Application of Cross-Validation Techniques to Handle Overfitting in a Case Study of Decision Tree Implementation for Lung Cancer Prediction

Faurika [1], Ahsanun Naseh Khudori [2], M. Syauqi Haris [3]
[1,2,3] Informatika, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW

INFORMASI ARTIKEL

Received 11 Juli 2024
Revised 16 Juli 2024
Published 19 Juli 2024

Kata kunci:

Lung cancer;
Machine learning;
Decision tree;
Rules;
Cross-validation

ABSTRACT

Lung cancer is a condition caused by cancer cells growing in the lungs. Lung cancer causes a weakened immune system, tumors, and other abnormalities that prevent the body from functioning properly. Lung cancer examination uses various technologies, namely CT Scan, X-ray, and others. However, the examination is relatively expensive and takes a long time. The use of machine learning makes it possible to support lung cancer diagnosis. With the large amount of medical data available today, machine learning can recognize patterns in the data so that it will help the process of diagnosing lung cancer more effectively. This study aims to correct overfitting in previous research which used the decision tree method to predict lung cancer with cross-validation techniques. In this research, we use a public dataset from Data World. This dataset consists of 25 data attributes and has 1000 data. The results of this research are rules obtained from decision trees which are then evaluated to produce 96.7% accuracy, 96.7% precision, 96.7% recall, and 96.7% f1-score. These results show that the decision tree method performs well in predicting lung cancer early and the cross-validation technique can overcome overfitting in decision trees with more general and stable results.

Corresponding Author:

M. Syauqi Haris, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW
Email: haris@itsk-soepraoen.ac.id

1. INTRODUCTION

Lung cancer is a condition caused by cancer cells growing in the lungs. Lung cancer causes a weakened immune system, tumors, and other abnormalities that prevent the body from functioning properly [1]. Based on cell origin, lung cancer is divided into two subtypes, namely the first non-small-cell lung cancer (NSCLC) which also has several subtypes and the most common are adenocarcinoma and squamous carcinoma [2]. second, small-cell lung cancer (SCLC), which is a neuroendocrine carcinoma that is often suffered by active smokers [3]. The prevalence of the NSCLC type is greater than SCLC and is classified as a type of lung cancer with a high level of malignancy [4]. Based on cancer statistics data, as many as 236,740 cases of lung cancer and as many as 130,180 will die from lung cancer in 2022 [5]. Congenital factors that are at risk of lung cancer include age where the average lung cancer sufferer is 70 years between men and women, gender where men are at greater risk of suffering from lung cancer compared to women, and heredity [6].

America has the second-highest number of deaths from lung cancer after China [7]. Not only in America but throughout the world, cancer continues to increase even though medical equipment is becoming more sophisticated and the number of medical personnel is increasing, the diagnosis of this disease is carried out at a late stage[8]. After prostate cancer in men and breast cancer in women, lung cancer is common not only in men but also in women [8]. In America, deaths caused by lung cancer are more numerous than deaths caused by prostate and breast cancer, with 236,740 cases of lung cancer and 130,180 cases of death [9]. In Indonesia, lung cancer cases reached 34.78 in 2020 [10].

Computer-based diagnosis systems have a very important role in faster detection and diagnosis of body disorders [11]. Lung cancer detection can use technology that is currently increasingly sophisticated such as X-ray, CT scan, MRI, and so on [12]. However, it is expensive, and medical personnel can detect the disease at an advanced stage [13]. Thus, early detection of lung cancer needs to be carried out in order to minimize the number of deaths and increase the life expectancy of patients with lung cancer. Machine learning (ML) can be applied to assist in early detection and prediction of disease [13]. ML helps analyze and process data or information to find patterns or problems that cause disease. Applying ML in the medical world helps doctors diagnose diseases better and faster thereby saving money and time [13]. However, ML algorithms need to be improved to assist medical personnel in making effective clinical decisions with good accuracy [13]. Several ML methods that are widely used to make predictions are linear regression, logistic regression, decision tree (DT), naive bayes, random forest, neural network, and so on [14]. Researchers chose to apply the decision tree method because this method is easy to implement, simple, and has good performance for prediction [15], [16].

