JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 7, No. 2, December 2024, pp. 232~238
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: 10.36378/jtos.v7i2.3999

Expert System for Lung Disease Prediction Using the C4.5 Algorithm

Tya Septiani Nurfauzia Koeswara[1], Eva Marsusanti[2], Rifa Nurafifah Syabaniyah[3], Rusli Nugraha[4], Resti Yulistria[5]
[1,2,3,4,5]Department of Information System, Bina Sarana Informatika University, Jakarta - Indonesian

Article Info

Article history:

Received 12 01, 2024
Revised 12 15, 2024
Accepted 12 30, 2024

Keywords:

C4.5 Algorithm
Expert System
Lung Disease
Prediction System
Predictive Performance

ABSTRACT

Lung diseases significantly impact public health, with smoking, age, and lifestyle being primary risk factors. This study develops an expert system using the C4.5 algorithm to predict lung diseases based on a dataset containing 30,000 records and 11 attributes. Data preprocessing included handling missing values, outlier removal, and attribute conversion. The results indicate smoking as the most influential factor, followed by employment status and insurance ownership. The model achieved an accuracy of 94.66% and an AUC score of 0.993, demonstrating excellent predictive performance. Furthermore, the system can assist healthcare professionals in early diagnosis, enabling timely intervention and improved patient outcomes. Additionally, the integration of such predictive systems may contribute to the development of more personalized healthcare strategies, ensuring that preventive measures are targeted and effective.

Corresponding Author:

Tya Septiani Nurfauzia Koeswara
Department of Information System
Bina Sarana Informatika University
Jakarta, Indonesia
Email: tya.tsf@bsi.ac.id

1. Introduction

The lungs are one of the vital organs in humans, responsible for the exchange of oxygen and carbon dioxide in the blood [1]. The occurrence of lung disease can affect the respiratory system, potentially causing short-term or long-term damage. Factors contributing to the development of lung disease include exposure to harmful substances such as active and passive cigarette consumption, which introduce toxins and carcinogens [2]. Smoking habits are prevalent in various age groups, including both teenagers and adults. This behavior often persists due to the belief that smoking provides pleasure and a sense of relaxation [3]. The number of lung cancer cases in Indonesia increased by 34,783 cases with a death rate of 30,843 people [4]. Steps that can be taken to prevent lung cancer are: by diagnosing or examining the early symptoms of the disease. This step can certainly help the public regarding lung cancer so that treatment or care can be carried out earlier prevent disease severity [5]. Expert systems are a form of human artificial intelligence learn how an expert thinks in solving a problem and make a decision [6]. Expert systems can also be utilized to diagnose diseases based on existing symptoms, enabling prompt examination and treatment. This serves as the foundation for research on the implementation of a lung disease prediction expert system using the C4.5 algorithm. By leveraging historical data and machine learning models, such systems can identify patterns and correlations, facilitating early detection and intervention. Ultimately, this contributes to better patient outcomes and more efficient healthcare management [7]

The C4.5 algorithm is one of the most widely recognized types of decision trees. Decision trees are highly useful in data analysis, as they help uncover hidden relationships between multiple input variables and the target variable [8]. The C4.5 algorithm, based on attribute selection, achieves an accuracy value of 96.81% using attribute selection methods such as Chi-squared, Information Gain, and Relief. The Criterion, Information Gain, provides the highest accuracy value. Thus, it can be concluded that the C4.5 algorithm primarily relies on attribute criteria, with Information Gain contributing the most to achieving high accuracy. By selecting the most informative attributes, the C4.5 algorithm effectively constructs a decision tree that maximizes predictive performance. This makes the C4.5 algorithm particularly suitable for applications where the relationship between input variables and the target variable needs to be clearly defined. Furthermore, the high accuracy achieved by the C4.5 algorithm ensures its effectiveness in scenarios such as disease prediction, where precise classification is crucial [9].

