Pneumonia Prediction System Using Classification and Regression Trees Algorithm

Taufik Ramadhan
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
tararmd123@gmail.com

Agung Wahana
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
wahana.agung@uinsgd.ac.id
https://orcid.org/0000-0002-6468-0741

Dian Sa'adillah Maylawati
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
diansm@uinsgd.ac.id
https://orcid.org/0000-0002-1193-3370

Nur Lukman
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
n.lukman@uinsgd.ac.id
https://orcid.org/0000-0003-2674-6460

Ichsan Taufik
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
ichsan@uinsgd.ac.id
https://orcid.org/0000-0001-5052-0635

Ichsan Budiman
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
ichsanbudiman@uinsgd.ac.id

***Abstract*— Pneumonia is an inflammation or chronic infection of lung tissue caused by various microorganisms, including parasites, viruses, and bacteria, as well as physical damage to the lungs and exposure to chemicals. The method used to calculate predictions uses the CART (Classification and Regression Trees) algorithm. The model is then implemented into a website-based prediction system. The purpose of this study was to determine the implementation of the CART algorithm to determine pneumonia and to determine the accuracy of the CART algorithm in predicting pneumonia. The average accuracy of the results of this study led to an accuracy value of 94%, r-square 87%, precision 95%, recall 94%, and f-1 score 94% of the total dataset of 283. The results of this study get the best r-square on the 5th test with an accuracy of 85%.**

Keywords*— CART, classification and regression tree, machine learning, pneumonia, prediction system.

I. Introduction

Pneumonia is an inflammation or chronic infection of the lung tissue caused by various microorganisms, including parasites, viruses, bacteria, physical damage to the lungs, or exposure to chemicals. This pneumonia disease can attack children, adolescents, and adults, but cases in toddlers and the elderly are more common. Pneumonia affects more than 450 million people every year and is often found in developing countries. According to *Riset Kesehatan Dasar* (Riskesdas, Basic Health Research) data in 2018, the prevalence of pneumonia based on the diagnosis of health workers was around 2%, while in 2013, it was 1.8% [1]. In addition, according to data from the Ministry of Health in 2014, the number of pneumonia sufferers in Indonesia in 2013 ranged from 23%-27%, and deaths from pneumonia were 1.19%. In 2010 in Indonesia, pneumonia was included in the top 10 hospitalizations with a CFR (crude mortality rate) or a specific mortality rate at a particular time period in which the number of cases was 7.6%. According to the Indonesian Health Profile, pneumonia causes 15% of under-five deaths, which is around 922,000 children under five in 2015. From 2015 to 2018, confirmed cases of pneumonia in children under five years old increased by about 500,000 per year, reaching 505,331 patients, with 425 patients dying.

Jakarta Health Office estimates 43,309 cases of pneumonia in children under five during 2019 [2]. Symptoms caused by this disease are coughing or difficulty breathing, such as rapid breathing and pulling in the chest wall. It is essential to early detection of pneumonia so that we can overcome and prevent this disease.

Several studies have been conducted using the CART algorithm before, such as that studied by Pungkas Subarkah, who compared the performance of the CART and Naïve Bayes algorithms which obtained CART results 76.93% higher than the results of the Naïve Bayes algorithm 73.75% [3]. Research conducted by Ulfa Khaira using the CART algorithm has an accuracy of 84% [4]. Moreover, research conducted by Ria Dhea Layla Nur using the CART algorithm has an accuracy of 92.9% [5]. Another research used CART to predict cervical cancer [6], to classify malaria complication [7], to predict self-efficacy and risk of persistent shoulder pain [8], and to predict coronary artery disease [9] that proven can perform well.

Based on the description above, it can be seen that the CART algorithm has classification results that have good accuracy compared to other algorithms. The CART algorithm has several advantages, including it being easier to interpret and the calculations being faster and more accurate. The CART algorithm is an algorithm that can be applied to large amounts of data with many variables and through binary selection procedures [10]. Based on this description, this study aims to predict pneumonia by using the CART algorithm in an effort to detect the disease early.

II. Research Method

This study uses the CART (Classification and Regression Tree) algorithm to predict Pneumonia. CART is a method or algorithm of the decision tree methodology, which is one of the data exploration strategies [11], [12]. The CART algorithm performs classification, referring to grouping with a binary decision tree model that is illustrated in Figure. In Figure 2 is the CART algorithm used in building a decision tree model to predict Pneumonia disease.

