DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM

Wildan Wiguna[1]; Dwiza Riana[2]

Program Studi Sistem Informasi Kampus Kota Tasikmalaya[1]
Universitas Bina Sarana Informatika
www.bsi.ac.id
wildan.wwg@bsi.ac.id

Program Studi Magister Ilmu Komputer[2]
Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
www.nusamandiri.ac.id
dwiza@nusamandiri.ac.id

Abstract— Coronavirus Disease 2019 (COVID-19) has become a pandemic in Indonesia as a non-natural disaster in the form of disease outbreaks which must be undertaken as a response. The Ministry of Health in the Republic of Indonesia published a guidebook for prevention and control of COVID-19 in its response efforts. This guideline is intended for health officials as a reference in preparing for COVID-19. This handbook contains early detection and response activities to identify conditions of PDP, ODP, OTG, or confirmed cases of COVID-19. The efforts made are adjusted to the world situation progress from COVID-19 which is monitored by the World Health Organization (WHO). From the results of a documentation study that has been carried out on the COVID-19 pandemic in Indonesia, there are several problems that must be resolved from the prevention of the disease outbreak COVID-19. Lack of knowledge and awareness of the general public in the prevention and control of COVID-19 is one of the factors increasing the spread of that virus in Indonesia. Furthermore, there are difficulties in carrying out surveillance, early detection, contact tracing, infection prevention or control, and risk communication or people empowerment. This is due to the lack of implementation and testing on artificial intelligence methods for COVID-19 diagnosis that can be used by the public. The purpose of this research is to make a diagnosis of surveillance classification which includes PDP, ODP, and OTG using the C4.5 algorithm. The results showed that the diagnosis of COVID-19 surveillance using the C4.5 algorithm was successfully modeled into a decision tree with PDP, ODP, and OTG classification. The testing process in a confusion matrix with 3 (three) classes presents an accuracy rate of 92.86% which is included in the excellent classification category.

Keywords: Artificial Intelligence, Coronavirus Disease 2019, COVID-19 Diagnosis, Decision Tree, C4.5 Algorithm.

Abstrak— *Coronavirus Disease 2019 (COVID-19) telah menjadi pandemi di Indonesia sebagai bencana non alam berupa wabah penyakit yang wajib dilakukan upaya penanggulangan. Kemenkes RI menerbitkan sebuah buku pedoman pencegahan dan pengendalian COVID-19 dalam upaya penanggulangannya. Pedoman ini ditujukan bagi petugas kesebahatan sebagai acuan dalam melakukan kesiapsiagaan menghadapi COVID-19. Pada buku pedoman tersebut berisi kegiatan deteksi dini dan respon untuk mengidentifikasi kondisi PDP, ODP, OTG, maupun kasus konfirmasi COVID-19. Upaya-upaya yang dilakukan disesuaikan dengan perkembangan situasi dari COVID-19 dunia yang dipantau dari WHO. Dari hasil studi dokumentasi yang telah dilakukan terhadap pandemi dari COVID-19 yang terjadi di Indonesia, terdapat beberapa permasalahan yang harus dibenahi dari penanggulangan wabah tersebut. Kurangnya pengetahuan dan kesadaran masyarakat umum dalam melaksanakan pencegahan dan pengendalian COVID-19 merupakan salah satu faktor peningkatan penyebaran virus tersebut di Indonesia. Kemudian terdapat kesulitan dalam melaksanakan surveilans, deteksi dini, pelacakan kontak, pencegahan dan pengendalian infeksi, serta komunikasi risiko maupun pemberdayaan masyarakat. Hal ini disebabkan kurangnya penerapan dan pengujian suatu metode kecerdasan buatan pada diagnosa COVID-19 yang dapat dimanfaatkan secara publik. Tujuan dari penelitian ini yaitu untuk melakukan diagnosa dari kategori surveilans yang meliputi PDP, ODP, dan OTG menggunakan algoritma C4.5. Hasil penelitian menunjukkan bahwa diagnosa terhadap kategori surveilans COVID-19 menggunakan algoritma C4.5 berhasil dimodelkan menjadi sebuah pohon keputusan dengan klasifikasi PDP, ODP, dan*

OTG. Kemudian proses pengujian pada confusion matrix dengan 3 (tiga) kelas menghasilkan tingkat akurasi sebesar 92,86% yang termasuk ke dalam klasifikasi unggul atau sangat baik.

***Kata Kunci**: Kecerdasan Buatan, Coronavirus Disease 2019, Diagnosa COVID-19, Pohon Keputusan, Algoritma C4.5.*

INTRODUCTION

Coronavirus is a large family of viruses that cause diseases ranging from mild to severe symptoms. There are at least two types of Coronavirus that are known to cause diseases with severe symptoms such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Coronavirus Disease 2019 (COVID-19) is a new type of disease that has never been identified before in humans. The virus that causes COVID-19 is named as Sars-CoV-2 which is zoonotic (transmitted between animals and humans). Research says that SARS is transmitted from Civet cats to humans and MERS from camels to humans. But the animals that are the main source of COVID-19 transmission are still not known in certain (Kemenkes RI, 2020).