2. Research Method

Systems Analysis

Existing lung disease prediction systems have weaknesses in identifying risk factors that may influence disease development [10]. Through analysis, it was found that the current prediction model it is less accurate in taking into account certain risk factors By Therefore, system updates are needed to improve accuracy and prediction accuracy [11]. This research will produce a prediction regarding risk factors for lung disease using Algorithm C.45. The system analysis required is as follows[12]: 1. Review the data used for testing. ensure the data includes relevant attributes related to lung disease, such as smoking history, age, exposure to pollution, and other risk factors. 2. Review the attributes that have the most significant influence in creating prediction of lung disease. Identify attributes that are risk factors main. 3. Analyze the decision tree produced by Algorithm C4.5. 4. Review the evaluation matrix used to measure model performance, such as accuracy, precision, recall, and AUC score. Consider whether the metrics meet your needs lung disease risk analysis. 5. Analyze errors made by the model, such as classification what is wrong with positive or negative cases 6. Analyze whether the prediction results have clinical validity according to medical knowledge. 7. Make conclusions about the extent to which Algorithm C4.5 is successful in identify risk factors for lung disease.

User Needs Analysis

In the process of identifying user needs, we conduct studies literature to determine risk factors for lung disease is necessary integrated[13]. These needs include expanding the data used, increased prediction accuracy, and better management of risk factors[14].

System Design

In predicting lung disease we use algorithm C.45. This algorithm was chosen because of its ability to construct decision trees that can identify risk factors effectively[15]. The design steps involve[16]: 1. Selection of attributes Selection of attributes helps determine the most risk factors significant in predicting lung disease. The attribute selected by C4.5 algorithm provides the most useful information in separating classes target, focusing attention on the variables most relevant to lung health. Selecting the right attributes can increase accuracy and accuracy of prediction of lung disease. Truly relevant attributes and informative can help the model understand patterns better in datasets. 2. Calculation of information gain Information Gain helps the C4.5 algorithm to judge which attributes are which provides the greatest information in splitting the dataset. Attributes with Gain The highest information is selected as the separating attribute, ensuring that the model focus on the most relevant attributes for lung disease prediction. Attributes that have high Information Gain tend to be risk factors significant in predicting lung disease. Hence, the calculation Information Gain helps improve our understanding of factors that contribute to the risk of the disease. 3. Decision Tree Formation Decision trees present predictive models in a simple form understood. Decision trees help identify disease risk factors the lungs are the most significant. Each branch on the tree represents a rule decisions based on certain attributes, helping to understand the relationships between risk factors and disease conditions.

3. Result and Discussion

3.1. Dataset Review

Dataset Review The dataset that will be modeled is taken from the Kaggle website. Datasets lung disease contains 30,000 data with 11 attributes. 1 attribute of data type integer, and the other 10 are nominal data types. Attribute name and explanation can be seen in the following table.

Table 1. Lung Disease Patient Data Attributes

Attributes Name	Attributes Explanation	Data Type
No	Patient Number	Integer
Age	Age (Old or Young)	Nominal
Gender	Patient Gender (Male or Female)	Nominal
Smoker	Smoker Status (Active or Passive)	Nominal
Work	Working Status (Yes or No)	Nominal
Household	Married Status (Yes or No)	Nominal
Activity_Staying Up Late	Frequently Staying Up Late (Yes or No)	Nominal
Sports_Activities	Frequency of Exercise (Frequent or Infrequent)	Nominal
Insurance	Have Insurance (Whether or Not There Is)	Nominal
Congenital_Disease	Have Congenital Diseases (Yes or Not)	Nominal
Results	Lung Disease Diagnosis (Yes or No)	Nominal

3.2. Attributes Review

Attribute Review Based on table 1 all nominal attributes only has 2 values. Therefore, the attribute data type needs to be changed be binomial. This can be done using operators “Nominal to Binominal” in the RapidMiner application[17]. The attribute filter is set so that all attributes nominal type changed everything, because we use the C4.5 algorithm as the processing model, then we have to determine the label of the dataset[18]. To determine the role, You can use the "Set Role" operator in the RapidMiner application.

