Collabits Journal
Vol. 1 No. 3 September 2024 : 146-151
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

Comparative Analysis of Performance Between KNN and C5.0 Algorithms in Lung Cancer Disease Detection

Siti Maesaroh[1], Ibka Anhar Fatcha[2*], Fikri Ramadhan[3]

[1,2,3] Computer Science study program, Universitas Mercu Buana, Jakarta, Indonesia

*Coressponden Author: ibkaanhar1@gmail.com

Abstract - Lung cancer is a disease characterized by the growth of abnormal cells in the lungs that can spread to other parts of the body. In practice, medical teams will usually evaluate a patient's symptoms conventionally, which is highly inefficient and time-consuming, especially if there are a large number of patients. This manual evaluation process can cause delays in diagnosis and treatment, and increase the risk of errors. Therefore, this research will discuss lung cancer detection using the K-Nearest Neighbors (KNN) algorithm and the C5.0 algorithm in order to solve the problems previously described. The use of the K-Nearest Neighbors (KNN) algorithm and the C5.0 algorithm was chosen because these two algorithms have the ability to process complex data and produce accurate models. The results of this study will show a comparison of which performance is much better used to accurately detect lung cancer based on the amount of training data available, and it can be known that the lung cancer detection process can be done more quickly and efficiently, using the K-Nearest Neighbors (KNN) or C5.0 algorithm to improve diagnosis accuracy. The results show that the KNN algorithm is superior to the C5.0 algorithm specifically for lung cancer detection.

Keywords :
Lung cancer;
K-Nearest Neighbors Algorithm;
C5.0 Algorithm;
Classification;

Article History:
Received: 05-07-2024
Revised: 26-08-2024
Accepted: 18-09-2024

Article DOI : 10.22441/collabits.v1i3.27285

1. INTRODUCTION

Lung cancer is a condition caused by the growth of abnormal cells in the lung tissue. When we inhale, the air will enter through the nostrils, then will pass through the filtering or filtration process on small particles carried out by the nose hairs which will then go to the trachea, then the air from the trachea will then enter the lungs through the bronchi and bronchioles which will then end in the alveolus.

Lung cancer itself is divided into two types, namely non-small cell lung cancer and small cell lung cancer [1]. For non-small cell lung cancer, which is a relatively common type of lung cancer. For the initial symptoms of its development, it comes from several cells that are responsible for lining the bronchial tubes. Then for small cell lung cancer, which is a type of lung cancer that is classified as uncommon, but this type of cancer spreads so quickly through blood vessels to various parts of the body. Early symptoms of lung cancer can be from some of the cells lining the alveolus.

Lung cancer itself is one of the types of cancer cases most commonly encountered by the medical team and is the most dangerous type of cancer. From the Global Cancer Observation (Globocan) data quoted in 2020, it can be explained that lung cancer is the type of cancer with the 3rd most cases in Indonesia with a total of 34,000 new cases. But even more concerning, lung cancer is the type of cancer that is classified as the highest casualty in Indonesia with a total death rate of more than 30,000 people, all of which are caused by this type of cancer.

The most common method used by medical or medical authorities to determine the level of development of this lung disease, namely by using the "TNM" system, which stands for "Tumor", "Nodus", and "Metastasis", here is a brief explanation related to the method or system, which is as follows:

1. Tumors
   Tumors are usually fixed by the size and location of the tumor in the lungs or body parts.

2. Nodus
   Nodus are usually involved with the lymph nodes, which allows whether it is true that the cancer in the human body has spread to the lymph nodes that are close to the lungs.

3. Metastasis
   Metastasis is a level of cancer spread to other organs in the human body, for example in addition to the lungs, it can also spread to the liver, bones, brain, kidneys, adrenal glands, and other organs.