Several researchers have applied the decision tree method to predict lung cancer, such as Saeed's (2019) research which predicted lung cancer using a dataset of lung cancer patients in hospitals in Karachi, Pakistan, and produced an accuracy of 60%. Kumar Moan and Bhraguram Thayyil (2023) also conducted research using a database of patients with lung cancer and produced an accuracy of 88%. Another researcher, namely Gupta (2023), predicted lung cancer using a dataset obtained from the cbioportal and produced an accuracy of 85.7%. Another research was conducted by Haris (2024) who used a dataset from world data and produced an accuracy of 100%. This research used a random sampling technique. In this technique, data is selected randomly to be used as training data and validation data [21]. The model created also uses unlimited parameters, including the minimum number of branches is not specified, and uses the minimum number of branches by default in the orange tools used. The depth of the decision tree built in this research is also not limited, which causes the decision tree to capture noise. The use of random sampling techniques to evaluate the performance of machine learning models does not use all the data to be used as training data and validation data because it only divides the training data by percentage and validation data by percentage [22] so that the results obtained are not general and unstable.

From the research above, the implementation of decision trees for lung cancer prediction produces good accuracy, but there are still gaps for improvement. Based on the results of this research, this research intends to modify decision tree research for lung cancer prediction with a larger and more diverse dataset and use cross-validation evaluation techniques to overcome overfitting in the ID3 decision tree model.

The statistical technique that is widely used to evaluate the performance of machine learning models is cross-validation (CV) that divide data into training sets called folds [23]. The use of CV in evaluating model performance can highlight problems in the data such as selection bias or overfitting [24]. CV selection can be done based on the size of the dataset used. The use of folds in CV is usually used to shorten computing time and stabilize the performance accuracy of the model [25].

This research aims to overcome overfitting in research that has applied decision trees for lung cancer prediction by creating a model and analyzing the performance produced by the ID3 decision tree model. Thus, the results obtained from this research can be used by medical personnel to support early detection of lung cancer so that diagnosis becomes more effective and does not take much time.

2. LITERATURE REVIEW

Pada bagian ini, Anda harus menjelaskan bagaimana penelitian dilakukan, termasuk desain penelitian, prosedur penelitian (dalam bentuk algoritma, Pseudocode, atau lainnya), cara memperoleh data, dan cara melakukan pengujian apa pun. Deskripsi program penelitian harus didukung dengan referensi, sehingga penjelasannya dapat diterima secara ilmiah.

2.1. Lung cancer

Lung cancer is a condition caused by cancer cells growing in the lungs. Lung cancer causes a weakened immune system, tumors, and other abnormalities that prevent the body from functioning properly [6]. Based on cell origin, lung cancer is divided into two subtypes, namely the first non-small-cell lung cancer (NSCLC) which also has several subtypes and the most common are adenocarcinoma and squamous carcinoma [2]. second, small-cell lung cancer (SCLC), which is a neuroendocrine carcinoma that is often suffered by active smokers [3]. Congenital factors that are at risk of lung cancer include age where the average lung cancer sufferer is 70 years between men and women, gender where men are at greater risk of suffering from lung cancer compared to women, and heredity [6].

2.2. Machine learning

Machine learning is a branch of artificial intelligence that focuses on developing algorithms that allow computers to learn from data and make predictions without being explicitly programmed. [26][27]. Techniques in machine learning are widely applied in many fields, namely computer vision, pattern recognition, spacecraft engineering, finance, biological computing, and health [28]. ML algorithms use input data to achieve tasks without being programmed just by learning from experience so that they are better at producing a particular output. This process is referred to as training where sample data is entered along with the desired results [29]. In ML, it is divided into two learning models that are commonly used, namely the first is supervised learning which is applied for classification, and generally the learning process uses labeled datasets [30]. secondly, unsupervised learning is applied for grouping and generally, the learning process uses unlabeled datasets [29]. There are many algorithms in both models. Several algorithms in ML that fall into the supervised learning category include linear regression, logistic regression, decision tree (DT), naive bayes, random forest, neural network, and so on. In this research, the ID3 decision tree method is used to predict lung cancer.