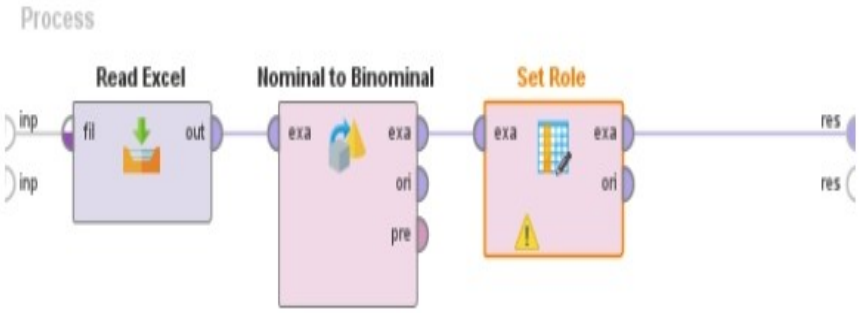

Figure 1. Process of set Nominal to Binominal and Set Role Operator in the Rapidminer Application

3.3. Modelling

The modeling algorithm that we created uses the C4.5 algorithm. Before modeling, the dataset must be split into a training dataset to train the model and testing dataset to test the model that has been created[19]. To make it more practical, the cross validation operator can be used.

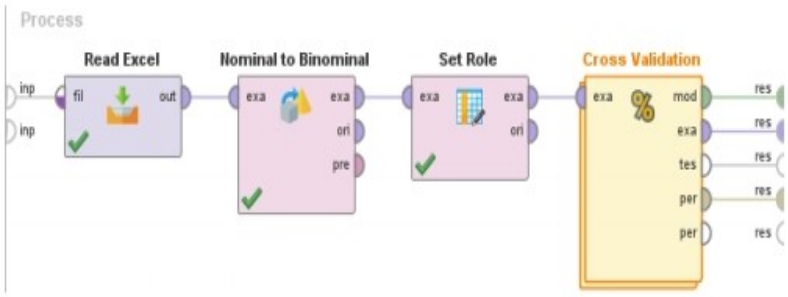

Figure 2. Used Cross Validation for Modelling with RapidMiner

This operator will automatically divide the dataset into a training dataset and testing datasets. In the model training process, the dataset will be connected with the algorithm that will be used for modeling[20].

Algorithm that used in our research is the decision tree algorithm (C4.5). In process model testing, modeling is applied to the testing dataset. Next model will be measured how accurately the performance is using the performance operator. There are 4 final results from cross validation, namely the results of the modeling created in the form of decision trees, dataset views, and model performance results.

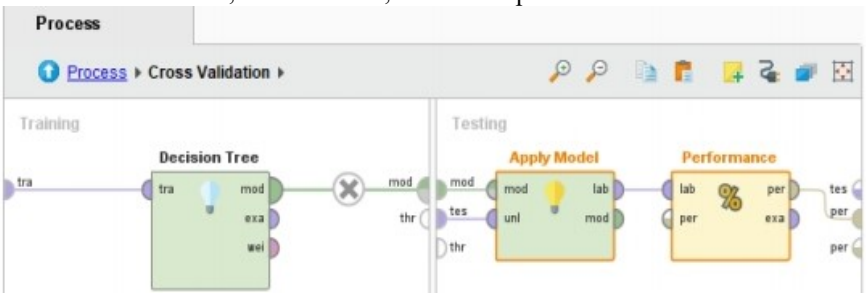

Figure 3. Process Accurately the Performance is Using Decission Tree Algorithm

These results help determine how well the model generalizes to unseen data, ensuring its reliability and effectiveness in making accurate predictions. The C4.5 algorithm, known for its high accuracy in classification tasks, plays a crucial role in this process by efficiently identifying patterns and relationships within the data. By leveraging the C4.5 algorithm, the model can extract meaningful insights from large datasets, reducing the risk of overfitting and improving the robustness of the predictions. This allows the model to provide more precise and consistent outcomes, especially in complex scenarios such as disease prediction or decision-making in healthcare applications

3.4. Research Result

Based on decision trees resulting from algorithm modeling C4.5, the Smoking attribute becomes the root of the decision tree. This proves it that the smoking attribute is the most influential predictive factor.

