Lung Cancer Risk Analysis Using Machine Learning: Logistic Regression

Catherine Olivia[1], Zevanaya Beverly Drew[2], Ronan Lie[3], Axl Ray Aditya[4]
[1, 2, 3]Department of Information System, Faculty of Engineering and Informatics, Multimedia Nusantara University, Banten, Indonesia 15810
[4]Department of Informatics, Faculty of Engineering and Informatics, Multimedia Nusantara University, Banten, Indonesia 15810
Email:[1] catherine.olivia1@student.umn.ac.id, [2]zevanaya.beverly@student.umn.ac.id, [3]ronan.lie@student.umn.ac.id, [4]axl.aditya@student.umn.ac.id

Abstract – ***This research uses machine learning models to predict the risk level of lung cancer, which has long been ranked third in the category of the deadliest cancer in the world. The dataset used consists of 1000 rows and 26 columns and is obtained through the open source website Kaggle. The research method is carried out through a preprocessing stage which includes feature pruning and data visualization. Then training and testing data using the Logistic Regression model. Validation is done using K-Fold Cross Validation where the K is 10. From the results of the research that has been done, the Logistic Regression model gives an accuracy result of 1.00 using 21 features in the dataset without any signs of underfitting or overfitting. Then tested again using a new dataset and the accuracy reached 0.97. Based on these results, it can be concluded that the model used can accurately analyze the risk of lung cancer.***

Keywords: Artificial Intelligence; Disease; Logistic Regression; Lung Cancer; Machine Learning

I. Introduction

Lung cancer represents a significant global health threat. According to data from the Global Burden of Cancer (GLOBOCAN) obtained from the International Agency Research on Cancer, lung cancer ranks second as one of the leading causes of death worldwide, accounting for 11.4% of cases and an estimated 1.8 million deaths (18%) [1]. In Indonesia, lung cancer ranks third with a prevalence of 8.8% and is predominantly found in men [1]. Previous research indicates that 20-50% of cancer cases are first detected when individuals visit the hospital due to emergency situations. However, due to limited data, individuals diagnosed with cancer after visiting the hospital in an emergency tend to have poorer treatment outcomes [2].

Previous studies have shown that patients with lung cancer can have a survival rate of 85.11% by undergoing thoracic surgery [3]. Although thoracic surgery has a positive impact on the treatment of lung cancer, this procedure carries risks and complications that can be fatal. Several factors contribute to the development of lung cancer, such as cigarette smoke (both active and passive smoking), air pollution, chest pain, genetic susceptibility, and chronic lung diseases [4]. Therefore, it is essential to further understand the factors that can cause lung cancer to effectively carry out preventive measures.

There are several methods for detecting lung cancer, one of which is chest X-ray. Chest X-rays are the primary method used to detect lung cancer. Through this process, lung cancer will be detected in the imaging as a grayish-white mass. However, this examination is not always accurate in detecting lung cancer, as there are other conditions that may have similar imaging results, such as lung abscess. According to previous studies, the accuracy of detecting lung cancer from chest X-ray images is 72.97% [5]. This is considered less effective because chest X-rays are not sufficient for a definitive diagnosis, thus requiring other more practical methods for detecting lung cancer. In this case, the use of artificial intelligence (AI) can help process and analyze large amounts of data to diagnose diseases more informatively and promptly. Machine learning algorithms used in

AI can learn complex patterns in medical data, allowing for early disease detection with higher accuracy compared to conventional methods [6] [7].

The study "Evaluating Country Level Lung Cancer" by Heechan Lee, Heidi A., and others provides insights into various factors affecting lung cancer at the country level, including RadNet, radon, smoking, and PM2.5. Although this research uses comprehensive data from various sources and advanced analytical methods such as Poisson regression and Poisson Random Forest, there are still research gaps that need to be filled. One gap is the lack of exploration of the interaction between these factors and how these interactions collectively affect lung cancer incidence. Additionally, the study primarily focuses on data from the US, so the results may not fully apply to a global context. Further research is needed to test the validity of these findings in other countries with different environmental and demographic characteristics. Future research could also explore the use of other analytical methods that might be more effective in capturing the complexity of lung cancer risk factors [8].

