JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine

Fita Sheila Gomiasti[1], Warto[2], Etika Kartikadarma[1], Jutono Gondohanindijo[3] and De Rosal Ignatius Moses Setiadi [1,*]

[1] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia; e-mail : sheilafita62@gmail.com, etika.kartikadarma@dsn.dinus.ac.id, moses@dsn.dinus.ac.id
[2] Informatics Department, Faculty of Dakwah, UIN Profesor Kiai Haji SaifuddinZuhri, Purwokerto, Indonesia; e-mail : warto@uinsaizu.ac.id
[3] Department of Informatics and Engineering, AKI University, Semarang, Indonesia; e-mail : jutono.gondohanindijo@unaki.ac.id
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: This research aims to improve the effectiveness of lung cancer classification performance using Support Vector Machines (SVM) with hyperparameter tuning. Using Radial Basis Function (RBF) kernels in SVM helps deal with non-linear problems. At the same time, hyperparameter tuning is done through Random Grid Search to find the best combination of parameters. Where the best parameter settings are C = 10, Gamma = 10, Probability = True. Test results show that the tuned SVM improves accuracy, precision, specificity, and F1 score significantly. However, there was a slight decrease in recall, namely 0.02. Even though recall is one of the most important measuring tools in disease classification, especially in imbalanced datasets, specificity also plays a vital role in avoiding misidentifying negative cases. Without hyperparameter tuning, the specificity results are so poor that considering both becomes very important. Overall, the best performance obtained by the proposed method is 0.99 for accuracy, 1.00 for precision, 0.98 for recall, 0.99 for f1-score, and 1.00 for specificity. This research confirms the potential of tuned SVMs in addressing complex data classification challenges and offers important insights for medical diagnostic applications.

Keywords: Hyperparameter Tuning; Lung cancer classification; Radial Basis Function Kernel; Random Grid Search; Support Vector Machine.

1. Introduction

Cancer is a deadly disease, with 19.3 million new cases and nearly 10 million deaths in 2020 globally[1]. Lung cancer is the main cause of death, with 34,783 new cases recorded in Indonesia in 2020, causing 30,843 deaths[2]. The lifetime risk for lung cancer is 6.2% in men and 5.8% in women, with a slightly higher prevalence in men [3]. About 80% of lung cancer deaths are caused by smoking, which increases the risk up to 25-fold. Other factors include passive cigarette exposure, radon, asbestos, air pollution, and arsenic in drinking water[4]. Symptoms of lung cancer include cough, chest pain, and shortness of breath, often diagnosed at an advanced stage[5]. Technological advances in various sectors are increasing, especially in the health sector. Data mining technology in the health sector can help diagnose various diseases. Data mining involves several techniques, such as classification, clustering, association, estimation, and prediction[6]–[8], and for identifying/detecting/recognizing diseases, the classification process is generally used. Features that support the classification process and extraction of important attributes from the dataset are needed, which can influence the results of the classification process[9]–[12]. Classification is a type of data mining used to categorize input data into classes or categories determined based on their features[13]–[16]. Specifically, in this study, classification was carried out to identify lung and non-lung cancer categories based on existing data.

Cancer classification has been carried out in various previous studies using machine learning algorithms, such as K-Nearest Neighbor (KNN)[17], [18]; Logistic Regression

Received: February, 4th 2024
Revised: March, 24th 2024
Accepted: March, 25th 2024
Published: March, 25th 2024

(LR)[19], [20]; Random Forest (RF)[21], [22]; Support Vector Machine (SVM)[23]–[25]; Gradient Boosting (GB)[23], [26]; and Light Gradient Boosting Machine (LGBM)[27], [28]. KNN is a simple and easy-to-implement classification method. KNN is susceptible to data outliers and is less efficient for large datasets. LR is a classification method that is simple and easy to interpret. LR is better for modeling binary outcomes, is efficient with small datasets, and provides probability outcomes for decision-making. However, LR is limited to categorical results, assumes linearity in variable relationships, is susceptible to overfitting on multi-feature datasets, and is less effective for capturing non-linear relationships between variables [29]. Research [30] on lung classification shows that SVM and LR have the highest accuracy. Another study [31] aims to predict lung cancer using SVM, KNN, and Convolutional Neural Network (CNN) methods. This research shows that SVM is superior by achieving the best accuracy, namely, 95.56%, while KNN reaches 88.40% and CNN reaches 92.11%. Another study [32] focused on lung cancer and used the dataset [33] by applying the SVM algorithm, achieving an accuracy of 95.4%.

Another study [34] that compared RF and SVM found that the accuracy of RF was 90%, while SVM was 95%. This shows that SVM accuracy is superior. RF has advantages, such as the ability to model non-linear relationships and interactions between variables, identify important features in the process, handle large datasets, reduce variation by taking the average of several decision trees, and overcome overfitting. However, RF has disadvantages in complexity and difficulty of interpretation, is computationally time-consuming, especially for large datasets, is prone to overfitting if not well adjusted, and uses significant memory[35]. Meanwhile, SVM is more effective in high-dimensional spaces, is able to handle complex data, and has many features. SVM produces an accurate and stable model in data classification and can handle non-linear classification problems through kernels[36]. However, SVM has disadvantages such as long model training times, the inability to handle data that has a lot of noise or outliers, and the need for appropriate parameter selection to optimize performance.