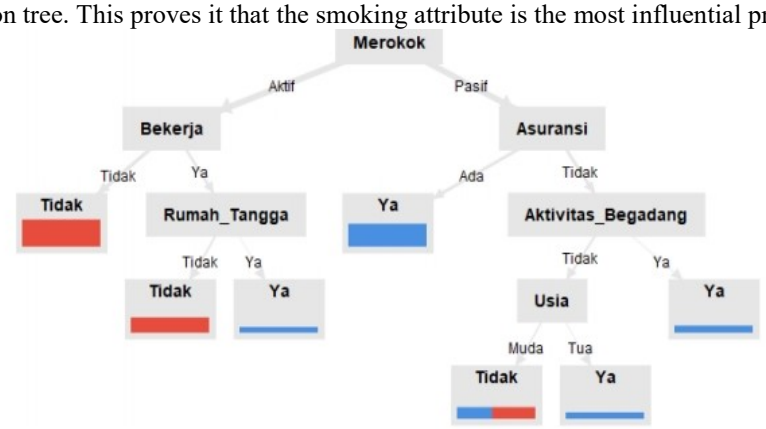

Figure 4. Result of Decision Tree Algorithm

Apart from smoking, other attributes have the most influence on predictions lung disease, namely Work and Insurance attributes. If someone is a smoker active, then the prediction of lung disease is based on work status. If an active smoker does not work, then that person is unpredictable have lung disease. But if an active smoker works, there are Another attribute that must be considered, namely the House_Household attribute. If an active smoker who works and is married, then someone is predicted to have lung disease. But if you're not married yet stairs, then the person is predicted not to suffer from lung disease. If someone becomes a passive smoker, then lung disease is predicted The result will be seen from the insurance ownership status. If one passive smokers have insurance, then that person is predicted to have it lung disease. But if a passive smoker doesn't have insurance, There are other attributes that must be considered, namely attributes Activity_Staying Up Late. If a passive smoker who does not have insurance If you often stay up late, a person is predicted to have lung disease. If you don't stay up late, you will look again at the age factor. If one passive smokers who do not have insurance and often stay up late are at an age old, then the person is predicted to have lung disease. But if being at a young age, it is predicted that they will not suffer from lung disease.

3.5. Model Evaluation

3.5.1. Confusion Matriks

Confusion matrix is an important tool in internal performance evaluation data mining classification that provides a comprehensive picture of the results model predictions [21]. In this matrix, there are main elements, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

	true Ya	true Tidak
pred. Ya	12750	0
pred. Tidak	1602	15648

Figure 5. Result of Confusion Matriks

True Positive (TP): is the number of correct guesses, where the model we guess someone has lung disease, and in fact It is true. TP is worth 12750, meaning there are 12750 successful model times correctly guessing someone who has lung disease, and in fact it is true if someone has lung disease.
True Negative (TN): is the number of correct guesses, where the model we guess someone doesn't have lung disease, and in fact it is true. TN is worth 15648, meaning there are 15648 times The model managed to correctly guess that someone was not affected by the disease lungs, and in fact it is true that a person is not affected lung disease.
False Positive (FP): is the number of guesses where our model is guessing someone has lung disease, but in reality it is wrong. In the confusion matrix, FP has a value of 0.
False Negative (FN): is the number of guesses in our model guess someone doesn't have lung disease, but the truth is That is wrong. FN has a value of 1602, meaning there are 1602 times the model guesses someone did not have lung disease, even though the results of the examination were actual have lung disease.

3.5.2. Accuracy

Model accuracy in predicting someone affected by lung disease is 94.66%. In this context, meaning the model succeeded in predicting 94.66% of the total lung disease correctly, whether affected or not affected by lung disease lung disease. So, in this case, the model was successful in predicting 28398 had lung disease (12750 had lung disease and 15648 did not affected by lung disease) correctly from a total of 30,000 existing data.

```
accuracy: 94.66% +/- 0.33% (micro average: 94.66%)
ConfusionMatrix:
True:   Ya      Tidak
Ya:     12750   0
Tidak:  1602    15648
```

Figure 6. Result of Accuracy Value in Prediction with Decision Tree Algorithm

3.5.3. Precission and ROC Curve

Model precision in predicting lung disease the actual positive is 90.72%. In this context, it means model managed to predict 90.72% of the total positive predictions correctly. So, in this case, the model successfully predicted 12750 people affected by the disease lungs correctly out of a total of 14352 positive predictions made by the model.