The World Health Organization (WHO) has stated that Coronavirus Disease 2019 (COVID-19) is a pandemic, and Indonesia has stated that COVID-19 is a non-natural disaster in the form of disease outbreaks that must be resolved so that there is no increase in cases. In the effort to overcome COVID-19, guidance is needed for the public society in making efforts to prevent the spread of the virus. The possibility of transmission can occur both for ourselves and for the people around us, including family (Menkes RI, 2020).

The Ministry of Health in The Republic of Indonesia or better known as "*Kementerian Kesehatan Republik Indonesia* (Kemenkes RI)" is a ministry within the Indonesian government in charge of health matters. The ministry seeks to make various preparations to deal with the Novel Coronavirus infection in Indonesia. One of the efforts is by issuing a circular (number HK.02.01/MENKES/202/2020) about the protocol for self-isolation in handling Coronavirus Disease (COVID-19). The ministry has Directorate General of Disease Prevention and Control or namely as "*Direktorat Jenderal Pencegahan dan Pengendalian Penyakit* (P2P)" which is an implementation element in the Indonesian Ministry of Health. The Directorate General of Disease Prevention and Control published a guidebook that had several revisions in the prevention and control of Coronavirus Disease (COVID-19).

From the results of a documentation study that has been performed on the COVID-19 pandemic that occurred in Indonesia, there are several problems that must be overcome in the prevention of this disease spread or transmission. Lack of knowledge and awareness from the general public in implementing the prevention and control of COVID-19 is one of the factors increasing the virus spread that causes this disease in Indonesia. Then there are difficulties in carrying out surveillance, early detection, contact tracing, infection prevention and control, as well as risk communication and people empowerment. Some of these problems are caused by the lack of implementation and testing of an artificial intelligence method to diagnose Coronavirus Disease (COVID-19) which can be used publicly.

The diagnosis of COVID-19 surveillance can be optimized, implemented, and tested using one of the artificial intelligence methods with the C4.5 algorithm. There is research on the C4.5 algorithm in diagnosing and classifying diseases, such as research on medical record analysis to determine patterns of disease groups using the C4.5 algorithm. The results of the C4.5 algorithm calculation are able to analyze the trends of disease experienced by public society (Rafiska et al., 2018). Furthermore, there is also research on the C4.5 algorithm for the diagnosis of Tuberculosis disease. The results of performance measurements from the model are known that the C4.5 algorithm has an accuracy rate of 84.56% which is included as good classification (Amrin et al., 2019).

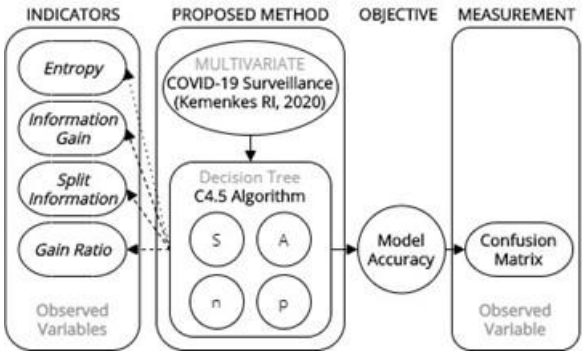
Related to Acute Respiratory Infections (ARI) there is research on the prediction of ARI disease prediction using C4.5 algorithm. In this research, the C4.5 algorithm is able to predict acute and non-acute respiratory infection sufferers (Tarigan et al., 2017). Then there is also research on the C4.5 algorithm for the diagnosis of pneumonia. The research showed that the C4.5 algorithm succeeded in modeling a decision tree with 10 rules of pneumonia (Mujahidin & Pribadi, 2017).

From the issues that have been described previously, this research aims to further discuss the early diagnosis of the COVID-19 surveillance categories for the general public in Indonesia. These surveillance categories include patient under supervision or "*Pasien Dalam Pengawasan* (PDP)", person in monitoring or "*Orang Dalam Pemantauan* (ODP)", and person without symptoms or "*Orang Tanpa Gejala* (OTG)". Meanwhile, the method to be used is the C4.5 algorithm to perform a diagnosis classified operational definitions of Coronavirus Disease (COVID-19). This algorithm is used to modeling a decision tree from the COVID-19 surveillance categories.

MATERIALS AND METHODS

Research Design

The methodology used in this research is an experimental research design with high internal causal validity in constructing research (Campbell & Stanley, 2015). The experimental research enables the causal relationship identification, it is able to identify the true cause of a phenomenon that allows researchers to manipulate a method and achieve the desired results (Lazar et al., 2017). The experimental design is the conceptual framework in the experiments performed (Ary et al., 2018). The framework of this research design is shown in Fig. 1.



Source: (Wiguna & Riana, 2020)

Figure 1. Research Framework

Fig. 1 shows that the data used in this research is the operational definition of COVID-19 surveillance from the Indonesian Ministry of Health, while the proposed method uses the C4.5 algorithm to classify symptoms on COVID-19 surveillance. Some indicators observed include Entropy, Information Gain, Split Information, and Gain Ratio in determining node of the decision tree. The objective is to test the C4.5 algorithm and find accuracy using a confusion matrix.

