Early Detection of COVID-19 Disease Based on Behavioral Parameters and Symptoms Using Algorithm-C5.0

[1]Joko Riyono, [2]Aina Latifa Riyana Putri, [3]Christina Eni Pujiastuti
[1,3]Industrial Technology Faculty, Universitas Trisakti
[2] Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Diponegoro
Email: [1]jokoriyono@trisakti.ac.id, [2]ainalatif47@gmail.com, [3]christina.eni@trisakti.ac.id

Article Info

Article history:
Received Jan 04th, 2023
Revised Feb 20th, 2023
Accepted Mar 24th, 2023

Keyword:
Algorithm C5.0
Eearly Detection
Classification
Confusion Matrix

ABSTRACT

The spread of COVID-19 disease has continued since it was first discovered at the end of 2019 until now. Transmission of COVID-19 is very fast, including through close contact through droplets and through the air. Therefore, early detection of COVID-19 is very important for patients and also those around them to be able to fight the COVID-19 pandemic because if patients get proper and fast treatment, then other people around them will be protected. In this study, an analysis of the classification of decision making for COVID-19 detection was carried out based on behavioral parameters and symptoms that could trigger exposure to COVID-19 using the C5.0 algorithm, followed by measuring the performance of the model using the Confusion Matrix. The C5.0 algorithm is a decision tree-based data mining method. The results of the C5.0 algorithm use a comparison of training data and test data of 70:30. After going through the Confusion Matrix test, an accuracy value of 98% is obtained which indicates that the resulting classification is very good, so that the resulting model can be used for early detection of COVID-19 patients.

Corresponding Author:
Joko Riyono,
Industrial Technology Faculty,
Universitas Trisakti,
Kyai Tapa Street, Grogol Jakarta 11440, Indonesia
Email: jokoriyono@trisakti.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v2i2.22074

1. INTRODUCTION

COVID-19 is an infectious disease caused by the Corona virus. This virus does not only attack animals but among them also attacks humans. In December 2019, a new type of corona virus was discovered in Wuhan China which was later named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2) and is known to attack humans. This virus is similar to or comes from the same family as the virus that caused (SARS) Severe Acute Respiratory Syndrome in 2002. The symptoms experienced by people with COVID-19 are usually mild and some may not even show any symptoms. The symptoms that sufferers can usually feel and experience include experiencing aches, nasal congestion, sore throat, diarrhea, loss of smell or skin rashes. In severe cases, sufferers will experience Acute Respiratory Distress Syndrome (ARDS), kidney failure, heart failure, which can result in death [1].

The total number of cases of COVID-19 in the world is still increasing, namely 504,571,336 on 24 June 2022. Because based on scientific evidence that the spread of COVID-19 is very fast and can be transmitted through close contact or droplets and through the air, the government and World Health Organization (WHO) have taken several precautions to be able to help reduce cases of COVID-19 such as preparing for COVID-19 treatment in infected patients, increasing surge capacity at health care facilities and arranging patient screening. Preventive measures have a major role in suppressing COVID-19 cases if protocol therapy (Protocol Therapy) is implemented from an early stage [2]. Early detection of COVID-19 is one way to help expedite action for patients whether they are healthy or need further testing regarding COVID-19. The COVID-19 early detection system is considered very important for patients and also the people around them to be able to fight the COVID-19 pandemic because if the patient gets proper and fast treatment, then other people around him will be protected. Several studies related to disease detection have been carried out, such as the study by Sisodia & Sisodia (2018) who designed a model for detecting diabetes in patients with the Naïve Bayes algorithm and obtained an accuracy of 76.30% [3].

In Li et al's study (2020) identified heart disease in patients using the FCMIM-SVM algorithm and the resulting model had good accuracy [4]. In the study Karthiyekan & Thangaraju (2013) analyzed hepatitis patients using the Naïve Bayes algorithm with an accuracy of 81 -84% [5]. In the study Ramana et al (2011) evaluated the detection of liver disease and obtained an accuracy of 71.59% using the Back propagation Neural Network algorithm [6]. The four studies were classification analyzes using various algorithms.