SVM has adjustable hyperparameter tuning. Adjusting and setting hyperparameter values is crucial to optimize performance. Hyperparameters are often tuned via Random Grid Search (RGM). RGM is a more efficient and random Grid Search (GM) version. In GM, a grid of hyperparameters is defined, and the system systematically goes through many possible combinations, training a model for each combination and evaluating performance. However, this can be very computationally expensive, especially when the defined grid is huge. However, RGM will randomly select a combination of hyperparameters to try from a given grid. This allows the control of the desired number of iterations and in this way, can significantly reduce the computational time required to find the optimal configuration[37], [38]. Kernel is a fundamental concept that allows SVM to work on data that is not linearly separable by mapping data into a higher dimensional space. Radial Basis Function (RBF) is a kernel that increases the linear separation between data classes by transforming input into a higher dimensional space. SVM searches for an optimal hyperplane to maximize the margin between data classes. Using the RBF kernel has proven effective in dealing with non-linear problems, performing complex data mapping, and optimizing SVM parameters [39]. After optimization, the evaluation of classification performance can be measured using a confusion matrix through a number of metrics, including Accuracy, Recall, Precision, F1-Score, Specificity, and Area Under-curve (AUC)[3], [40].

Based on the description above, SVM is a superior and efficient algorithm that maintains a focus on lung cancer detection with high sensitivity. In accordance with medical interests related to early detection. This research aims to improve the effectiveness of lung cancer classification using SVM with hyperparameter tuning to optimize SVM performance. By combining random oversampling and hyperparameter tuning. Next, SVM performance is measured based on tools such as accuracy, precision, recall, F1-score, specificity, and AUC. The remainder of this paper is presented in four parts, namely related literature in the second part, methodology in the third part, followed by results and discussion in the fourth part, which ends with a conclusion section.

2. Related Works

Several previous studies using the ML approach have researched lung cancer classification. A study [32] focused on lung cancer using a dataset [33] with the Rotation Forest (RoTF)

method. Experimental results show that the model was successfully built, achieving an AUC rate of 99.3%, with F-Measure, precision, recall, and accuracy reaching 97.1%.

Study [41] also used a dataset [33] but applied the Genetic Folding Strategy (GFS) to enhance the kernel function in SVM classification to classify lung cancer. Performance evaluation and comparison were conducted with three types of SVM kernels on the actual lung cancer dataset, and the results showed an accuracy rate of 96.2%, which is the highest compared to the other kernels. Another study [42] aimed at early diagnosis using significantly accurate classification methods to increase the success of lung cancer diagnosis. Applying the Decision Tree (DT) classifier in lung cancer classification significantly increased accuracy, reaching 95.16% at a model depth (max_depth) of 15, tested in 40 experimental iterations.

The study [43] compared several ML algorithms, such as LR, KNN, GB, LGBM, and SVM. The metric evaluation results showed that the RF algorithm achieved the best accuracy of 97%, LR reached 93%, LGBM at 91%, and KNN at 73%. Another study [44] also tested the dataset [33]. Since this dataset has imbalanced classes, a random oversampling technique was used to balance it. Based on the theory, an imbalanced dataset generally reduces classification performance. Additionally, the Shapley Additive Explanation (SHAP) algorithm was used to provide useful insights that can assist in the feature selection process, and the classification was carried out with a GB classifier. The results showed that the robustness of this method achieved an accuracy of 98.76%, precision of 98.79%, recall of 98.76%, F-Measure of 98.76%, and an error rate of 0.16%.

Based on the several studies outlined above, it is explained that they all use the same dataset. This dataset is imbalanced. Hence, some also carried out the class balancing process by random oversampling. Various ML model approaches have provided diverse metric evaluation results. This study retests several models on the same dataset but focuses more on the SVM model with hyperparameter tuning. It also tests the effectiveness of the random oversampling technique.

3. Proposed Method

This section explains the proposed method, the stages of which are illustrated in Figure 1. The proposed method has four main stages: dataset collection, data pre-processing, classification using SVM, and evaluation, which are presented in detail in subsections 3.1 to 3.4.

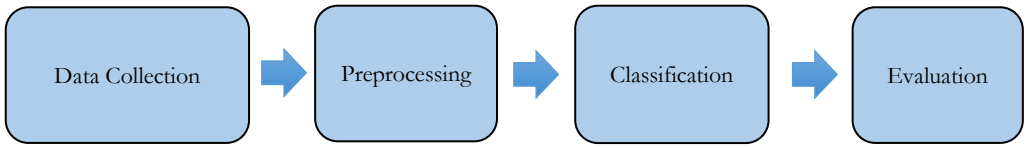

Figure 1. Research stages

3.1. Data Collection

This research uses a secondary dataset, namely the Lung Cancer Survey[33]. This dataset was chosen because it is a dataset that has been widely used so it is easier to compare. This structured data is stored in a CSV file with 309 rows and 16 columns. Each column describes a feature, while each row represents one respondent. Of the 16 columns in Table 1, there are 15 columns as features and 1 as a target or label where features in the lung cancer dataset are used to identify patterns and relationships between various risk factors, habits, and symptoms associated with lung cancer.