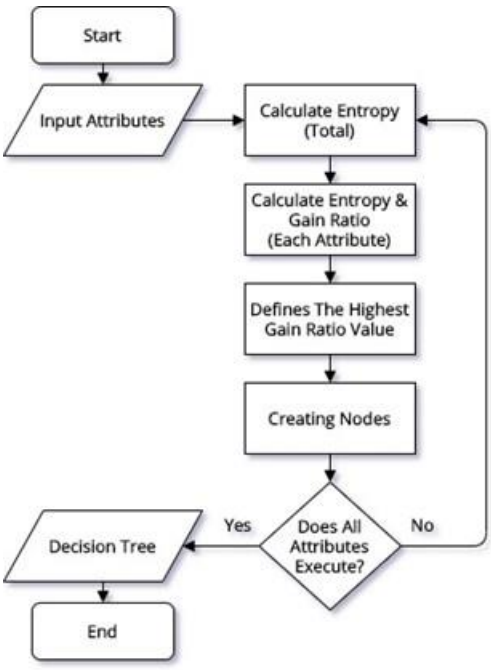
Data Collection Methods

This research uses secondary data that has been collected from government publications or even from some official websites of health agencies (Sekaran & Bougie, 2016). Some data collection techniques in this research include:

1. Literature Study
   The literature study performed in search of theories and empirical evidence or the results of scientific research that support and direct the research (Muharto & Ambarita, 2016). The literature review is used in the systematic examination of scientific literature on a particular topic (Efron & Ravid, 2018). Literature survey is a comprehensive study of technical and official content related to research keywords (Bairagi & Munot, 2019). A literature review was carried out by collecting and studying some journals and books relating to COVID-19 diagnosis using the C4.5 algorithm with Rapidminer Studio implementation.
2. Documentation Study
   Information derived from important records, both from institutions or individuals to strengthen research results (Anggito & Setiawan, 2018). Official and public documents were collected from the Indonesian Ministry of Health, the National Disaster Management Agency, and the Indonesian Lung Doctors Association to determine the preparedness process to face COVID-19 along with the proper rules for carrying out the diagnosis.
3. Judgmental Sampling
   A decision sampling in accordance with a decision that makes sense by the sample taker (Mulyani, 2017). Samples of documents and forms were taken with the aim to study the process of prevention and control of COVID-19 which is expected to be able to represent all documents in Indonesian government.

C4.5 Algorithm

The decision tree method is derived from the learning systems concept. One of the methods developed is the C4.5 algorithm which can deal with attributes continuous value (Li et al., 2017). The C4.5 algorithm can be used to handle data classification problems (Santoso & Azis, 2020). The flowchart of the decision tree modeling steps can be seen in Fig. 2.



Source: (Anwar et al., 2018)

Figure 2. Flowchart of C4.5 Algorithm

There are several elements that must be looked for in the decision tree modeling using the C4.5 algorithm (Buulolo, 2020), including:

1. Entropy(S) is a parameter used to measure the diversity of each attribute value against the decision attribute.

$$Entropy(S) = \sum_{i=n}^{n} -pi * log_2(pi) \quad \text{.....(1)}$$

Information:
S = Sum of case samples (sampling)
n = Number of partitions for S
pi = Proportion of Si to S

2. Gain(S,A) is the gain value that used as the basis for forming nodes or roots and branches of a decision tree.

$$Gain(S,A) = E(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * E(Si) \quad \text{.....(2)}$$

Information:
E = *Entropy*
S = Sum of case samples (*sampling*)
A = *Attribute*
n = Number of partitions for S
|Si|= Sum of cases on i-partition
|S| = Sum of cases in S

3. The C4.5 algorithm uses different sizes in selecting attributes to be split. C4.5 algorithm uses the Gain Ratio in the calculation process instead of the Information Gain obtained (Castaño, 2018). Split Information is first calculated with the following formula:

$$SplitInfo.(S,A) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} * log_2 \frac{|Si|}{|S|} \quad \text{.....(3)}$$

Therefore, the Gain Ratio value can be calculated with the following equation:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S.A)} \quad \text{.....(4)}$$

Information:
S = Sum of case samples (*sampling*)
A = *Attribute*
n = Number of partitions for S
|Si|= Sum of cases on i-partition
|S| = Sum of cases in S

4. The steps in the process of decision tree formation using the C4.5 algorithm as follows:
   a. Find the Entropy(S) value for total cases and each attribute value.
   b. Find the Gain(S,A), SplitInformation(S,A), and GainRatio(S,A) for each attribute value.
   c. Create root, node, and branches based on the highest Gain Ratio value.
   d. Repeat the process for each branch.

RESULTS AND DISCUSSION

Case Study of COVID-19 Surveillance

There is an operational definition of surveillance and response provided by the Indonesian government to implement prevention and control in dealing with COVID-19 outbreak cases with the following categories:

1. Patient under Supervision (PUS) or "*Pasien Dalam Pengawasan* (PDP)".
2. Person in Monitoring (PIM) or "*Orang Dalam Pemantauan* (ODP)".
3. Person without Symptoms (PWS) or "*Orang Tanpa Gejala* (OTG)"

Some of the symptoms along with their respective codes from the COVID-19 surveillance categories include:

1. Fever or history of fever (G01).
2. Symptoms and signs of respiratory distress (cough, cold, sore throat, etc.) (G02).
3. Severe Pneumonia or Acute Respiratory Infections (ARI) (G03).
4. There are no other causes based on convincing clinical descriptions (G04).
5. In the last 14 days before the symptoms have a history of travel or living abroad who reported local transmission (G05).
6. In the last 14 days before the symptoms have a history of travel or stay in the local transmission area in Indonesia (G06).
7. Contact with Coronavirus Disease 2019 (COVID-19) confirmation cases in the last 14 days before symptoms (G07).