978-1-6654-5122-2/22/$31.00 ©2022 IEEE

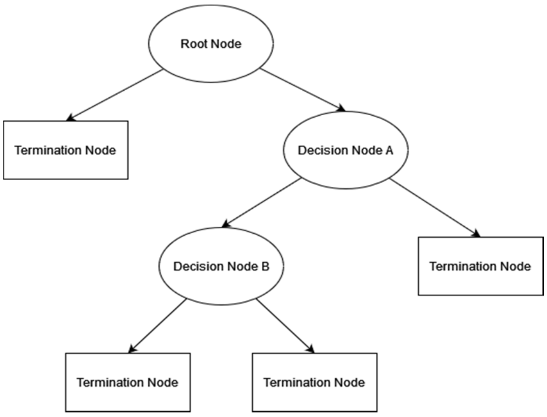


Fig. 1. Decision Tree Model

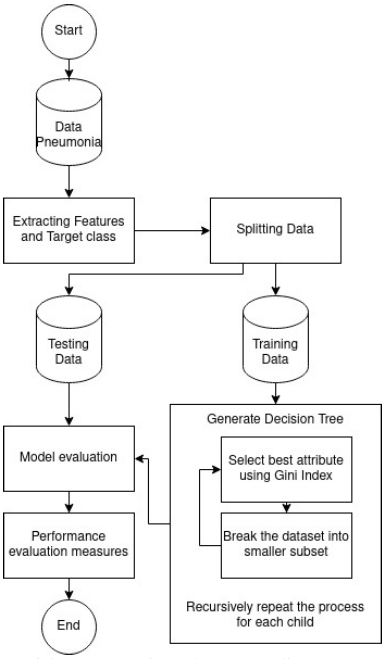

Start
Data Pneumonia
Extracting Features and Target class
Splitting Data
Testing Data
Training Data
Generate Decision Tree
Select best attribute using Gini Index
Break the dataset into smaller subset
Recursively repeat the process for each child
Model evaluation
Performance evaluation measures
End

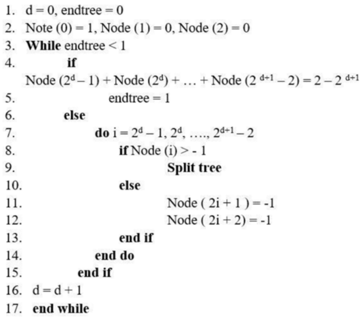
Fig. 2. Flowchart of the CART algorithm to predict pneumonia disease

In this study, using data obtained from the Limbangan Community Health Center, with the data collected is medical record data for pneumonia patients in 2019 with a total of 283 patients. In predicting pneumonia disease, symptom determinants are used, namely coughing with phlegm, body heat or fever, lack of appetite, weakness and weakness, respiratory frequency 18-20 times/minute, severe shortness of breath, cyanosis (bluing), chest wall indrawing, breathing nostrils and respiratory rate 24-30 breaths/minute. Prediction results will be categorized into three categories, namely no pneumonia, mild pneumonia, and severe pneumonia.

This study began by collecting data on pneumonia patients, which was then continued with the data selection stage. At this stage, several variables were intuitively selected from the pneumonia patient data to be used for two things, namely: variables to predict and predictive target variables. The symptom variable is selected to be used to make predictions, and the predictive target variable will use the status variable. For other variables such as village variables, population, pneumonia estimates, age, and gender were not used.

The next stage is data preparation. This study used 283 pneumonia patient data obtained from the UPT Puskesmas (Community Health Center, Integrated Service Unit) Limbangan. Of the 283 data from Limbangan residents, 75% of the data will be used as training data, and 25% will be used as test data. So, the amount of the division is 212 data as training data and 71 data as test data. The training data serves to form a decision tree, while the test data is data to test the built model.