2.3. Decision tree ID3

Decision trees are a method in machine learning for carrying out various classification, regression, and prediction tasks [32]. Decision trees are popular because they are easy to understand and not too complex. In general, decision trees are used as prediction models to predict target variable values[33]. Rules or rules models usually use if-then rules[34]. Entropy and information gain are used in ID3 classification. Information gain to determine the features that will be used to build a decision tree[33]. Making a decision tree starts from the root node to predict the class label of the decision tree. From the root node, the attribute information gain is then compared until the largest information gain is obtained to be used as an internal node or branch until a leaf node is obtained[35]. The stages in ID3 are as follows. The stages in ID3 are as follows.

1) Calculate the entropy and information gain values
2) Calculation of the entropy value using (1).

$$Entropy\ (S) = -p_a\ log_2\ p_a - \ p_b\ log_2\ p_b \tag{1}$$

Where S represents the universe, p_a represents the number of first-class attributes, and p_b represents the number of second-class attributes.

After calculating the entropy value, then calculate the information gain attribute value using (2).

$$Gain\ (S, A) = Entropy(S) - \sum_{v\ \epsilon\ nilai\ (A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

Where S represents the universe, A represents the attributt, Entropy (S) represents the entropy of the entire class, S_v represents the number of attributes of each class, S the proportion of the universe, and Entropy (S) represents the entropy of the entire class.

3) Specifies the root

The root is determined based on the information gain attribute from the results of previous calculations. The largest information gain will be the root node.

4) Forms internal node

Internal nodes are determined from branching root nodes. When the root node's branch entropy value is 0, there are no further branches. If the entropy is not 0, then the information gain is recalculated to form the internal node.

5) Forming leaf nodes

The final node represents the target class of classification using a decision tree.

2.4. Cross-validation

The statistical technique that is widely used to evaluate the performance of machine learning models is cross-validation (CV) that divide data into training sets called folds [36][37]. The dataset will be divided by the specified k-fold and iterated by k-fold as well. One subset of folds is used as validation data and for each iteration, training data is taken from the remaining folds [38]. Cross-validation techniques can be used to optimize hyperparameters of statistical and machine learning models, prevent overfitting of models, and estimate model generalization errors [39].

3. METHODS

Fig. 1 shows the methodology used in this research.

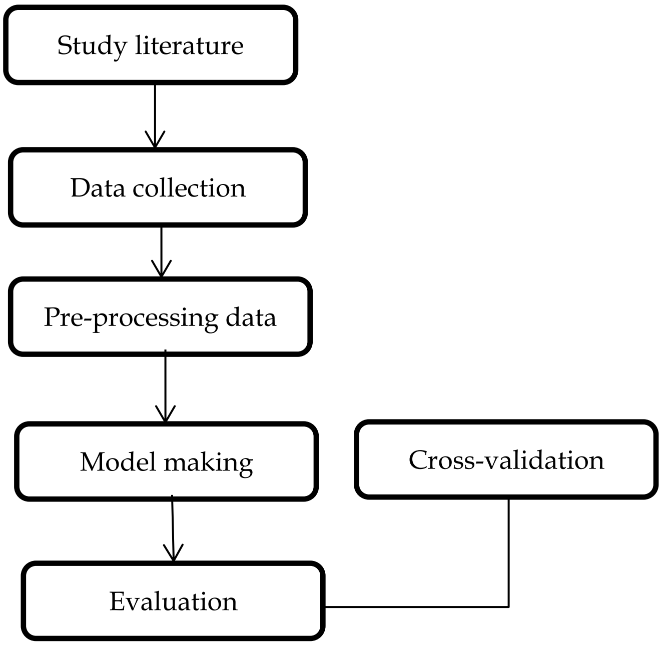


Fig. 1. Methodology used in this research

A more detailed description of the research methodology is shown in Fig. 1 as follows.