The surveillance activities on PDP along for 14 days from the start of symptoms, the PDP category details can be seen in Table 1.

Table 1. PDP Category Details

PDP							
No.	G01	G02	G03	G04	14 days		G07
					G05	G06	
1	+	+	+	+	+	-	-
2	+	+	-	+	+	-	-
3	+	+	+	+	-	+	-
4	+	+	-	+	-	+	-
5	+	-	-	-	-	-	+
6	+	+	+	-	-	-	+
7	+	+	-	-	-	-	+
8	+	+	+	+	-	-	-
Action	▪ Treatment: - Mild: Self-isolation at home. - Moderate: Hospitalized in an emergency Hospital. - Severe: Hospitalized in a referral hospital. ▪ Specimen Examination.						

Source: (Kemenkes RI, 2020)

The surveillance activities on ODP along for 14 days from the start of symptoms, the ODP category details can be seen in Table 2.

Table 2. ODP Category Details

ODP							
No.	G01	G02	G03	G04	14 days		G07
					G05	G06	
1	+	-	-	+	+	-	-
2	-	+	-	+	+	-	-
3	+	-	-	+	-	+	-
4	-	+	-	+	-	+	-
5	-	+	-	-	-	-	+
Action	▪ Self-isolation at home. ▪ Specimen examination. ▪ Healthcare facilities supervise the patient's condition every day for approximately 2 weeks. If the patient's condition is getting worse according to the PDP criteria, or the laboratory test is positive, then patient must be taken to the Emergency Hospital (moderate) or Referral Hospital (mild).						

Source: (Kemenkes RI, 2020)

Surveillance on OTG for 14 days since the last contact with COVID-19 positive cases, the OTG category details shown in Table 3.

Table 3. OTG Category Details

OTG							
No.	G01	G02	G03	G04	14 days		G07
					G05	G06	
1	-	-	-	-	-	-	+
Action	▪ Self-quarantine. ▪ Specimen examination. ▪ Public health center supervise the patients' condition every day, for approximately 2 weeks. If they experience symptoms, then: - Mild: Self-isolation at home. - Moderate: Hospitalized in an emergency hospital. - Severe: Hospitalized in a referral hospital.						

Source: (Kemenkes RI, 2020)

Decision Tree of COVID-19 Surveillance

The process of the decision tree modeling will be carried out based on the decision rules.

A. Node 1

There are several COVID-19 surveillance categories including PDP, ODP, and OTG with decision combinations in Table 4.

Table 4. COVID-19 Surveillance Categories

No.	G01	G02	G03	G04	G05	G06	G07	Cat.
1	+	+	+	+	+	-	-	PDP
2	+	+	-	+	+	-	-	PDP
3	+	+	+	+	-	+	-	PDP
4	+	+	-	+	-	+	-	PDP
5	+	-	-	-	-	-	+	PDP
6	+	+	+	-	-	-	+	PDP
7	+	+	-	-	-	-	+	PDP
8	+	+	+	+	-	-	-	PDP
9	+	-	-	+	+	-	-	ODP
10	-	+	-	+	+	-	-	ODP
11	+	-	-	+	-	+	-	ODP
12	-	+	-	+	-	+	-	ODP
13	-	+	-	-	-	-	+	ODP
14	-	-	-	-	-	-	+	OTG

Source: (PDPI, 2020)

The steps for completing a case from Table 4 are as follows:

a. Calculating the sum of cases for PDP decisions, ODP decisions, OTG decisions, and Entropy of all cases as well as cases divided based on symptoms from COVID-19 surveillance.
b. Perform Entropy, Information Gain, Split Information and Gain Ratio calculations for each available symptom. All the attributes and labels of the research case must be included.

The node 1 or root node calculation results can be shown in Table 5.

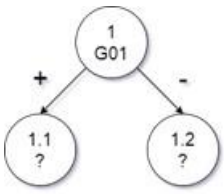
Table 5. Node 1 (Root Node) Calculation

		Cases (S)	PDP (S1)	ODP (S2)	OTG (S3)	Entropy	Gain Ratio
Total		14	8	5	1	1.264	
G01							1.686
	+	11	8	2	0	0	
	-	3	0	3	1	0	
G02							0.968
	+	10	7	3	0	0	
	-	4	1	2	1	1.5	
G03							0.338
	+	4	4	0	0	0	
	-	10	4	5	1	1.361	
G04							0.823
	+	9	5	4	0	0	
	-	5	3	1	1	1.371	
G05							0.392
	+	4	2	2	0	0	
	-	10	6	3	1	1.295	
G06							0.392
	+	4	2	2	0	0	
	-	10	6	3	1	1.295	
G07							0.823
	+	5	3	1	1	1.371	
	-	9	5	4	0	0	

Source: (Wiguna, 2020)

From the calculation results in table 5, it can be seen that the highest Gain Ratio value is G01 symptom of 1,686 which will be used as the root node. In G01 there are two attribute values, namely positive (+) and negative (-) which still need to find the next nodes (node 1.1 & node 1.2) using reduced data. This is because G01+ still has two categories between PDP and ODP, as well as G01- still has two categories between ODP and OTG.

