Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 2, April 2024, pp: 125-136; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received February 10, 2024; revised March 18, 2024; accepted March 20, 2024; date of publication March 27, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i2.382

How to cite: Yra Fatria Zamzam, Triando Hamonangan Saragih, Rudy Herteno, Muliadi, Dodon Turianto Nugrahadi, and Phuoc-Hai Huynh. Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 2, pp. 125-136, April 2024.

Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based

Yra Fatria Zamzam[1], Triando Hamonangan Saragih[1], Rudy Herteno[1], Muliadi[1], Dodon Turianto Nugrahadi[1], and Phuoc-Hai Huynh[2]

[1] Computer Science Department, Lambung Mangkurat University, Banjarbaru, South Kalimantan, Indonesia
[2] Computer Science Department, An Giang University, Vietnam National University, Ho Chi Minh City, Vietnam
Corresponding author: Triando Hamonangan Saragih (e-mail: triando.saragih@ulm.ac.id)

ABSTRACT Lung Cancer is a disease that has a high mortality rate and is often difficult to detect until it reaches a very severe stage. Data indicates that lung cancer cases are typically diagnosed late, posing significant challenges to effective treatment. Early detection efforts offer the potential for better recovery chances. Therefore, this research aims to develop methods for the identification and classification of lung cancer in the hope of providing further knowledge on effective ways to detect this condition at an early stage. One approach under scrutiny involves employing machine learning classification techniques, anticipated to serve as a pivotal tool in early disease detection and enhancing patient survival rates. This study involves five stages: data collection, data preprocessing, data partitioning for training and testing using 10-fold cross validation, model training, and analysis of evaluation results. In this research, four experiments consist of applying two classification methods, CatBoost and Random Forest, each tested using default hyperparameter and hyperparameter tuning using Bayesian Optimization. It was found that the Random Forest model using hyperparameter tuning Bayesian Optimization outperformed the other models with accuracy (0.97106), precision (0.97339), recall (0.97185), f-measure (0.97011), and AUC (0.99974) for lung cancer data. These findings highlight that Bayesian Optimization for hyperparameter tuning in classification models can improve clinical prediction of lung cancer from patient medical records. The integration of Bayesian Optimization in hyperparameter tuning represents a significant step forward in refining the accuracy and effectiveness of classification models, thus contributing to the ongoing enhancement of medical diagnostics and healthcare strategies.

INDEX TERMS Lung Cancer, CatBoost, Random Forest, Bayesian Optimization

I. INTRODUCTION

Cancer is one of the most common causes of death in the world, accounting for nearly 10 million deaths in 2020. The statistics show that nearly 1 in 6 deaths in the world are caused by cancer [1]. Data from the International Agency for Research on Cancer (IARC) estimates that the number of cancer cases worldwide will increase by 28.4 million by 2040 [2]. There are many types of cancer that can be diagnosed in both men and women, such as lung cancer, skin cancer, liver, colon, and rectal cancer. The highest percentage of deaths reaching 19.4% occurs in lung cancer [3]. Lung cancer can originate from organs within the lung (primary) or from outside the lung (metastasis) [4]. Smoking is the biggest risk factor for lung cancer. Men are more often affected by lung cancer than women because smoking mostly occurs in men [5]. In addition to direct smoking, inhaled cigarette smoke also increases the risk of developing lung cancer. Additional factors include genetics, occupation, family history of cancer, coffee consumption of more than six cups per day, meat consumption, fresh vegetable/fruit consumption, preserved or fried food consumption, chronic lung disease, alcohol consumption, air pollution, and chemical exposure [6].

The reason lung cancer is a deadly disease is because it is difficult to detect before it becomes a severe disease. About 85% of lung cancers are only detected after they are in the final stage [7]. Early detection of lung cancer can increase the chances of recovery from this disease. Therefore, early diagnosis of lung cancer is important [8]. A better

understanding of the risk factors for lung cancer symptoms can help in the prevention of the disease. The key to improving survival rates is early detection using classification techniques in machine learning [9].

CatBoost and Random Forest are machine learning algorithms that can be applied to maintain the accuracy of prediction results based on patient medical record data in solving classification tasks. CatBoost (Categorical Boosting) is an open-source machine learning library implemented with the principle of gradient boosting for classification, regression, and ranking tasks. Utilizing the Gradient Boosted Decision Tree (GBDT) algorithm framework, CatBoost exhibits superior performance in handling categorical features and substantially enhances feature dimensionality. Additionally, CatBoost efficiently mitigates overfitting concerns, thereby enhancing prediction accuracy [10]. Research conducted by [11] compared eight machine learning methods, and the results showed the CatBoost model to be the best classifier compared to other classification models. With an accuracy rate of 97.8%, the researchers also mentioned that CatBoost has a very fast prediction ability, which makes it an efficient choice for tasks that require real-time prediction.

Random Forest (RF) is an ensemble machine learning technique applicable to both classification and regression tasks. Random forest is a collection of decision trees, and class determination is based on the majority of votes from all the trees formed [12]. This methodology harnesses the collective wisdom of diverse trees, offering resilience against biases, resilience against data outliers, and guarding against overfitting. By leveraging a diverse set of classifiers, the Random Forest algorithm adeptly navigates complex datasets, ensuring reliable and accurate predictions. In essence, Random Forest emerges as a powerful tool for addressing various challenges encountered in predictive modeling [13].

In this study, to differentiate it from previous studies [14], [15], tuning is used to optimize the performance of the model and tune the hyperparameters in CatBoost and Random Forest to get the optimal hyperparameter value so that the model can provide the best results. Tuning the hyperparameters of machine learning algorithms has a positive effect on the final results. However, the effect of this tuning differs depending on the algorithm used [16]. A study [17] found that the Gradient Boosted Decision Tree algorithm had the most effect on hyperparameter tuning, with an average improvement of 8-11%. Another study conducted by [18] applied the Hyperparameter-Tuning Bayesian Optimization method to adjust the hyperparameters of the Random Forest algorithm, resulting in a more efficient Random Forest evaluation model with a high level of accuracy.

This study seeks to contrast the efficacy of CatBoost and Random Forest methods in assessing their accuracy for lung cancer classification. The aim is to gain insights into the most suitable approach for this specific task, with or without the implementation of Hyperparameter Tuning Bayesian Optimization. This comparison is anticipated to shed light on the advantages and disadvantages of each method, aiding in the development of a more effective classification model for early-stage detection of lung cancer. The examination of these two methods aims to improve the precision and reliability of lung cancer diagnosis, potentially leading to better patient outcomes and survival rates. Ultimately, this research endeavors to contribute to the refinement of lung cancer classification techniques, enhancing the overall quality of healthcare for individuals diagnosed with this disease. The findings of this study are anticipated to offer contributions such as:

a. Provide knowledge on the application of classification techniques and hyperparameter tuning using patient medical records.
b. Introduce Hyperparameter Tuning Bayesian Optimization as a novel approach to optimize the performance of CatBoost and Random Forest methods, especially in lung cancer classification tasks.
c. Aid medical professionals in refining decision-making processes through data analytics.

