e-ISSN: 2614-8404
p-ISSN:2776-3234

Decision Tree for Determining Hospital Treatment for Covid-19 Patients Based on Hematology Parameters Using the C5.0 Algorithm

Joko Riyono[1], Christina Eni Pujiastuti[2], Supriyadi[3], Dody Prayitno[4], Aina Latifa Riyana Putri[5]
[1,2,3,4]Teknik Mesin, Fakultas Teknologi Industri, Universitas Trisakti
[5]Sains Data, Fakultas Informatika, Telkom University
Email: jokoriyono@trisakti.ac.id[1],christina.eni@trisakti.ac.id[2], supri@trisakti.ac.id[3], dodyprayitno@trisakti.ac.id[4], ainaqp@telkomuniversity.ac.id[5]

Abstract – The rapid spread of the COVID-19 disease, which occurred globally from late 2019 to the early 2020s, significantly impacted communities worldwide, requires early detection of COVID-19 which is very important for patients and also the people around them to be able to fight the COVID-19 pandemic. Therefore, a classification analysis will be carried out to make decisions regarding determining COVID-19 patients who do not require hospitalization or who require Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in hospitals based on hematology parameters from the Machine Learning Repository. Kaggle Dataset uses the C5.0 algorithm assisted by Rstudio software. It is also known that because the data contains missing data, it is also necessary to handle missing data using the Mean Method assisted by SPSS software. Performance evaluated using the Confusion Matrix method produces an accuracy value of 78% which is considered quite good, where testing with the C5.0 Algorithm uses a training and testing data ratio of 40:60. This research simplifies and speeds up medical decision-making, improving patient management. With COVID-19 declining, the method can be applied to enhance healthcare systems' accuracy and efficiency in handling other diseases or emergencies, ensuring better preparedness for future challenges.

Keywords – Early Detection, Classification, Missing Data, Confusion Matrix

I. INTRODUCTION

Coronavirus is part of a group of viruses that generally infect animals but can adapt and eventually transmit to humans. In humans, this virus typically attacks the respiratory system, ranging from mild symptoms like the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS), which emerged in 2012, and Severe Acute Respiratory Syndrome (SARS) in 2002. The latest type of coronavirus was discovered in humans in Wuhan, China, in December 2019. This virus is named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and causes the disease COVID-19. Due to its rapid global spread, the World Health Organization (WHO) declared this disease a pandemic on March 11, 2020.Clinical symptoms of COVID-19 infection can vary from asymptomatic (no symptoms) to fever, cough, runny nose, fatigue, sore throat, and severe conditions (e.g., acute respiratory distress syndrome [ARDS], acute heart injury, and kidney injury) [1]. As of April 19, 2022, the number of COVID-19 cases globally reached 504,571,336. Preventive measures have also been implemented by governments and WHO to help reduce COVID-19 cases, such as requiring activities to be accompanied by strict health protocols as part of daily life, travel restrictions between countries and cities, requiring vaccination, and RT-PCR (Real Time PCR COVID-19) tests as tools to detect the presence of the COVID-19 virus in the body. Furthermore, healthcare systems have been established to detect, test, isolate, treat each case, and track every contact. Preventive measures play a major role in reducing COVID-19 cases when protocol therapy is applied from the early stages [2]. Therefore, early detection of COVID-19 is crucial for patients and those around them to help prevent a resurgence of the pandemic. When patients receive timely and appropriate care, those around them are also protected.

COVID-19 is a systemic infection that significantly impacts the hematopoietic and hemostasis systems, leading to several cardiovascular complications [1], [3], [4]. Research indicates [1] that hematological indices are associated with disease severity and can contribute to decision-making for predicting whether a COVID-19 patient will require ICU admission upon hospital entry. Several hematological abnormalities have also been identified, with significant changes in hematological parameters observed in patients with severe COVID-19 requiring hospitalization and ICU care [5]. The hematological consequences of COVID-19 infection must be utilized by researchers and medical personnel to initiate new treatment approaches or breakthroughs in managing COVID-19 infection.

Therefore, a classification analysis will be conducted to make decisions regarding which COVID-19 patients do not require hospitalization or need care in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital, using a Kaggle Dataset and based on hematological parameters and the C5.0 algorithm. Since the dataset contains missing data, a method to handle missing data using the Mean Method will also be applied. The goal is to facilitate and accelerate the work of medical

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

personnel, ensuring that COVID-19 patients receive timely and appropriate treatment, ultimately reducing COVID-19 cases in a population.

A. C5.0 Algorithm

The C5.0 algorithm is a data mining method that operates using a decision tree structure. This algorithm is a further development of the ID3 and C4.5 algorithms, with improved efficiency and better capability in categorizing data into appropriate groups. C5.0 is known for its superior performance in solving classification problems [6]. This algorithm has already been applied in various ways [7], such as to analyze the perceived stress levels of healthcare workers treating COVID-19 patients during the early stages of the pandemic in northeastern Mexico. The aim was to understand and categorize their stress levels, producing a visualization model to help identify stress risks and potential mental health issues; [8] as a decision-making tool and comparing performance in breast cancer diagnosis using C5.0 Algorithm and Boosting method.

The Mean method is the most common imputation technique, where missing data in a variable is replaced with the average value of all available data for that variable [9].

A decision tree is a structure resembling a flow diagram shaped like a tree, where each internal node represents a test on an attribute, each branch indicates the test result, and each leaf node represents a class or class distribution that is the final outcome of the test. The C5.0 algorithm is an enhancement of earlier decision tree algorithms developed by Ross Quinlan in 1987, specifically ID3 and C4.5. ID3 evolved into C4.5, which can handle both discrete and continuous attributes. C4.5 was further developed into C5.0 to address certain weaknesses, such as overlapping when handling large amounts of data, which increases decision-making time. C5.0 offers higher accuracy, faster decision-making, and more efficient memory usage compared to its predecessors.

The tree-building process in the C5.0 algorithm is similar to that of the C4.5 algorithm. However, while the C4.5 algorithm stops after calculating information gain, the C5.0 algorithm continues by calculating the gain ratio using the obtained information gain and entropy. Therefore, the calculations in the C5.0 algorithm involve several attributes, including entropy, information gain, and gain ratio. The C5.0 algorithm can select attributes based on the highest gain ratio.