The increasing number of lung cancer cases worldwide is the main driver for intensified research on the factors causing this disease. Early detection of lung cancer allows for earlier treatment, which automatically increases the chances of patient recovery. This research aims to predict the risk factors for lung cancer early on to enable more effective prevention and treatment using the Machine Learning model: Logistic Regression. Logistic Regression is an algorithm used in machine learning for classification. Logistic regression can model with regression analysis of variables that can predict outcomes with two or more possibilities, such as yes and no, in the context of categorical data [9].

Based on the above explanation, this study aims to investigate the correlation of lung cancer risk factors using a Logistic Regression model. A specific algorithm will be designed to enable the prediction of lung cancer risk for each individual. This study will use features with high correlation to outcomes as labels in the dataset, selected features close to the researchers' environment, and all features present in the dataset. The results of this study are expected to provide new information beneficial to the health sector, particularly in cases of lung cancer.

II. Research Method

In this research, four main processes were carried out. These processes include:

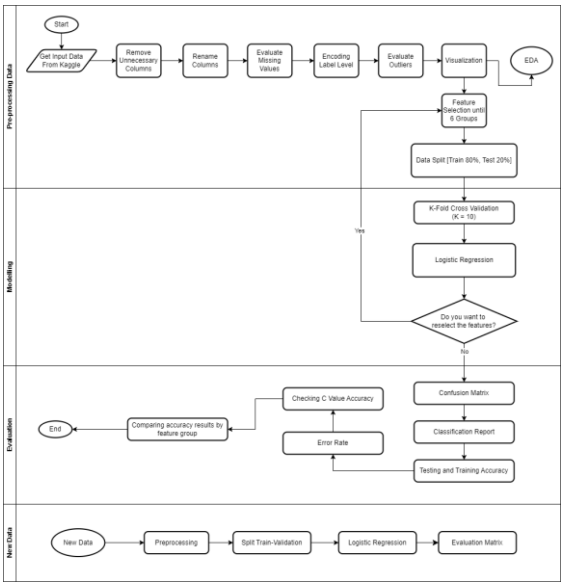

Fig. 1. Research Flowchart

Data Source

This research started with determining the dataset to be used. After brainstorming and discussion, the research team found a dataset relevant to their research title, found on March 6, 2023. The dataset used in this research comes from the open source website Kaggle, titled “Lung Cancer Prediction”. This dataset contains information about patients with lung cancer, including index, Patient Id, Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, OccuPational Hazards, Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Coughing of Blood and Level. This dataset consists of 1000 rows and 26 columns, with 25 columns of integer data type and 1 column of object data type. This dataset will be analyzed through several subsequent processes.

Data Pre-Processing

In this research, the dataset that has been obtained will be processed. Data processing

includes deleting unnecessary columns, renaming columns, identifying missing values, encoding LevelCode columns which will later become labels, identifying outliers, and performing visualization. The following is the pre-processing process performed.

1. Deleting irrelevant columns. In this case, the research team deleted 4 columns that were considered irrelevant to this study, namely: index, Patient Id, Age, and Gender.
2. Renaming columns. In this case, the researchers renamed some of the remaining columns to make the research easier.
3. Evaluating missing values. The researcher conducted an evaluation by using the isnull() function. After the evaluation, no missing values were found in this dataset.
4. Encoding the LevelCode feature. We changed the LevelCode feature from category to binary value to facilitate prediction in this research. After encoding, we delete the LevelCode features that have not been encoded.
5. Outlier evaluation. Researchers evaluated the presence of outliers by making visualizations using Boxplot. After visualizing, no outliers were detected in this dataset.
6. Creating visualizations. After going through the previous steps, researchers created several visualizations to deepen the understanding of this dataset.

Feature Selection

The study aims to predict the likelihood of lung cancer in patients based on the triggering factors. Modeling in the study uses the Logistic Regression algorithm to predict whether the patient is at the Low, Medium, or High risk level. The algorithm selection is based on binary features and labels, as well as previous research journals. The purpose of this study is to find which features have the highest accuracy and accuracy in triggering the risk of lung cancer. Accuracy discovery is done by comparing groups of features that have been made.