II. MATERIALS AND METHODS

The research process generally involves comparing the results of two classification methods: CatBoost and Random Forest. Each method undergoes testing under two conditions: using default hyperparameters and through hyperparameter tuning via Bayesian Optimization. This study is structured using five sequential stages: data collection using a lung cancer dataset, data preprocessing, data partitioning for training and testing using 10-fold cross validation, model training, and analysis of evaluation results. The research flow carried out in this study can be seen in FIGURE 1 as follows.

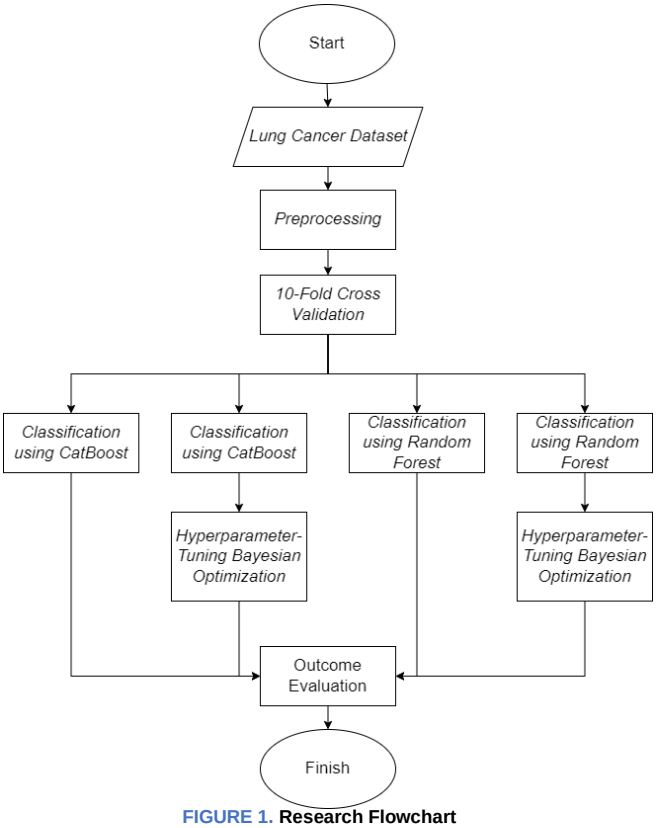


FIGURE 1. Research Flowchart

A. DATA COLLECTION

The dataset used in this study is the Survey Lung Cancer Dataset taken from the Kaggle Repository site, can be seen at

https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. The dataset contains information about what attributes are used as features to classify patients with lung cancer. This dataset consists of 309 instances and 16 attributes, with 15 attributes representing the clinical status of patients used as predictive variables and one attribute designated as the target variable. The attribute description, as outlined in prior research [14], [15], is cataloged in TABLE 1, supplying a comprehensive overview of the dataset attributes and their implications in lung cancer diagnosis.

TABLE 1
Lung Cancer Data Attribute Description

No	Attribute	Description	Value
1	Gender	Participant's gender	M, F
2	Age	Age in years	21 - 87
3	Smoking	Smoker or not	1, 2
4	Yellow Fingers	Has yellow fingers	1, 2
5	Anxiety	Anxious or not	1, 2
6	Peer_pressure	Feels peer pressure	1, 2
7	Chronic Disease	Suffers from a chronic disease	1, 2
8	Fatigue	Suffers from fatigue	1, 2
9	Allergy	Has an allergy	1, 2
10	Wheezing	Suffers from wheezing	1, 2
11	Alcohol	Consumes alcohol	1, 2
12	Coughing	Suffers from coughing	1, 2
13	Shortness of Breath	Has shortness of breath	1, 2
14	Swallowing Difficulty	Has difficulty swallowing	1, 2
15	Chest Pain	Has chest pain	1, 2
16	Lung Cancer	Diagnosed with lung cancer	YES, NO

B. PREPROCESSING

The final prediction may be affected by noise in the raw data and/or missing values. Sometimes, the dataset from secondary sources is not prepared for use in machine learning algorithms. Dataset are pre-processed to prepare them for algorithmic processing. At this stage, null values are checked and corrected, and the data is balanced. Anything that impacts the performance of the machine learning model at this point can be handled more skillfully [15].

First, check the dataset for null values or missing values. It was found that the Survey Lung Cancer dataset does not contain null values or missing values, so no data imputation or replacement of missing values is required. Furthermore, label encoding is used to convert category label values into numerical form so that they can be processed more effectively by machine learning algorithms [19]. For example, in the Survey Lung Cancer dataset, in the target variable which has categorical variables with values {NO and YES} and in the Gender feature with a value set {M = Male and F = Female}, then after the label encoder process it becomes {0, 1} to simplify the modeling process.

The Survey Lung Cancer dataset used in this study shows a significant imbalance in the target variable column, with 270 rows having "YES" values and only 39 rows having "NO" values. To ensure the accuracy and balance of predictions and results, such imbalanced data must be controlled. In addition, 33 duplicate entries in the dataset have also been removed. Thus, after the removal of duplicate entries, the dataset contains 276 entries, with 238 entries indicating cancer and 38 entries indicating non-cancer. The Random Oversampling technique was applied to overcome the uneven distribution of data between the majority and minority classes [20], where "cancer" (the majority class) was oversampled. FIGURE 2 shows the distribution of lung cancer risk before and after balancing the dataset using random oversampling.

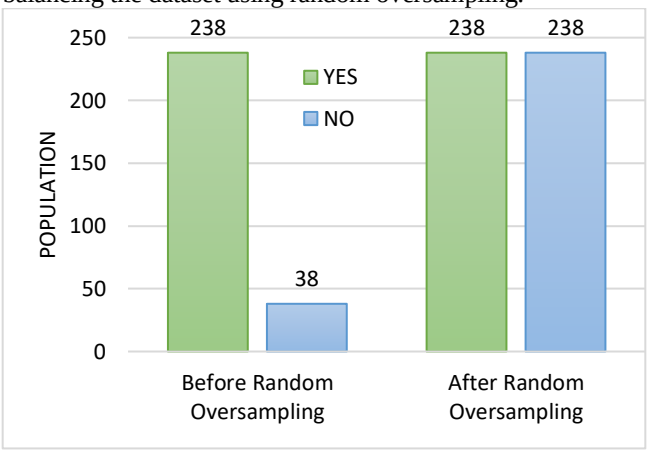


FIGURE 2. Dataset of lung cancer survey before and after Random Oversampling

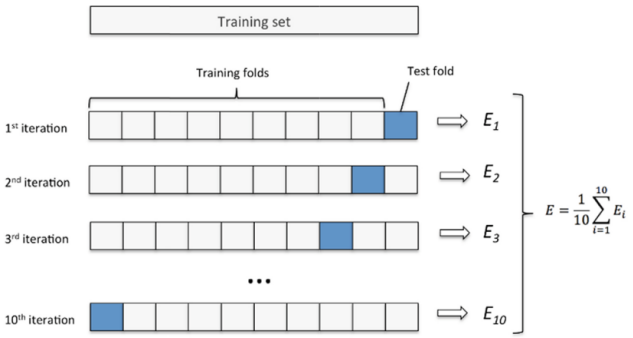


FIGURE 3. Repeated 10-Fold Cross Validation

C. DATA SHARING

In data sharing, this research uses the 10-Fold Cross Validation technique. Cross-validation partitions the initial dataset into training and validation sets. The training set is utilized to train the classification model, enabling the evaluation of its performance. In 10-fold cross-validation, the value of K is set to 10, meaning the dataset is divided into ten subsets. One subset serves as the validation set, while the remaining K-1 subsets are employed as the training set for model testing [21], [22]. The 10-fold cross validation technique is employed to provide an unbiased estimation of the prediction model's performance, facilitating comparison and mitigating the risk of overfitting [23]. A visual representation of the repetition of this data division featuring

the application of 10-fold cross validation according to [24] can be seen in FIGURE 3.