The ROC curve shows the visualization between true positives rate (TPR) and false positive rate (FPR)[22]. Classifier that provides curves getting closer to the top left corner (perfect classifier) shows performance which is getting better [23]. Furthermore, as a basis for evaluating model performance, a random classifier is created which gives the points located along the diagonal (FPR = TPR)[24]. The closer the curve is to the diagonal 45 degrees of ROC space, the less accurate the classifier[25].

```
precision: 90.72% +/- 0.52% (micro average: 90.71%) (positive class: Tidak)
ConfusionMatrix:
True:   Ya      Tidak
Ya:     12750   0
Tidak:  1602    15648
```

Figure 7. Result of Precission in Prediction with Decision Tree Algorithm

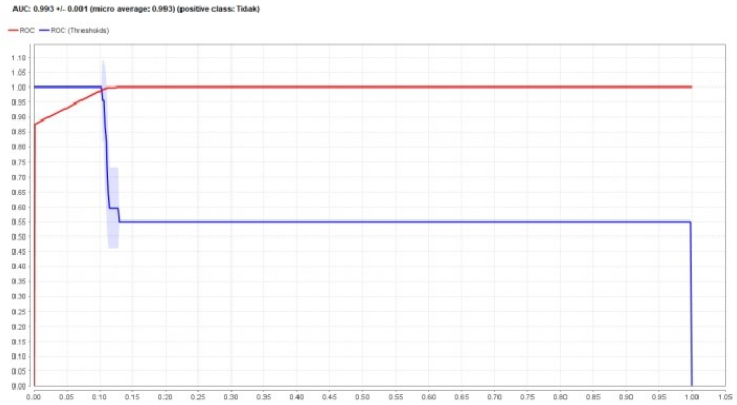

Figure 8. Result of ROC Curve in Prediction with Decision Tree Algorithm

From the ROC curve above, we can conclude that our model to provide excellent performance with AUC (Area Under the Curve) 0.993 (close to 1). The model is said to have perfect prediction accuracy if the AUC value is 1, which means 100% of the area under the curve. The ability of the model we created is expressed with a value of 0.993 meaning that the model we create has 99% area under the curve. Meanwhile, the blue curve represents random prediction where we just filled it with the value 0. This curve illustrates The ROC curve that will be generated if the predicted probability is generated is 0 or in other words the model has a confidence value of '0' when states the prediction result '1'.

4. Conclusion

Based on the findings of the research conducted, the following conclusions can be drawn: Smoking emerges as the most significant factor in predicting lung disease. Additionally, factors such as employment status, insurance ownership, living situation, sleep patterns, and age also have a considerable influence on lung disease. The model achieved an accuracy of 94.66% and an AUC value of 0.993, indicating excellent model performance. These results highlight the importance of considering multiple variables in lung disease prediction, which can contribute to more effective prevention and early diagnosis strategies.

Acknowledgement

We extend our sincere gratitude to all those who contributed to the completion of this research. Our heartfelt appreciation goes to our institution, Bina Sarana Informatika University, for providing the facilities and support necessary for this study. We also thank our colleagues and peers for their valuable insights and constructive feedback throughout the research process. Lastly, we express our gratitude to our families and loved ones for their unwavering support and encouragement.

References

[1] Yopento, J., Ernawati, & Coastera, F. F. "Identifikasi Pneumonia Pada Citra X-Ray Paru-Paru Menggunakan Metode Convolutional Neural Network (CNN) Berdasarkan Ekstraksi Fitur Sobel". Rekursif: Jurnal Informatika, 10(1), 40–47. 2022

[2] Meila Azzahra Sofyan, F., Voutama, A., & Umaidah, Y. "Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer". JATI (Jurnal Mahasiswa Teknik Informatika), 7(2). 2022.

[3] Goldman, Ian. and Pabari, M. “Gambaran pengetahuan siswa tentang bahaya merokok". 3, 71–77. 2023

[4] Buchori, A., Khotijah, S., & Ramdan, A. S. "Penerapan Algoritma C45 untuk Deteksi Penyakit Hepatitis". Jurnal Larik, ISSN: 5645-10629, Hal. 1–7. 2022

[5] Liana, C. F., & Sinaga, B. "Sistem Pakar Diagnosa Penyakit Dyslexia Pada Anak Dengan Metode Naive Bayes Berbasis Web". Jurnal Ilmu Komputer Dan Bisnis, 12(2a), 173–183. 2021

[6] Hutasuhut, M., Ginting, E. F., & Nofriansyah, D. "Sistem Pakar Mendiagnosa Penyakit Osteochondroma Dengan Metode Certainty Factor". JURIKOM (Jurnal Riset Komputer), 9(5), 2022.