Each variable of the acronym has its own indicators, so that the medical team or medicine decides on the staging of this cancer, which is from a scale of 0 to 4 based on this TNM method. For the initial symptoms of lung cancer itself, it is often not obvious, but if the cancer begins to spread or develop, then the following symptoms

can be categorized as the initial symptoms that cause lung cancer, which are as follows:

- Cough whose condition is getting worse day by day
- Hoarseness of voice
- Difficulty breathing, such as shortness of breath
- Feeling pain in the chest area constantly
- Coughing up blood
- Feeling tired or weak all the time
- Having an infection in the lungs, such as pneumonia
- Weight loss without doing a diet program or for no apparent reason.

For the main cause of lung cancer itself is smoking, although not all lung cancer is caused by smoking. Then, inhaling cigarette smoke from active smokers, or what we know as passive smoking, it can also be a cause of lung cancer for some people who do not smoke. The percentage of lung cancer cases is associated with smoking, with a percentage of 85%.

The way to prevent lung cancer is to adopt a healthy lifestyle and avoid factors that can cause lung cancer, such as not getting used to smoking, exercising regularly, increasing the consumption of fruits and vegetables, and other healthy foods, and having regular medical check-ups.

In this journal, we will discuss in detail the classification of lung cancer detection. But before that, we need to know about classification. Classification in general is a process of grouping or categorizing objects, data, or other information into much smaller groups or categories based on the similarity of certain properties or characteristics, with the aim of making it easier when we do an analysis, and organizing information. This classification is also often used to organize various elements, such as living things in the context of biology, data in computer science, and many others.

Therefore, in this study two classification algorithms were used with the aim of detecting lung cancer, namely by using the K-Nearest Neighbors (KNN) algorithm and the C5.0 algorithm. The use of the K-Nearest Neighbors (KNN) algorithm and the C5.0 algorithm was chosen because these two algorithms have the ability to process complex data and produce accurate models. The results of this study will show a comparison of which performance is much better used to accurately detect lung cancer based on the amount of training data available.

2. RESEARCH METHOD

2.1 Literature Review

To support this research, researchers conducted a literature review of previous existing research. The following is the research concerned:

1. The K-Nearest Neighbor method has been widely used in research that has the aim of classifying various types of datasets. Such as research on sentiment surveys regarding the relocation of Indonesia's capital city in 2023 using the K-Nearest Neighbor and Naïve bayes classification methods applied to a collection of 800 data obtained by crawling twitter with the keyword capital city transfer. KNN gets the results of Accuracy 88.12%, Precision 93.98%, and Recall 81.53%. While NB Accuracy 82.27%, Precision 86.36% and Recall 76.93% [2].
2. Other research in discussing the performance of the KNN method on breast cancer datasets using 569 data and tested using K-Fold Cross Validation, the results obtained are high Accuracy values of 93% on K3, K4, First K5, Precision 97% on K3 4th, Recall 98% K3 3rd and F-Measure 94% K5 Third [3].
3. The method of the C5.0 algorithm can be used to accurately classify graduation rates. For example, the C5.0 algorithm is used for classification of graduation data for the Tanjungpura University Statistics Study Program Batch 1 of the 2017/2018 academic year to batch 2 of the 2022/2023 academic year with a total of 140 data, and with a description of ten independent attributes and one dependent attribute. Where this algorithm shows that the influence on graduation rates with a percentage accuracy of 70% [4].
4. Then the C5.0 algorithm is also used in discussions related to classifying potential customers at a post office in the city of Cimahi using training data divided by 70% and test data divided by 30%. The goal is to find out the best information gain value for potential customers with those that are not which will then be used as the root. For the results obtained, namely in the form of a total entropy of 0.999717128, the gain ratio of the number of transactions is 0.828709376, and the accuracy of the calculation using the C5.0 algorithm, which is 96% [5].

2.2 Dataset

The dataset used in this study is public data taken from kaggle [6]. The dataset has a total of 309 rows of data with 16 attributes which after passing the data preprocessing process will consist of 276 rows, the dataset contains data from patients who do not have and who have lung cancer and also the habits, body conditions, and other diseases of each patient with lung cancer. The amount of data that has the class “YES” for patients with cancer is 270, the data that has the class “NO” is 39. The purpose of classification with this dataset as mentioned earlier, is to detect whether the patient has lung cancer or not, based on the symptoms possessed by the patient. The attributes of the dataset can be seen at

Table 1.