A temporary decision tree is illustrated from the results of the previous calculation contained in Fig. 3.



Source: (Wiguna & Riana, 2020)

Figure 3. Node 1 of Decision Tree

B. Node 1.1

The process of finding another node (node 1.1) uses reduced data by filtering the G01 with a positive value (G01+) as shown in Table 6.

Table 6. Node 1.1 of COVID-19 Surveillance

No.	G01	G02	G03	G04	G05	G06	G07	Cat.
1	+	+	+	+	+	-	-	PDP
2	+	+	-	+	+	-	-	PDP
3	+	+	+	+	-	+	-	PDP
4	+	+	-	+	-	+	-	PDP
5	+	-	-	-	-	-	+	PDP
6	+	+	+	-	-	-	+	PDP
7	+	+	-	-	-	-	+	PDP
8	+	+	+	+	-	-	-	PDP
9	+	-	-	+	+	-	-	ODP
11	+	-	-	+	-	+	-	ODP

Source: (Wiguna, 2020)

The calculation is performed to find a branch node of G01+ without involving G01 values because it has become a root node. The results of node 1.1 calculation can be seen in Table 7.

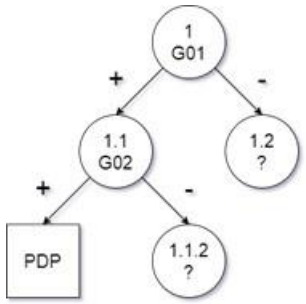
Table 7. Node 1.1 (G01+) Calculation

		Cases (S)	PDP (S1)	ODP (S2)	Entropy	Gain Ratio
Total		10	8	2	0.722	
G02						0.507
	+	7	7	0	0	
	-	3	1	2	0.918	
G03						0.176
	+	4	4	0	0	
	-	6	4	2	0.918	
G04						0.134
	+	7	5	2	0.863	
	-	3	3	0	0	
G05						0.037
	+	3	2	1	0.918	
	-	7	6	1	0.592	
G06						0.037
	+	3	2	1	0.918	
	-	7	6	1	0.592	
G07						0.134
	+	3	3	0	0	
	-	7	5	2	0.863	

Source: (Wiguna, 2020)

From the calculation results in table 7, it can be seen that the highest Gain Ratio value is G02 of 0.507 which will be used as a branch node of G01+. In G02 there are two attribute values which are positive (+) and negative (-). G02 symptom with a positive value (G02+) is known to have 0 cases in the ODP category, so it can be ascertained that all G02+ decisions will generate a diagnosis category with the PDP label and no further calculation is needed. But G02 symptom with a negative value (G02-) still needs to be calculated again.

A temporary decision tree is illustrated from the results of node 1.1 calculation shown in Fig. 4.



Source: (Wiguna & Riana, 2020)

Figure 4. Node 1.1 of Decision Tree

C. Node 1.1.2

The recalculation is performed to find out the next node (node 1.1.2) of the internal branch in G02 with a negative value (G02-). The filter results (reduced data by filtering G01+, G01-, and G02+) from the previous table obtained a new table as shown in Table 8.

Table 8. Node 1.1.2 of COVID-19 Surveillance

No.	G01	G02	G03	G04	G05	G06	G07	Cat.
5	+	-	-	-	-	-	+	PDP
9	+	-	-	+	+	-	-	ODP
11	+	-	-	+	-	+	-	ODP

Source: (Wiguna, 2020)

With the same process or method as before, a calculation is performed to find a branch node of G02- without calculating G01 and G02 values because both symptoms have become nodes. The results of the calculation are shown in Table 9.

Table 9. Node 1.1.2 (G02-) Calculation

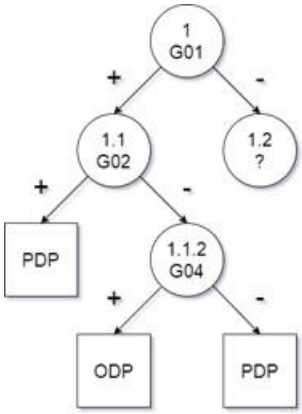
		Cases (S)	PDP (S1)	ODP (S2)	Entropy	Gain Ratio
Total		3	1	2	0.918	
G03						0
	+	0	0	0	0	
	-	3	1	2	0.918	
G04						1
	+	2	0	2	0	
	-	1	1	0	0	
G05						0.274
	+	1	0	1	0	
	-	2	1	1	1	
G06						0.274
	+	1	0	1	0	
	-	2	1	1	1	
G07						1
	+	1	1	0	0	
	-	2	0	2	0	

Source: (Wiguna, 2020)

From the calculation results in table 9 shows that there are two highest Gain Ratio values on G04 and G07 of 1. In this research, the G04 symptom was taken (Rapidminer Studio confirmed) as a branch node of the G02 symptom with negative value (G02-).