1) Study literature, At this stage, conducting a literature review of previous research relevant to this research topic.
2) Data collection, The dataset used in this research uses a dataset that is publicly available from Data World.
3) Pre-processing data, checked and ensured that the dataset complied with the data quality criteria specified in this research. The data quality criteria set in this research are complete data, using appropriate data types, and no duplication of data.
4) Model making, dataset that meets the data quality criteria in this research then a decision tree model with various parameters. The parameters used are first, induce binary tree, this parameter is used to build a binary model on a decision tree. Second, namely the min number of instances in leaves, this parameter is used to determine the minimum number of branches and is determined as 15. Third, namely the parameter limits the maximal tree depth, this parameter is to determine the maximum depth of the decision tree and is determined as deep as 20. Fourth namely stop when the majority reaches (%) to determine the learning threshold of 100%.
5) Evaluation, At this stage, the performance evaluation of the decision tree model which has been created using a cross-validation technique. This study used 10-fold cross-validation. The results of this evaluation are seen from several evaluation metrics, including:
   a. Accuration, is the ratio of the total correctly predicted data to the total of all data. Calculated using (3):

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

   b. Precision, is the ratio of the proportion of the number of correct positive predictions to all positive prediction data. Calculated using (4):

$$APrecision = \frac{TP}{TP + FP} \tag{4}$$

   c. Recall, is the ratio of the proportion of the number of correct positive predictions to the amount of actually positive data. Calculated using (5):

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

   d. F1-score, is a average of precision and recall. Calculated using (6):

$$F1 - score = \frac{presisi * recall}{presisi + recall} \tag{6}$$

Where TP means the facts and the predicted results are positive, TN means the facts and the predicted results are negative, FP means the predicted data is correct but the facts are negative, and finally FN means the predicted data is negative but the facts are positive.

4. RESULTS AND DISCUSSION

Data from Data World (https://data.world/cancerdatahp/lung-cancer-data) consists of 1000 data records and has 25 attributes, each attribute contains a range of risk severity levels on a scale of 1-9. The attributes in the dataset are described in Table 1.

Table 1. Attribute dataset

Attribute	Value
Patient id	Patient id
Age	14-73 years old
Gender	1, 2
Air pollution	1-8
Alcohol use	1-8
Dust allergy	1-8
Occupational hazards	1-8
Genetic risk	1-7
Chronic lung disease	1-7
Balanced diet	1-7
Obesity	1-7
Smoking	1-8
Passive smokers	1-8
Chest pain	1-9
Coughing of blood	1-9
Fatigue	1-9
Weight loss	1-8
Shortness of breatch	1-9
Wheezing	1-8
Swallowing difficulty	1-8
Clubbing of finger nail	1-9
Frequent cold	1-7
Dry coughing	1-7
Snoring	1-7
Level	Low, medium, high

At the pre-processing stage, the dataset met the data quality criteria in this research. However, there is 1 feature that will not be used in this research, namely the patient_id feature. This is because the patient_id feature has no effect on the learning process. The final result of this pre-processing stage is 24 features and 1000 data which will be used for the learning process.

4.1. Model making

The implementation of decision trees in this research produces rules obtained from the decision tree. The model is built using several parameters, namely:

a. Induce binary tree, to build a binary model on a decision tree.

b. Min number of intances in leaves, to determine the minimum number of branches of 15.

c. Limit the maximal tree depth, to determine the maximum depth of the decision tree as 20.

d. Stop when majority reaches (%), to determine the learning threshold of 100%.

From the model built using these parameters, the model then produces a visual binary decision tree shown in Fig. 2.

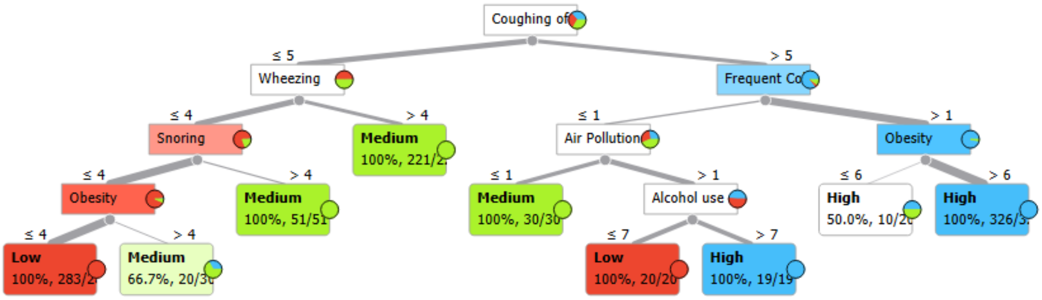


Fig. 2. Visual decision tree

From Fig. 2, rules were then created as shown in Table 2.