The equation for calculating entropy is:

$$Entropy(\mathrm{S})=\sum_{j=1}^{k} -p_j \, log_2(p_j) \qquad (1)$$

Where:

S= Set of cases

k = Number of partitions of S

p_j = Propotions of S_j to S

Next, to obtain the Information Gain calculation, the following equation is used:

$$\text{Information Gain(S, A)} = \text{Entropy(S)-}\sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i) \qquad (2)$$

Where:

S = Set of cases

A = Attribute

m = Number of categories in variable A

|Si| = Number of cases in partition i

|S| = Number of cases in S

Finally, to determine an attribute as a node in the C5.0 algorithm, the Gain Ratio is calculated using the formula:

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

Where:

Gain(S,A) = Gain value of a variable

Si = Entropy value in a variable

The Gain Ratio calculation simplifies the decision tree produced by C5.0 compared to the C4.5 algorithm. The tree is built continuously until no further sample subsets can be split.

B. Mean Method for Handling Missing Data

Missing data is a frequent issue in most research studies, typically arising from non-sampling errors. These errors can include:

Interviewer recording errors, where questions may be skipped during data collection.

Respondent inability errors, where participants fail to provide accurate responses due to misunderstanding the question, experiencing fatigue, or losing interest.

Respondent unwillingness errors, where individuals choose not to answer sensitive questions related to topics such as income, age, weight, or legal history, leading to incomplete responses or abandonment of the survey.

One of the most straightforward and commonly used approaches to address missing data is the Mean imputation method. This technique involves replacing missing values in a dataset with the average of the available values. However, this method is only suitable for numerical data.

The formula used to calculate the mean for imputing missing data is as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (4)$$

Where:

$\bar{x}$ = Mean (average value)

n = Total number of data points

x_i = Individual data points

By applying this formula, missing data points are replaced with the calculated mean of the existing values.

C. The Confusion Matrix

The Confusion Matrix is a useful tool for assessing the accuracy of a classification model by comparing predicted values with actual outcomes. It is applicable to both binary and multi-class classification problems and consists of four key values:

True Positive (TP): The number of cases that are correctly predicted as positive.

True Negative (TN): The number of cases that are correctly predicted as negative.

False Positive (FP): The number of cases incorrectly predicted as positive when they are actually negative.

False Negative (FN): The number of cases incorrectly predicted as negative when they are actually positive.

These values allow for the calculation of Accuracy, which measures how well the model's predictions match the actual values.

The formula for Accuracy is:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (5)$$

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

II. Research Methodology

In this study, the data to be used is secondary data from the Kaggle Dataset, "Diagnosis of COVID-19 and its clinical spectrum", https://www.kaggle.com/datasets/einsteindata4u/covid19 [10], created by Hospital Israelita Albert Einstein in São Paulo, Brazil. This dataset contains anonymous data from 5,644 patients at Hospital Israelita Albert Einstein in São Paulo, Brazil, including information such as patient ID, patient age, and other details. Additionally, samples collected from each patient who underwent RT-PCR testing for SARS-CoV-2 (both positive and negative for COVID-19), supplementary laboratory tests during hospital visits (such as hematocrit, hemoglobin, urine-urobilinogen, etc.), as well as the patients' COVID-19 care outcomes, are included. The latter includes whether the patient required care in the Regular Ward, Semi-Intensive Unit, Intensive Care Unit (ICU), or did not require hospitalization at all.

The Regular Ward is a hospital room with more than two beds, intended for patients who no longer require close monitoring and have a low level of dependency, where patients typically begin mobilizing in preparation for discharge.

The Semi-Intensive Unit is designated for patients who still require close monitoring but at a lower level than in the ICU.

The Intensive Care Unit (ICU) is designed for patients who require intensive monitoring and medical support.

In this study, classification analysis will be conducted to determine whether COVID-19 patients do not require hospitalization or if they need care in the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) based on hematological parameters using the C5.0 algorithm with a focus on accuracy. Since the dataset contains missing data**,** the Mean Method will be applied to handle the missing values. Therefore, several relevant variables that have a correlation with the study's objectives will be selected from the entire dataset.

The variables used in this study are as follows:

Table 1. Variable Identification

No.	Variable	Type
1	Hematocrit	Numerical
2	Hemoglobin	Numerical
3	Platelets	Numerical
4	Mean Platelet Volume (MPV)	Numerical
5	*Red Blood Cells*	Numerical
6	Lymphocytes	Numerical
7	Mean Corpuscular Hemoglobin Concentration (MCHC)	Numerical
8	Leukocytes	Numerical
9	Basophils	Numerical
10	Mean Corpuscular Hemoglobin (MCH)	Numerical
11	Eosinophils	Numerical
12	Mean Corpuscular Volume (MCV)	Numerical
13	Monocytes	Numerical
14	Red Blood Cell Distribution Width (RDW)	Numerical

In Table 1, the selection of variables is based on several previous studies. For example, in [11], hematocrit can significantly predict the risk of ICU admission in COVID-19 patients in Iran using multivariable analysis. Study [12] discusses red blood cells less frequently in the pathogenesis of COVID-19, but some studies have considered hemoglobin levels, which are also a major constituent of red blood cells. [13] shows that a decrease in hemoglobin levels in COVID-19 patients is associated with the severity of the disease. There is also a relationship between platelet levels in hospitalized patients and high severity of COVID-19 [14]. The MPV value was found to increase by 6.3% in COVID-19 patients with high severity [15]. Study [16] supports the hypothesis that lymphopenia (a condition when lymphocyte levels are low) can be a prognostic factor in determining the clinical course and severity of the disease in patients hospitalized due to COVID-19. High MCV or low MCHC was found in COVID-19 patients with high or critical severity [17]. Leukopenia (a condition when leukocyte levels are low) has also been reported in several studies, ranging from 28.1% to 68.1%, depending on the severity of the disease and underlying conditions, indicating a possible relationship between the severity of leukopenia and the severity of COVID-19 [12]. Severe COVID-19 cases are typically characterized by low lymphocyte counts, high leukocyte counts, increased neutrophil-lymphocyte ratio (NLR), as well as decreased percentages of monocytes, eosinophils, and basophils [18]. Study [19] found that RDW (Red Cell Distribution Width) could serve as an indicator for predicting the prognosis of COVID-19 patients experiencing severe conditions. A total of 96.4% and 90% of all COVID-19 patients showed low MCH and hemoglobin levels [20].