Table 1. Dataset Attributes

Table 1

Attribute	Data Type
GENDER	Category
AGE	Numeric

Attribute	Data Type
SMOKING	Numeric
YELLOW_FINGERS	Numeric
ANXIETY	Numeric
PEER_PRESSURE	Numeric
CHRONIC_DISEASE	Numeric
FATIGUE	Numeric
ALLERGY	Numeric
WHEEZING	Numeric
ALCOHOL_CONSUMING	Numeric
COUGHING	Numeric
SHORTNESS_OF_BREATH	Numeric
SWALLOWING_DIFFICULTY	Numeric
CHEST_PAIN	Numeric
LUNG_CANCER	Category

A column that is of numeric data type and has a value of "2" is equal to "YES", and if the column has a value of "1" it is equal to "NO".

2.3 Pre-processing

Pre-processing is the process of preparing raw data that will be processed into data that is ready to be used as input for the machine learning model. This stage is very important because the pre-processing stage can make the performance of the machine learning model better or vice versa [2]. In this stage, researchers conducted several pre-processing stages for the dataset used, namely:

1. Deduplication: The process of removing data that has the same value (duplication) from the dataset.
2. Label Encoding: The process of converting data from categorical data to numerical data.
3. Normalization: The process of converting data to a smaller scale.
4. Data Splitting: The process of dividing the dataset with a scale of 2:1, 2 for training and 1 for test.

2.4 *K-Nearest Neighbor* (KNN)

KNN is a classification method for a set of data used for classification of the distance from an object to the data that has been classified. This algorithm requires previously classified data (labels) for model building, because the KNN algorithm is a supervised learning type machine learning algorithm [3], so KNN requires a label for learning the algorithm and creating a model. The formula used for distance calculation in KNN which is most commonly used is the Euclidean distance calculation, the formula can be seen at Equation 1:

$$euc = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad 1$$

$n =$ The number of data
i = Index of the data
$p_i =$ the i training data
$q_i =$ the i test data

The steps to calculate using the KNN algorithm in general are:

1. Determine the K parameter
2. Calculate the distance from training data to testing data using the Euclidean formula in Equation 1.
3. Sort the distance obtained.
4. Determine the closest distance with parameter K.
5. Find the closest number of classes and assign the class for evaluation.

2.5 C5.0

C5.0 is an algorithm in data mining that can be used to build a decision tree. C5.0 itself is a development algorithm of the ID3 and C4.5 algorithms. When compared to other classification algorithms, this algorithm has an advantage, namely in a much shorter execution time and a relatively good accuracy rate. The C5.0 algorithm starts by becoming all the root data and attributes used from the classification tree into a divisor of the sample. the following stages are used in the C5.0 Algorithm, which are as follows:

1. Entropy Value Calculation

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad 2$$

S = The number of data
N = The number of classes of the dependent variable.
P_i = The ratio of the number of i class data from the data set.

Entropy is used to measure the uncertainty of a data set. The entropy value can be calculated from the comparison of the amount of data in each class.

2. Calculate the split info from each class on the variable and gain using Equation 3 and 4:

$$\text{SplitInfo}(S,A_i) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \qquad 3$$

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times \text{Entropy}(S) \qquad 4$$

S = Total number of samples in the data set
A_i = The i independent variable.
S_i = The number of samples for the i category
k = Number of categories in independent variable A

Split Info is used in order to measure how well an attribute can divide data into different classes. Then Gain will also measure how much information is obtained from the attribute. These two values can be used with the aim of selecting the attribute that is classified as the best, which will then be used as the parent node.

3. Then each attribute will have its gain ratio value calculated. Parent nodes will be selected from attributes that have a relatively high gain ratio value, which will then be made branches from each

parent node category. And the gain ratio will be calculated using Equation 5.

$$\text{Gain Ratio} = \frac{\text{Gain(S,A)}}{\text{SplitInfo}(S,A_i)} \qquad 5$$

Gain(S,A) = Represents gain-related information on independent variable A
$\text{SplitInfo}(S,A_i)$ = Represents the split information related to the independent variable A of the i class

Gain ratio itself is the ratio between gain and split info. The attribute that has the highest gain ratio will be chosen to be the parent node.