Classification analysis is carried out with the aim of classifying an object based on the characteristics or characteristics possessed by an individual. Algorithm C5.0 is one of the data mining methods for decision tree-based classification techniques, a refinement of the ID3 and C4.5 algorithms. This algorithm has also been widely applied, such as in the study of Bujlow & Pedersen (2012) to distinguish various types of traffic in computer networks with an average accuracy of 99.3-99.9% [7]. In Pang & Gong's research (2009) concluded the decision tree model for individual loans from commercial banks using the C5.0 Algorithm has high accuracy [8]. In the research of Kurniawan et al (2019) the C5.0 Algorithm classification model for forecasting rainfall in Bandung with an accuracy of 92% [9]. The C5.0 Algorithm is able to classify by Good [10].

In this study, based on the studies mentioned above, specifically related to the symptoms felt by COVID-19 patients and the use of the C5.0 algorithm, an analysis of the classification of decision making for early detection of COVID-19 will be carried out based on the symptoms felt, contact history, and patient mobility history. Considering the accuracy of the results, it will try to be analyzed for several values of the ratio of training data and test data. Separation between training data and test data is assisted by the Rstudio software. This research was conducted with the aim of facilitating and accelerating the performance of medical personnel so that COVID-19 patients receive fast and appropriate treatment to help reduce COVID-19 cases.

2. RESEARCH METHOD

The research method used in this study is a quantitative method using literacy studies where data is collected using a measuring instrument and then analyzed statistically and quantitatively. As for the data sources and data analysis methods used as in the following explanation.

2.1. Data Source

The data to be used is secondary data from the Kaggle Dataset, Symptoms and Presence of COVID [11]. This data set contains anonymous data of 5434 people in India who tested for COVID-19 (positive for COVID-19) along with their symptoms, contact history and mobility history as presented in Table 1.

Table 1. Research Variable

Variable	Data Type
Difficulty Breathing (BP)	Categorial (Yes/No)
Fever (FE)	Categorial (Yes/No)
Dry Cough (DC)	Categorial (Yes/No)
Sore Throat (ST)	Categorial (Yes/No)
Runny Nose (RN)	Categorial (Yes/No)
Headache (HE)	Categorial (Yes/No)
Fatigue (FA)	Categorial (Yes/No)
International Trip (AT)	Categorial (Yes/No)
Contact With COVID-19 Patients (CW)	Categorial (Yes/No)
Visiting Large Meeting (AL)	Categorial (Yes/No)
Visiting Public Open Places (VP)	Categorial (Yes/No)
Exposed To COVID-19 (COVID-19)	Categorial (Yes/No)

2.2. Data Analysis Method

The following are the experimental stages carried out in the research that can represent by the flowchart graphic image below:

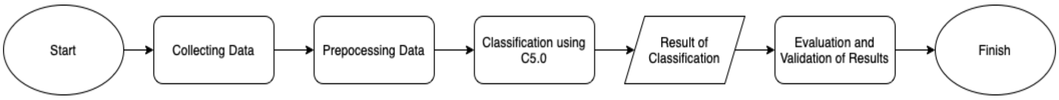


Figure 1. Research Flowchart

1. Data collection
   The data used will be downloaded and saved as a file with an .xlsx extension.

2. Processing the previous data
   Initial data processing will include Data Selection, namely the selection (selection) of data from a set of data. The data selected from the selection results, as in Table 1, will be used for the data mining process, namely classification. Second, the variables in Table 1 at Rstudio were modified for ease of execution. The target variable in this study used was the patient's status category whether or not they were infected with COVID-19, consisting of 2 categories, namely:
   a. Y = (1), if the patient is not infected with COVID-19 (No)
   b. Y = (2), if the patient is infected with COVID-19 (Yes)

   Likewise for the predictor variable which consists of 2 categories named the value X = (1), if the variable has a value of Yes and is given a value of X = (2), if the variable has a value of No. At this stage it will also be carried out frequently dividing the data ratio between training data and test data to obtain the highest accuracy in a model can be seen in Table 2.