Modeling

After the data has gone through the preprocessing process, the divided data will be used in the research to perform modeling using the Logistic Regression algorithm. But before that, the model will be validated using K-Fold Cross Validation. Logistic regression is a simple yet effective data analysis technique in solving classification problems with two classes.

$$\square(\square) = \frac{1}{1 + \square^{-\square}}$$

The features that have been selected in the previous stage will be used in this modeling to find the highest accuracy of features against labels. Furthermore, the data is divided into each group of features as variable X and labels as variable y will be separated using train test split. The data will be divided into two parts, namely training data and testing data. The division percentage used is a percentage commonly used by many other researchers, namely 80% of the data for training data and the other 20% for testing data.

Evaluation

The Logistic Regression model that has been proposed in the previous stage will be evaluated using Classification Report, Testing and Validation Accuracy, K-Fold Cross Validation, Error Rate, and Checking C Value Accuracy. The evaluation conducted will be very useful in seeing the quality of the model. Furthermore, the quality of the model will be tested using the following types of model evaluation and validation. This is done so that the model can be ensured to run well at the next stage, namely during the deployment stage.

1. Confusion Matrix & Classification Report: Confusion Matrix is an evaluation table that shows the performance of the classification model by distinguishing classes. While the Classification Report displays the precision matrix, recall, and F1 score for each class in the classification [11] [12].
2. Testing & Training Accuracy: An evaluation matrix to measure model performance, measured by dividing the data into training and testing sets. Training Accuracy refers to the accuracy of the model when tested with training data,

while Testing Accuracy refers to the accuracy when tested with separate testing data [13].

3. K-Fold Cross Validation: A model evaluation approach that divides the data into k subsets and trains the model with k-1 subsets while the remaining ones are tested. Model performance is measured by taking the average of the test results from all iterations [14].
4. Error Rate: A measure that indicates how often errors occur in digital circuits. Error Rate predicts the error rate in the output bits of a customized circuit [15].
5. Checking C Value Accuracy: An evaluation of the C parameter in SVM that controls the trade-off between margin and misclassification. A high C value tends to result in tighter decision boundaries but can lead to overfitting if the data is not linearly distributed [16].

New Data

At this stage, we run the same process on a new dataset to evaluate whether it has similar accuracy or experiences over/underfitting. In this case, the research team used a similar dataset, the "Lung Cancer Survey". We tested by building a model on this dataset to assess its accuracy.

III. Results and Discussion

Preprocessing Data, data that has been found must be preprocessed first before proceeding to the next stage. Data preprocessing starts with deleting columns, renaming columns, evaluating missing values, encoding labels, evaluating outliers, and finally, visualizing the data. The Lung Cancer Patient dataset has 26 columns and 1000 rows. The dataset consists of 24 columns with integer data types, including index, Age, Gender, Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, and Coughing of Blood, and 2 columns with object data types, namely Patient Id and Level. The first preprocessing step performed by the researcher is deleting columns that are not needed and relevant to the study. The researcher deleted the columns index, Patient Id, Age, and Gender, leaving 22 columns, with 21 integer data type columns and 1 object data type column. The second preprocessing step involved renaming several column names to facilitate their use in the subsequent processes. The third preprocessing step is evaluating missing values, where the researcher assessed the null values present in the dataset. The researcher found that the Lung Cancer Patient dataset did not have any missing values, allowing the study to proceed with the dataset as is. Next, the fourth preprocessing step performed by the researcher was encoding the remaining categorical column, which is the Level column. The researcher encoded this column to facilitate modeling using the Logistic Regression algorithm in the next stage. The encoding process involved converting the Level values where Level High becomes binary 0, Level Low becomes binary 1, and Level Medium becomes binary 2. The results of the Level encoding are stored in a new column called LevelCode. After that, the fifth preprocessing step is the evaluation of outliers. The researcher has evaluated outliers through the visualization shown in Figure 2. Based on this figure, the Cancer Patient dataset does not have any outlier values, allowing the study to proceed without addressing outlier values.

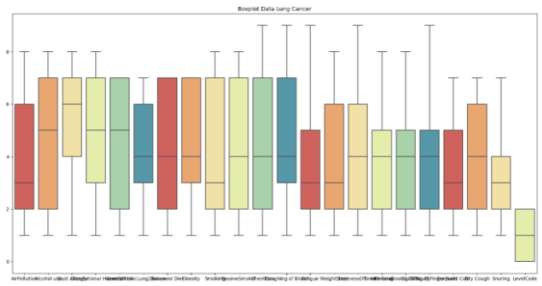

Fig. 2. Boxplot Data Lung Cancer Patient

Next, the final preprocessing step performed by the researcher is data visualization. Visualization is conducted to strengthen the understanding of the dataset. Here are some data visualizations that have been performed by the researcher:

The first visualization in Figure 3 shows a bar plot distribution of the frequency count for each LevelCode, along with their respective percentages. The LevelCode column has been previously encoded as follows: 0 = High, 1 = Low, 2 = Medium. Based on this visualization,

the researcher can determine that the most frequent LevelCode in the dataset is at the High level, with a percentage of 36.5%.