In the G04 symptom, there are two attribute values which are positive (+) and negative (-). From these values, the G04 symptom with positive value (G04+) can be seen that those who have 0 cases are in the PDP category, so it can be ascertained that all G04+ decisions will generate the ODP label. While G04 symptom with a negative value (G04-) can be seen that those who have 0 cases are in the ODP category, so it can be ascertained that all G04- decisions will generate the PDP label. Therefore, these two attribute values do not need to be calculated anymore.

A temporary decision tree is illustrated from the results of node 1.1.2 calculation as shown in Fig. 5.



Source: (Wiguna & Riana, 2020)

Figure 5. Node 1.1.2 of Decision Tree

From Fig. 5 it can be seen that the process of finding the next node (node 1.2) uses the reduced data, by filtering G01-.

D. Node 1.2

In a negative branch (-) of G01 symptom must contain another node, so the process of finding another node (node 1.2) will use the data that has been reduced by filtering G01 which is a negative value (G01-) as shown in Table 10.

Table 10. Node 1.2 of COVID-19 Surveillance

No.	G01	G02	G03	G04	G05	G06	G07	Cat.
10	-	+	-	+	+	-	-	ODP
12	-	+	-	+	-	+	-	ODP
13	-	+	-	-	-	-	+	ODP
14	-	-	-	-	-	-	+	OTG

Source: (Wiguna, 2020)

The calculation is performed to find an internal branch node of G01 symptom with a negative value (G01-) without calculating G01 because will be the parent node of node 1.2. Then look for the highest Gain Ratio value to be used as a branch node of G01-. The results of the calculation can be shown in Table 11.

Table 11. Node 1.2 (G01-) Calculation

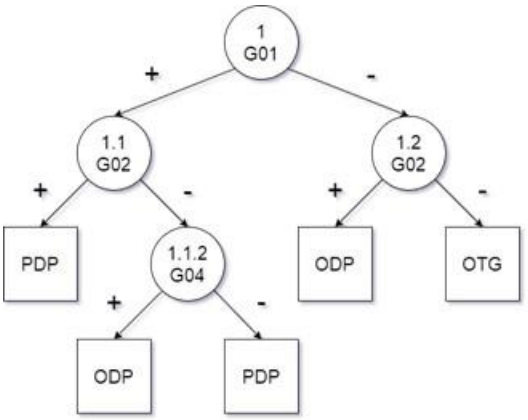
		Cases (S)	ODP (S2)	OTG (S2)	Entropy	Gain Ratio
Total		4	3	1	0.811	
G02						1
	+	3	3	0	0	
	-	1	0	1	0	
G03						0
	+	0	0	0	0	
	-	4	3	1	0.811	
G04						0.311
	+	2	2	0	0	
	-	2	1	1	1	
G05						0.151
	+	1	1	0	0	
	-	3	2	1	0.918	
G06						0
	+	1	1	0	0	
	-	3	2	1	0.918	
G07						0
	+	2	1	1	1	
	-	2	2	0	0	

Source: (Wiguna, 2020)

From the calculation results in table 11 shows that the highest Gain Ratio value is the G02 symptom with a value of 1. So that G02 is an internal branch node of G01 symptom with a negative value (G01-).

G02 symptom has two attribute values which are positive (+) and negative (-). From these values, G02 symptom with a positive value (G02+) can be seen that those who have 0 cases are in the OTG category, so it is ascertained that all G02+ decisions will generate the ODP label. While the G02 symptom with a negative value (G02-) can be seen that those who have 0 cases are in the ODP category, so it is ascertained that all G02- decisions will generate the OTG label. Therefore, these two attribute values do not need recalculating anymore.

The decision tree is illustrated from node 1.2 calculation results, at the same time this is the final process of the decision tree modeling for COVID-19 surveillance categories shown in Fig. 6.



Source: (Wiguna & Riana, 2020)

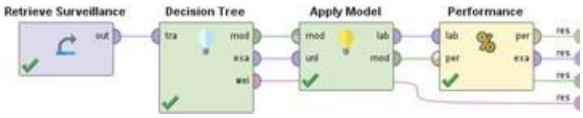
Figure 6. Node 1.2 of Decision Tree

Fig. 6 is a decision tree for the diagnosis of COVID-19 surveillance using the C4.5 algorithm. This model can be explained with the following statements:

1. If G01 is positive and G02 is positive, then the diagnosis is included in the PDP category.
2. If G01 is positive, G02 is negative, and G04 is positive, then the diagnosis is in the ODP category.
3. If G01 is positive, G02 is negative, and G04 is negative, then the diagnosis is in the PDP category.
4. If G01 is negative and G02 is positive, then the diagnosis is in the ODP category.
5. If G01 is negative and G02 is negative, then the diagnosis is in the OTG category.

C4.5 Algorithm Modeling

The actual data classification process from COVID-19 surveillance with the C4.5 algorithm is modeled using Rapidminer Studio shown in Fig. 7.