Table 2. Ratio Separation Data

Ratio	Number of Training Data	Number of Testing Data
90:10	4891	543
80:20	4347	1087
70:30	3804	1630
60:40	3261	2173
50:50	2716	2718
40:60	2173	3261
30:70	1630	3804
20:80	1087	4347
10:90	543	4891

3. Classification with the C5.0 algorithm
   At this stage classification is carried out using data that has previously been processed at the Data Preprocessing stage and an experiment to select each data separation ratio in Table 2 with the C5.0 Algorithm assisted by Rstudio Software. This algorithm is a refinement of the previous algorithm created by Ross Quinlan in 1987, namely the ID3 and C4.5 algorithms. The ID3 algorithm is developed into the C4.5 algorithm where the algorithm is able to handle attributes with discrete and continuous types. This C4.5 algorithm was also developed into the C5.0 algorithm because there are still various weaknesses in the C4.5 algorithm. calculations using the C5.0 algorithm use several attributes, namely entropy, information gain, and gain ratio. Whereas in Algorithm C4.5 the calculation stops until the information is obtained. C5.0 algorithm can choose the attribute based on the highest gain ratio. The equation used for entropy conclusion is:

$$Entropy\ (S) = \sum_{j=1}^{k} -\pi_i log_2(\pi_i) \tag{1}$$

   With S = Case Set ; k = Number of Partitions S; π_i= Proportion of $S_i$and S. The next step is to get the Information Gain Calculation value with the following equation:

$$Information\ Gain(\mathrm{S}, \mathrm{A})\ = Entropy(\mathrm{S})\ - \sum_{i=1}^{m} \frac{|S_i|}{|S|} xEntropy(S_i) \tag{2}$$

   With S = Case Set ; A = Attribute ; m = Number of Categories in Variable A; $|S_i|$= Number of Cases on The i-th Partition; |S|= Number of Cases In S. The final step, calculates the Gain Ratio as the selection of attributes used as nodes based on the highest Gain Ratio with the following equation:

$$Gain\ ratio\ = \frac{Information\ Gain\ (S,A)}{\sum_{i=1}^{m} Entropy(S_i)} \tag{3}$$

   With S_i= Total entropy value in a variable. With this Gain Ratio Calculation, it is what makes the tree builder in C5.0 more concise than the tree in the C4.5 Algorithm. The process is carried out until the subset sample cannot be split.

4. Evaluation and validation of results
   At this stage each model will be evaluated using the Confusion Matrix measurement. The Confusion Matrix is a table with four different combinations of predicted values and actual values to measure the performance of classifying problems. Furthermore, the value can be calculated as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{4}$$

   In determining the best C5.0 Algorithm model, it will be selected based on the highest accuracy value in the test data. Furthermore, the best model also obtained the following values:

$$Precision = \frac{TP}{(TP+FP)} \tag{5}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{6}$$

$$F-1\,Score \;= \frac{(2*Recall*Precision)}{(Recall+Precision)} \tag{7}$$

   With TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. The results of the research will be obtained later on a decision tree using the best model for detecting COVID-19 based on perceived symptoms, contact history, and mobility.

3. RESULTS AND ANALYSIS

In selecting the best C5.0 algorithm model, it is selected based on the accuracy of the complexity matrix in the experiment for each data refinement ratio. Table 3 shows a comparison of the accuracy values of each model. While Figure 2 shows the accuracy value of each model in graphical form.

Table 3. Accuracy

Ratio	Accuracy
90:10	0.9797
80:20	0.9742
70:30	0.9816
60:40	0.9751
50:50	0.975
40:60	0.9779
30:70	0.9703
20:80	0.9696
10:90	0.9401