D. CLASSIFICATION

1. CATBOOST CLASSIFICATION

Developed in 2017 by Yandex Researchers and Engineers [25], CatBoost is an open-source machine learning library that uses a type of boosting algorithm. Binary decision trees are used for base predictors in CatBoost. This algorithm has the ability to handle classification features more rationally and efficiently so as to reduce the possibility of overfitting [10]. Proposed by [25], [26] CatBoost demonstrates superior performance and shorter execution time compared to the XGBoost and LightGBM algorithms. CatBoost distinguishes itself from other gradient boosting algorithms by employing ordered boosting, a modification of the gradient boosting algorithm designed to address the target leakage issue efficiently [26]. CatBoost is useful for small datasets and is suitable for handling categorical features [27]. The estimation results described by [26] can be seen in equation (1) as follows.

$$Z = H(x_i) = \sum_{j=1}^{J} c_j 1_{\{x \in R_j\}} \quad (1)$$

where $H(x_i)$ represents the decision tree function of the explanatory variable x_i and R_j is the disjoint region associated with a leaf of the tree. To overcome the problem of prediction error in gradient scaling, [25] developed a new method that involves generating pseudocode shown in TABLE 2.

TABLE 2
CatBoost Algorithm

Ordered Boosting

input: $\{(X_k, y_k)\}_{k=1}^{n}, I$;

$\sigma \leftarrow$ random permutation of [1, n];
$M_i \leftarrow 0$ **for** $i = 1..n$;
for $t \leftarrow 1$ **to** I **do**
  for $i \leftarrow 1$ **to** n **do**
    $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_1)$
  for $i \leftarrow 1$ **to** n **do**
    $\Delta M \leftarrow LearnModel\left((x_j, r_j) : \sigma(j) \le i\right)$;
    $M_i \leftarrow M_i + \Delta M$;
return M_n

2. RANDOM FOREST CLASSIFICATION

Random Forest was first published by [28] officially in 2001. Random Forest was developed to improve decision tree methods that often experience overfitting. The methodology of Random Forest revolves around the creation of numerous decision trees, with the final prediction result being determined through majority voting from all the individual prediction outcomes. This approach effectively overcomes the problems that may arise when performing classification with only one decision tree, which is often not optimal [29].

The ensemble model encompasses two variations: Bagging and Boosting. Random Forest belongs to the Bagging techniques, alternatively referred to as bootstrap aggregation methods[30]. This method operates on two core principles: row sampling and voting classifiers. Initially, the dataset is resampled and fed into the subsequent base learner model for training. Following the training phase, aggregation, or voting classifier, is employed, where the test data's output is determined by the class receiving the highest votes from the base learner models [31]. The schematic representation of the random forest's general model is depicted in FIGURE 4.

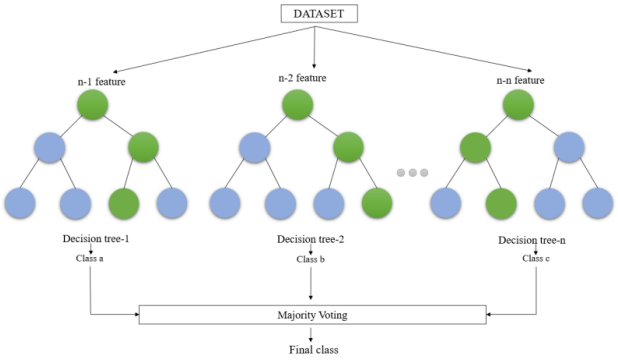


FIGURE 4. Generalized Structure for Random Forest [31]

The Random Forest method starts with the formation of trees, where each decision tree is formed by applying the gini index defined in equation (2) below [32].

$$Gini\ Index\ (D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2)$$

Where P_i is the proportion of the number of attributes in each class, and m is the number of each attribute. The feature that has the lowest total Gini Index value will be the root node in the tree. The total Gini Index at an internal node (e.g., K) is calculated in the following equation (3).

$$Tot.Gini\ Index\ (K) = \frac{T_1}{T} Gini\ Index\ (D_1) + \frac{T_2}{T} Gini\ Index\ (D_2) \quad (3)$$

Where T_1 is the total records belonging to the first class, T_2 is the total records belonging to the second class, and T is the total records of all classes. This process continues with the formation of child nodes until all nodes in the tree cannot be split. After the entire tree is formed, the classification stage continues using the voting method. The stages of completion with the Random Forest algorithm are as follows [33]:

1. Determine the number of trees (k) selected from a total of m features, where k < m.
2. Then random samples are taken as many as N in the dataset for each tree.
3. In each tree, a random subset of predictors is taken, where m < p, p is the number of predictor variables.
4. Then, the process in the second and third steps is repeated for k trees.
5. Prediction results are obtained from the most votes from the classification results of as many trees.

E. BAYESIAN OPTIMIZATION

Bayesian optimization is a popular method in hyperparameter tuning. This method was chosen because of its ability to obtain the optimal value quickly. It utilizes a Gaussian Process (GP) that comprehensively understands

prior knowledge. The algorithm depends on fitting a probability model to the observed value of the target to be optimized. By analyzing the predictive distribution, Bayesian optimization techniques direct the search towards areas of the input space anticipated to offer the most informative insights into solving the optimization problem [34].

Bayesian optimization uses a Gaussian process to update previous values by considering previous parameter information. The Bayes formula is employed to compute the posterior probability distribution along with the mean and variance of the accuracy for each hyperparameter value. A higher average accuracy signifies strong model performance. During subsequent iterations, the process of choosing points with high average accuracy is termed exploitation, whereas selecting points with significant accuracy variance is termed exploration. In the Bayesian optimization procedure, exploration takes precedence in the initial stages, transitioning to exploitation in the later stages [35]. Equation (4), which states that model A and observation B, is the basis for the optimization process based on Bayesian theory [34], [36].

$$P\,(A|B) = \frac{P\,(B|A)\,P(A)}{P(B)} \tag{4}$$

Where P(A) is the prior probability and P(B) represents the probability of a value of variables A and B respectively. P(A|B) and P(B|A) are conditional probabilities, i.e., the posterior of variable A if the value of B is known, and vice versa, the likelihood of variable B if the value of A is known. Equation (4) can be simplified by ignoring the normalization factor P(B), resulting in a simpler formula as shown in Eq. (5) below.

$$P\,(A|B) = P\,(B|A)\,P(A) \tag{5}$$

The steps in operational Bayesian Optimization are as follows [36]:

1. Form a probability model as a surrogate of the objective function.
2. Find hyperparameters that produce optimal performance.
3. Implement these hyperparameters in the actual objective function.
4. Update the surrogate model by incorporating information from the new results.
5. Repeating steps 2-4 until reaching the maximum number of iterations or a set time limit.