Next is modeling activity. To make the model using the Classification and Regression Trees algorithm. From previous studies, this algorithm was able to obtain good accuracy. Therefore, this algorithm was chosen to predict pneumonia. Figure 3 presents the pseudocode of the CART algorithm to illustrate the decision tree. The classification process that has been carried out to predict pneumonia will be evaluated using a confusion matrix. The data in this confusion matrix is test data, totaling 71 data with 21 decision tree rules that have been formed.



```
d = 0, endtree = 0
Note (0) = 1, Node (1) = 0, Node (2) = 0
While endtree < 1
    if
    Node ($2^d - 1$) + Node ($2^d$) + … + Node ($2^{d+1} - 2$) = $2 - 2^{d+1}$
        endtree = 1
    else
        do i = $2^d - 1, 2^d, \dots, 2^{d+1} - 2$
            if Node (i) > - 1
                Split tree
            else
                Node ( 2i + 1 ) = -1
                Node ( 2i + 2) = -1
            end if
        end do
    end if
d = d + 1
end while
```

Fig. 3. Pseudocode CART algorithm

III. Result and Discussion

Based on the CART algorithm, the classification of data sets as candidates for left and right branches are determined. The data consists of several symptoms of pneumonia, namely coughing up phlegm, body heat or fever, decreased appetite, weak body, respiratory frequency 18-20 times/minute, severe shortness of breath, cyanosis (bluish), chest wall traction, nostril breathing, respiratory frequency 24 - 30 times/minute. Table I is the result of dividing the data for each candidate for the left and right branches. There are three labels: no symptoms, symptoms (low level), and severe symptoms (high level).

TABLE I. Result of Dividing the Data for Each Candidate for The Left and Right Branches

No.	Left & Right Branch Candidate
1.	**L**: Cough with phlegm **R**: Cough with phlegm
2.	**L**: Body heat or fever **R:** Body heat or fever
3.	**L**: Decreased appetite **R**: Decreased appetite
4.	**L**: Weak body **R**: Weak body
5.	**L**: Respiratory rate 18 – 20 x / minute **R**: Respiratory rate 18 – 20 x / minute
6.	**L**: Severe shortness of breath **R**: Severe shortness of breath
7.	**L**: Cyanosis (bluish) **R**: Cyanosis (bluish)
8.	**L**: Chest wall traction **R**: Chest wall traction
9.	**L**: Nostril breathing **R**: Nostril breathing
10.	**L**: Respiratory rate 24 - 30 x / minute **R**: Respiratory rate 24 - 30 x / minute

To form a decision tree based on each predetermined branch candidate can be calculated using the equation (1), in which $\Phi(s|t)$ in prediction, t_L is candidate left branch of the decision node, t_R is candidate right branch of the decision node, P_L is a number of data records on candidate left branch t_L divided by total number of data records, and P_R is a number of data records on candidate left branch t_R divided by total number of data records.

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{n\ category} |P(j|t_L) - P(j|t_R)| \qquad (1)$$

The overall data used in this study were 283 data on patients with pneumonia symptoms. From this data, 75% was taken to be used as training data. The number of distributions is 212 data as training data. The total symptoms of cough with phlegm are 85, for the training data 212, so that the P_L = 85/212 or P_L = 0.4009. The total did not get cough with phlegm 127, for the training data 212, so that the P_R = 127/212 or P_R = 0.5990. Furthermore, it is calculated that $P(j|t_L)$ for the unaffected status is 7 out of 85, so that $P(j|t_L)$ = 7/85. For mild status there are 39 out of 85, so $P(j|t_L)$ = 39/85. For weight status there are 39 out of 85, so $P(j|t_L)$ = 39/85. Then calculated $P(j|t_R)$ for the unaffected status, there are 51 out of 127, so $P(j|t_R)$ = 51/127. For light status there are 40 out of 127, so $P(j|t_R)$ = 40/127. For weight status there are 36 out of 127, so $P(j|t_R)$ = 36/127. From equation 3.1 we get $\Phi(s|t)$ = | 0.0823-0.4015 | + | 0.4588 – 0.3149 | + | 0.4588 – 0.2834 | = 0.6385. From equation (1), the value $\Phi(s|t)$ = 2 * 0.4009 * 0.5990 * 0.6385 = 0.3066. Calculations were performed for all data sets. Based on the results of calculations on the entire data, the largest value obtained is candidate no. 8, namely "Chest wall traction", so the decision tree of iteration 1 CART algorithm is shown in Figure 4.