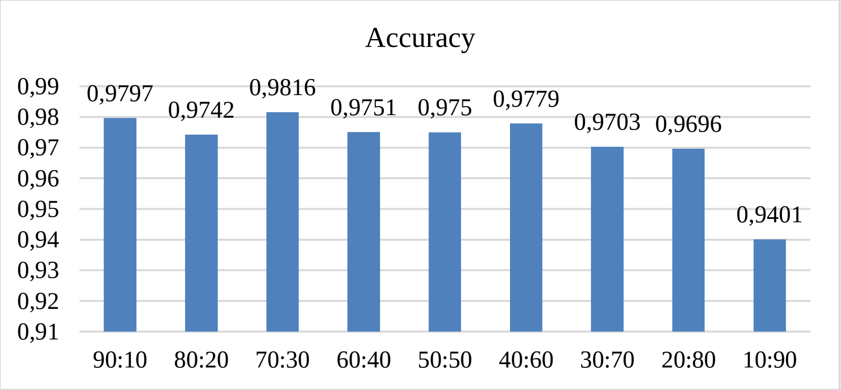
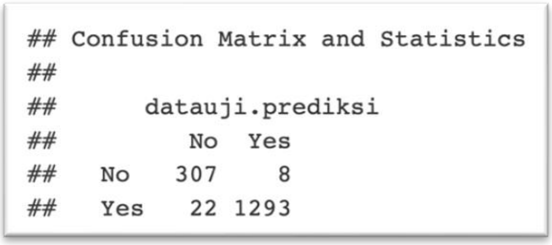


Figure 2. Accuracy Graphic

Based on Table 3 and Figure 2, it was found that the highest accuracy value of 98% was the C5.0 model with a data splitting ratio of 70:30, therefore this model was chosen to be the best model to use for the COVID-19 early detection model based on the symptoms felt , contact history, and patient mobility . From this model it is also possible to obtain other classification performance measuring values such as precision, recall, and F1-Score values in Table 4 based on the matrix confusion table in Figure 3.



```
## Confusion Matrix and Statistics
##
##       datauji.prediksi
##          No  Yes
##   No    307    8
##   Yes    22 1293
```

Figure 3. Best Confusion Matrix Model

Table 4. Best Model Performance

	Precision	Recall	F1 Score
Score	0,933	0,975	0,954

Based on table 4, a precision value of 93% can be obtained, meaning that there are 93% of patients who are truly negative for COVID-19 out of all patients who are predicted to be negative for COVID-19. a recall value of 97% means that there are 97% of patients who are predicted to be negative for COVID-19 compared to all patients who are negative for COVID-19. An F1-Score value of 95% is also obtained. It can be concluded that the C5.0 model with a data separation ratio of 70:30 is very good to use as a model for detecting COVID-19 based on perceived symptoms, contact history, and mobility history.

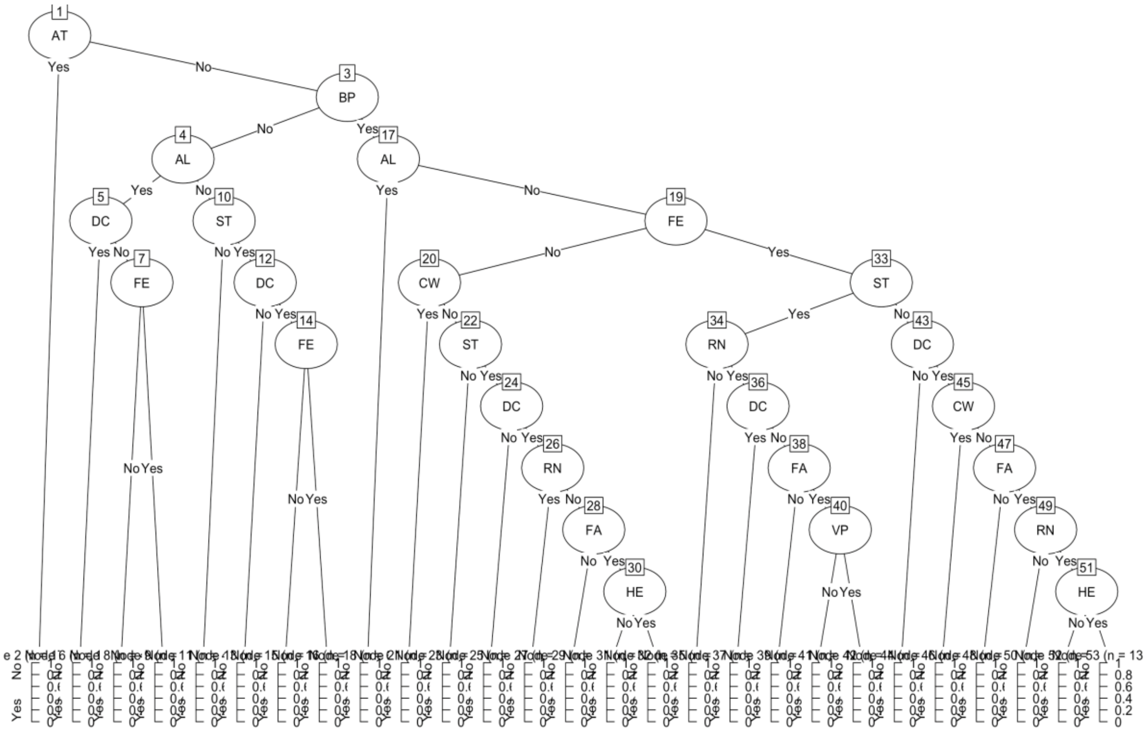
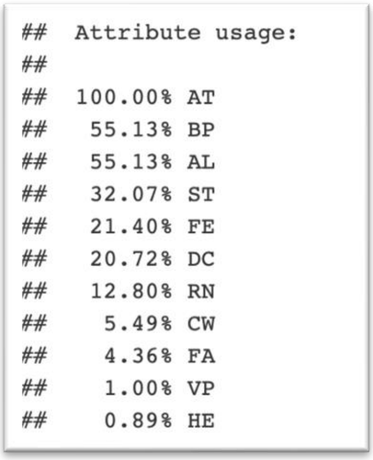