Table 1. Attribute Details of Dataset Survey Lung Cancer

No	Attribute	Type	Note
1	Gender	Object	M = Male, F = Female
2	Age	Int	Age range 21 to 87 years old
3	Smoking	Int	1 = No, 2 = Yes
4	Yellow Finger	Int	1 = No, 2 = Yes
5	Anxiety	Int	1 = No, 2 = Yes
6	Peer Pressure	Int	1 = No, 2 = Yes
7	Chronic Disease	Int	1 = No, 2 = Yes

No	Attribute	Type	Note
8	Fatigue	Int	1 = No, 2 = Yes
9	Allergy	Int	1 = No, 2 = Yes
10	Wheezing	Int	1 = No, 2 = Yes
11	Alcohol	Int	1 = No, 2 = Yes
12	Coughing	Int	1 = No, 2 = Yes
13	Shortness of Breath	Int	1 = No, 2 = Yes
14	Swallowing Difficulty	Int	1 = No, 2 = Yes
15	Chest Pain	Int	1 = No, 2 = Yes
16	Lung Cancer	Object	Yes or No (Target/Label)

3.2 Pre-processing

In this stage, data pre-processing is carried out on raw data to improve data quality and ensure that the data is ready to be used by the ML model. The initial dataset consists of 309 rows and 16 columns that will be utilized. The initial step in pre-processing is to ensure that no values are lost to avoid negative influences on subsequent processing. In trials using the df.isnull().sum() function, no missing values were found in the dataset. The next pre-processing step is checking for duplicate data, which means copies of identical data. The existence of duplicate data can affect the analysis results; therefore, duplicate data is removed. In this dataset, 33 duplicate data were identified. After deleting, 276 rows remain for the next process. Then, data encoding is carried out to convert categorical data into numerical form. This study uses binary encoding in the target variable where 'YES' becomes 1 and 'NO' becomes 0. In the gender category variable, change 'M' to 1 and 'F' to 0. Then change '1' to 0 and ' 2' becomes 1 to make it easier to understand the dataset. Figures 2 and 3 present the sampling dataset before and after encoding.

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
5	F	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES
6	M	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES
7	F	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES
8	F	68	2	1	2	1	1	2	1	1	1	1	1	1	1	NO
9	M	53	2	2	2	2	2	1	2	1	2	1	1	2	2	YES

Figure 2. Sample Data Before Encoding

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	1	69	0	1	1	0	0	1	0	1	1	1	1	1	1	1
1	1	74	1	0	0	0	1	1	1	0	0	0	1	1	1	1
2	0	59	0	0	0	1	0	1	0	1	0	1	1	0	1	0
3	1	63	1	1	1	0	0	0	0	0	1	0	0	1	1	0
4	0	63	0	1	0	0	0	0	0	1	0	1	1	0	0	0
5	0	75	0	1	0	0	1	1	1	1	0	1	1	0	0	1
6	1	52	1	0	0	0	0	1	0	1	1	1	1	0	1	1
7	0	51	1	1	1	1	0	1	1	0	0	0	1	1	0	1
8	0	68	1	0	1	0	0	1	0	0	0	0	0	0	0	0
9	1	53	1	1	1	1	1	0	1	0	1	0	0	1	1	1

Figure 3. Sample Data After Encoding

The next step is data normalization to balance the values to produce a range of values with a low-value range, where the values are from 0 to 1. Normalization helps prevent certain features from dominating the data process. Data normalization in this study used standardization (z-score normalization). The function used in this processing is StandardScaler(). To see a sample of the results of this processing stage, see Figure 4.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.987473	0.778486	-1.008439	1.016950	1.139964	-0.829423	-0.877221	0.862316	-0.762493	1.200427	1.221535	1.092496	0.851257	1.445998	1.060660
1	0.987473	1.324711	0.991632	-0.983332	-0.877221	-0.829423	1.139964	0.862316	1.311488	-0.833037	-0.818642	-0.915335	0.851257	1.445998	1.060660
2	-1.012685	-0.313965	-1.008439	-0.983332	-0.877221	1.205657	-0.877221	0.862316	-0.762493	1.200427	-0.818642	1.092496	0.851257	-0.691564	1.060660
3	0.987473	0.123015	0.991632	1.016950	1.139964	-0.829423	-0.877221	-1.159667	-0.762493	-0.833037	1.221535	-0.915335	-1.174734	1.445998	1.060660
4	-1.012685	0.123015	-1.008439	1.016950	-0.877221	-0.829423	-0.877221	-1.159667	-0.762493	1.200427	-0.818642	1.092496	0.851257	-0.691564	-0.942809
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
471	0.987473	-0.750945	0.991632	-0.983332	-0.877221	-0.829423	-0.877221	0.862316	1.311488	-0.833037	-0.818642	-0.915335	0.851257	-0.691564	1.060660
472	0.987473	0.669241	0.991632	-0.983332	1.139964	-0.829423	-0.877221	0.862316	-0.762493	-0.833037	-0.818642	-0.915335	-1.174734	-0.691564	-0.942809
473	0.987473	-1.624905	0.991632	1.016950	-0.877221	-0.829423	1.139964	-1.159667	-0.762493	-0.833037	-0.818642	-0.915335	-1.174734	-0.691564	1.060660
474	-1.012685	-0.532455	0.991632	1.016950	-0.877221	1.205657	-0.877221	-1.159667	-0.762493	-0.833037	-0.818642	-0.915335	-1.174734	-0.691564	1.060660
475	-1.012685	-0.532455	0.991632	1.016950	-0.877221	-0.829423	-0.877221	-1.159667	-0.762493	-0.833037	-0.818642	-0.915335	0.851257	-0.691564	-0.942809