F. EVALUATION

1. CONFUSION MATRIX

Confusion Matrix is a method used to evaluate the accuracy and performance of classification algorithms, whether they are used for classifying or predicting attributes. It is designed as an evaluation method for machine learning algorithms used in solving classification problems[37]. The confusion matrix comprises data comparing the system's classification outcomes with the expected classification results [38]. The confusion matrix is a data matrix that juxtaposes the system's classification outcomes with the anticipated classification results. It includes False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP), each of which is precisely detailed in the accompanying table, depicted as TABLE 3. This matrix offers a comprehensive view of the system's performance, enabling a nuanced understanding of classification accuracy and error rates [22], [32].

TABLE 3
Confusion Matrix

Actual Class	Predicted Class	
	Class = Yes	Class = No
Class = Yes	True Positif (TP)	False Negatif (FN)
Class = No	False Positif (FP)	True Negatif (TN)

By using a confusion matrix, it can calculate various evaluation matrices such as accuracy, precision, recall, and f-measure. The following is the calculation formula [14], [15], [32], [39].

1. Accuracy
   Accuracy is the percentage of accuracy of the model in classifying data correctly on the test, with both positive and negative results. Accuracy can be calculated in Eq. (6) as follows.
   $$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$
2. Recall
   Recall is the proportion of true positive estimates to the total true positive data. Recall can be calculated in Eq. (7) as follows.
   $$Recall = \frac{TP}{TP + FN} \tag{7}$$
3. Precision
   Precision is the ratio of the original positive estimate to the overall estimate of the positive result prediction. Precision can be calculated in Eq. (8) as follows.
   $$Precision = \frac{TP}{FP + TP} \tag{8}$$
4. F - Measure
   F-Measure is a metric that combines Precision and Recall into a single value that presents a balance between the two. F-Measure can be calculated in Eq. (9) as follows.
   $$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

2. AREA UNDER CURVE (AUC)

The Area Under Curve (AUC) method quantifies the area beneath the Receiver Operating Characteristic (ROC) curve, serving as a measure to evaluate classification performance[40]. AUC assesses the likelihood that the classification system will assign a higher value to a positive sample than a negative one when randomly selected[41]. Therefore, a higher AUC value signifies the superior quality of the classification method utilized [42]. The categories for the range of AUC values can be seen in TABLE 4.

TABLE 4
Category AUC Value

Category	AUC Value
Excellent Classification	0.90 - 1.00

Good Classification	0.80 - 0.90
Fair Classification	0.70 - 0.80
Poor Classification	0.60 - 0.70
Failure Classification	0.50 - 0.60

AUC has a range from 0 to 1, in this study, AUC is used as a metric to assess the performance of machine learning classification models in classifying between Lung Cancer and Non-Lung Cancer. When the AUC value is close to 1, it indicates that the model has a perfect ability to distinguish two different class distributions [14]. The AUC calculation formula can be seen in Eq. (10) below [38].

$$AUC = \frac{\left(\frac{TP}{TP + FN}\right) * \left(\frac{TN}{TN + FP}\right)}{2} \quad (10)$$

III. RESULTS

This study will present the results of evaluating model performance in lung cancer classification using the CatBoost, CatBoost using Hyperparameter Tuning Bayesian Optimization, Random Forest, and Random Forest using Hyperparameter Tuning Bayesian Optimization. Data that has gone through the preprocessing and data balancing stages is used for testing. The distribution of both training and test data is assessed through the implementation of the 10-Fold Cross Validation method. This entails dividing the dataset into ten equal segments, with one segment allocated as test data and the remaining nine segments utilized for training purposes. This procedure is iterated until each segment has been employed as test data, ultimately yielding an average performance value across all iterations.

A. THE RESULTS OF THE CATBOOST METHOD

This section unveils the experimental findings derived from the utilization of the CatBoost classification model employing default hyperparameters. The performance metrics of this assessment model are comprehensively detailed in TABLE 5.

TABLE 5
CatBoost Results

Set	CatBoost				
	Accuracy	Precision	Recall	F-M	AUC
1	0.9737	0.9737	0.975	0.9744	1
2	0.8684	0.8913	0.875	0.8571	0.991
3	0.9737	0.9762	0.9722	0.9756	1
4	0.9474	0.95	0.95	0.9474	0.9976
5	0.9737	0.975	0.9737	0.973	1
6	0.9474	0.9524	0.9474	0.9444	0.9858
7	0.9211	0.9318	0.9211	0.9143	0.983
8	0.9737	0.975	0.9737	0.973	1
9	0.9474	0.9524	0.9474	0.9444	0.995
10	0.9737	0.975	0.9737	0.973	0.995
Average	**0.95002**	**0.95528**	**0.9509**	**0.9476**	**0.9947**

TABLE 5 exhibits the outcomes derived from executing of the CatBoost model utilizing default hyperparameters, which underwent ten segments through a cross validation process. Subsequent to these segments, it is discerned that the average performance of the model yielded commendable metrics: an accuracy of 0.95002, a precision score of 0.95528, a recall rate of 0.9509, an F-measure of 0.9476, and an AUC value of 0.9947. These results underscore the efficacy of the CatBoost algorithm in achieving high levels of accuracy and reliability in lung cancer classification tasks.

B. THE RESULTS OF THE CATBOOST METHOD USING HYPERPARAMETER TUNING BAYESIAN OPTIMIZATION

This section unveils the experimental findings derived from the utilization of the CatBoost classification model with hyperparameter tuning using the Bayesian Optimization method. The hyperparameters of the Catboost model used for the tuning process in this study are comprehensively depicted in TABLE 6, which provides insight into the description of the hyperparameters and their use in the model.

TABLE 6
List of CatBoost Hyperparameters Used in the Tuning Process

Hyperparameter	Description
learning_rate	Controls the contribution rate of each tree at each iteration
depth	Defines the maximum depth of each tree
iteration	Defining the maximum number of trees in the ensemble
12_leaf_reg	Adjusting tree node weights
border_count	Defining the number of bins in feature quantization
subsample	Setting the data fraction
colsample_bylevel	Setting the feature fraction

Subsequently, the hyperparameter configuration is presented in TABLE 7, which provides a comprehensive overview of the refined hyperparameter settings achieved through the optimization procedure.

TABLE 7
Hyperparameter Setup for CatBoost Method

Hyperparameter	Value
learning_rate	0.01, 1.0
depth	1, 10
iterations	10, 100
l2_leaf_reg	0.1, 10.0
border_count	1, 255
subsample	0.1, 1.0
colsample_bylevel	0.1, 1.0
Best Hyperparameter	**Value**
learning_rate	0.61185806
depth	10
iterations	100
l2_leaf_reg	0.1
border_count	255
subsample	0.1
colsample_bylevel	0.1
Train Accuracy	**0.9974**
Test Accuracy	**0.9792**

After obtaining the best combination value of the hyperparameters from the tuning process, which we can see in TABLE 7, then insert the best hyperparameters into the

CatBoost model. The performance of the CatBoost model using Bayesian Optimization can be seen in TABLE 8.