Figure 4. Decision Tree

In addition, a decision tree can also be obtained using this model as shown in Figure 4. For example, if a patient has a history of traveling abroad, the patient can be predicted to be positive for COVID-19. Opposed, if you have no history of traveling abroad, no complaints of difficulty breathing, no history of visiting large gatherings, and no complaints of dry cough, then the patient is predicted to be negative for COVID-19 and so on. Figure 5 also shows the 11 attributes that are considered the most influential in the formation of this C5.0 Algorithm decision tree. With the attribute "Travel Overseas" which is the root or the most influential attribute with 100% usage and so on. So the further down, the influence of the attribute on a model will be smaller.



```
## Attribute usage:
##
## 100.00% AT
##  55.13% BP
##  55.13% AL
##  32.07% ST
##  21.40% FE
##  20.72% DC
##  12.80% RN
##   5.49% CW
##   4.36% FA
##   1.00% VP
##   0.89% HE
```

Figure 5. Attribute Effect Order

The research results obtained are shown in Figure 5, in accordance with studies that have been carried out by several previous researchers such as the study by Huang et al (2020) of 41 hospital patients in Wuhan China, there were 40 patients with fever symptoms, 31 patients with cough symptoms, 3 patients with dizziness symptoms, and 18 patients with fatigue symptoms [12]. According to a study by Hui et al (2020) out of 41 patients who had been diagnosed with COVID-19, 20% had symptoms of difficulty breathing [13]. According to Mahase's research (2021) people infected with the new variant of COVID-19 (B.1.1.7) in the UK tend to have symptoms of fatigue and tiredness [14]. According to Iacobucci's research (2021) the top symptoms reported in the COVID-19 variant (Omicron) are runny nose, headache, fatigue, and sore throat [15].According to the National Incident Chamber Surveillance Team for COVID-19 (2020) in Australia [16], the highest rates of COVID-19 among 65-79 year olds reported that three-quarters of cases were associated with overseas travel. In Wilson et al's (2020) study, a qualitative and quantitative approach was used for behavior that affects the risk of exposure to COVID-19, one of which is the Social Gathering in Winnebago [17].