2.6 Metrics Evaluation

At this stage, after the KNN and C5.0 models have been trained and created, the evaluation process of the KNN and C5.0 model results will be compared using the Confusion Matrix method. Confusion Matrix is one of the specialized evaluation methods of classification algorithms [7].

Table 2. Confusion Matrix

		Predicted	
		False	*True*
Actual	*False*	TN	FP
	True	FN	TP

TN (True Negative) = The model predicts that the data is Negative, and the actual data is Negative, or in other words the model successfully predicts.
TP (True Positive) = The model predicts that the data is Positive, and the actual data is Positive, or in other words the model successfully predicts.
FN (False Negative) = The model predicts that the data is Negative, but the actual data is Positive, or in other words the model failed to predict.
FP (False Positive) = The model predicts that the data is Positive, but the actual data is Negative or in other words the model fails to predict.

The process of measuring the performance of the algorithm model will be carried out using several parameters that will be obtained from the confusion matrix, including accuracy, recall and precision [8].

Accuracy is the percentage value of closeness between predicted values and data, for calculating the accuracy of the results of each model can use the formula in Equation 6.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad 6$$

Precision is the ratio of the match between the correct prediction result and the actual value, to calculate precision the formula in Equation 7 can be used.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad 7$$

And then there is recall, recall is a measure of how well our model predicts, to get the recall value of the model the formula in Equation 8 can be used.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad 8$$

After getting accuracy, precision and recall, F-Measure will be calculated to determine the harmonic mean weight of the percentage of precision and recall values, Equation 9 is the formula for calculating the F-Measure value.

$$\text{F1 Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad 9$$

3. RESULT AND DISCUSSION

3.1 Pre-processing

As explained earlier that pre-processing is very important, because this stage can affect the performance of the algorithm used.

1. Deduplication

   Deduplication is a pre-processing step that removes duplicate data, because there are 33 rows with the same content then the data will be removed. Therefore, the total rows of the dataset will be reduced to 276.

2. Label Encoding

   Label Encoding is the process of changing data that has a categorical data type into numeric. As mentioned earlier that there are 2 attribute columns with categorical data types, namely "GENDER" and "LUNG_CANCER", because the KNN and C5.0 algorithms only accept numeric input, data that has a categorical data type needs to be converted into numeric. So the "GENDER" column is split into 2 columns, namely "MALE" and "FEMALE", then if the data from the "GENDER" column is "M" then the value of the "MALE" column data will be 1 and for the "FEMALE" column it will be 0 and vice versa. Then for the "LUNG_CANCER" column, only the value of "YES" is changed to 1, and the value of "NO" becomes 0.

3. Normalization

   Normalization is the process of changing data into a small scale so that the model to be created has excellent performance. The Normalization process is only done for the "AGE" column using the StandardScaler class provided by the scikit library in python.

4. Data Splitting

   Data Splitting is the process of dividing the dataset into a 2:1 scale, 2 for training and 1 for test. Data Splitting is done to avoid overfitting the dataset, so

the dataset will be divided into 80% training data and 20% test data.

3.2 Evaluation Results

After modeling and predicting the data that has been split into test data, researchers evaluate using the previously mentioned metrics.