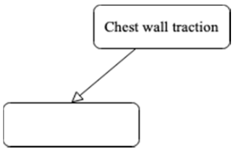


Fig. 4. First iteration decision tree

The total symptoms of cough with phlegm are 60, while the total data is now 150, so that P_L = 60/150 or P_L = 0.4. The total was not exposed to cough with phlegm 90, for the training data was 150 so that the P_R = 90/150 or P_R = 0.6. Furthermore, it is calculated that $P(j|t_L)$ for the unaffected status is 7 out of 60, so $P(j|t_L)$ = 7/60. For mild status there are 39 out of 60, so $P(j|t_L)$ = 39/60. For weight status there are 14 out of 60, so $P(j|t_L)$ = 14/60. Then calculated $P(j|t_R)$ for the unaffected status, there are 44 out of 90, so $P(j|t_R)$ = 44/90. For mild states there are 40 out of 90, so $P(j|t_R)$ = 40/90. For weight status there are 6 out of 90, so $P(j|t_R)$ = 6/90. From equation (1) we get $\Phi(s|t)$ = | 0.117-0.49 | + | 0.65 – 0.4444 | + | 0.2333 – 0.07 | = 0.7419. From equation 3.2, the value $\Phi(s|t)$ = 2 * 0.4 * 0.6 * 0.7419 = 0.3561. The results of calculations are carried out for all 150 data sets. From the results of these calculations, the largest value is candidate no. 10, namely "Respiratory rate 24 - 30 x / minute", so that the decision tree of the 2nd iteration CART algorithm is formed as shown in Figure 5. If all branch value calculations are completed, then the CART algorithm tree is formed, as shown in Figure 6. Left branch is “No” and right branch is “Yes”.

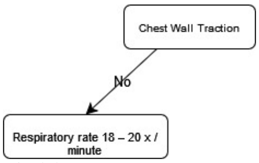


Fig. 5. Second iteration decision tree

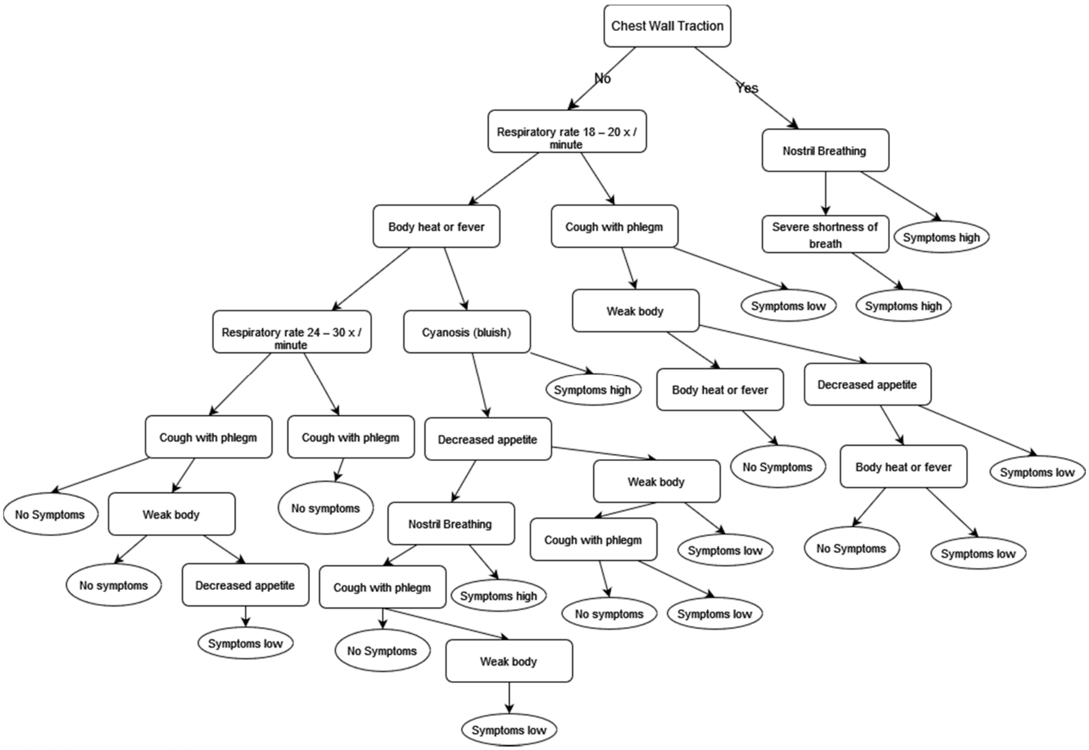
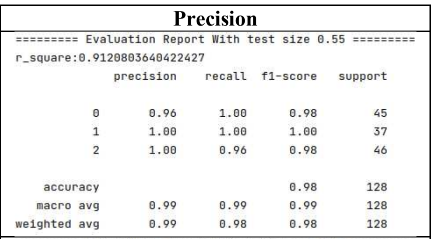


