data is used to build the model, while the test data serves to test the model that has been made while evaluating its performance. The division ratio used is 70% for train data and

 ISSN:2302-8149
e-ISSN:2540-9719

30% for test data, which is a common approach, especially for small datasets. This resulted in 338 train data and 145 test data, as shown in Figure 4.

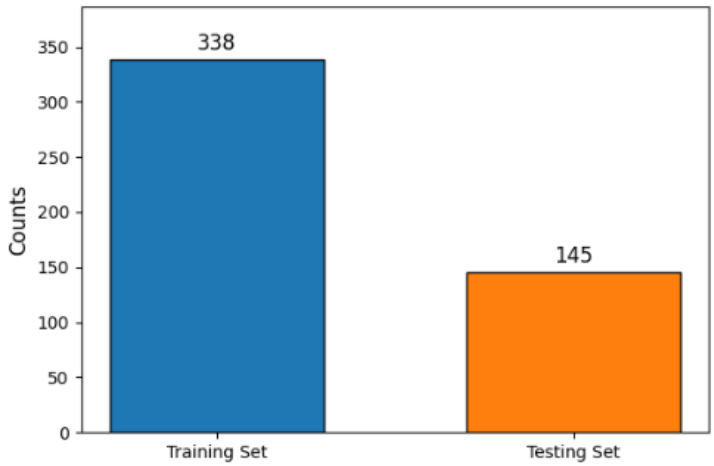

Figure 4. Training and Testing Data Distribution

Tabel 5. Random forest algorithm testing results

Class	Precision	Recall	F1-Score
0	0.97	0.93	0.95
1	0.93	0.97	0.95
Accuracy			0.95

According to Table 5, the Random Forest algorithm used in this study produces quite good performance with accuracy reaching 0.95, with balanced performance between the two classes. Class 0 has a higher precision (0.97) but a lower recall (0.93), while Class 1 is the opposite with a precision of 0.93 and recall of 0.97. The F1-Score is consistently 0.95 for both classes, showing a good balance between precision and recall. Although these results show good performance, there is still room for improvement through hyperparameter optimization using Optuna.

The trial process in hyperparameter optimization using Optuna aims to identify the hyperparameter combination that produces the best performance for the Random Forest model.

Tabel 6. Hyperparameter search space and best configuration

Hyperparameter	**Search Space**	**Best Configuration**
n_estimators	10, 200, log=True	19
max_depth	2, 32	23
min_samples_split	2, 10	6
min_samples_leaf	1, 10	6

Table 6 details the hyperparameter search space and the optimal configuration used to train the Random Forest model. This configuration played a critical role in achieving optimal performance, as demonstrated by the evaluation results in Table 7 and further supported by the confusion matrix and ROC graphs shown in Figures 5 and 6.

Tabel 7. Results of random forest model evaluation and optuna optimization

Class	Precision	Recall	F1-Score
0	1.00	0.97	0.99
1	0.97	1.00	0.99
Accuracy			0.9862

Based on the evaluation results of the Random Forest model optimized using Optuna, the model demonstrates excellent performance in lung cancer classification. The achieved accuracy of 98.62% indicates that the model is highly effective in predicting both classes. A precision of 100% for the negative class shows the absence of false positives in that class. Meanwhile, a recall of 100% for the positive class confirms that the model correctly identifies all instances of the positive class. An average F1-Score of 99% combines the model's precision and sensitivity, indicating that the model is well-balanced in terms of precision and recall.

 ISSN:2302-8149
e-ISSN:2540-9719

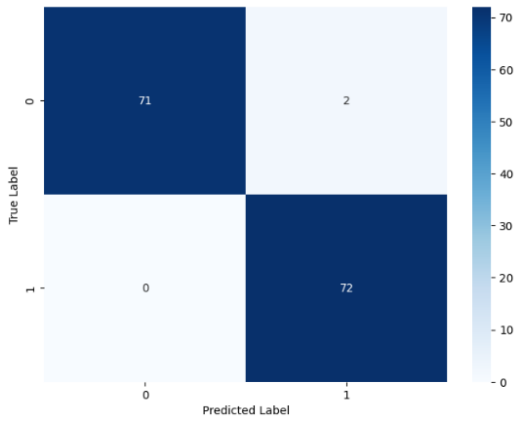

Figure 5. Confusion matrix

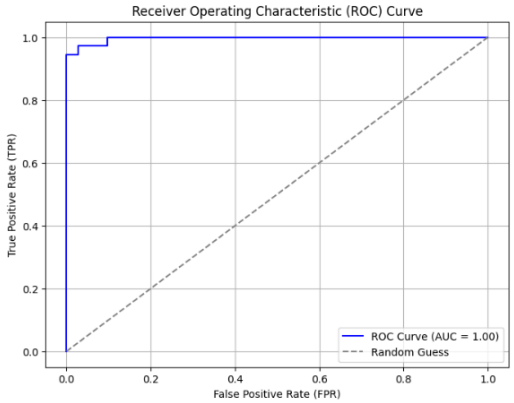

Figure 6. Plot ROC

Overall, these results show that this Random Forest model optimized with Optuna hyperparameters is well suited for the lung cancer classification task, with a good balance between precision, recall, and accuracy, and a strong ability to generalize performance to new data.