TABLE 8
CatBoost with Bayesian Optimization Results

Set	CatBoost + Bayesian Optimization				
	Accuracy	Precision	Recall	F-M	AUC
1	0.9737	0.9737	0.975	0.9744	0.984
2	0.8684	0.8913	0.875	0.8571	0.981
3	1	1	1	1	1
4	0.9474	0.95	0.95	0.9474	0.995
5	0.9737	0.975	0.9737	0.973	0.997
6	0.9474	0.9524	0.9474	0.9444	0.975
7	0.9737	0.975	0.9737	0.973	0.974
8	0.9737	0.975	0.9737	0.973	1
9	0.9474	0.9524	0.9474	0.9444	1
10	0.9737	0.975	0.9737	0.973	0.992
Average	**0.95791**	**0.96198**	**0.9589**	**0.9559**	**0.9901**

In TABLE 8, the findings of the CatBoost model enriched with hyperparameter tuning using Bayesian optimization are carefully detailed. By employing a 10-fold cross-validation strategy and obtaining the average, the CatBoost model shows significant performance improvements. Specifically, after applying hyperparameter tuning with Bayesian optimization, the CatBoost model showed significant progress, achieving a commendable accuracy of 0.95791, precision score of 0.96198, recall rate of 0.9589, F-measure of 0.9559, and AUC value of 0.9973. These results underscore the efficacy and substantial potential of using Bayesian optimization techniques to improve the performance of the CatBoost algorithm in lung cancer classification tasks.

C. THE RESULTS OF THE RANDOM FOREST METHOD

This section unveils the experimental findings derived from the utilization of the Random Forest classification model employing default hyperparameters. The performance metrics of this assessment model are comprehensively detailed in TABLE 9.

TABLE 9
Random Forest Results

Set	Random Forest				
	Accuracy	Precision	Recall	F-M	AUC
1	0.9737	0.9737	0.975	0.9744	1
2	0.8684	0.8913	0.875	0.8571	1
3	1	1	1	1	1
4	0.9474	0.95	0.95	0.9474	1
5	0.9737	0.975	0.9737	0.973	1
6	0.9474	0.9524	0.9474	0.9444	0.995
7	0.9474	0.9524	0.9474	0.9444	0.989
8	0.9737	0.975	0.9737	0.973	1
9	0.9737	0.975	0.9737	0.973	1
10	0.9737	0.975	0.9737	0.973	0.989
Average	**0.95791**	**0.96198**	**0.9589**	**0.9559**	**0.9973**

In TABLE 9, the findings from the Random Forest model execution, which includes ten segment iterations using cross validation, are well presented. It can be seen that on average the Random Forest model using the default hyperparameters achieves good performance metrics. Specifically, the model demonstrated an accuracy of 0.95791, precision of 0.96198, recall rate of 0.9589, F-measure of 0.9559, and AUC of 0.9973. These results underscore the efficacy and reliability of the Random Forest algorithm in the context of lung cancer classification tasks.

D. THE RESULTS OF THE RANDOM FOREST METHOD USING HYPERPARAMETER TUNING BAYESIAN OPTIMIZATION

This section unveils the experimental findings derived from the utilization of the Random Forest classification model with hyperparameters tuning using the Bayesian Optimization method. The hyperparameters of the Random Forest model used for the tuning process in this study are comprehensively depicted in TABLE 10, which provides insight into the description of the hyperparameters and their use in the model.

TABLE 10
List of Random Forest Hyperparameters Used in the Tuning Process

Hyperparameter	Description
n_estimators	Number of decision trees to be created in the ensemble
criterion	Criteria for selecting the best features to split nodes
max_features	Number of randomly drawn features at each split
max_depth	Maximum depth of the tree
min_samples_split	Minimum number of samples required to split nodes
min_samples_leaf	The Minimum number of samples required to be a leaf (last) node
bootstrap	Determine whether to use bootstrap samples when building the tree

Subsequently, the hyperparameter configuration is presented in TABLE 11, which provides a comprehensive overview of the refined hyperparameter settings achieved through the optimization procedure.

TABLE 11
Hyperparameter Setup for Random Forest Method

Hyperparameter	Value
n_estimators	10, 1000
criterion	'gini', 'entropy'
max_features	0.1, 1.0
max_depth	1, 20
min_samples_split	2, 20
min_samples_leaf	1, 10
bootstrap	True, False
Best Hyperparameter	**Value**
n_estimators	1000

criterion	gini
max_features	0.1
max_depth	13
min_samples_split	2
min_samples_leaf	1
bootstrap	False
Train Accuracy	**0.9974**
Test Accuracy	**0.9792**

After obtaining the best combination value of the hyperparameters from the tuning process, which we can see in TABLE 11, then insert the best hyperparameters into the CatBoost model. The performance of the CatBoost model using Bayesian Optimization can be seen in TABLE 12.

TABLE 12
Random Forest with Bayesian Optimization Results

Set	Random Forest + Bayesian Optimization				
	Accuracy	Precision	Recall	F-M	AUC
1	0.9737	0.9737	0.975	0.9744	1
2	0.8947	0.9091	0.9	0.8889	1
3	1	1	1	1	1
4	0.9737	0.9737	0.975	0.9744	1
5	1	1	1	1	1
6	0.9737	0.975	0.9737	0.973	1
7	0.9474	0.9524	0.9474	0.9444	1
8	1	1	1	1	1
9	0.9737	0.975	0.9737	0.973	1
10	0.9737	0.975	0.9737	0.973	0.997
Average	**0.97106**	**0.97339**	**0.9718**	**0.9701**	**0.99986**

In TABLE 11, the findings of the Random Forest model enriched with hyperparameter tuning using Bayesian optimization are carefully detailed. By employing a 10-fold cross-validation strategy and obtaining the average, the Random Forest model shows significant performance improvements. Specifically, after applying hyperparameter tuning with Bayesian optimization, the Random Forest model showed significant progress, achieving a commendable accuracy of 0.97106, precision score of 0.97339, recall rate of 0.9718, f-measure of 0.97011, an AUC value of 0.999. These results underscore the efficacy and substantial potential of employing Bayesian optimization techniques to enhance the performance of the Random Forest algorithm in lung cancer classification tasks.

IV. DISCUSSION

This study conducted four experiments based on the previously described research results. The experiments include the application of two classification methods, namely CatBoost and Random Forest, first using default hyperparameters and then using hyperparameters that have been set up through a tuning process using Bayesian Optimization with the aim of improving the prediction performance of the model. It can be seen in TABLE 7 and TABLE 11 that each method shows the best hyperparameter results after the tuning process. However, since some of these parameters, such as *border_count*, *depth*, and *iterations* for the CatBoost method, and *n_estimators* for the Random Forest method are near the extreme of their range, there is still room for improvement. Thus, these values can still be searched and explored for optimal combinations using Bayesian Optimization.

The comparison of assessment outcomes for CatBoost utilizing default hyperparameters and those optimized through Bayesian Optimization hyperparameter tuning is depicted in FIGURE 5, based on the results provided in TABLE 5 and TABLE 8. This visualization offers a comprehensive overview of the performance disparities between the two approaches, highlighting the efficacy of employing Bayesian Optimization in refining the model hyperparameters.