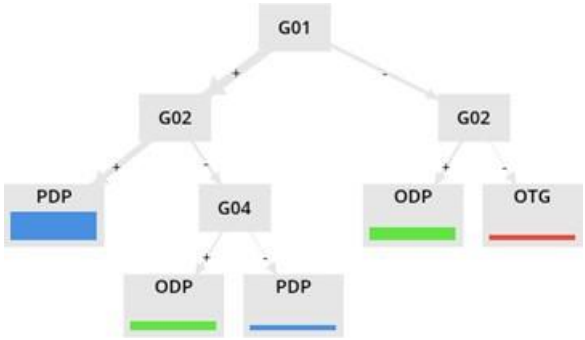


Source: (RapidMiner GmbH, 2019)

Figure 7. Rapidminer Modeling

Fig. 7 shows that the surveillance data was retrieved using the operator retrieve. This data is fed to the decision tree operator by connecting the output port of retrieve to the input port of the decision tree operator. Furthermore, the apply model operator is used to display the visualization of the decision tree results in graph form, while the performance operator will be used for the accuracy testing process.

Rapidminer Studio displays the results of the decision tree in a graph form. Graphs basically mean all visualizations which show nodes and their relationships. These can be nodes within a hierarchical clustering or the nodes of a decision tree for COVID-19 surveillance as in Fig. 8.



Source: (RapidMiner GmbH, 2014)

Figure 8. COVID-19 Surveillance Decision Tree

Meanwhile, the most fundamental form of visualization is that in text form. The decision tree model with the C.45 algorithm in the COVID-19 surveillance is shown in text form as in Table 12.

Table 12. Decision Tree Description

Tree

```
G01 = +
|   G02 = +: PDP {PDP=7, ODP=0, OTG=0}
|   G02 = -
|   |   G04 = +: ODP {PDP=0, ODP=2, OTG=0}
|   |   G04 = -: PDP {PDP=1, ODP=0, OTG=0}
G01 = -
|   G02 = +: ODP {PDP=0, ODP=3, OTG=0}
|   G02 = -: OTG {PDP=0, ODP=0, OTG=1}
```

Source: (Wiguna & Riana, 2020)

The text view form in Table 12 shows that the decision tree description provides information that corresponds to the 5 (five) statements in the final results from the research case.

C4.5 Algorithm Testing

This research has three classes and deals with multi-class problems, so the performance vector calculation is performed while calculating true-negative (TN), true-positive (TP), false-positive (FP), and false-negative (FN) values of each class separately (Ali et al., 2017).

The confusion matrix calculation with three classes in this research are presented in Table 13.

Table 13. Confusion Matrix with 3 Classes

Confusion Matrix		Actual			FP
		Class 1	Class 2	Class 3	
Predicted	Class 1	A = 7	B = 0	C = 0	B+C = 0
	Class 2	D = 1	E = 5	F = 0	D+F = 1
	Class 3	G = 0	H = 0	I = 1	G+H = 0
FN		D+G = 1	B+H = 0	C+F = 0	

▪ True Positive ▪ True Negative ▪ Misclassified

Source: (Ali et al., 2017)

Table 13 that has been formed obtained the performance vector calculation details as follows:

Class 1:

$$Precision = \frac{TP}{TP+FP} = \frac{A}{A+B+C} = \frac{7}{7} = 1$$

$$Recall = \frac{TP}{TP+FN} = \frac{A}{A+D+G} = \frac{7}{7+1+0} = 0.875$$

Class 2:

$$Precision = \frac{TP}{TP+FP} = \frac{E}{E+D+F} = \frac{5}{5+1+0} = 0.833$$

$$Recall = \frac{TP}{TP+FN} = \frac{E}{E+B+H} = \frac{5}{5+0+0} = 1$$

Class 3:

$$Precision = \frac{TP}{TP+FP} = \frac{I}{I+G+H} = \frac{1}{1+0+0} = 1$$

$$Recall = \frac{TP}{TP+FN} = \frac{I}{I+C+F} = \frac{1}{1+0+0} = 1$$

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{A + E + I}{A + E + I + D + G + B + C} = \frac{7 + 5 + 1}{7 + 5 + 1 + 1 + 0 + 0 + 0} = \frac{13}{14} = 0.9286$$

The accuracy testing process of the C4.5 algorithm for the COVID-19 surveillance categories in this research uses Rapidminer Studio with the confusion matrix presented in Table 14.

Table 14. Confusion Matrix of C4.5 Algorithm

accuracy: 92.86%

	true PDP	true ODP	true OTG	precision
pred. PDP	7	0	0	100%
pred. ODP	1	5	0	83.33%
pred. OTG	0	0	1	100%
recall	87.50%	100%	100%	

▪ True Positive ▪ True Negative ▪ Misclassified Cases

Source: (Wiguna & Riana, 2020)

The accuracy value of the C4.5 algorithm calculation for the COVID-19 surveillance categories shown in Table 14 is obtained from the accuracy value percentage in the performance vector that is 0.9286 x 100 = 92.86%.

CONCLUSION

The diagnosis process of the COVID-19 surveillance categories using the C4.5 algorithm shows that the algorithm is able to do classification based on PDP, ODP, and OTG. The classification results of the three categories were successfully modeled into a decision tree. Furthermore, the resulting description tree is in accordance with several diagnosis statements of the decision rules in the research case.