Figure4. Sample Data Normalization Results

After checking the proportion of labels on the target variable, namely LUNG CANCER, as seen in Figure 5, it was found that the proportion of labels (NO:YES) or (0:1) showed a significant imbalance in the dataset labels, both before and after pre-processing. . This imbalanced proportion can affect classification performance, tending to provide a bias towards the majority class [45], so measures are needed to balance the data. Therefore, as in research, the Random Oversampling technique was applied to overcome this imbalance[44]. The Random Oversampling technique works by randomly adding samples from the minority class to the dataset to increase the number of samples in the minority class in the hope that the machine learning model can learn better from that class[46]. After the balancing process, it can be seen in Figure 5 After Oversampling.

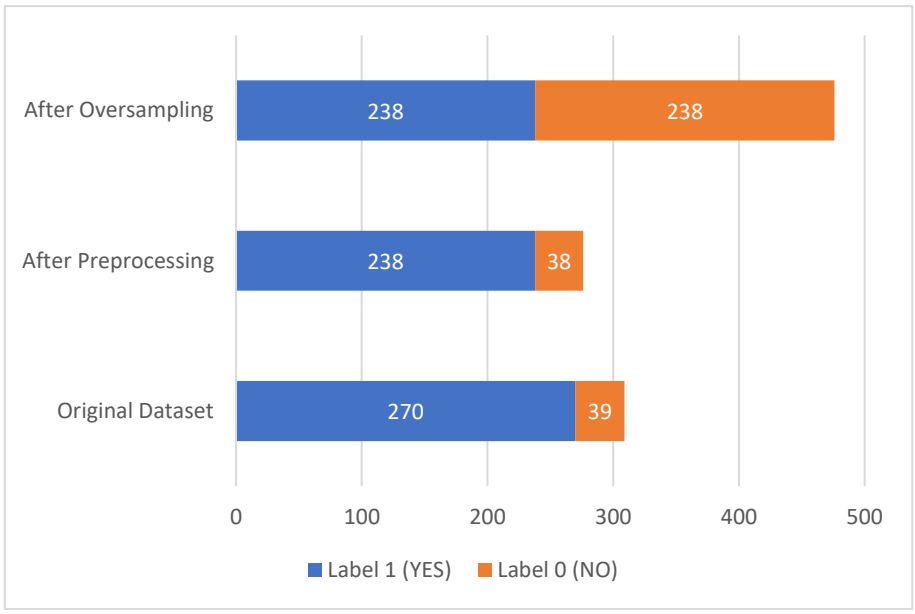

Figure 5. Comparison Label Lung Cancer

The final step is to divide the data into train and test data, which is used to build the model and test data to evaluate model performance. Test data and training data are divided using the 5-fold cross-validation technique. Where the data used for the next stage is data after oversampling and data after pre-processing.

3.3 Classification Stage

SVM is proposed as a classifier at this stage because it is relatively good at handling complex dataset problems, has many features, and is capable of handling imbalanced data. SVM can overcome non-linear problems and adjust flexible parameters to adapt to lung cancer datasets. The parameters used are C (Regularization Parameter), Gamma (kernel Coefficient), and Probability. Hyperparameter C, which applies to all SVM kernels, divides the weights between incorrectly classified training examples and the hyperplane surface. A low C

value makes the hyperplane surface smoother, while a high C value tries to identify all training samples accurately, but a small C value will provide a wider margin with a higher error tolerance. Meanwhile, gamma determines how big the effect of a single training example is; the higher the gamma value, the closer the additional instances must be to be affected [41]. Another parameter is Probability, which produces probability estimates for each class, and the kernel used is the Radial Basis Function (RBF).

Random Grid Search and Grid Search are general methods for finding optimal parameters for tuning machine learning models. In Grid Search, combinations of parameter values are evaluated systematically in a grid, but the disadvantage is that it consumes significant time and resources, especially in large parameter spaces. As an efficient alternative, Random Grid Search selects random combinations of parameters from the entire parameter space, overcoming time and resource constraints, especially when only a small portion of the parameter space significantly impacts model performance. Therefore, Random Grid Search is a more computationally efficient choice in large parameter spaces and is faster in finding optimal parameter combinations[38].

3.4 Evaluation

In this research, the evaluation of the calcification method was carried out using two evaluation metrics, namely:

1. Confusion Matrix

Accuracy, precision, recall, F1-score, and specificity can be measured based on the confusion matrix. Accuracy measures the overall correctness of the model in predicting classes, which can be calculated using Equation (1). Precision focuses on the level of correct positive predictions, which can be calculated using Equation (2). Meanwhile, Recall assesses the model's ability to identify true positive instances from all existing positive instances, which can be calculated by Equation (3). F1-score, as a metric that creates a balance between precision and recall, provides a balanced assessment of model performance, which can be calculated in Equation (4). Then, Equation (5) is the specificity to correctly identify the negative class without misclassifying it as a positive class[40], [47].

$$\text{acc} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$\text{precision} = \frac{TP}{(TP + FP)} \tag{2}$$

$$\text{recall} = \frac{TP}{(TP + FN)} \tag{3}$$

$$\text{F1} = 2 * \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{Recall})} \tag{4}$$

$$\text{specificity} = \frac{TN}{(TN + FP)} \tag{5}$$

Where TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative.