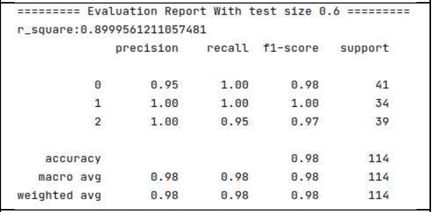
Fig. 6. Complete decision tree model

The next stage is evaluation, the evaluation process uses the confusion matrix method, namely the table used to determine the performance of a classification model. The data in this confusion matrix is test data, totaling 71 data with 21 decision tree rules that have been formed. The test was carried out five times with a comparison of the first test data to the fifth test, as follows [55:45, 60:40, 80:20, 70:30, 75:25]. Based on the results of the tests carried out, the following results were obtained. Table II presents the result of experiments with splitting data scenarios. Then, Table III provides summary of experiment result using confusion matrix.

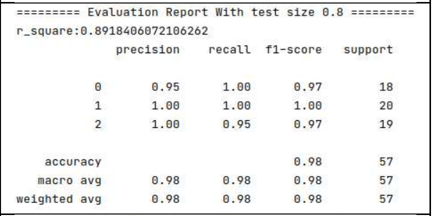
TABLE II. SPLIT TEST RESULTS

<table>
<tr><th>Split Data</th><th>Precision</th></tr>
<tr><td>55:45</td><td>
<pre>
========= Evaluation Report With test size 0.55 =========
r_square:0.9120803640422427
              precision   recall f1-score   support

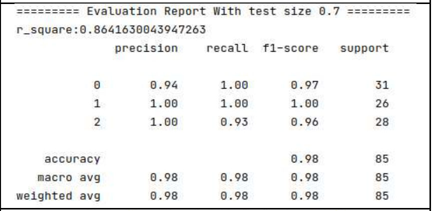
           0   0.96   1.00   0.98   45
           1   1.00   1.00   1.00   37
           2   1.00   0.96   0.98   46

    accuracy   0.98   128
   macro avg   0.99   0.99   0.99   128
weighted avg   0.99   0.98   0.98   128
</pre></td></tr>
<tr><td>60:40</td><td>
<pre>
========= Evaluation Report With test size 0.6 =========
r_square:0.8999561211057481
              precision   recall f1-score   support

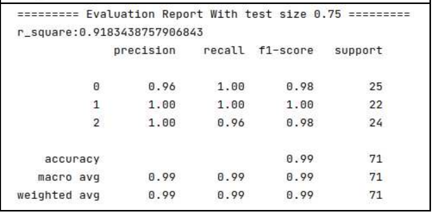
           0   0.95   1.00   0.98   41
           1   1.00   1.00   1.00   34
           2   1.00   0.95   0.97   39

    accuracy   0.98   114
   macro avg   0.98   0.98   0.98   114
weighted avg   0.98   0.98   0.98   114
</pre></td></tr>
<tr><td>80:20</td><td>
<pre>
========= Evaluation Report With test size 0.8 =========
r_square:0.8918406072106262
              precision   recall f1-score   support

           0   0.95   1.00   0.97   18
           1   1.00   1.00   1.00   20
           2   1.00   0.95   0.97   19

    accuracy   0.98   57
   macro avg   0.98   0.98   0.98   57
weighted avg   0.98   0.98   0.98   57
</pre></td></tr>
<tr><td>70:30</td><td>
<pre>
========= Evaluation Report With test size 0.7 =========
r_square:0.8641630043947263
              precision   recall f1-score   support

           0   0.94   1.00   0.97   31
           1   1.00   1.00   1.00   26
           2   1.00   0.93   0.96   28

    accuracy   0.98   85
   macro avg   0.98   0.98   0.98   85
weighted avg   0.98   0.98   0.98   85
</pre></td></tr>
<tr><td>75:25</td><td>
<pre>
========= Evaluation Report With test size 0.75 =========
r_square:0.9183438757906843
              precision   recall f1-score   support