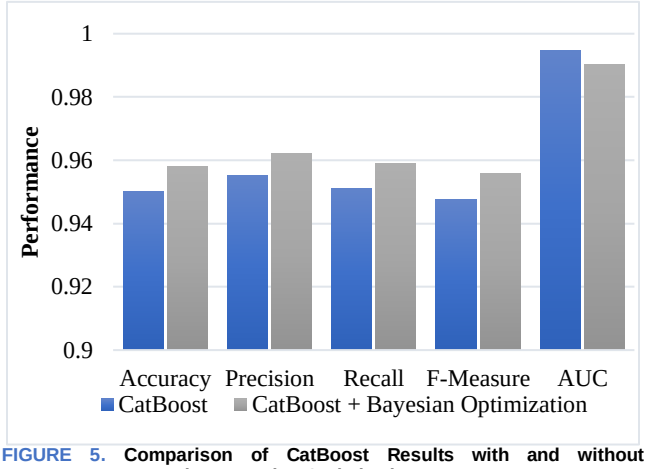


FIGURE 5. Comparison of CatBoost Results with and without Hyperparameter Tuning Bayesian Optimization

FIGURE 5 shows that there is an improvement in the performance of several evaluation metrics, such as accuracy, precision, recall, and F-measure, when the CatBoost model uses Bayesian Optimization for hyperparameter tuning. These results illustrate that hyperparameter tuning can have a real positive impact on improving model performance. The relatively small decrease in the AUC value after hyperparameter tuning may indicate that the tuning has found a more consistent configuration or fit to the data used, resulting in more stable results. The decrease in AUC value from 0.99475 to 0.99014 in the model after hyperparameter tuning indicates an improvement in the model's performance in distinguishing between positive and negative classes. It can be interpreted that the model before and after tuning is stable and reliable in distinguishing between positive and negative classes despite the small numerical change. In other words, hyperparameter tuning has successfully improved the overall performance of the CatBoost model. The comparison of assessment outcomes between the random forest model utilizing default hyperparameters and those refined through hyperparameter tuning via Bayesian Optimization, as delineated in the findings presented in TABLE 9 and TABLE 12 respectively, is depicted in FIGURE 6. This visual representation encapsulates the performance disparities

between the two approaches, providing a comprehensive overview of the effectiveness of hyperparameter tuning in optimizing the random forest model. The graphical illustration serves as a valuable tool for discerning the impact of parameter optimization on key evaluation metrics, such as accuracy, precision, recall, and F-measure. Through this comparative analysis, researchers gain deeper insights into the efficacy of employing Bayesian Optimization to enhance the predictive capabilities of the random forest algorithm.

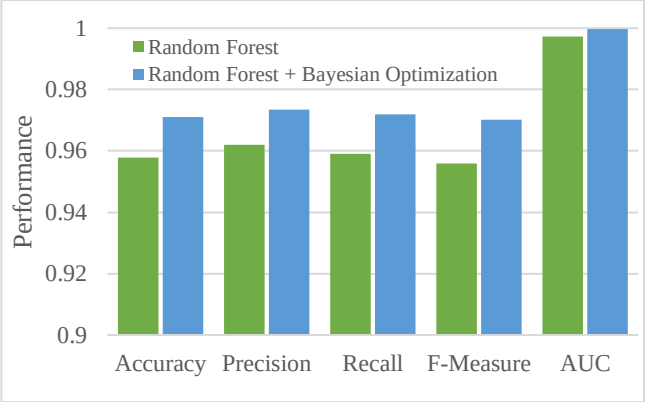
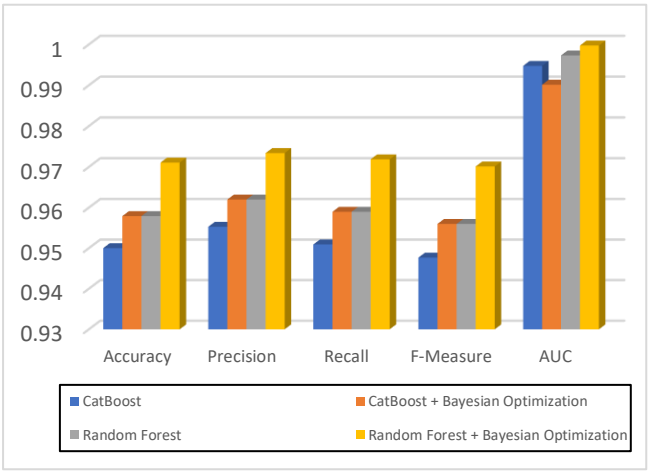


FIGURE 6. Comparison of Random Forest Results with and without Hyperparameter Tuning Bayesian Optimization

FIGURE 6 visually demonstrates a noteworthy enhancement in performance across all evaluation metrics when the Random Forest model integrates Bayesian Optimization for hyperparameter tuning. These findings vividly illustrate the tangible positive effects of hyperparameter tuning on elevating the model's performance within the Random Forest framework. The discernible improvement suggests that the hyperparameter configuration derived through the Bayesian Optimization process aligns more closely with the inherent characteristics of the training data. Consequently, the Random Forest model exhibits heightened efficacy in discerning underlying patterns within the dataset, thereby facilitating more precise predictions. This empirical evidence underscores the pivotal role of hyperparameter tuning in optimizing model performance. It underscores the importance of tailored parameter settings in enhancing the predictive capabilities of machine learning algorithms.

The comprehensive findings of all the evaluations performed on all the models in this study using the lung cancer dataset are carefully presented in FIGURE 7. This figure summarizes the culmination of the analysis and assessment performed on the performance of the models, providing a concise overview of their efficacy in handling the complexity of the lung cancer dataset.

As seen in FIGURE 7, the Random Forest model optimized with Bayesian Optimization provides superior performance compared to other models. In comparison with other studies, the findings of this research indicate that the use of CatBoost and Random Forest combined with Bayesian Optimization hyperparameter tuning is more effective in classifying lung cancer data. Specifically, this methodology elevates the evaluation metric value compared to prior studies that employed alternative classification algorithms or did not use Bayesian Optimization hyperparameter tuning. Additional testing was conducted through comparative analysis of the



evaluation results obtained in this research with the results of others research using the same dataset. This comparison aims to assess the performance achieved in this research in relation to previous research efforts. The comparative results are presented in TABLE 13.

FIGURE 7. Comparison of Evaluation Results

TABLE 13
Comparison of Accuracy Results with Previous Research

Algorithms	Accuracy (%)
NB [15]	91.6
SVM [15]	92.6
k-NN [15]	90.5
AdaBoost [15]	90.5
J48 [15]	90.5
LR [15]	94.7
CatBoost	95
CatBoost + BO	95.7
RF	95.7
RF + BO	97.1

As evidenced in TABLE 13, it can be concluded that the methods employed in this study are superior to previous research. Despite the absence of hyperparameter tuning using Bayesian optimization, both the CatBoost and Random Forest models demonstrate commendable performance when applied to the lung cancer dataset compared to previous research. However, it is noteworthy that more favorable outcomes can be achieved through the implementation of hyperparameter tuning using Bayesian optimization techniques. By using Bayesian Optimization as a hyperparameter tuning method, this research reveals the improved performance of lung cancer classification when two machine learning methods, CatBoost and Random Forest are used. This method helps in finding better hyperparameter configurations that optimize the classification model, allowing the model to classify data more accurately and efficiently.