2. Area Under Curva (AUC)

AUC is used to see how well the model can separate between two classes. AUC ranging between zero and one is used to determine the machine learning model with the best performance in differentiating lung cancer cases from non-lung cancer. The higher the AUC, the better the model's ability to differentiate between two class distributions.

4. Results and Discussion

This study is implemented using Python, utilizing various libraries for data pre-processing, classification, and visualization. In the SVM modeling process for lung cancer classification, a hyperplane is used to separate two classes in the feature space, with SVM parameters such as C, Gamma, and Probability playing a role in determining the position and orientation of the hyperplane. The values for the parameters are obtained after experimentation,

and the optimal parameters are presented in Table 3, which shows the SVM parameter settings used. By adjusting these parameters, SVM aims to find the best hyperplane that maximizes the margin (the distance between the hyperplane and the nearest point of each class). The margin is a safe area between the two classes, and SVM strives to find the hyperplane that maximizes this margin. SVM parameters will be optimized using the Random Grid Search method, where a random combination of parameters is evaluated to find the best configuration. Random Grid Search is more efficient and faster in finding the optimal parameter combination.

Table 2. SVM Parameter Settings

Models	Original Parameter	Tuned Parameters
SVM	C =1	C = 10
	Gamma = 0.1	Gamma = 10
	Probability = False	Probability = True
	Kernel type = RBF	Kernel type = RBF

After combining the parameters with Random Grid Search, we get the best parameters, namely C=10, Gamma=10, and Probability=True. Large C values indicate a model tendency for moderate complexity, while smaller values tend to produce simpler models. Meanwhile, a high Gamma value indicates the model's desire to create sharp decision boundaries, while a lower value can produce more diffuse decision boundaries. With C = 10 and gamma = 10, the model selects a moderate level of complexity and sharp decision boundaries, resulting from the optimal performance evaluation in cross-validation. Then the Radial Basis Function (RBF) Kernel became one of the most commonly used kernels for disease prediction and classification. The RBF kernel can handle non-linear problems, map complex data, and work optimally in determining the best parameters for the SVM model [39]. The non-linear concept refers to the SVM's ability to handle non-linear relationships between input and output variables. In the features in the dataset used, there are non-linear problems because the relationships between features are more complex.

This research involves several models, including KNN, LR, RF, GB, and LGBM. Each model is adjusted to the training data, and the target variables in the test data are classified. Its performance is measured using evaluation metrics such as accuracy, precision, recall, F1-score, and AUC. The test results were analyzed to understand the model's ability to classify lung cancer by conducting a sensitivity analysis to evaluate the model's response to parameter changes. This approach provides a solid basis for evaluating the quality of research methods. Specifically, Table 4 evaluates model performance after SVM achieved random oversampling with 5-fold cross-validation and of all the best performance.

Table 3. Comparative Model Performance Evaluation

Models	Accuracy	Precision	Recall	F1-score	Specificity	AUC
LR	0.93	0.96	0.90	0.93	0.87	0.94
KNN	0.95	1.00	0.90	0.95	0.86	0.95
RF	0.95	1.00	0.90	0.95	0.90	0.88
SVM (tuned)	0.99	1.00	0.98	0.99	1.00	0.99
SVM (non-tuned)	0.95	1.00	0.91	0.91	0.90	0.97
GB	0.95	1.00	0.90	0.95	0.90	0.95
LGBM	0.95	1.00	0.90	0.95	0.89	0.95

Table 4 shows that SVM-tuned performs best with 99% accuracy, 98% recall, 100% precision, 100% specificity, 99% F1, and 99% AUC in classifying the Lung Cancer dataset. Accuracy measures the model's ability to predict positive and negative classes. However, in the case of class imbalance, where the number of patients with non-lung cancer is less, the accuracy may be unrepresentative. Precision assesses the percentage of correct positive predictions, which is especially important in predicting diseases such as lung cancer to avoid

serious consequences of false positives. Recall or sensitivity indicates how well the model can detect all true positive cases, which is very important in the context of disease prediction to prevent false negative errors that have serious consequences. Specificity is measured to detect true negative classes. AUC measures how well a model is at separating two classes.

A confusion matrix is a table used to explain the performance of classification models based on prediction results from actual data. This is useful in calculating evaluation metrics so that you can analyze the extent to which the model is classifying data well and identifying errors made by the model. The positive ('1') and negative ('0') prediction results in the confusion matrix can be seen in Figure 7. Based on Figure 7, the model apparently succeeded in predicting 72 cases as class '0' which were actually class '0' (TN). There were no cases of being incorrectly predicted as class '1' when it was actually class '0' (FP). There was one case where it was incorrectly predicted as class '0' when it was actually class '1'(FN). And the model succeeded in predicting 70 cases as class '1' which were actually class '1'(TP). This research also conducted an ablation study to compare the effects of hyperparameter tuning and random oversampling, the results of which are presented in Figure 7.