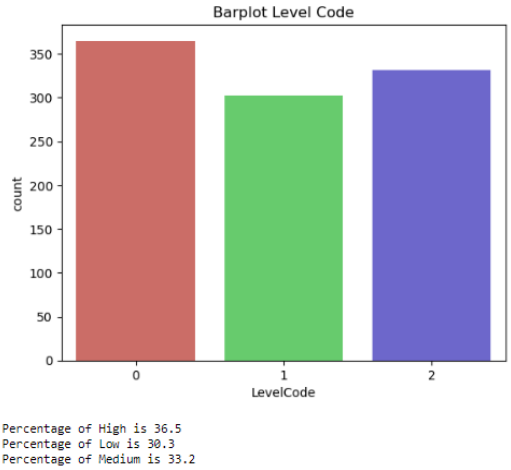

Fig. 3. Barplot Level Code

The second visualization in Figure 4 shows the distribution of the Lung Cancer Patient data. Based on the results, all features have more than one peak, indicating that the data follows a non-normal distribution.

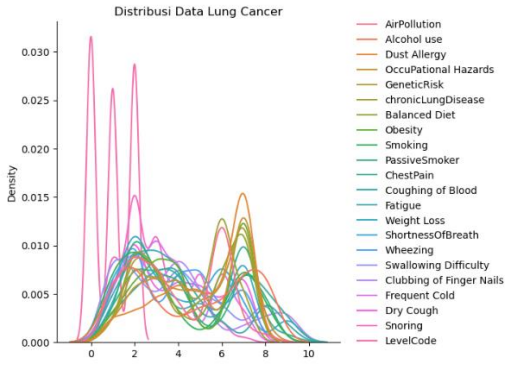

Fig. 4. Distribusi Data Lung Cancer Patient

The next visualization in Figure 5 shows the distribution of lung cancer levels based on the Air Pollution index. Based on the results, the researcher can determine that many patients with a high-risk level come from an index level of 6.

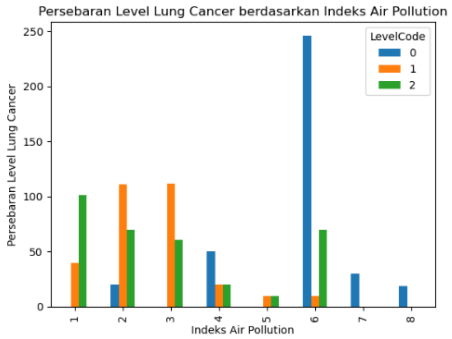

Fig. 5. Crosstab Level Lung Cancer Patient

The next visualization shows a histogram of the frequency of the Passive Smoker attribute in the dataset. The researcher can determine that many patients exposed to passive smoking are at level 2.

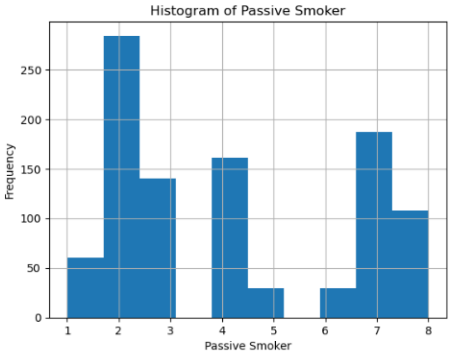

Fig. 6. Histogram Passive Smoker

Feature Selection, in this stage, the researcher will select features to be used in the next modeling phase. In line with the research objective of finding the features that have the highest accuracy, the researcher creates feature groups containing features based on their similarities. The researcher forms 6 groups of features to achieve the research goal. Each feature group will become variable X, with the LevelCode feature as variable y.

The first group is a group of features that have a high correlation with the variable y, LevelCode. These features are selected based on the results of heatmap in Figure 7, which are in the correlation value range of 1 and -1, and are relevant to LevelCode. This group has 11 features as variable X, namely Air Pollution, Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease,

Balanced Diet, Obesity, Smoking, Passive Smoker, and Coughing of Blood.