4. CONCLUSION

Based on the results of previous research, especially regarding the symptoms experienced by sufferers of COVID-19 and the results of analysis using the C5.0 Algorithm used in this study on anonymous data of 5434 people in India who were tested for COVID-19 (positive and negative for COVID-19) from The Kaggle Dataset machine learning repository uses a data separation ratio of 7 0 : 3 0. It was found that the performance evaluated using the Confusion Matrix method resulted in accuracy, precision, recall, and F1 scores of 9 8%, 93%, 97% and 95%, respectively. With these results, it can be concluded that the classification in detecting Covid-19 positivity produced by the C5.0 Algorithm is very good, so that existing patients can be predicted using this pattern to determine the results of the COVID-19 test. It is hoped that this research can simplify and speed up the performance of medical personnel so that COVID-19 patients receive prompt and appropriate treatment to help reduce COVID-19 cases in the community.

REFERENCES

[1] Hafeez, A., Ahmad, S., Siddqui, SA, Ahmad, M., & Mishra, S. (2020). A review of COVID-19 (Coronavirus Disease-2019) diagnosis, treatments and prevention. Ejmo, 4(2), 116-125.

[2] Khishe, M., Caraffini, F., & Kuhn, S. (2021). Evolving deep learning convolutional neural networks for early detection of COVID-19 in chest X-ray images. Mathematics, 9(9), 1002.

[3] Sisodia, D., & Sisodia, DS (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[4] Li, JP, Haq, AU, Din, SU, Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.

[5] Karthikeyan, T., & Thangaraju, P. (2013). Analysis of classification algorithms applied to hepatitis patients. International Journal of Computer Applications, 62(15).

[6] Ramana, BV, Babu, MSP, & Venkateswarlu, NB (2011). A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems, 3(2), 101-114.

[7] Bujlow , T., Riaz, T., & Pedersen, JM (2012, January). A method for classification of network traffic based on C5. 0 Machine Learning Algorithms. In *2012 international conference on computing, networking and communications (ICNC)* (pp. 237-241). IEEE.

[8] Pang, SL, & Gong, JZ (2009). C5. 0 classification algorithm and application on individual credit evaluation of banks. Systems Engineering-Theory & Practice, 29(12), 94-104.

[9] Kurniawan, E., Nhita, F., Aditsania, A., & Saepudin, D. (2019, July). C5. 0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in Bandung regency. In 2019 7th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.
[10] Aesyi, US, Diwangkara, TW, & Kurniawan, RT (2020). Diagnosis of Hernia Disk Disease and Spondylolisthesis Using the C5 Algorithm. Telematics: Journal of Informatics and Information Technology, 16(2), 81-86.
[11] https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence
[12] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506.
[13] Hui, DS, Azhar, EI, Madani, TA, Ntoumi, F., Kock, R., Dar, O., ... & Petersen, E. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. International journal of infectious diseases, 91, 264-266.
[14] Mahase, E. (2021). Covid-19: Sore throat, fatigue, and myalgia are more common with the new UK variant.
[15] Iacobucci, G. (2021). Covid-19: Runny nose, headache, and fatigue are the commonest symptoms of omicron, early data show.
[16] COVID-19 National Incident Room Surveillance Team. (2020). COVID-19, Australia: Epidemiology Report 19 (Fortnightly reporting period ending 21 June 2020). Communicable diseases intelligence (2018), 44.
[17] Wilson, RF, Sharma, AJ, Schluechtermann, S., Currie, DW, Mangan, J., Kaplan, B., ... & Gieryn, D. (2020). Factors influencing risk for COVID-19 exposure among young adults aged 18–23 years—Winnebago County, Wisconsin, March–July 2020. Morbidity and Mortality Weekly Report, 69(41), 1497.

BIBLIOGRAPHY OF AUTHORS

Joko Riyono is an active Mechanical Engineering Department, Faculty Of Industrial Technology, Trisakti University. He received Bachelor's Degree in Diponegoro University and Master's Degree in Gajah Mada University.

Aina Latifa Riyana Putri is an active postgraduate student from the Mathematic Department, Diponegoro University. She received Bachelor's Degree in Semarang State University

Christina Eni is an active Lecturer at the Mechanical Engineering Department, Faculty Of Industrial Technology, Trisakti University. She received Bachelor's Degree and Master's Degree in Gajah Mada University.