Figure 1. KNN Confusion Matrix Result

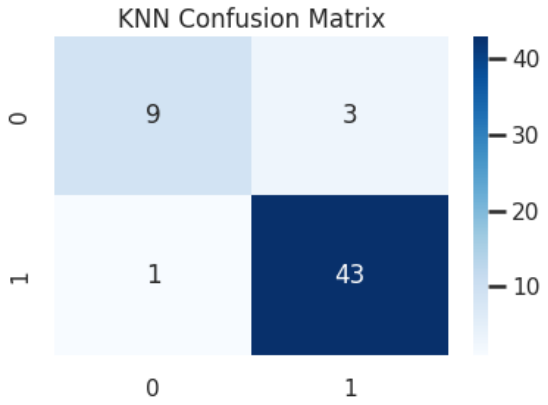
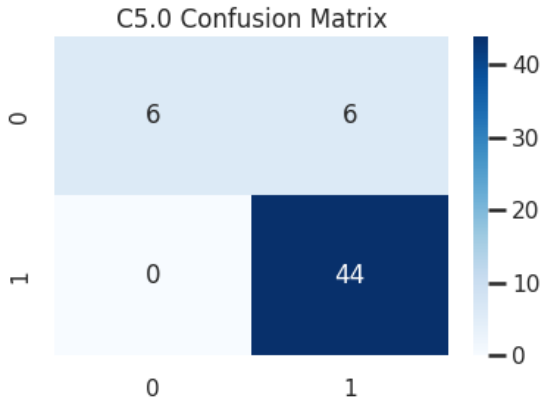
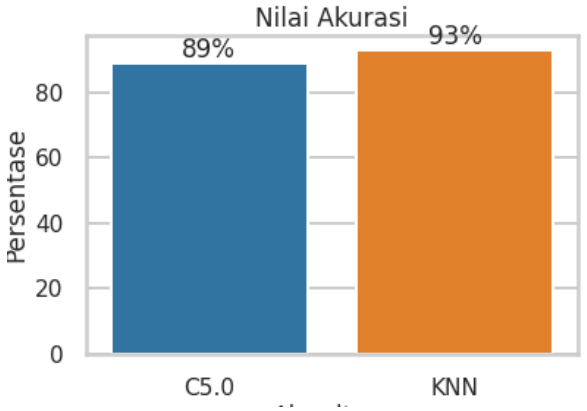


Figure 2. C5.0 Confusion Matrix Results



From the Confusion Matrix results in Figure 1 and Figure 2 obtained after classifying the 2 algorithms used, it shows that KNN has 9 True Negative, 43 True Positive, 3 False Positive, and 1 False Negative value. Followed by C5.0 which has 6 True Negative, 44 True Positive, 6 False Positive, and 0 False Negative values.

Figure 3. Accuracy Results from KNN and C5.0



The accuracy results listed in

Figure 3 are obtained from 2 algorithm models that have been made, showing that the accuracy of the KNN algorithm model is superior with a value of 93% followed by the C5.0 algorithm model which only has 89% accuracy.

Figure 4. Precision Results of KNN and C5.0

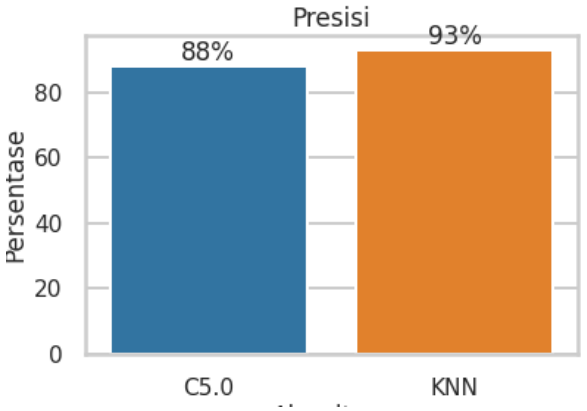
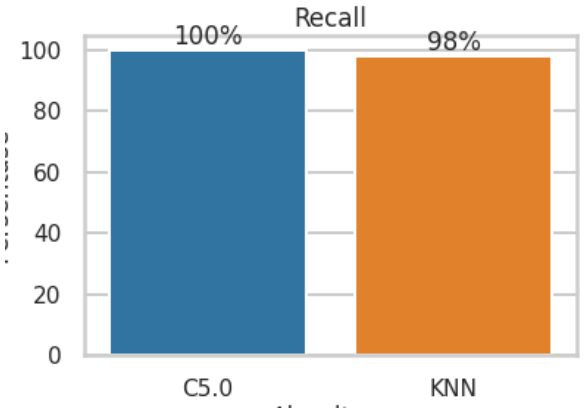