The stages of data analysis in this study are as illustrated in the following flowchart:

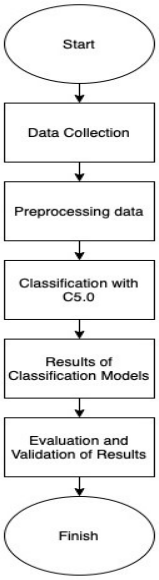

Start
Data Collection
Preprocessing data
Classification with C5.0
Results of Classification Models
Evaluation and Validation of Results
Finish

Figure 1. Data Analysis Method Flowchart

As shown in Figure 1, this study begins with the following stages:

Data Collection

This stage describes how and from where the data for this research is obtained. The data will be saved in files with a

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

.xlsx extension.

Data Preprocessing

The initial data processing will include Data Selection, which involves selecting data from a larger dataset. The selected data will be used for data mining processes, specifically classification. After that, corrections will be made for errors in the data; in this study, missing values will first be filled using the Mean Method with the assistance of SPSS software. Furthermore, before classification, the data will also be divided into Training and Testing datasets.

Classification with C5.0

This stage involves classifying the data that has been processed in the Data Preprocessing stage using the C5.0 Algorithm with RStudio software.

Classification Results

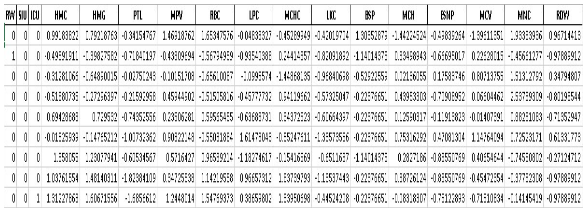
The classification results will take the form of a decision tree to determine whether COVID-19 patients require hospitalization or need to be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters.

Evaluation and Validation of Results

In this stage, the classification results will be evaluated using Confusion Matrix measurements

III. RESULTS AND DISCUSSION

The data mentioned in the Research Method section will undergo Data Selection by choosing several attributes that are relevant to this study. Therefore, from 5.644 patients, 83 COVID-19 positive patients who underwent blood tests upon hospital admission were selected. It is also known which ward these patients were placed in at the hospital. Below is the data used in this study.



RW	SIU	ICU	HMC	HMG	PTL	MPV	RBC	LPC	MCHC	LKC	BSP	MCH	ESNP	MCV	MNC	RDW
0	0	0	0.99183822	0.79218763	-0.34154767	1.46918762	1.65347576	-0.04838327	-0.45289949	-0.42019704	1.30352879	-1.44224524	-0.49839264	-1.39611351	1.93333936	0.96714413
1	0	0	-0.49591911	-0.39827582	-0.71840197	-0.43809694	-0.56794959	-0.93540388	0.24414857	-0.82091892	-1.14014375	0.33498943	-0.66695017	0.22628015	-0.45661277	-0.97889912
0	0	0	-0.31281066	-0.64890015	-0.02750243	-0.10151708	-0.65610087	-0.0995574	-1.44868135	-0.96840698	-0.52922559	0.02136055	0.17583746	0.80713755	1.51312792	0.34794807
0	0	0	-0.51880735	-0.27296397	-0.21592958	0.45944902	-0.51505816	-0.45777732	0.94119662	-0.57325047	-0.22376651	0.43953303	-0.70908952	0.06604462	2.53739309	-0.80198544
0	0	0	0.69428688	0.729532	-0.74352556	0.23506281	0.59565455	-0.63688731	0.34372523	-0.60664397	-0.22376651	0.12590317	-0.11913823	-0.01407391	0.88281083	-0.71352947
0	0	0	-0.01525939	-0.14765212	-1.00732362	0.90822148	-0.55031884	1.61478043	-0.55247611	-1.33573556	-0.22376651	0.75316292	0.47081304	1.14764094	0.72523171	0.61331773
0	0	0	1.358055	1.23077941	-0.60534567	0.5716427	0.96589214	-1.18274617	-0.15416569	-0.6511687	-1.14014375	0.2827186	-0.83550769	0.40654644	-0.74550802	-0.27124712
0	0	0	1.03761554	1.48140311	-1.82384109	0.34725538	1.14219558	0.96657312	1.83739793	-1.13537443	-0.22376651	0.38726124	-0.83550769	-0.45472354	-0.37782308	-0.97889912
0	0	1	1.31227863	1.60671556	-1.6856612	1.2448014	1.54769373	0.38659802	1.33950698	-0.44524208	-0.22376651	-0.08318307	-0.75122893	-0.71510834	-0.14145419	-0.97889912

Figure 2. Data of COVID-19 Positive Patients

The description for Figure 2 can be found in Table 2. The next step is to convert the attributes into a format that can be processed by the program.

Table 2. Description of Figure 2

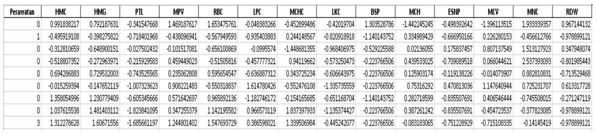
Notation	Description
RW	Regular Ward (1=yes, 0=no)
SIU	Semi-Intensive Unit (1=yes, 0=no)
ICU	Intensive Care Unit (1=yes, 0=no)
HMC	Hematocrit
HMG	Hemoglobin
PTL	Platelets
MPV	Mean Platelet Volume (MPV)
RBC	Red Blood Cells
LPC	Lymphocytes
MCHC	Mean Corpuscular Hemoglobin Concentration (MCHC)
LKC	Leukocytes
BSP	Basophils
MCH	Mean Corpuscular Hemoglobin (MCH)
ESNP	Eosinophils
MCV	Mean Corpuscular Volume (MCV)
MNC	Monocytes
RDW	Red Blood Cell Distribution Width (RDW)

The attributes RW, SIU, and ICU had their values converted from discrete to numeric (label) format in the program, resulting in numeric labels as shown in Table 3.