Table 2. Rules

No	Rules
1	If Coughing of blood > 5 ^ Frequent cold > 1 ^ Obesity > 6, Then High
2	If Coughing of blood > 5 ^ Frequent cold > 1 ^ Obesity <= 6, Then High
3	If Coughing of blood > 5 ^ Frequent cold <= 1 ^ Air pollution > 1 ^ Alcohol use > 7, Then High
4	If Coughing of blood > 5 ^ Frequent cold <= 1 ^ Air pollution > 1 ^ Alcohol use <= 7, Then Low
5	If Coughing of blood > 5 ^ frequent cold <= 1 ^ Air pollution <= 1, Then Medium
6	If Coughing of blood <= 5 ^ Wheezing > 4, Then Medium
7	If Coughing of blood <= 5 ^ Wheezing <= 4 ^ Snoring > 4, Then Medium
8	If Coughing of blood <= 5 ^ Wheezing <= 4 ^ Snoring <= 4 ^ Obesity > 4, Then Medium
9	If Coughing of blood <= 5 ^ Wheezing <= 4 ^ Snoring <= 4 ^ Obesity <= 4, Then Low

The rules in Table 2 are then explained in more detail as follows.

Rules 1 : If the severity of coughing blood is above 5 and the severity of fever (frequent cold) is above 1 and the severity of obesity is above 6, then a person is categorized as having a high risk of lung cancer.

Rules 2 : If the severity of coughing blood is above 5 and the severity of fever (frequent cold) is above 1 and the severity of obesity is less than or equal to 6, then a person is categorized as having a high risk of lung cancer.

Rules 3 : If the severity of coughing blood is above 5 and the severity of fever (frequent cold) is less or equal to 1 and the severity of environmental air pollution is above 1 and the level of alcohol addiction is above 7, then a person is categorized as being at risk of lung cancer High.

Rules 4 : If the severity of coughing blood is above 5 and the severity of fever (frequent cold) is less or equal to 1 and the severity of environmental air pollution is above 1 and the level of alcohol addiction is below or equal to 7, then a person is categorized as at risk of cancer Low.

Rules 5 : If the severity of coughing blood is above 5 and the severity of fever (frequent cold) is less than or equal to 1 and the severity of environmental air pollution is below or equal to 1, then a person is categorized as being at medium risk of lung cancer.

Rules 6 : If the severity of the coughing blood is below or equal to 5 and the severity of the wheezing is above 4, then a person is categorized as being at medium risk of lung cancer.

Rules 7 : If the severity of coughing blood is below or equal to 5 and the severity of wheezing is below or equal to 4 and the severity of snoring is above 4, then a person is categorized as being at medium risk of lung cancer.

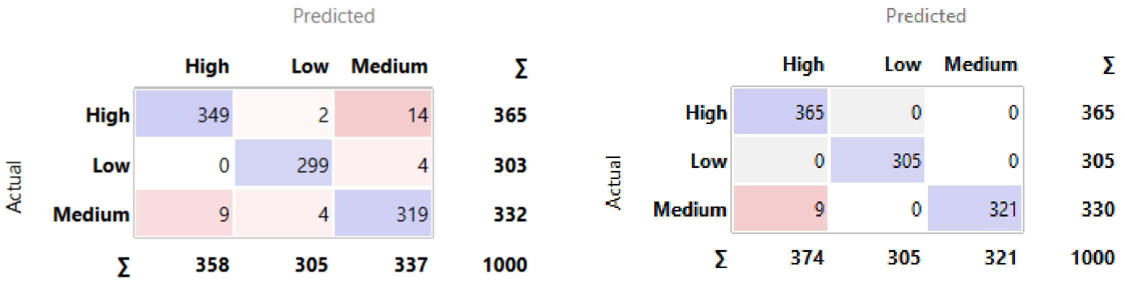
Rules 8 : If the severity of coughing blood is below or equal to 5 and the severity of wheezing is below or equal to 4 and the severity of snoring is below or equal to 4 and the severity of obesity is above 4, then a person is categorized as being at risk of lung cancer medium.