While the experiments conducted in this study demonstrate significant improvements in model performance through hyperparameter tuning using Bayesian Optimization, there remain certain limitations and areas for further exploration. Despite achieving optimized hyperparameter settings, some parameters, such as *border_count*, *depth*, and *iterations* for the CatBoost method, and *n_estimators* for the Random Forest method, approach the extremes of their respective ranges. This proximity to the boundary suggests potential limitations in further optimizing these hyperparameters, as pushing them beyond their current range could lead to overfitting or computational inefficiencies. Additionally, despite the improvements in performance metrics, the impact of hyperparameter tuning may vary depending on the specific characteristics and distribution of the dataset. Therefore, generalizing the effectiveness of Bayesian Optimization across different datasets and problem domains may require further investigation and validation.

The findings presented in FIGURE 5 and FIGURE 6, alongside the comprehensive evaluation results summarized in FIGURE 7, hold several implications for both research and practical applications in lung cancer classification. Firstly, the significant performance improvements observed in the Random Forest model optimized with Bayesian Optimization underscore the potential of this approach in enhancing the accuracy and efficiency of lung cancer classification models. By leveraging advanced optimization techniques like Bayesian Optimization, researchers and practitioners can effectively navigate the complex landscape of hyperparameter configuration, leading to more robust and reliable predictive models. Furthermore, the comparison with previous studies highlights the superiority of the CatBoost and Random Forest models combined with Bayesian Optimization hyperparameter tuning in classifying lung cancer data. This suggests that adopting advanced machine learning methods and optimization techniques can yield substantial advancements in medical diagnosis and treatment planning. Overall, the findings of this study emphasize the importance of methodological rigor and innovation in developing predictive models for medical applications, ultimately contributing to improved patient outcomes and healthcare delivery.

V. CONCLUSION

In this study, machine learning algorithms were used to identify and classify lung cancer. The approach involves five steps, including collecting data, data preprocessing, dividing the data into training and testing sets using 10-fold cross validation, training the model, and comparing the results. The results illustrate four experiments that include applying two classification methods, CatBoost and Random Forest, using default hyperparameters and a Bayesian Optimization tuning process. Data analysis shows the most effective hyperparameter combinations for each method after hyperparameter tuning with Bayesian Optimization. Nevertheless, further enhancements are possible by exploring their optimal configurations, considering that some hyperparameters may still be at the edge of their value ranges.

Evaluation of the results is done by considering various performance parameters such as accuracy, recall, precision, F-measure, and AUC. After completing the training of all models, it was found that the Random Forest model using Bayesian Optimization hyperparameter tuning outperformed the other models with accuracy 0.97106, precision 0.97339, recall 0.97185, f-measure 0.97011 and AUC 0.99974 for the lung cancer dataset. Thus, this study can help improve the clinical prediction of lung cancer from patients medical records. The findings obtained by the CatBoost model show consistent performance, characterized by stability both before and after hyperparameter tuning. Although it does not surpass the Random Forest model in all evaluation matrices, the CatBoost model shows a fairly good level of accuracy overall. This stability shows that the CatBoost model maintains robustness and reliability at various stages, indicating its potential as a reliable choice for classification tasks.

Nevertheless, in order to further enhance the performance of the approach in predicting lung cancer, future research needs to pay attention to several key aspects. One important thing that needs to be considered in future research is that the dataset used should be larger and diverse in the features included. This will allow the model to learn more complex patterns and provide more accurate predictions. Furthermore, future research could also consider other classification methods that may be more suitable or effective when combined with Bayesian Optimization. Lastly, further exploration of the optimal hyperparameter combination using Bayesian Optimization on the classification method to be used. By making improvements to these aspects, it is hoped that future research can attain more precise and comprehensive outcomes in the classification of lung cancer data.

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] S. V. S. Deo, J. Sharma, and S. Kumar, "GLOBOCAN 2020 Report on Global Cancer Burden: Challenges and Opportunities for Surgical Oncologists," *Ann Surg Oncol*, 2022, doi: 10.1245/s10434-022-12151-6.

[3] K. Tuncal, B. Sekeroglu, and C. Ozkan, "Lung cancer incidence prediction using machine learning algorithms," *Journal of Advances in Information Technology*, vol. 11, no. 2, pp. 91–96, May 2020, doi: 10.12720/jait.11.2.91-96.

[4] H. H. Popper, "Progression and metastasis of lung cancer," *Cancer and Metastasis Reviews*, vol. 35, no. 1, pp. 75–91, Mar. 2016, doi: 10.1007/s10555-016-9618-0.

[5] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung cancer statistics," *Adv Exp Med Biol*, vol. 893, pp. 1–19, 2016, doi: 10.1007/978-3-319-24223-1_1.

[6] W. Li, L. Ah Tse, J. S. K Au, F. Wang, H. Qiu, and I. Tak-sun Yu, "Secondhand smoke enhances lung cancer risk in male smokers: an interaction," 2016. [Online]. Available: http://ntr.oxfordjournals.org/

[7] X. X. Li, B. Li, L. F. Tian, and L. Zhang, "Automatic benign and malignant classification of pulmonary nodules in thoracic computed tomography based on RF algorithm," *IET Image Process*, vol. 12, no. 7, pp. 1253–1264, Jul. 2018, doi: 10.1049/iet-ipr.2016.1014.

[8] S. B. Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, "Progress and prospects of early detection in lung cancer," *Open Biol*, vol. 7, no. 9, 2017, doi: 10.1098/rsob.170070.

[9] P. R. Radhika, R. A. Nair, and G. Veena, "A Comparative Study of Lung Cancer Detection Using Machine Learning Algorithms," in *In 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*, 2019, pp. 1–4.

[10] N. Ke, G. Shi, and Y. Zhou, "Stacking model for optimizing subjective well-being predictions based on the cgss database," *Sustainability (Switzerland)*, vol. 13, no. 21, Nov. 2021, doi: 10.3390/su132111833.

[11] H. Gupta *et al.*, "CATEGORY BOOSTING MACHINE LEARNING ALGORITHM FOR BREAST CANCER PREDICTION," *Rev. Roum. Sci. Techn.-Électrotechn. et Énerg*, vol. 66, pp. 201–206, 2021.

[12] D. Dhiyaussalam, A. Wibowo, F. A. Nugroho, E. A. Sarwoko, and I. M. A. Setiawan, "Classification of Headache Disorder Using Random Forest Algorithm," in *ICICoS 2020 - Proceeding: 4th International Conference on Informatics and Computational Sciences*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ICICoS51170.2020.9299105.

[13] T. L. Octaviani and Z. Rustam, "Random forest for breast cancer prediction," in *AIP Conference Proceedings*, American Institute of Physics Inc., Nov. 2019. doi: 10.1063/1.5132477.

[14] E. Dritsas and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," *Big Data and Cognitive Computing*, vol. 6, no. 4, Dec. 2022, doi: 10.3390/bdcc6040139.

[15] T. R. Ojha, "Machine Learning based Classification and Detection of Lung Cancer," *Journal of Artificial Intelligence and Capsule Networks*, vol. 5, no. 2, pp. 110–128, Jun. 2023, doi: 10.36548/jaicn.2023.2.003.

[16] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, Dec. 2021, doi: 10.3390/informatics8040079.

[17] A. Callens, D. Morichon, S. Abadie, M. Delpey, and B. Liquet, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Applied Ocean Research*, vol. 104, Nov. 2020, doi: 10.1016/j.apor.2020.102339.

[18] D. Sun, H. Wen, D. Wang, and J. Xu, "A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm," *Geomorphology*, vol. 362, Aug. 2020, doi: 10.1016/j.geomorph.2020.107201.