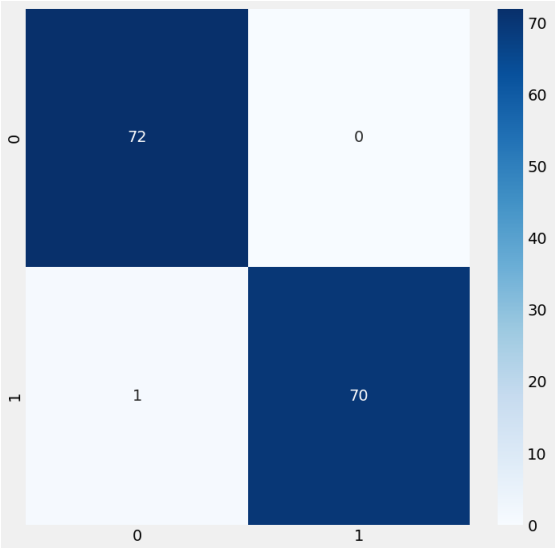

Figure 6. Confusion Matrix Results

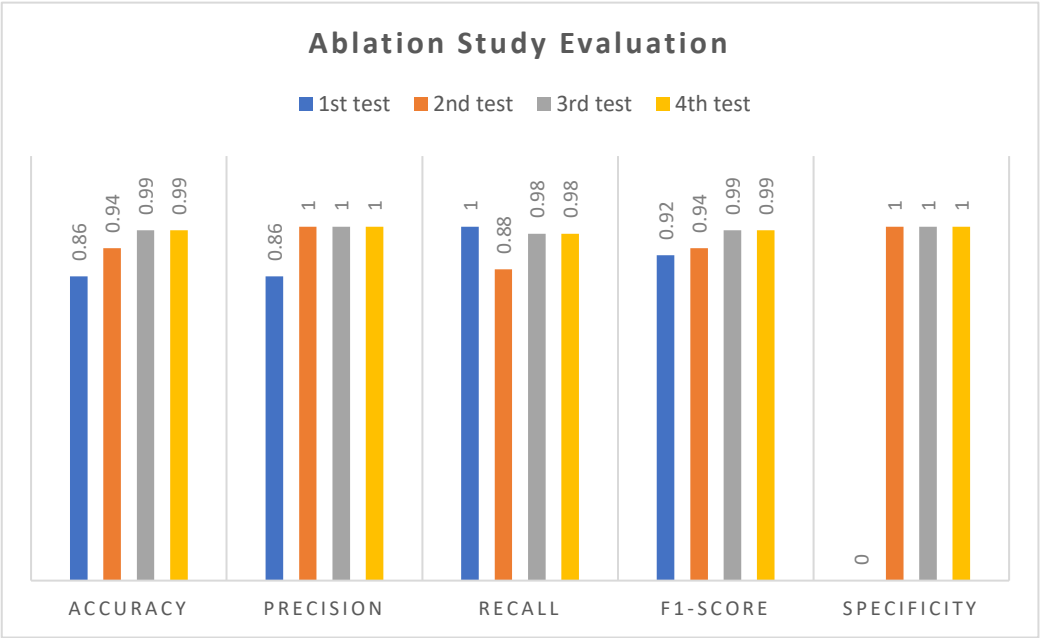

Figure 7. Ablation Study Evaluation

The following is an explanation of Figure 7:
1st test: SVM (defaults) with pre-processed dataset
2nd test: SVM (defaults) with random oversampling dataset
3rd test: SVM (tuned) with random oversampling dataset
4th test: SVM (tuned) with pre-processed dataset

Based on the graph presented in Figure 7, it appears that hyperparameter tuning in the SVM model has a very significant effect. In contrast, the application of random oversampling does not provide any performance improvement. The results in the fourth and third experiments were the same. This is possible because the random oversampling method works by randomly multiplying samples from the minority class, so it does not provide maximum effect. Even at the pre-processing stage, duplicate datasets are removed. So, tuning the parameters has a much better effect because it directly affects the classifier's performance. Furthermore, in Table 5, a comparison is presented with related research that uses the same dataset, namely dataset [33].

Table 4. Comparison with prior art

Methods	Accuracy	Precision	Recall	F1-score	Specificity
Method [32] Rotation Forest (RoTF)	0.97	0.97	0.97	0.97	-
Method [32] SVM	0.95	0.95	0.95	0.95	-
Method [41] SVM	0.96	-	-	-	-
Method [42] Decision Tree (DT)	0.95	-	-	-	-
Method [43] RF	0.97	0.95	1.00	0.98	-
Method [44] GB	0.98	0.98	0.98	0.98	-
Proposed SVM	0.99	1.00	0.98	0.99	1.00

After looking at the model comparison with other research, it can be concluded that SVM achieved the best performance with a percentage of 0.99 for accuracy, 1.00 for precision, 0.98 for recall, 0.99 for f1-score, and 1.00 for specificity. SVM is proven to be able to handle complex data and has many features, thus providing accurate predictions. However, research [14] said that recall in disease classification is considered very important to prevent false negative errors which have serious consequences, especially on datasets. Specificity also plays an important role in avoiding misidentifying negative cases[40]. Based on these considerations, recall, and specificity are used as metrics for evaluating effective models in classifying diseases.