Rules 9 : If the severity of coughing blood is below or equal to 5 and the severity of wheezing is below or equal to 4 and the severity of snoring is below or equal to 4 and the severity of obesity is below or equal to 4, then a person is categorized as in a low risk of lung cancer.

From these rules, a person can know whether they are at risk of suffering from lung cancer or not. From the model that has been built, then an evaluation is carried out of the performance of the model that has been created.

4.2. Evaluation

The evaluation stage is carried out using orange tools. The cross-validation technique was used to overcome the overfitting that occurred in previous research. The number of folds used in this research was determined to be 10. After evaluating using the 10-fold cross-validation technique, accuracy, precision, recall, and f1-score were obtained with the proportion of predicted data and actual data presented in the confusion matrix table as in Fig. 3.



Actual \ Predicted	High	Low	Medium	Σ
High	349	2	14	365
Low	0	299	4	303
Medium	9	4	319	332
Σ	358	305	337	1000

(a). Using cross-validation

Actual \ Predicted	High	Low	Medium	Σ
High	365	0	0	365
Low	0	305	0	305
Medium	9	0	321	330
Σ	374	305	321	1000

(b). Using random sampling

Fig. 3. Confusion matrix

The results obtained from the evaluation stage using the cross-validation technique produced an accuracy of 96.7%.This accuracy result means that the probability of a correct prediction for new data is 96.7%. These accuracy results show that the decision tree model has very good accuracy. However, model accuracy can be influenced by various things, including the quality used and many new data features are used. The better data quality, the greater the possibility that the resulting accuracy will be higher. The more features in the dataset used, the higher the resulting accuracy.

The results of the precision and recall evaluation using the cross-validation technique obtained a result of 96.7%, which means that the prediction accuracy will reach 96.7% when the model is used to predict new data.

The results of the f1-score evaluation using the cross-validation technique produced an f1-score value of 96.7%, which means the performance of the model built was very good.

As for previous research, the model experienced overfitting using random sampling techniques. His research resulted in an accuracy of 100%, precision of 100%, recall of 100%, and f1-score of 100% [20]. These results indicate that there is overfitting here and special treatment is needed so that the results are better.

By using cross-validation techniques the resulting accuracy, precision, recall, and f1-score were lower than in previous research, but the results were more stable. This is because in evaluating the model using the cross-validation technique, all data is used for training data and validation data according to the k-fold value, namely 10. The 1000 data used are divided into 10 segments and in each iteration, these segments are used as validation data so that All data is useful as training data and validation data. Based on this, the evaluation carried out is more general and the results are also more stable so that it can overcome overfitting in the decision tree model.

5. CONCLUSION

Implementing a decision tree for lung cancer prediction produces a rules model that can be used to predict lung cancer. After the model rules were created, the model was evaluated using cross-validation techniques to overcome overfitting in previous research with a k-fold value = 10. This cross-validation technique is better compared to the random sampling technique because the results obtained from the cross-validation technique are more stable and with the cross-validation technique the prediction error can be estimated. The results of overfitting improvements in previous research were successful in reducing accuracy, precision, recall, and f1-score by 3.3%. In this study, the resulting accuracy was 96.7%; precision of 96.7%; recall of 96.7%; and f1-score of 96.7%. Based on these results, it can be concluded that the cross-validation technique can overcome overfitting in the decision tree model with more stable and general results.

DAFTAR PUSTAKA

[1] S. S. A.-N. Ibrahim M. Nasser, “Lung Cancer Detection Using Artificial Neural Network,” Int. J. Eng. Inf. Syst., vol. 3, no. 3, pp. 17–23, 2019.

[2] P. Chen, Y. Liu, Y. Wen, and C. Zhou, “Non-small cell lung cancer in China,” Cancer Commun., vol. 42, no. 10, pp. 937–970, 2022, doi: 10.1002/cac2.12359.

[3] C. M. Rudin, E. Brambilla, C. Faivre-Finn, and J. Sage, “Small-cell lung cancer,” Nat. Rev. Dis. Prim., vol. 7, no. 1, 2021, doi: 10.1038/s41572-020-00235-0.