Table 3. Programming Label Results

Label	Description	Numeric
Care	No hospital care	0
	Patient admitted to Regular Ward	1
	Patient admitted to Semi-Intensive Unit	2
	Patient admitted to Intensive Care Unit	3

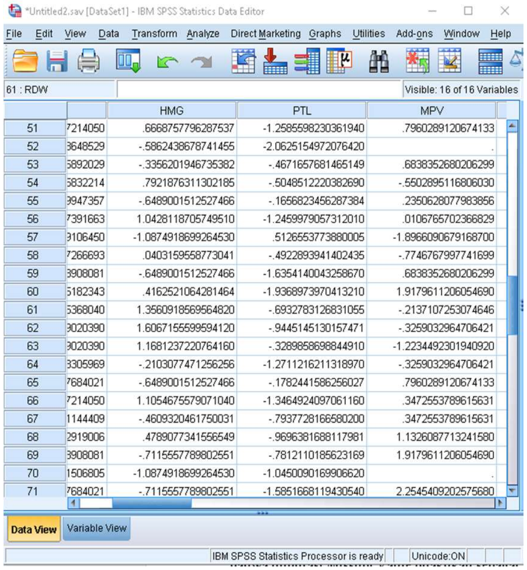
Figure 3 below shows that after converting the discrete attribute values into labels with numeric values, the data appears simpler and easier to process by machine learning algorithms. The next step is to search for missing values.



Perawatan	HMC	HMG	PTL	MPV	RBC	LPC	MCHC	LKC	BSP	MCH	ESNP	MCV	MNC	RDW
0	0.991838217	0.792187631	-0.341547668	1.469187617	1.653475761	-0.048383266	-0.452899486	-0.42019704	1.303528786	-1.442245245	-0.498392642	-1.396113515	1.933339357	0.967144132
1	-0.495919108	-0.398275822	-0.718401968	-0.438096941	-0.567949593	-0.935403883	0.244148567	-0.820918918	-1.140143752	0.334989429	-0.666950166	0.226280153	-0.456612766	-0.978899121
0	-0.312810659	-0.648900151	-0.027502432	-0.101517081	-0.656100869	-0.0995574	-1.448681355	-0.968406975	-0.529225588	0.02136055	0.175837457	0.807137549	1.513127923	0.347948074
0	-0.518807352	-0.272963971	-0.215929583	0.459449023	-0.51505816	-0.457777321	0.94119662	-0.573250473	-0.223766506	0.439533025	-0.709089518	0.066044621	2.537393093	-0.801985443
0	0.694286883	0.729532003	-0.743525565	0.235062808	0.595654547	-0.636887312	0.343725234	-0.606643975	-0.223766506	0.125903174	-0.119138226	-0.014073907	0.882810831	-0.713529468
0	-0.015259394	-0.147652119	-1.007323623	0.908221483	-0.550318837	1.614780426	-0.552476108	-1.335735559	-0.223766506	0.75316292	0.470813036	1.147640944	0.725231707	0.613317728
0	1.358054996	1.230779409	-0.605345666	0.571642697	0.965892136	-1.182746172	-0.154165685	-0.651168704	-1.140143752	0.282718599	-0.835507691	0.406546444	-0.745508015	-0.271247119
0	1.037615538	1.481403112	-1.823841095	0.347255379	1.142195582	0.966573119	1.837397933	-1.135374427	-0.223766506	0.387261242	-0.835507691	-0.454723537	-0.377823085	-0.978899121
3	1.312278628	1.60671556	-1.685661197	1.244801402	1.547693729	0.386598021	1.339506984	-0.445242077	-0.223766506	-0.083183065	-0.751228929	-0.715108335	-0.14145419	-0.978899121

Figure 3. Latest Data of COVID-19 Positive Patients

It can be seen that several data entries have missing values, specifically in the MPV variable for patients 52 and 70, as shown in Figure 4. To address this issue, missing value imputation will be performed, considering that missing value imputation is a treatment for outliers in an effort to improve data quality [22. BPS, 2017].



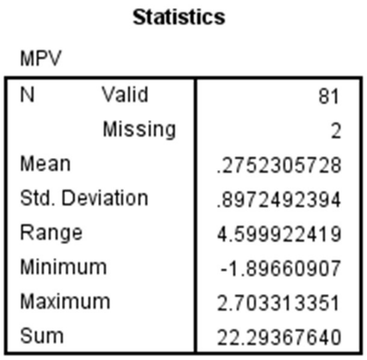
*Untitled2.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
61 : RDW
Visible: 16 of 16 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

		HMG	PTL	MPV
51	7214050	.6668757796287537	-1.2585598230361940	.7960289120674133
52	3648529	-.5862438678741455	-2.0625154972076420	.
53	5892029	-.3356201946735382	-.4671657681465149	.6838352680206299
54	5832214	.7921876311302185	-.5048512220382690	-.5502895116806030
55	3947357	-.6489001512527466	-.1656823456287384	.2350628077983856
56	7391663	1.0428118705749510	-1.2459979057312010	.0106765702366829
57	3106450	-1.0874918699264530	.5126553773880005	-1.8966090679168700
58	7266693	.0403159558773041	-.4922893941402435	-.7746767997741699
59	3908081	-.6489001512527466	-1.6354140043258670	.6838352680206299
60	5182343	.4162521064281464	-1.9368973970413210	1.9179611206054690
61	5368040	1.3560918569564820	-.6932783126831055	-.2137107253074646
62	3020390	1.6067155599594120	-.9445145130157471	-.3259032964706421
63	3020390	1.1681237220764160	-.3289858698844910	-1.2234492301940920
64	3305969	-.2103077471256256	-1.2711216211318970	-.3259032964706421
65	7684021	-.6489001512527466	-.1782441586256027	.7960289120674133
66	7214050	1.1054675579071040	-1.3464924097061160	.3472553789615631
67	1144409	-.4609320461750031	-.7937728166580200	.3472553789615631
68	2919006	.4789077341556549	-.9696381688117981	1.1326087713241580
69	3908081	-.7115557789802551	-.7812110185623169	1.9179611206054690
70	1506805	-1.0874918699264530	-1.0450090169906620	.
71	7684021	-.7115557789802551	-1.5851668119430540	2.2545409202575680

Figure 4. Missing Value in Data of COVID-19 Positive Patients

From the process carried out, the results are shown in Figure 5. In the figure below, it can be seen that the MPV variable has two missing values, with a valid count of 81, along with the Mean, Standard Deviation, Range, Minimum Value, Maximum Value, and Sum of the MPV variable.

under Creative Commons Attribution-ShareAlike 4.0 International License.