[7] Ramdhani, Lis Saumi., "Sistem Pakar Diagnosis Penyakit Dalam". Jurnal Swabumi, 2(2), 45-53. 2022

[8] Girsang, R., Ginting, E. F., & Hutasuhut, M. "Penerapan Algoritma C4.5 Pada Penentuan Penerima Program Bantuan Pemerintah Daerah". Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), 1(4), 449. 2022

[9] Hoiriyah, H. "Algoritma C4.5 Berbasis Seleksi Atribut Untuk Menentukan Kemungkinan Pengunduran Diri Mahasiswa". Technologia: Jurnal Ilmiah, 9(1), 67. 2018

[10] Abas, W. "Analisa Kepuasan Mahasiswa Terhadap Website Universitas Negeri Yogyakarta (UNY)". Publikasi Ilmiah Unwahas, 1–6. 2021

[11] Dhika, H., Isnain, N., & Tofan, M. "Manajemen Villa Menggunakan Java Netbeans Dan Mysql" IKRA-ITH INFORMATIKA : Jurnal Komputer Dan Informatika, 3(2), 104–110. 2021
[12] Ganda Anggara, Gede Pramayu, A. W. "Membangun sistem pakar menggunakan teorema bayes untuk mendiagnosa penyakit paru-paru". Seminar Nasional Teknologi Informasi Dan Multimedia, 79–84. 2016
[13] Mauli, D. "Tanggung Jawab Hukum Dokter Terhadap Kesalahan Diagnosis Penyakit Kepada Pasien". Cepalo, 2(1), 33. 2019
[14] Rahmawati, E. "Sistem Pakar Diagnosis Penyakit Paru-Paru Menggunakan Metode Forward Chaining". Jurnal Teknik Elektro, 8(2), 64–69. 2019
[15] Rikhiana, E. D., & Fadlil, A. "Penyakit Dalam Pada Manusia Menggunakan Metode". Jurnal Sarjana Teknik Informatika, 1(1), 1–10. 2016
[16] Ritonga, E. R., & Irawan, M. D. "Sistem Pakar Diagnosa Penyakit ParuParu". Journal Of Computer Engineering, System And Science, 2(1), 39–47. 2017
[17] Rizqifaluthi, H., & Yaqin, M. A. "Process Mining Akademik Sekolah Menggunakan RapidMiner". Matics, 10(2), 47. 2019
[18] Romadhon, M. H., Yudhistira, Y., & Mukrodin, M. "Sistem Informasi Rental Mobil Berbsasis Android Dan Website Menggunakan Framework Codeigniter 3 Studi Kasus : CV Kopja Mandiri". Jurnal Sistem Informasi dan Teknologi Peradaban (JSITP), 2(1), 30–36. 2020
[19] Saputro, A. H. "Membuat Kurva dalam Python". Jakarta: PT. Elex Media Komputindo. 2021
[20] Soetarmono, A. N. D. "Perancangan Sistem Pakar Dalam Mendiagnosa Penyakit Pada Balita". Teknika, 2(1), 28–39. 2018
[21] Tobin, M. J. Asthma, "Airway Biology, and Nasal Disorders". AJRCCM: American Journal of Respiratory and Critical Care Medicine, 169(2), 265–276. 2014
[22] Trimarsiah, Y., & Arafat, M. "Analisis Dan Perancangan Website Sebagai Sarana Informasi Pada Lembaga Bahasa Kewirausahaan Dan Komputer Akmi Baturaja". Jurnal Ilmiah Matrik, 19, 1–10. 2019
[23] Yulianti, Ita., Rizal Amegia Saputra., Ami Rahmawati. "Sistem Diagnosis Penyakit Hepatitis Menggunakan Algoritma C45". Jurnal Swabumi, 2(1), 76-89. 2019
[24] Zahedi, B., Nahid-Mobarakeh, B., Pierfederici, S., & Norum, L. E. "A robust active stabilization technique for dc microgrids with tightly controlled loads". Proceedings - 2016 IEEE International Power Electronics and Motion Control Conference, PEMC 2016, VI(1), 254–269. 2019
[25] Sutarman, Jaya, R.H Abigail. "Expert System of Diagnosis". Jurnal Tematik Vol. 1, No. 2 Desember 2018.