5. Conclusions

This study successfully implemented SVM with hyperparameter tuning for lung cancer classification, achieving significant performance with high evaluation metrics. Hyperparameter tuning and the use of RBF kernels in SVM effectively improve the capacity of the model to differentiate between lung cancer and non-lung cancer cases, which is important in a medical context for early detection. Compared with previous studies using the same dataset, the SVM model tuned in this study showed superior performance, confirming its effectiveness in disease classification. Another finding worth underlining is that the use of random over-sampling did not provide a significant effect because it worked by doubling minority data. In future research, other oversampling techniques should be used that do not duplicate minority data. This research makes an important contribution to improving lung cancer diagnostics, demonstrating the importance of hyperparameter tuning in improving classification accuracy in medical use.

Author Contributions: Conceptualization: F.S.G. and D.R.I.M.S.; methodology: F.S.G. and D.R.I.M.S; software: F.S.G. and D.R.I.M.S; validation: W., E.K. and J.G.; formal analysis: W., E.K. and J.G.; investigation: All.; resources: F.S.G. and D.R.I.M.S.; writing—original draft preparation: F.S.G.; writing—review and editing: All.; visualization: W. and J.G.; supervision: D.R.I.M.S, W., E.K. and J.G.; project administration: E.K.; funding acquisition: All.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

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] G. C. Observatory and Global Cancer Observatory, "Global Cancer Observatory," Nov. 2021. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf

[3] M. Vedaraj, C. S. Anita, A. Muralidhar, V. Lavanya, K. Balasaranya, and P. Jagadeesan, "Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6s, pp. 838–848, 2023.

[4] A. C. Society, "Cancer Facts & Figures 2023." Nov. 16, 2023. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cff-special-section-lung-cancer.pdf

[5] World Health Organization, "Cancer." Nov. 16, 2023. [Online]. Available: https://www.who.int/health-topics/cancer

[6] W.-T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Mil. Med. Res.*, vol. 8, no. 1, p. 44, Aug. 2021, doi: 10.1186/s40779-021-00338-z.

[7] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 2, no. 1, 2024, doi: 10.62411/jcta.10057.

[8] M. S. Sunarjo, H. Gan, and D. R. I. M. Setiadi, "High-Performance Convolutional Neural Network Model to Identify COVID-19 in Medical Images," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 19–30, Aug. 2023, doi: 10.33633/jcta.v1i1.8936.

[9] Z. Rustam and S. A. A. Kharis, "Comparison of Support Vector Machine Recursive Feature Elimination and Kernel Function as feature selection using Support Vector Machine for lung cancer classification," *J. Phys. Conf. Ser.*, vol. 1442, no. 1, p. 12027, Jan. 2020, doi: 10.1088/1742-6596/1442/1/012027.

[10] T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, "Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 138–147, Mar. 2024, doi: 10.62411/jcta.10129.

[11] S. Ali, A. Hashmi, A. Hamza, U. Hayat, and H. Younis, "Dynamic and Static Handwriting Assessment in Parkinson ' s Disease : A Synergistic Approach with C-Bi-GRU and VGG19," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 151–162, 2023, doi: 10.33633/jcta.v1i2.9469.

[12] F. Omoruwou, A. A. Ojugo, and S. E. Ilodigwe, "Strategic Feature Selection for Enhanced Scorch Prediction in Flexible Polyurethane Form Manufacturing," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 126–137, Mar. 2024, doi: 10.62411/jcta.9539.

[13] M. Siraj-Ud-Doula and M. A. Alam, "Ecological Data Analysis Based on Machine Learning Algorithms," p. 18, Dec. 2018, [Online]. Available: https://arxiv.org/abs/1812.09138

[14] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 41–48, Jan. 2023, doi: 10.33633/jcta.v1i1.9190.

[15] H. T. Adityawan, O. Farroq, S. Santosa, H. M. M. Islam, M. K. Sarker, and D. R. I. M. Setiadi, "Butterflies Recognition using Enhanced Transfer Learning and Data Augmentation," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 115–128, Nov. 2023, doi: 10.33633/jcta.v1i2.9443.

[16] N. N. Wijaya, D. R. I. M. Setiadi, and A. R. Muslikh, "Music-Genre Classification using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 13–26, Jan. 2024, doi: 10.62411/jcta.9655.

[17] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," *J. Robot. Control*, vol. 2, no. 3, Jan. 2021, doi: 10.18196/jrc.2363.

[18] H. Karamti *et al.*, "Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach," *Cancers (Basel).*, vol. 15, no. 17, p. 4412, Sep. 2023, doi: 10.3390/cancers15174412.

[19] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," *J. Thorac. Dis.*, vol. 11, no. S4, pp. S574–S584, Mar. 2019, doi: 10.21037/jtd.2019.01.25.

[20] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," *Comput. Biol. Med.*, vol. 134, p. 104431, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104431.

[21] S. Dasariraju, M. Huo, and S. McCalla, "Detection and Classification of Immature Leukocytes for Diagnosis of Acute Myeloid Leukemia Using Random Forest Algorithm," *Bioengineering*, vol. 7, no. 4, p. 120, Oct. 2020, doi: 10.3390/bioengineering7040120.

[22] B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in African women using Random Forest Classifier," *Cancer Treat. Res. Commun.*, vol. 28, p. 100396, Jan. 2021, doi: 10.1016/j.ctarc.2021.100396.