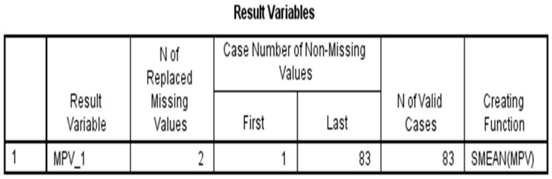
Statistics

MPV

N	Valid	81
	Missing	2
Mean		.2752305728
Std. Deviation		.8972492394
Range		4.599922419
Minimum		-1.89660907
Maximum		2.703313351
Sum		22.29367640

Figure 5. Output of Step 2: Missing Value Imputation

Figure 6 explains that the new variable “MPV_1” will perform imputation on the 2 missing values from 83 data entries using the function “SMEAN(MPV).”

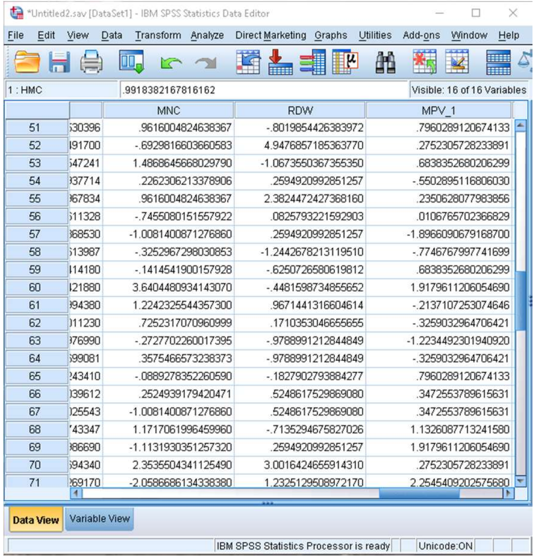


Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	MPV_1	2	1	83	83	SMEAN(MPV)

Figure 6. Output of Step 4: Missing Value Imputation

As shown in Figure 7, the MPV values for patients 52 and 70 have been filled or replaced with the value 0.2752306, which is the mean of the MPV variable inserted into the cells containing the missing values.



*Untitled2.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

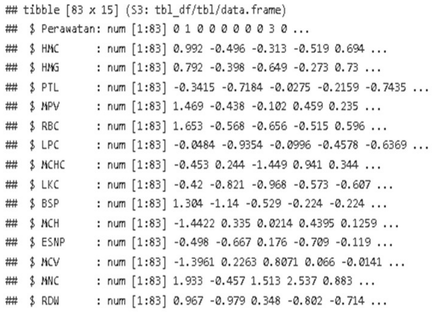
1 : HMC .9918382167816162 Visible: 16 of 16 Variables

		MNC	RDW	MPV_1
51	530396	.9616004824638367	-.8019854426383972	.7960289120674133
52	491700	-.6929816603660583	4.9476857185363770	.2752305728233891
53	547241	1.4868645668029790	-1.0673550367355350	.6838352680206299
54	937714	.2262306213378906	.2594920992851257	-.5502895116806030
55	967834	.9616004824638367	2.3824472427368160	.2350628077983856
56	511328	-.7455080151557922	.0825793221592903	.0106765702366829
57	868530	-1.0081400871276860	.2594920992851257	-1.8966090679168700
58	513987	-.3252967298030853	-1.2442678213119510	-.7746767997741699
59	414180	-.1414541900157928	-.6250726580619812	.6838352680206299
60	421880	3.6404480934143070	-.4481598734855652	1.9179611206054690
61	994380	1.2242325544357300	.9671441316604614	-.2137107253074646
62	011230	.7252317070960999	.1710353046655655	-.3259032964706421
63	976990	-.2727702260017395	-.9788991212844849	-1.2234492301940920
64	599081	.3575466573238373	-.9788991212844849	-.3259032964706421
65	243410	-.0889278352260590	-.1827902793884277	.7960289120674133
66	039612	.2524939179420471	.5248617529869080	.3472553789615631
67	025543	-1.0081400871276860	.5248617529869080	.3472553789615631
68	743347	1.1717061996459960	-.7135294675827026	1.1326087713241580
69	886690	-1.1131930351257320	.2594920992851257	1.9179611206054690
70	594340	2.3535504341125490	3.0016424655914310	.2752305728233891
71	269170	-2.0586686134338380	1.2325129508972170	2.2545409202575680

Figure 7. Results of Missing Value Imputation

Before the classification process begins, the data is typically divided into two parts: the training set and the testing set, according to a specific proportion. This separation uses the Train-Test Split method, which serves to evaluate the performance of the machine learning model. The training data is used to build and train the model, while the testing data is used to measure how well the model can predict data it has never seen before. This is important to ensure that the model has good generalization and can perform well on new data. Therefore, the training and testing data will be divided using RStudio software with the following steps:

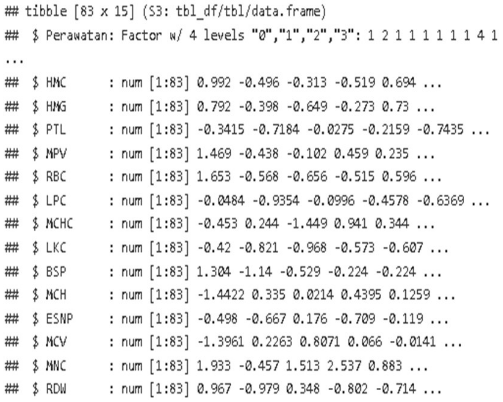
- The data that has previously had missing values replaced will be imported in .xlsx file format using the “readxl” library, which has been downloaded beforehand, along with specifying the file name and the folder where the file is stored.



```
## tibble [83 x 15] (S3: tbl_df/tbl/data.frame)
##  $ Perawatan: num [1:83] 0 1 0 0 0 0 0 0 3 0 ...
##  $ HMC      : num [1:83] 0.992 -0.496 -0.313 -0.519 0.694 ...
##  $ HMG      : num [1:83] 0.792 -0.398 -0.649 -0.273 0.73 ...
##  $ PTL      : num [1:83] -0.3415 -0.7184 -0.0275 -0.2159 -0.7435 ...
##  $ MPV      : num [1:83] 1.469 -0.438 -0.102 0.459 0.235 ...
##  $ RBC      : num [1:83] 1.653 -0.568 -0.656 -0.515 0.596 ...
##  $ LPC      : num [1:83] -0.0484 -0.9354 -0.0996 -0.4578 -0.6369 ...
##  $ MCHC     : num [1:83] -0.453 0.244 -1.449 0.941 0.344 ...
##  $ LKC      : num [1:83] -0.42 -0.821 -0.968 -0.573 -0.607 ...
##  $ BSP      : num [1:83] 1.304 -1.14 -0.529 -0.224 -0.224 ...
##  $ MCH      : num [1:83] -1.4422 0.335 0.0214 0.4395 0.1259 ...
##  $ ESNP     : num [1:83] -0.498 -0.667 0.176 -0.709 -0.119 ...
##  $ MCV      : num [1:83] -1.3961 0.2263 0.8071 0.066 -0.0141 ...
##  $ MNC      : num [1:83] 1.933 -0.457 1.513 2.537 0.883 ...
##  $ RDW      : num [1:83] 0.967 -0.979 0.348 -0.802 -0.714 ...
```

Figure 8. Data Structure

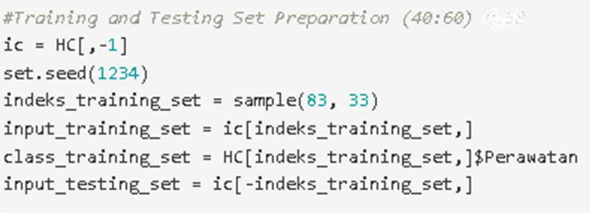
In Figure 8, it can be seen that this dataset consists of 83 entries with 15 variables. Based on the data structure results above, some variable types need to be adjusted according to their characteristics, as described in the previous section. Specifically, the “Care” variable is converted from numeric to factor type with levels “No hospital care,” “Patient admitted to Regular Ward,” “Patient admitted to Semi-Intensive Unit,” and “Patient admitted to Intensive Care Unit” using the following syntax. After changing the “Care” variable to factor data type, Figure 9 shows the structure of the dataset after adjustments have been made to the data structure.



```
## tibble [83 x 15] (S3: tbl_df/tbl/data.frame)
##  $ Perawatan: Factor w/ 4 levels "0","1","2","3": 1 2 1 1 1 1 1 1 4 1
...
##  $ HMC      : num [1:83] 0.992 -0.496 -0.313 -0.519 0.694 ...
##  $ HMG      : num [1:83] 0.792 -0.398 -0.649 -0.273 0.73 ...
##  $ PTL      : num [1:83] -0.3415 -0.7184 -0.0275 -0.2159 -0.7435 ...
##  $ MPV      : num [1:83] 1.469 -0.438 -0.102 0.459 0.235 ...
##  $ RBC      : num [1:83] 1.653 -0.568 -0.656 -0.515 0.596 ...
##  $ LPC      : num [1:83] -0.0484 -0.9354 -0.0996 -0.4578 -0.6369 ...
##  $ MCHC     : num [1:83] -0.453 0.244 -1.449 0.941 0.344 ...
##  $ LKC      : num [1:83] -0.42 -0.821 -0.968 -0.573 -0.607 ...
##  $ BSP      : num [1:83] 1.304 -1.14 -0.529 -0.224 -0.224 ...
##  $ MCH      : num [1:83] -1.4422 0.335 0.0214 0.4395 0.1259 ...
##  $ ESNP     : num [1:83] -0.498 -0.667 0.176 -0.709 -0.119 ...
##  $ MCV      : num [1:83] -1.3961 0.2263 0.8071 0.066 -0.0141 ...
##  $ MNC      : num [1:83] 1.933 -0.457 1.513 2.537 0.883 ...
##  $ RDW      : num [1:83] 0.967 -0.979 0.348 -0.802 -0.714 ...
```

Figure 9. Data Conversion Structure

To perform the splitting of Training Data and Testing Data from the 83 observation data, this study will allocate 40% as Training Data and 60% as Testing Data randomly. Below is the command to select 33 rows to be stored in the Training Data, with the remaining rows stored in the Testing Data.

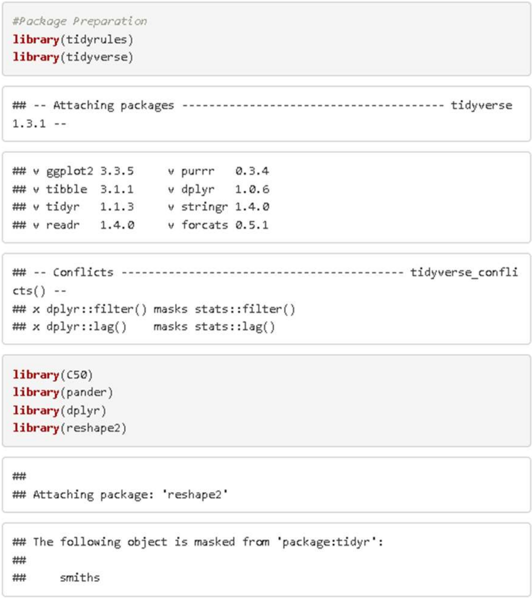


```
#Training and Testing Set Preparation (40:60)
ic = HC[,-1]
set.seed(1234)
indeks_training_set = sample(83, 33)
input_training_set = ic[indeks_training_set,]
class_training_set = HC[indeks_training_set,]$Perawatan
input_testing_set = ic[-indeks_training_set,]
```

Figure 10. Data Splitting Syntax

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.