[19] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00305-w.

[20] V. García, J. S. Sánchez, A. I. Marqués, R. Florencia, and G. Rivera, "Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data," *Expert Syst Appl*, vol. 158, Nov. 2020, doi: 10.1016/j.eswa.2019.113026.

[21] H. Wei, C. Hu, S. Chen, Y. Xue, and Q. Zhang, "Establishing a software defect prediction model via effective dimension reduction," *Inf Sci (N Y)*, vol. 477, pp. 399–409, Mar. 2019, doi: 10.1016/j.ins.2018.10.056.

[22] R. T. Yunardi, R. Apsari, and M. Yasin, "Comparison of Machine Learning Algorithm For Urine Glucose Level Classification Using Side-Polished Fiber Sensor," *Journal of Electronics, Electromedical, and Medical Informatics (JEEEMI)*, vol. 2, no. 2, pp. 33–39, 2020, doi: 10.35882/jeeemi.v2i2.1.

[23] C. M. Lynch *et al.*, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *Int J Med Inform*, vol. 108, pp. 1–8, Dec. 2017, doi: 10.1016/j.ijmedinf.2017.09.013.

[24] S. A. Sontakke, J. Lohokare, R. Dani, and P. Shivagaje, "Classification of cardiotocography signals using machine learning," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 439–450. doi: 10.1007/978-3-030-01057-7_35.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features." [Online]. Available: https://github.com/catboost/catboost

[26] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.11363

[27] P. S. Kumar, K. Anisha Kumari, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost ensemble approach for diabetes risk prediction at early stages," in *1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology, ODICON 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ODICON50556.2021.9428943.

[28] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[29] I. Yoo, J. Bi, X. Hu, National Science Foundation (U.S.), and Institute of Electrical and Electronics Engineers, *Proceedings, 2019 IEEE International Conference on Bioinformatics and Biomedicine : November 18-21, 2019, San Diego, CA, USA*.

[30] N. Hamid Arif, M. Reza Faisal, A. Farmadi, D. Turianto Nugrahadi, F. Abadi, and U. Ali Ahmad, "A Comparative Study of Machine Learning Models An Approach to ECG-based Gender Recognition Using Random Forest Algorithm," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 107–115, 2024, doi: 10.35882/jeeemi.v6i2.363.

[31] H. B. Kibria and A. Matin, "The Severity Prediction of The Binary And Multi-Class Cardiovascular Disease -- A Machine Learning-Based Fusion Approach," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.04921

[32] S. Bhanumathi and Dr. S. N. Chandrashekara, "Impute, Select, Decision Tree and Naïve Bayes (ISE-DNC): An Ensemble Learning Approach to Classify the Lung Cancer," 2020. [Online]. Available: https://ssrn.com/abstract=3667438

[33] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2022. doi: 10.1088/1742-6596/2161/1/012033.

[34] D. Sun, J. Xu, H. Wen, and D. Wang, "Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest," *Eng Geol*, vol. 281, Feb. 2021, doi: 10.1016/j.enggeo.2020.105972.

[35] 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, Sep. 2020, doi: 10.1016/j.powtec.2020.07.065.

[36] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042085.

[37] M. Fawwaz Akbar, M. I. Mazdadi, H. Saragih, and F. Abadi, "Implementation of Information Gain Ratio and Particle Swarm Optimization in the Sentiment Analysis Classification of Covid-19 Vaccine Using Support Vector Machine," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 4, pp. 261–270, 2023, doi: 10.35882/jeemi.v5i4.328.

[38] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput Oper Res*, vol. 152, Apr. 2023, doi: 10.1016/j.cor.2022.106131.

[39] N. Banerjee and S. Das, "Prediction Lung Cancer– In Machine Learning Perspective," *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132913.

[40] S. Napi, T. Hamonangan Saragih, D. Turianto Nugrahadi, D. Kartini, and F. Abadi, "Implementation of Monarch Butterfly Optimization for Feature Selection in Coronary Artery Disease Classification Using Gradient Boosting Decision Tree," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 4, pp. 314–323, 2023, doi: 10.35882/jeemi.v5i4.331.

[41] M. R. Ansyari, M. I. Mazdadi, F. Indriani, D. Kartini, and T. H. Saragih, "Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease Using Particle Swarm Optimization Feature Selection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 4, pp. 250–260, 2023, doi: 10.35882/jeemi.v5i4.322.

[42] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the Best Classification Threshold in Imbalanced Classification," *Big Data Research*, vol. 5, pp. 2–8, Sep. 2016, doi: 10.1016/j.bdr.2015.12.001.

AUTHORS BIOGRAPHY

Yra Fatria Zamzam originally from Banjarbaru, South Kalimantan. After graduating from high school, she continued her education to the university level. Since 2020, she has been pursuing her education as a student of the Computer Science Study Program at Lambung Mangkurat University. Her current area of research is in the field of data science. This study program offers her the opportunity to develop her interest in data science. She chose this special interest because of her interest in data science and his deep interest in this field. In addition, her final project was to conduct research centered on the classification of Lung Cancer Disease.

Triando Hamonangan Saragih is a lecturer in the Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. He completed his bachelor's degree in Informatics at Brawijaya University, Malang in 2016. After that, he pursued a master's degree in Computer Science at Brawijaya University, Malang in 2018. The research field he is involved in is Data Science.

Rudy Herteno was born in Banjarmasin, South Kalimantan. After completing high school, he pursued his undergraduate studies in the Computer Science Department at Lambung Mangkurat University and graduated in 2011. Following his undergraduate program, he gained experience as a software developer for several years, particularly focusing on developing software for local governments. In 2017, he obtained his master's degree in Informatics from STMIK Amikom University. Currently, he serves as a lecturer in the Faculty of Mathematics and Natural Sciences at Lambung Mangkurat University. His research interests encompass software engineering, software defect prediction, and deep learning.

Muliadi is a lecturer in the Department of Computer Science at Lambung Mangkurat University, where he specializes in Artificial Intelligence, Decision Support Systems, and Data Science. His academic journey began with a bachelor's degree in Informatics Engineering from STMIK Akakom in 2004, followed by the attainment of a master's degree in Computer Science from Gadjah Mada University in 2009. With expertise in Data Science, he also brings valuable skills in Start-up Business Development, Digital Entrepreneurship, and Data Management Staff.

Dodon Turianto Nugrahadi is a lecturer in the Department of Computer Science at Lambung Mangkurat University. His research interests are focused on Data Science and Computer Networking. He obtained his bachelor's degree in Informatics Engineering from Petra University, Surabaya in 2004. Subsequently, he pursued a master's degree in Information Engineering at Gajah Mada University, Yogyakarta in 2009. Currently, his research areas include Network, Data Science, Internet of Things (IoT), and network Quality of Service (QoS).

Phuoc-Hai Huynh is Lecturer at An Giang University, Vietnam. He have a Ph.D in Computer Science from Can Tho University, Vietnam, and he have published several papers on machine learning applications for biomedical data. His research interests include developing novel algorithms and models for analyzing complex and high-dimensional data, such as genomic sequences, medical images and electronic health records. His skills and expertise: Data Mining and Knowledge Discovery, Machine Learning Classification, Neural Networks, Artificial Intelligence, Pattern Recognition, Supervised Learning