5 Conclusion

This study utilized the Survey Lung Cancer dataset. The performance of the Random Forest algorithm was evaluated by applying hyperparameter optimization using Optuna. The optimized model demonstrated superior performance compared to the non-optimized Random Forest model, as reflected in higher scores across metrics such as Accuracy, Recall, Precision, and AUC. The optimized Random Forest algorithm achieved an Accuracy of 98.6%, significantly surpassing the non-optimized version. Notably, it recorded a Recall of 100% for the positive class and 97% for the negative class, delivering perfect predictions as evidenced by an AUC value of 1. These findings highlight the importance of employing optimized Random Forest algorithms in developing early detection systems for lung cancer. Such advancements can not only improve treatment success rates but also significantly reduce lung cancer-related mortality rates.

Reference

[1] S. Alfarisa, E. Mitra, and S. Wahyuni, "Karakteristik Pasien Kanker Paru di RSUP Dr. M. Djamil Padang Tahun 2021," *SCI. J.*, vol. 2, no. 6, pp. 141–149, 2023, doi: 10.56260/sciena.v2i6.116.

[2] Hajiar Yuliana, "*Hyperparameter Optimization of Random Forest for 5G Coverage Prediction*," *Bul. Pos dan Telekomun.*, vol. 22, no. 1, pp. 75–90, 2024, doi: 10.17933/bpostel.v22i1.390.

[3] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "*Optuna: a Next-Generation*

 ISSN:2302-8149
e-ISSN:2540-9719

Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery \& data mining*, 2019, pp. 2623–2631.

[4] S. Chintakindi, A. Alsamhan, M. H. Abidi, and M. P. Kumar, "*Annealing of Monel 400 Alloy using Principal Component Analysis, Hyper-Parameter Optimization, Machine Learning Techniques, and Multi-Objective Particle Swarm Optimization*," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, 2022, doi: 10.1007/s44196-022-00070-z.

[5] R. B. Sinaga, D. Widiyanto, and B. T. Wahyono, "Deteksi Dini Penyakit Kanker Paru dengan Gabungan *Algoritma Adaboost* dan *Random Forest*," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 1–10, 2022, [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

[6] B. F. Sitanggang and P. Sitompul, "Deteksi Awal Kelangsungan Hidup Pasien Gagal Jantung menggunakan *Machine Learning Metode Random Forest*," *Innov. J. Soc. Sci. ...*, vol. 4, pp. 3347–3357, 2024, [Online]. Available: http://j-innovative.org/index.php/Innovative/article/view/8189%0Ahttps://j-innovative.org/index.php/Innovative/article/download/8189/6657

[7] D. Juliani and M. Soleh, "*Implementasi Machine Learning* untuk Klasifikasi Penyakit Kanker Paru menggunakan Metode *Naïve Bayes* dengan Tambahan Fitur *Chatbot* (*Implementation of Machine Learning for Lung Cancer Classification using Naïve Bayes Method with Additional Chatbot Features*," 2020.

[8] L. Sari, A. Romadloni, and R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru menggunakan *Algoritma Random Forest*," *Infotekmesin*, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.

[9] S. Hanifi, A. Cammarono, and H. Zare-Behtash, "*Advanced Hyperparameter Optimization of Deep Learning Models for Wind Power Prediction*," *Renew. Energy*, vol. 221, no. November 2023, p. 119700, 2024, doi: 10.1016/j.renene.2023.119700.

[10] M. Sipper, "*High Per Parameter: A Large-Scale Study of Hyperparameter Tuning for Machine Learning Algorithms*," *Algorithms*, vol. 15, no. 9, 2022, doi: 10.3390/a15090315.

[11] M. Banurea, D. Betaria Hutagaol, and O. Sihombing, "Klasifikasi Penyakit Stunting dengan menggunakan *Algoritma Support Vector Machine* dan *Random Forest*," *J. TEKINKOM*, vol. 6, no. 2, pp. 540–549, 2023, doi: 10.37600/tekinkom.v6i2.927.

[12] Jan Melvin Ayu Soraya Dachi and Pardomuan Sitompul, "Analisis Perbandingan *Algoritma XGBoost* dan Algoritma *Random Forest Ensemble Learning* pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. Dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: 10.55606/jurrimipa.v2i2.1470.

[13] J. Brandt and E. Lanzén, "*A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification*," *2021*, p. 42, 2020, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1519153

[14] M. Schonlau and R. Y. Zou, "*The Random Forest Algorithm for Statistical Learning*," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[15] T. Kurniawan, L. Hermawanti, and A. N. Safriandono, "*Interpretable Machine Learning with SHAP and XGBoost for Lung Cancer Prediction Insights*," vol. 8, no. 2, pp. 296–303, 2024.