After splitting the data into training and testing sets, the classification process will be carried out using all variables with R software.
The first step is to prepare the packages that will be used during the classification process by installing R packages such as tidyrules, tidyverse, C50, pander, dplyr, and reshape2 using the function "install.packages( )." Once the installation process is complete, load the packages into the R session using the function "library( )" as shown in the figure below.



```
#Package Preparation
library(tidyrules)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_confli
cts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

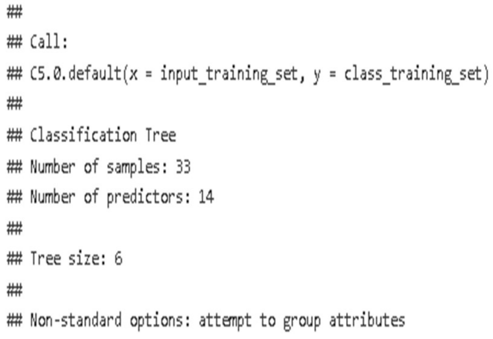
```
library(C50)
library(pander)
library(dplyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

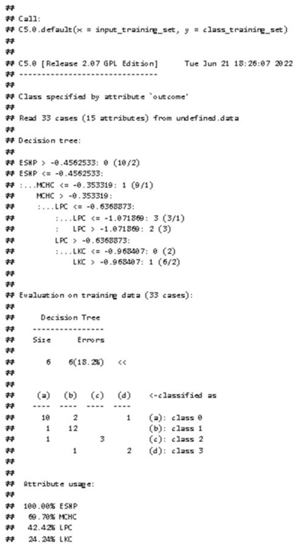
Figure 11. Loading Packages in R

And the packages are ready to use.
In Figure 12, "Number of samples" depicts the amount of data used, specifically the Training Data consisting of 33 entries. Meanwhile, "Number of predictors" refers to the number of attributes used, which includes 14 variables with "Care" as the Class in this classification. Using the C5.0 Algorithm, a decision tree with 6 branches will be generated.



```
##
## Call:
## C5.0.default(x = input_training_set, y = class_training_set)
##
## Classification Tree
## Number of samples: 33
## Number of predictors: 14
##
## Tree size: 6
##
## Non-standard options: attempt to group attributes
```

Figure 12. Number of Samples



```
##
## Call:
## C5.0.default(x = input_training_set, y = class_training_set)
##
##
## C5.0 [Release 2.07 GPL Edition]      Tue Jun 21 18:26:07 2022
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 33 cases (15 attributes) from undefined.data
##
## Decision tree:
##
## ESNP > -0.4562533: 0 (10/2)
## ESNP <= -0.4562533:
## :...MCHC <= -0.353319: 1 (9/1)
##     MCHC > -0.353319:
##     :...LPC <= -0.6368873:
##         :...LPC <= -1.071869: 3 (3/1)
##         :   LPC > -1.071869: 2 (3)
##         LPC > -0.6368873:
##         :...LKC <= -0.968407: 0 (2)
##             LKC > -0.968407: 1 (6/2)
##
##
## Evaluation on training data (33 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       6    6(18.2%)   <<
##
##
##     (a)   (b)   (c)   (d)    <-classified as
##    ----  ----  ----  ----
##      10     2           1    (a): class 0
##       1    12                (b): class 1
##       1           3          (c): class 2
##             1           2    (d): class 3
##
##
##  Attribute usage:
##
##  100.00% ESNP
##   69.70% MCHC
##   42.42% LPC
##   24.24% LKC
```

Figure 13. Output of C5.0 Algorithm Classification

The decision tree obtained from the C5.0 Algorithm is used for determining whether COVID-19 patients require hospitalization or if they should be placed in a Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters, as shown in the "Decision Tree" section below.

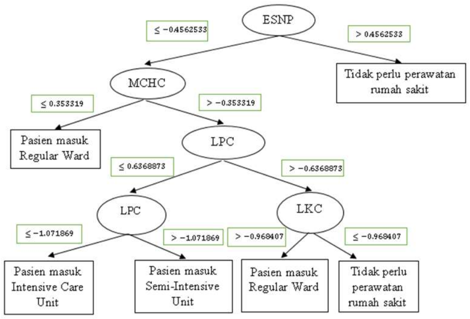
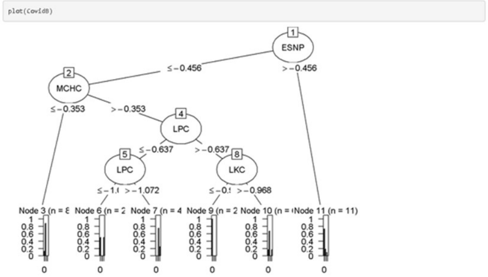

ESNP
≤ −0.4562533
> 0.4562533
MCHC
Tidak perlu perawatan rumah sakit
≤ 0.353319
> −0.353319
Pasien masuk Regular Ward
LPC
≤ 0.6368873
> −0.6368873
LPC
LKC
≤ −1.071869
> −1.071869
> −0.968407
≤ −0.968407
Pasien masuk Intensive Care Unit
Pasien masuk Semi-Intensive Unit
Pasien masuk Regular Ward
Tidak perlu perawatan rumah sakit

Figure 14. Decision Tree of C5.0 Algorithm Classification

In Figure 14, the decision tree of the C5.0 Algorithm can be interpreted as follows: for example, if a patient undergoes a blood test in the hospital and the ESNP result is > 0.4562533, then the patient can be predicted not to require hospitalization. If both ESNP and MHC values are considered, the patient can be predicted to be admitted to the Regular Ward, and so on. Figure 15 shows a plot of the decision tree from the C5.0 Algorithm.



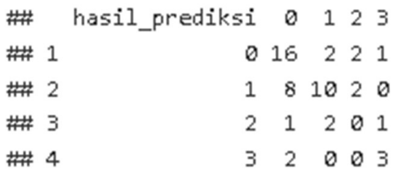
```
plot(CovidB)
```

Figure 15. Plot of C5.0 Algorithm Decision Tree

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

In Figure 13, the "Evaluation on training data" shows an error rate of 18.2% in the classification using the C5.0 Algorithm with the Training Data. Regarding "Attribute usage," it can be observed that there are 4 attributes considered influential in forming the decision tree of the C5.0 Algorithm in this study, with the ESNP attribute being the root or the most important attribute, accounting for 100% usage and so on. Thus, as you move down the tree, the usage of attributes decreases.