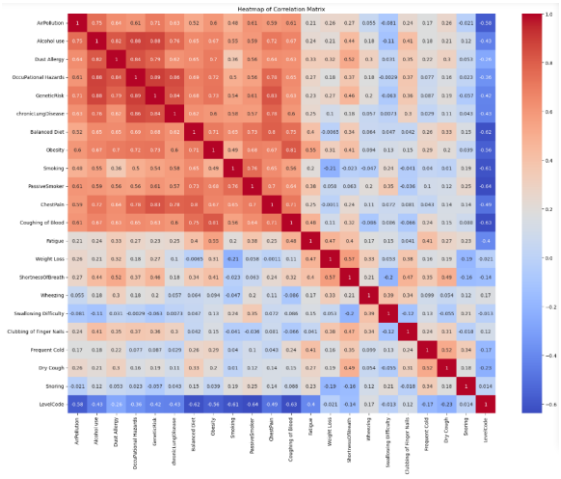

Fig. 7. Heatmap Dataset Lung Cancer Patient

The second group is a set of features related to the surrounding environment. The researcher selected 6 relevant features based on the environmental conditions in Indonesia and their proximity to the researcher. These features are Air Pollution, Smoking, Passive Smoker, Chronic Lung Disease, ChestPain, and ShortnessOfBreath. Furthermore, the researcher divided these 6 features into 3 other groups, which will be the order of the next groups. The third group consists of smoking-related features, namely Smoking and Passive Smoker. The fourth group consists of air pollution-related features, with Air Pollution. Lastly, the fifth group comprises diseases closely related to lung cancer, including Chronic Lung Disease, ChestPain, and ShortnessOfBreath. The researcher subdivided this large group to further evaluate their accuracies.

The last group is Group Six, consisting of all features in the dataset. The features that will serve as variable X are Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, and Snoring. In the next stage, each group will be called according to the names outlined in this stage, namely Group One, Two, Three, Four, Five, and Six.

Modeling, the researcher divides the data in each feature group using a train-test split. The data split ratio will be divided into two parts, with 80% of the data used for training and 20% for testing. After that, the researcher will perform validation using K-Fold Cross Validation (K=10), which means the data will be divided into 10 parts, where 9 parts are used for training and the remaining 1 part is used for validation. Then, the researcher will model the data using one type of regression algorithm, namely Logistic Regression.

Evaluation, each feature group represents a solution to each research problem. The first feature group consists of features highly correlated with LevelCode. The model using the first group achieved high testing accuracy of 91% and training accuracy of 87%. This aligns with the low Error Rate value produced, which is 0.9%. To validate the high accuracy results and check for overfitting, the researcher employed K-Fold Cross Validation. This model obtained non-overfitting results as the min, average, and max cross-validation scores were close, at 0.83, 0.873, and 0.9, respectively. In contrast to the first group, the second group used features closely related to the researcher's life. The model that used the features from the second group as X also achieved high accuracy results, with 89% for both training and testing. Along with these results, the error rate of this model is also quite low, at 11%. Model validation using K-Fold provided information that the model is not overfitting, as evidenced by the close range of the min, average, and max cross-validation scores, which are 0.8, 0.856, and 0.9, respectively. Next, the model using the features from the third group, which consists of smoking-related features, achieved a relatively high accuracy of 76% for testing and 69% for training, with a sizable error rate of 24%. Similarly, the model using features from the fourth group, comprising AirPollution features, yielded accuracy rates of 53% for both training and testing, with an error rate of 48%. The significant error rates indicate that the model using features from the fourth group has poor accuracy. The last small group of environmental features is the fifth group,

consisting of features related to lung diseases. The model using these features achieved testing and training accuracies of 77% and 79%, respectively, with a relatively large error rate of 22%. Lastly, the sixth group encompasses all features in the dataset as variable X. The model using these features achieved perfect accuracy of 100% with an error rate of 0%. This perfect accuracy value also applies to the min, average, and max cross-validation scores. Based on this explanation, Table I is formed, showing the comparison of accuracy and validation scores for each feature group.