The training data samples used for decision tree modeling in this research have been verified. In the test with a confusion matrix of 3 (three) classes from the COVID-19 surveillance categories, its calculation results in an accuracy of 0.9286 (92.86%). The accuracy of this diagnosis is in the excellent classification category.

For further research, the C4.5 algorithm can be compared with the ID3 algorithm which only uses Entropy and Information Gain values in its calculations (without Split Information & Gain Ratio). In future research, it can even be improved using the J48 algorithm which is a derivative of the C4.5 algorithm in WEKA implementation. In addition, there is also the latest method, namely the C5 algorithm which is oriented towards decision tree modeling.

REFERENCES

Ali, M., Son, D.-H., Kang, S.-H., & Nam, S.-R. (2017). An Accurate CT Saturation Classification Using a Deep Learning Approach Based on Unsupervised Feature Extraction and Supervised Fine-Tuning Strategy. *Energies*, *10*, 1830. https://doi.org/10.3390/en10111830

Amrin, A., Satriadi, I., & Rosanto, O. (2019). ALGORITMA C4. 5 UNTUK DIAGNOSA PENYAKIT TUBERKULOSIS. *Jurnal Khatulistiwa Informatika*, *7*(2).

Anggito, A., & Setiawan, J. (2018). *Metodologi penelitian kualitatif*. CV Jejak (Jejak Publisher).

Anwar, N., Pranolo, A., & Kurnaiwan, R. (2018). Grouping the community health center patients based on the disease characteristics using C4. 5 decision tree. *IOP Conference Series: Materials Science and Engineering*, *403*(1), 12084.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Bairagi, V., & Munot, M. V. (2019). *Research methodology: A practical and scientific approach*. CRC Press.

Buulolo, E. (2020). *Data Mining Untuk Perguruan Tinggi*. Deepublish.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio Books.

Castaño, A. P. (2018). *Practical Artificial Intelligence: Machine Learning, Bots, and Agent Solutions Using C*. Apress.

Efron, S. E., & Ravid, R. (2018). *Writing the literature review: A Practical Guide*. Guilford Publications.

Kemenkes RI. (2020). *Pedoman Pencegahan dan Pengendalian Coronavirus Disease (COVID-19)* (L. Aziza, A. Aqmarina, & M. Ihsan (eds.); Revisi Ke4). Kementerian Kesehatan RI; Direktorat Jenderal Pencegahan dan Pengendalian Penyakit (P2P). https://infeksiemerging.kemkes.go.id/

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

Li, F., Zhang, L., & Zhang, Z. (2017). *Dynamic Fuzzy Machine Learning*. Walter de Gruyter GmbH & Co KG.

Menkes RI. (2020). *Surat Edaran No. HK.02.01/MENKES/2020 tentang Protokol Isolasi Diri Sendiri dalam Penanganan Coronavirus Diseases (COVID-19)*. Kementerian Kesehatan RI. https://covid19.kemkes.go.id/

Muharto & Ambarita, A. (2016). *Metode Penelitian Sistem Informasi: Mengatasi Kesulitan Mahasiswa Dalam Menyusun Proposal Penelitian*. Deepublish. Yogyakarta.

Mujahidin, A., & Pribadi, D. (2017). Penerapan Algoritma C4. 5 Untuk Diagnosa Penyakit Pneumonia Pada Anak Balita Berbasis Mobile. *Jurnal Swabumi*, *5*(2), 155–161.

Mulyani, S. (2017). *Metode Analisis dan Perancangan Sistem*. Abdi Sistematika.

PDPI. (2020). *Pnemonia Covid-19. Diagnosis & Penatalaksanaan di Indonesia*. https://www.klikpdpi.com/

Rafiska, R., Defit, S., & Nurcahyo, G. W. (2018). Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4. 5. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *2*(1), 391–396.

RapidMiner GmbH. (2014). *RapidMiner Studio: Manual*. RapidMiner GmbH. https://doi.org/http://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf

RapidMiner GmbH. (2019). *RapidMiner 9: Operator Reference Manual*. RapidMiner GmbH. https://docs.rapidminer.com/latest/studio/operators/rapidminer-studio-operator-reference.pdf

Santoso, B., & Azis, A. I. S. (2020). *Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner*. Deepublish.

Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach*. John Wiley & Sons.

Tarigan, D. M., Rini, D. P., & Puspita, V. (2017). Perancangan Data Mining untuk Klasifikasi Prediksi Penyakit ISPA dengan Algoritma C4. 5. *Annual Research Seminar (ARS)*, *3*(1), 179–182.

Wiguna, W. (2020). *Decision Tree of Coronavirus Disease (COVID-19) Surveillance*. IEEE Dataport. https://doi.org/10.21227/remc-6d63

Wiguna, W., & Riana, D. (2020). *Laporan Penelitian Dosen Yayasan: Diagnosis of Coronavirus Disease 2019 (COVID-19) Surveillance Using C4.5 Algorithm*. Universitas Adhirajasa Reswara Sanjaya. https://doi.org/10.21227/9anj-0p64