[4] Z. Bing, Z. Zheng, and J. Zhang, “Risk factors influencing chemotherapy compliance and survival of elderly patients with non-small cell lung cancer,” Afr. Health Sci., vol. 23, no. 3, pp. 291–300, 2023, doi: 10.4314/ahs.v23i3.35.

[5] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” CA. Cancer J. Clin., vol. 72, no. 1, pp. 7–33, 2022, doi: 10.3322/caac.21708.

[6] L. Wheless, J. Brashears, and A. J. Alberg, “Epidemiology of lung cancer,” Lung Cancer Imaging, pp. 1–15, 2021, doi: 10.1007/978-1-60761-620-7_1.

[7] M. B. Schabath and M. L. Cote, “Cancer progress and priorities: Lung cancer,” Cancer Epidemiol. Biomarkers Prev., vol. 28, no. 10, pp. 1563–1579, 2019, doi: 10.1158/1055-9965.EPI-19-0221.

[8] F. Verhaegen, C. Caruana, and P. Allisy-roberts, Lung Cancer and Imaging. 2019. doi: 10.1088/978-0-7503-2540-0.

[9] American Lung Association, “State of Lung Cancer: Texas,” 2021, [Online]. Available: https://www.lung.org/research/state-of-lung-cancer/states/texas

[10] Globocan, “Cancer in Indonesia,” JAMA J. Am. Med. Assoc., vol. 247, no. 22, pp. 3087–3088, 2020, doi: 10.1001/jama.247.22.3087.

[11] N. S. Deepa et al., “Prediction of skin cancer using convolutional neural network (cnn),” Neuromorphic Comput. Syst. Ind. 4.0, vol. 9, no. 3, pp. 117–143, 2023, doi: 10.4018/978-1-6684-6596-7.ch005.

[12] S. Senthil and B. Ayshwarya, “Lung Cancer Prediction using Feed Forward Back Propagation Neural Networks with Optimal Features,” Int. J. Appl. Eng. Res., vol. 13, no. 1, pp. 318–325, 2018, [Online]. Available: http://www.ripublication.com

[13] R. Patra, Prediction of lung cancer using machine learning classifier, vol. 1235 CCIS. Springer Singapore, 2020. doi: 10.1007/978-981-15-6648-6_11.

[14] R. A. Sowah, A. A. Bampoe-Addo, S. K. Armoo, F. K. Saalia, F. Gatsi, and B. Sarkodie-Mensah, “Design and Development of Diabetes Management System Using Machine Learning,” Int. J. Telemed. Appl., vol. 2020, 2020, doi: 10.1155/2020/8870141.

[15] S. Ghosh, A. Dasgupta, and A. Swetapadma, “A study on support vector machine based linear and non-linear pattern classification,” Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2019, no. Iciss, pp. 24–28, 2019, doi: 10.1109/ISS1.2019.8908018.

[16] A. Septhiani, “Analisis Perbandingan Algoritma Supervised Learning untuk Prediksi Kasus Covid-19 di Jakarta,” J. Sains Komput. Inform., vol. 7, no. September, pp. 583–594, 2023, doi: http://dx.doi.org/10.30645/j-sakti.v7i2.668.

[17] S. Saeed, A. Abdullah, N. Jhanjhi, and T. Malaysia, “Analysis of the Lung Cancer patient’s for Data Mining Tool,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 7, p. 90, 2019.

[18] Kumar Mohan and Bhraguram Thayyil, “Machine Learning Techniques for Lung Cancer Risk Prediction using Text Dataset,” Int. J. Data Informatics Intell. Comput., vol. 2, no. 3, pp. 47–56, 2023, doi: 10.59461/ijdiic.v2i3.73.

[19] T. Gupta, T. Qawasmeh, and S. McCalla, “Predictions of Programmed Cell Death Ligand 1 Blockade Therapy Success in Patients with Non-Small-Cell Lung Cancer,” BioMedInformatics, vol. 3, no. 4, pp. 1060–1070, 2023, doi: 10.3390/biomedinformatics3040063.

[20] Faurika, A. N. Khudori, and M. S. Haris, “Implementasi Decision tree Untuk Prediksi Kanker Paru-

Paru," J. Ris. Sist. Inf. dan Tek. Inform., vol. 9, pp. 94–106, 2024, doi: http://dx.doi.org/10.30645/jurasik.v9i1.717.g692.