[23] M. R. Abbasniya, S. A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, "Classification of Breast Tumors Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods," *Comput. Electr. Eng.*, vol. 103, no. 1, p. 108382, Jan. 2022, doi: 10.1016/j.compeleceng.2022.108382.

[24] P. Arunachalam *et al.*, "Synovial Sarcoma Classification Technique Using Support Vector Machine and Structure Features," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 1241–1259, Jan. 2022, doi: 10.32604/iasc.2022.022573.

[25] B. A. Akinnuwesi *et al.*, "Application of support vector machine algorithm for early differential diagnosis of prostate cancer," *Data Sci. Manag.*, vol. 6, no. 1, pp. 1–12, Mar. 2023, doi: 10.1016/j.dsm.2022.10.001.

[26] H. Tabrizchi, M. Tabrizchi, and H. Tabrizchi, "Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree," *SN Appl. Sci.*, vol. 2, no. 4, p. 752, Apr. 2020, doi: 10.1007/s42452-020-2575-9.

[27] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021, doi: 10.3390/diagnostics11091714.

[28] W. Bakasa and S. Viriri, "Light Gradient-Boosting Machine Edge Detection With Cropping Layer for Semantic Segmentation of Pancreas," *Adv. Artif. Intell. Mach. Learn.*, vol. 03, no. 03, pp. 1274–1294, Jan. 2023, doi: 10.54364/AAIML.2023.1175.
[29] F. Su *et al.*, "Prognostic models for breast cancer: based on logistics regression and Hybrid Bayesian Network," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 120, Jul. 2023, doi: 10.1186/s12911-023-02224-1.
[30] M. Pyingkodi, R. Mahalakshmi, and M. Gowthami, "Performance Evaluation Of Machine Learning Algorithm For Lung Cancer," vol. 12, no. 03, p. 11, 2021.
[31] D. Mustafa Abdullah, A. Mohsin Abdulazeez, and A. Bibo Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques," *Qubahan Acad. J.*, vol. 1, no. 2, pp. 141–149, May 2021, doi: 10.48161/qaj.v1n2a58.
[32] E. Dritsas and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 139, Nov. 2022, doi: 10.3390/bdcc6040139.
[33] M. A. Bhat, "Lung Cancer Classification Dataset." Nov. 05, 2023. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[34] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, p. 815, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14785.
[35] H. Naik, K. Yashwanth, S. P, and N. Jayapandian, "Machine Learning based Food Sales Prediction using Random Forest Regression," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, Dec. 2022, pp. 998–1004. doi: 10.1109/ICECA55336.2022.10009277.
[36] B. Yassin, C. Mohamed, and A.-A. Yassine, "A Nonlinear Support Vector Machine Analysis Using Kernel Functions for Nature and Medicine," *E3S Web Conf.*, vol. 319, p. 01103, Nov. 2021, doi: 10.1051/e3sconf/202131901103.
[37] D. A. Anggoro and S. S. Mukti, "Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 6, pp. 198–207, Dec. 2021, doi: 10.22266/ijies2021.1231.19.
[38] R. Akbarinia, "Parallel Techniques for Big Data Analytics," Université de Montpellier, 2019. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/tel-02169414
[39] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *Int. J. Inf. Technol.*, vol. 15, no. 2, pp. 965–980, Feb. 2023, doi: 10.1007/s41870-019-00409-4.
[40] Muljono, S. A. Wulandari, H. Al Azies, M. Naufal, W. A. Prasetyanto, and F. A. Zahra, "Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI," *IEEE Access*, vol. 12, pp. 9292–9307, Jan. 2024, doi: 10.1109/ACCESS.2024.3353788.
[41] M. A. Mezher, A. Altamimi, and R. Altamimi, "A Genetic Folding Strategy Based Support Vector Machine to Optimize Lung Cancer Classification," *Front. Artif. Intell.*, vol. 5, p. 826374, Jun. 2022, doi: 10.3389/frai.2022.826374.
[42] W. Setiawan, J. Banjarnahor, M. F. Shandika, A. -, and M. Radhi, "Analysis of Classification of Lung Cancer using The Decision Tree Classifier Method," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 7, no. 1, pp. 121–131, Aug. 2023, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v7i1.4136.
[43] N. Devihosur and R. K. M. G, "Enhancing Precision in Lung Cancer Diagnosis Through Machine Learning Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, Jan. 2023, doi: 10.14569/IJACSA.2023.01408116.
[44] S. T. Rikta, K. M. M. Uddin, N. Biswas, R. Mostafiz, F. Sharmin, and S. K. Dey, "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer," *J. Pathol. Inform.*, vol. 14, p. 100307, Jan. 2023, doi: 10.1016/j.jpi.2023.100307.
[45] D. Dablain, B. Krawczyk, and N. V Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.
[46] M. M. Pushpalatha and N. Indira, "Application and Comparison of Majority Weighted Minority Oversampling Techniques and Random OverSampling Examples Data Balancing Methods on the Vertebral Column Dataset," vol. 8, no. 3, 2021, [Online]. Available: https://www.jetir.org/papers/JETIR2103328.pdf
[47] S. B. Imanulloh, A. R. Muslikh, and D. R. I. M. Setiadi, "Plant Diseases Classification based Leaves Image using Convolutional Neural Network," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 1–10, Aug. 2023, doi: 10.33633/jcta.v1i1.8877.