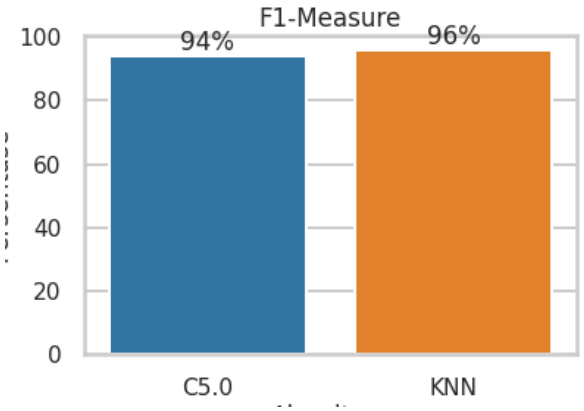
Figure 4 shows that the KNN precision value is still superior with a 5% difference, the precision value of the KNN algorithm is 93% compared to the C5.0 algorithm which has a value of 88%.

Figure 5. Recall Results of KNN and C5.0



From Figure 5 above, it can be seen that the recall value of the C5.0 algorithm is superior by 2% with a value of 100% and for the KNN value has a value of 98%.

Figure 6. F1-Measure Results from KNN and C5.0



The value for F1-Measure that has been attached in

Figure 6 shows that the F1-Measure value for the KNN algorithm is 96% while the C5.0 algorithm has a value of 94% which indicates that the KNN algorithm is 2% superior in F1-Measure value.

4. CONCLUSION

Overall, from the evaluations that have been carried out, it shows that the KNN algorithm is superior in almost every metric evaluation, it's just that there are a number of things that need to be considered if we want to use the model that has been made as a cancer detector because if a model incorrectly predicts a patient who actually has cancer but the prediction results show that he does not have cancer it is a fatal error. Therefore, researchers will compare the performance of 2 algorithms using the F1-Measure value which is the harmonic mean of precision and recall, and also accuracy which are all obtained from the confusion matrix. From the overall performance evaluation of the 2 algorithms used, it shows that KNN is almost always superior in every category, from F1-Measure which has a value of 96%, and accuracy with a value of 93%. It's just that in the confusion matrix results, the KNN algorithm shows that there is 1 data included in the True Negative, but because the KNN metric value is superior to C5.0, therefore 1 data will not argue that the KNN algorithm is better than C5.0. So it can be concluded that overall the KNN algorithm is superior and suitable for lung cancer detection.

BIBLIOGRAPHY

[1] "Types of lung cancer." Accessed: May 19, 2024. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types

[2] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, pp. 1–7, Jan. 2023, doi: 10.35134/komtekinfo.v10i1.330.

[3] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, Jul. 2020, doi: 10.33096/ijodas.v1i2.13.

[4] Y. Crismayella, N. Satyahadewi, and H. Perdana, "Algoritma Adaboost pada Metode Decision Tree untuk Klasifikasi Kelulusan Mahasiswa," *Jambura J. Math.*, vol. 5, no. 2, pp. 278–288, Aug. 2023, doi: 10.34312/jjom.v5i2.18790.

[5] N. H. Harani and F. S. Damayanti, "Implementasi Algoritma C5.0 Untuk Menentukan Pelanggan Potensial Di Kantor Pos Cimahi," *J. SITECH Sist. Inf. Dan Teknol.*, vol. 4, no. 1, pp. 69–76, Jun. 2021, doi: 10.24176/sitech.v4i1.6281.

[6] "Lung Cancer." Accessed: May 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer

[7] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 3, pp. 504–510, Jun. 2021, doi: 10.29207/resti.v5i3.3067.

[8] E. Wulandari, "KLASIFIKASI KANKER PARU-PARU MENGGUNAKAN METODE NAIVE BAYES," *Int. Res. Big-Data Comput. Technol. - Robot*, vol. 6, no. 2, pp. 20–24, Sep. 2022, doi: 10.53514/ir.v6i2.325.