           0   0.96   1.00   0.98   25
           1   1.00   1.00   1.00   22
           2   1.00   0.96   0.98   24

    accuracy   0.99   71
   macro avg   0.99   0.99   0.99   71
weighted avg   0.99   0.99   0.99   71
</pre></td></tr>
</table>

TABLE III. SUMMARY OF CONFUSION MATRIX TEST RESULTS

Split Data	Precision	Recall	Accuracy	R-Square
55:45	0.99	0.99	0.98	0.91
60:40	0.98	0.98	0.98	0.89
80:20	0.98	0.98	0.98	0.89
70:30	0.98	0.98	0.98	0.86
75:25	0.99	0.99	0.99	0.92
Average	**0.984**	**0.984**	**0.982**	**0.89**

CONCLUSION

Based on the final project research entitled "Implementation of the CART (Classification and Regression Trees) Algorithm to Predict Pneumonia Disease" it can be concluded that the implementation of the CART (Classification and Regression Trees) algorithm as a classification method in the system in predicting pneumonia has been applied and can predict outcomes into 3 classes, namely no symptom, low-level symptom and high-level symptom. The best model is using 75% training data and 25% testing data with an r-square accuracy value of 92%. The accuracy of the CART (Classification and Regression Trees) algorithm in predicting pneumonia has an average accuracy value of 98.2%, r-square 86%, precision 95%, recall 98.4% from all experiment scenarios. For further research, it is necessary to try other methods or algorithms that can support the level of accuracy of the prediction results, add training data that must be reproduced again so that the accuracy of predicting the recovery of pneumonia patients can increase, and more data are needed that represent all the possibility of pneumonia cases so that the learning system is better.

ACKNOWLEDGMENT

We would like to thank the Center for Research and Publications, Institute for Research and Community Service at UIN Sunan Gunung Djati Bandung for supporting and financing this publication.

REFERENCES

[1] N. Wati, O. Oktarianita, A. Ramon, H. Husin, and J. Harsismanto, "Determinants of the Incident of Pneumonia in Toddlers in Bengkulu City in 2020," *KEMAS J. Kesehat. Masy.*, vol. 17, no. 2, pp. 180–186, 2021.

[2] Perhimpunan Dokter Paru Indonesia (PDPI), "Press Release Perhimpunan Dokter Paru Indonesia (PDPI): Outbreak Pneumonia di Tiongkok," Jakarta, 2019. [Online]. Available: https://infeksiemerging.kemkes.go.id/download/Press_Release_Outbreak_pneumonia_Pneumonia_Wuhan-17_Jan_2020.pdf.

[3] P. Subarkah, "Perbandingan Kinerja Algoritma CART dan Naive Bayesian untuk Mendiagnosis Penyakit Diabetes Mellitus," 2016.

[4] U. Khaira, "Prediksi Tingkat Fertilitas Pria Dengan Algoritma Pohon Keputusan Cart," *Jakiyah J. Ilm. Umum dan Kesehat. Aisyiyah*, vol. 5, no. 1, pp. 35–42, 2020.

[5] R. D. L. N. Karisma and B. W. Otok, "Model Machine Learning CART Diabetes Melitus," in *Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika Dan Nilai-Nilai Islami)*, 2017, vol. 1, no. 1, pp. 485–491.

[6] T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," *Creat. Inf. Technol. J.*, vol. 4, no. 2, pp. 83–93, 2018.

[7] R. Irmanita, S. S. Prasetiyowati, and Y. Sibaroni, "Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naive Bayes," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 5, no. 1, pp. 10–16, 2021.

[8] R. Chester, M. Khondoker, L. Shepstone, J. S. Lewis, and C. Jerosch-Herold, "Self-efficacy and risk of persistent shoulder pain: results of a Classification and Regression Tree (CART) analysis," *Br. J. Sports Med.*, vol. 53, no. 13, pp. 825–834, 2019.

[9] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 366–374, 2008.

[10] F. E. Pratiwi and I. Zain, "Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara," *J. Sains dan Seni ITS*, vol. 3, no. 1, pp. D54–D59, 2014.

[11] W. Y. Loh, "Classification and regression trees," *Wiley Interdiscip. Rev. data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 14–23, 2011.

[12] R. Timofeev, "Classification and regression trees (CART) theory and applications," *Humboldt Univ. Berlin*, vol. 54, 2004.