TABLE I
Results Comparison Accuracy and Error Rate

Feature Group	Testing Accuracy	Training Accuracy	Error Rate
1.	91%	87%	0,9%
2.	89%	89%	11%
3.	76%	69%	24%
4.	53%	53%	48%
5.	77%	79%	22%
6.	100%	100%	0%

Based on Table I, each feature group has different accuracy values. The highest accuracy is obtained by the sixth feature group at 100%. This sixth group consists of all features in the dataset, namely Alcohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, and Snoring. The 100% accuracy result indicates that all features in the dataset significantly contribute to the model. Conversely, the fourth group has the lowest accuracy at 53%. This fourth group consists of only one feature, AirPollution, suggesting that the AirPollution feature contributes less significantly to the model.

In addition to Table I, the comparison between the minimum, average, and maximum values of K-Fold Cross Validation for each of the six groups is also presented in Table II below.

TABLE II
Results Comparison K-Fold Cross Validation

Feature Group	MIN Score	MAX Score	AVG Score
1.	81%	93%	87%
2.	79%	91%	86%
3.	62%	81%	72%
4.	43%	62%	52%
5.	67%	87%	79%
6.	100%	100%	100%

Based on the K-Fold Cross Validation results in Table II, all feature groups do not show significant differences in their outcomes, indicating that each feature group did not experience underfitting or overfitting.

Additionally, it can be concluded that the feature group which has the most significant influence on the model and does not experience overfitting or underfitting is the sixth feature group. With accuracy and K-Fold Cross Validation values of 100%, it indicates that the sixth feature group contributes significantly to the model and is reliable in making predictions.

New Data, the researcher has conducted testing with new data using the same model. Based on Figure 8 below, the researcher can observe that the Checking C Value results show a difference of 10% between training data accuracy and validation accuracy. With training accuracy reaching 96% and validation accuracy at 86%, it indicates that the model can operate stably and optimally but may potentially suffer from overfitting. Therefore, the researcher takes additional steps to validate this statement.

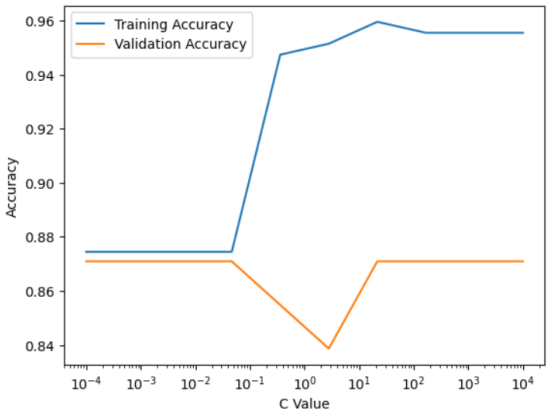

Fig. 8. Checking C Value New Dataset

Next, the researcher performs K-Fold Cross Validation using the same data to ensure whether the model suffers from overfitting or not. The researcher obtains cross-validation scores divided into three parts: the minimum score is 87%, the maximum score is 96%, and the average score is 92%. Based on these scores, which do not exhibit significant differences, the researcher can conclude that the model does not suffer from overfitting.

IV. Conclusion

Detection of lung cancer risk, typically conducted through physical examination using X-rays, is a time-consuming process. Moreover, it is costly, making it inaccessible for many to undergo early cancer detection. However, with artificial intelligence and machine learning, this process can be facilitated. This study utilizes a Logistic Regression machine learning model to analyze the risk level of lung cancer based on data. The objective of this research is to compare models with features that can most accurately predict lung cancer based on its causal factors. Therefore, all features are divided into 6 feature groups by the researcher. Based on the research findings, the sixth group, which incorporates all 21 features, yields the highest accuracy among others. The accuracy of the model using the Logistic Regression algorithm reaches 100% without overfitting. This model can be applied to predict lung cancer in patients based on the 21 causal factors. However, the researchers acknowledge the limitations of this study. Achieving a 100% accuracy rate in the model is often too optimistic. Therefore, the researchers recommend conducting further evaluation checks on the model to ensure its validation. Additionally, the researchers are aware that this study was conducted within time constraints, hence they hope for future researchers to delve deeper into lung cancer topics based on its causal factors.