After using the Training Data for modeling, the next step is to perform predictions using the C5.0 Algorithm on the Testing Data.



```
##   hasil_prediksi  0  1 2 3
## 1               0 16  2 2 1
## 2               1  8 10 2 0
## 3               2  1  2 0 1
## 4               3  2  0 0 3
```

Figure 16. Output of Syntax 4 for C5.0 Algorithm Predictions

In Figure 16, the confusion matrix from the obtained model is shown. For the 0 category (Patients who do not require hospitalization), there were 0 correct predictions (Patients who do not require hospitalization) based on the previously described hematological parameters, totaling 16 cases. For the 0 category (Patients who do not require hospitalization) predicted as 1 (Patients admitted to the Regular Ward), there were 2 cases, and so on.

To measure the performance of the classification model, the Accuracy will be calculated. Using the formula in equation (5), an Accuracy of 0.78 is obtained. This means that 78% of the patients were correctly predicted as not needing hospitalization or requiring admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU). Therefore, the classification using the C5.0 Algorithm is considered quite good.

IV. Conclusion

In this study, a classification analysis was conducted to determine whether COVID-19 patients require hospitalization or admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) in the hospital based on hematological parameters using the C5.0 algorithm. A decision tree was obtained for identifying COVID-19 patients who do not require hospitalization or those needing admission to the Regular Ward, Semi-Intensive Care Unit, or Intensive Care Unit (ICU) with an accuracy of 78%. The classification using the C5.0 Algorithm is considered quite good.

It is suggested that future research could improve the accuracy by increasing the sample size. It is also hoped that this study can facilitate and expedite the work of medical personnel, enabling COVID-19 patients to receive prompt and appropriate care to help reduce COVID-19 cases in a population.

References

[1] A. Asan *et al.*, "Do initial hematologic indices predict the severity of covid-19 patients?," *Turk J Med Sci*, vol. 51, no. 1, 2021, doi: 10.3906/sag-2007-97.

[2] M. Khishe, F. Caraffini, and S. Kuhn, "Evolving deep learning convolutional neural networks for early covid-19 detection in chest x-ray images," *Mathematics*, vol. 9, no. 9, 2021, doi: 10.3390/math9091002.

[3] B. Debuc and D. M. Smadja, "Is COVID-19 a New Hematologic Disease?," *Stem Cell Rev Rep*, vol. 17, no. 1, 2021, doi: 10.1007/s12015-020-09987-4.

[4] E. Terpos *et al.*, "Hematological findings and complications of COVID-19," 2020. doi: 10.1002/ajh.25829.

[5] A. Rahman, R. Niloofa, U. Jayarajah, S. De Mel, V. Abeysuriya, and S. L. Seneviratne, "Hematological abnormalities in COVID-19: A narrative review," 2021. doi: 10.4269/ajtmh.20-1536.

[6] U. S. Aesyi, T. W. Diwangkara, and R. T. Kurniawan, "DIAGNOSA PENYAKIT DISK HERNIA DAN SPONDYLOLISTHESIS MENGGUNAKAN ALGORITMA C5," *Telematika*, vol. 16, no. 2, 2020, doi: 10.31315/telematika.v16i2.3181.

[7] E. R. Jorda and A. R. Raqueno, "Predictive model for the academic performance of the engineering students using CHAID and C 5.0 algorithm," *International Journal of Engineering Research and Technology*, vol. 12, no. 6, 2019.

[8] P. N. Patil, R. Lathi, and V. Chitre, "Comparison of C5 . 0 & CART Classification algorithms using pruning technique," *International Journal of Engineering Research & Technology*, vol. 1, no. 4, 2012.

[9] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in *Classification, Clustering, and Data Mining Applications*, 2004. doi: 10.1007/978-3-642-17103-1_60.

[10] EINSTEIN DATA4U, "Diagnosis of COVID-19 and its clinical spectrum." Accessed: Sep. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/einsteindata4u/covid19

[11] A. Sadeghi *et al.*, "COVID-19 and ICU admission associated predictive factors in Iranian patients," *Caspian J Intern Med*, vol. 11, 2020, doi: 10.22088/cjim.11.0.512.

[12] M. Karimi Shahri, H. R. Niazkar, and F. Rad, "COVID-19 and hematology findings based on the current evidences: A puzzle with many missing pieces," 2021. doi: 10.1111/ijlh.13412.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.

[13] Y. Pan *et al.*, “Can routine laboratory tests discriminate SARS-CoV-2-infected pneumonia from other causes of community-acquired pneumonia?,” *Clin Transl Med*, vol. 10, no. 1, 2020, doi: 10.1002/ctm2.23.

[14] Y.-P. Liu *et al.*, “Combined use of the neutrophil-to-lymphocyte ratio and CRP to predict 7-day disease severity in 84 hospitalized patients with COVID-19 pneumonia: a retrospective cohort study,” *Ann Transl Med*, vol. 8, no. 10, 2020, doi: 10.21037/atm-20-2372.

[15] G. Lippi, B. M. Henry, and E. J. Favaloro, “Mean Platelet Volume Predicts Severe COVID-19 Illness,” 2021. doi: 10.1055/s-0041-1727283.

[16] J. Wagner, A. DuPont, S. Larson, B. Cash, and A. Farooq, “Absolute lymphocyte count is a prognostic marker in Covid-19: A retrospective cohort review,” *Int J Lab Hematol*, vol. 42, no. 6, 2020, doi: 10.1111/ijlh.13288.

[17] J. Mao, R. Dai, R. C. Du, Y. Zhu, L. P. Shui, and X. H. Luo, “Hematologic changes predict clinical outcome in recovered patients with COVID-19,” *Ann Hematol*, vol. 100, no. 3, 2021, doi: 10.1007/s00277-021-04426-x.

[18] C. Qin *et al.*, “Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China,” *Clin Infect Dis*, vol. 71, no. 15, pp. 762–768, Aug. 2020, doi: 10.1093/CID/CIAA248.

[19] C. Wang *et al.*, “Red cell distribution width (RDW): a prognostic indicator of severe COVID-19,” *Ann Transl Med*, vol. 8, no. 19, 2020, doi: 10.21037/atm-20-6090.

[20] S. M. Attiyah, H. M. Elsayed, J. A. Al Mughales, A. B. Moharram, and M. A. Fattah, “Critical cases of COVID-19 patients can be predicted by the biomarkers of complete blood count,” *Indian J Sci Technol*, vol. 13, no. 48, pp. 4739–4745, Jan. 2020, doi: 10.17485/IJST/V13I48.2033.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi under Creative Commons Attribution-ShareAlike 4.0 International License.