[21] X. Wang and Z. Cheng, "Cross-Sectional Studies: Strengths, Weaknesses, and Recommendations," Chest, vol. 158, no. 1, pp. S65–S71, 2020, doi: 10.1016/j.chest.2020.03.012.

[22] C. An, Y. W. Park, S. S. Ahn, K. Han, H. Kim, and S. K. Lee, "Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results," PLoS One, vol. 16, no. 8 August, pp. 1–13, 2021, doi: 10.1371/journal.pone.0256152.

[23] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," Comput. Mater. Sci., vol. 171, no. July 2019, p. 109203, 2020, doi: 10.1016/j.commatsci.2019.109203.

[24] Nti Isaac Kofi, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," Int. J. Inf. Technol. Comput. Sci., vol. 13, no. 6, pp. 61–71, 2021, doi: 10.5815/ijitcs.2021.06.05.

[25] D. Normawati and D. P. Ismi, "K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining," Signal Image Process. Lett., vol. 1, no. 2, pp. 23–35, 2019, doi: 10.31763/simple.v1i2.3.

[26] Q. Aini, N. Lutfiani, H. Kusumah, and M. S. Zahran, "Deteksi dan Pengenalan Objek Dengan Model Machine Learning: Model Yolo," CESS (Journal Comput. Eng. Syst. Sci., vol. 6, no. 2, p. 192, 2021, doi: 10.24114/cess.v6i2.25840.

[27] N. Wiranda, H. S. Purba, and R. A. Sukmawati, "Survei Penggunaan Tensorflow pada Machine Learning untuk Identifikasi Ikan Kawasan Lahan Basah," IJEIS (Indonesian J. Electron. Instrum. Syst., vol. 10, no. 2, p. 179, 2020, doi: 10.22146/ijeis.58315.

[28] G. Carleo et al., "Machine learning and the physical sciences," Rev. Mod. Phys., vol. 91, no. 4, p. 45002, 2019, doi: 10.1103/RevModPhys.91.045002.

[29] Ö. Günaydin, M. Günay, and Ö. Şengel, "Comparison of lung cancer detection algorithms," 2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019, 2019, doi: 10.1109/EBBT.2019.8741826.

[30] R. R. Pratama, "Analisis Model Machine Learning Terhadap Pengenalan Aktifitas Manusia," J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 19, no. 2, pp. 302–311, 2020, doi: https://doi.org/10.30812/matrik.v19i2.688.

[31] A. Saputra, U. Nahdlatul, U. Sidoarjo, and K. Sidoarjo, "KLASIFIKASI PENGENALAN BUAH MENGGUNAKAN ALGORITMA NAIVE," J. Rekayasa Sist. Komput., vol. 2, no. 2, pp. 83–88, 2019, doi: https://doi.org/10.31598/jurnalresistor.v2i2.434.

[32] A. Rajeshkanna and K. Arunesh, "ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 787–790, 2020, doi: 10.1109/ICESC48915.2020.9155578.

[33] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," Int. J. Adv. Eng. Res. Sci., vol. 7, no. 4, pp. 514–521, 2020, doi: 10.22161/ijaers.74.60.

[34] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," Ann. Appl. Stat., vol. 9, no. 3, pp. 1350–1371, 2015, doi: 10.1214/15-AOAS848.

[35] P. Pavithra, T. Nadu, and T. Nadu, "DETERMINE THE ENTROPY IN ID3 ALGORITHM TO SPLITTING A DECISION," Int. J. Mech. Eng., vol. 6, no. 3, pp. 2447–2450, 2021.

[36] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," Comput. Stat., vol. 36, no. 3, pp. 2009–2031, 2021, doi: 10.1007/s00180-020-00999-9.

[37] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," Ecol. Monogr., vol. 93, no. 1, pp. 1–24, 2023, doi: 10.1002/ecm.1557.

[38] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," Clean. Eng. Technol., vol. 15, no. July, p. 100664, 2023, doi: 10.1016/j.clet.2023.100664.

[39] D. Berrar, "Cross-validation 1," Encycl. Bioinforma. Comput. Biol., pp. 1–13, 2024.