References

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

[2] "Direktorat Jenderal Pelayanan Kesehatan." https://yankes.kemkes.go.id/view_artikel/1550/bagaimana-kanker-paru-dapat-diketahui-lebih-awal-sebelum-stadium-lanjut

[3] N. Pettit, A. Al-Hader, and C. A. Thompson, "Emergency department associated lung cancer diagnosis: Case series demonstrating poor outcomes and opportunities to improve cancer care," *Current Problems in Cancer. Case Reports*, vol. 3, p. 100059, Mar. 2021, doi: 10.1016/j.cpccr.2021.100059.

[4] R. T. Prasetio and S. Susanti, "Prediksi harapan hidup pasien kanker paru pasca operasi bedah Toraks menggunakan Boosted K-Nearest Neighbor," *ejurnal.ars.ac.id*, Aug. 2019, doi: 10.51977/jti.v1i1.66.

[5] H. Chen, "A review of the recent research progress on risk factors of lung cancer," *Theoretical and Natural Science*, vol. 21, no. 1, pp. 291–295, Dec. 2023, doi: 10.54254/2753-8818/21/20230905.

[6] L. Listyalina, E. L. Utari, and D. E. Puspaningtyas, "PENENTUAN PENYAKIT PARU DENGAN MENGGUNAKAN JARINGAN SARAF TIRUAN," *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer/Simetris*, vol. 11, no. 1, pp. 233–240, Apr. 2020, doi: 10.24176/simet.v11i1.3667.

[7] Baraka, "Pengaruh AI dalam Industri Kesehatan : Inovasi untuk Kesejahteraan Manusia," *Biro Perencanaan Sumber*

Daya Manusia Dan Karir, Feb. 29, 2024. https://baraka.uma.ac.id/pengaruh-ai-dalam-industri-kesehatan-inovasi-untuk-kesejahteraan-manusia

[8] "Artificial Intelligence Techniques and Methodology Available for Lung Cancer Detection," *IEEE Conference Publication | IEEE Xplore*, Jun. 14, 2023. https://doi.org/10.1109/ICSCSS57650.2023.10169510

[9] Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., & Putra, A. D. P. (2022). Diabetes Risk Prediction using Logistic Regression Algorithm. journal.literasisains.id. https://journal.literasisains.id/index.php/jomlai/article/view/598 Sci. Educ., Chattanooga, TN, USA, 2009, Rev. 47, pp. 777–780.

[10] N. A. C. Putri and D. B. Arianto, "Komparasi Penggunaan Information Gain Pada Machine Learning untuk Memprediksi Harga Rumah di Jabodetabek," *Jurnal Sains Dan Teknologi*, vol. 5, no. 3, pp. 756–762, Feb. 2024, doi: 10.55338/saintek.v5i3.2052.

[11] M. F. Amin, "Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial," Journal of Engineering Research - Egypt/Journal of Engineering Research, vol. 6, no. 5, p. 0, Dec. 2022, doi: 10.21608/erjeng.2022.274526.

[12] Akram, Sharjeel. (2021). CLASSIFICATION REPORT.

[13] I. Hammad and K. El-Sankary, "Practical Considerations for accuracy Evaluation in Sensor-Based Machine Learning and Deep Learning," Sensors, vol. 19, no. 16, p. 3491, Aug. 2019, doi: 10.3390/s19163491.

[14] J. M. Górriz, F. Segovia, J. L. Ramírez, A. Ortíz, and J. Suckling, "Is K-fold cross validation the best model selection method for Machine Learning?," arXiv (Cornell University), Jan. 2024, doi: 10.48550/arxiv.2401.16407.

[15] D. Ma and X. Jiao, "A Machine Learning-Based Error Model of Voltage-Scaled Circuits," A Machine Learning-Based Error Model of Voltage-Scaled Circuits Dongning Ma, Xun Jiao, Jun. 2020, doi: 10.1109/dsn-s50200.2020.00046.

[16] F. Budiman, "SVM-RBF Parameters testing optimization using cross validation and grid search to improve multiclass classification," Naučnaâ Vizualizaciâ, vol. 11, no. 1, Jan. 2019, doi: 10.26583/sv.11.1.07.

[17] N. A. C. Putri and D. B. Arianto, "Komparasi Penggunaan Information Gain Pada Machine Learning untuk Memprediksi Harga Rumah di Jabodetabek," *Jurnal Sains Dan Teknologi*, vol. 5, no. 3, pp. 756–762, Feb. 2024, doi: 10.55338/saintek.v5i3.2052.