Comparative Analysis of Machine Learning in Predicting the Treatment Status of COVID-19 Patients

Anthony Anggrawan*, Mayadi, Christofer Satria, Bambang Krismono Triwijoyo, and Ria Rismayati
Universitas Bumigora, Mataram, Indonesia; Email: {mayadi.yadot, chris, bkrismono, riris}@universitasbumigora.ac.id
*Correspondence: anthony.anggrawan@universitasbumigora.ac.id

***Abstract*—COVID-19 has become a global pandemic that causes many deaths, so medical treatment for COVID-19 patients gets special attention, whether hospitalized or self-isolated. However, the problem in medical action is not easy, and the most frequent mistakes are due to inaccuracies in medical decision-making. Meanwhile, machine learning can predict with high accuracy. For that, or that's why this study aims to propose a data mining classification method as a machine learning model to predict the treatment status of COVID-19 patients accurately, whether hospitalized or self-isolated. The data mining method used in this research is the Random Forest (RF) and Support Vector Machine (SVM) algorithm with Confusion Matrix and k-fold Cross Validation testing. The finding indicated that the machine learning model has an accuracy of up to 94% with the RF algorithm and up to 92% with the SVM algorithm in predicting the COVID-19 patient's treatment status. It means that the machine learning model using the RF algorithm has more accurate accuracy than the SVM algorithm in predicting or recommending the treatment status of COVID-19 patients. The implication is that RF machine learning can help/replace the role of medical experts in predicting the patient's care status.**

***Keywords*—data mining, random forest, support vector machine, prediction, COVID-19, machine learning**

I. Introduction

The COVID-19 disease is currently a world pandemic [1–3]. It is the cause of the global health crisis [4–6], which is not only because of its high-speed transmission [5, 7], but more than 100 million people have died infected worldwide, and more than two million people have died from it [5]. COVID-19 is a highly contagious viral disease that requires special care and follow-up predictive analytics for better treatment of the disease [8]. However, the COVID-19 pandemic poses a significant challenge to providing health care and services for patients [7]. So it is not surprising that researchers use many research methods to control the COVID-19 pandemic, including the research methods that have received the most attention: prediction, statistical, and epidemiological [6]. Generally, medical actions taken for COVID-19 patients are isolated [9], namely hospitalization or self-isolation. However, these hospitalized COVID-19 patients are receiving intensive medical care from doctors.

In essence, the care status for COVID-19 patients is self-isolation for patients with non-severe illness status and hospitalization for patients who are seriously ill and at critical risk or cause death. Hospitals or medical doctors take various ways to reduce the number of deaths of COVID-19 patients, including by regulating the status of services in hospitals. Expert doctors recommend that patients self-isolate or should be hospitalized after analyzing the patient's medical data. Determining the level of care for COVID-19 patients is a form of medical treatment or treatment for COVID-19 patients to get proper treatment or care.

Errors in decision-making often occur because decision-makers consider several criteria as the basis for decision-making [10]. So it is not surprising that previous researchers emphasized that the errors most often occur due to inaccuracies in decision making [11], and decision making is a difficult task because of the impact of the decisions made [11]. Likewise, in recommending whether a COVID-19 patient should be hospitalized or self-isolated, several criteria from the disease symptoms and the results of medical tests are the basis for considering whether a patient should be hospitalized or self-isolated. In essence, it is difficult to accurately determine the treatment status of COVID-19 patients, both inpatient and self-isolation.

Meanwhile, Machine Learning is a rapidly growing part of computer science today [12]. Although in most scientific studies, machine learning is popular, it is still very limited in health studies [13]. Machine learning helps mining data to predict mining results accurately [14]. Machine learning is a helpful technique for finding correlations based on cases to predict [15]. With the availability of big data, it is possible to develop various solutions using machine learning [16, 17]; moreover, with advances in information and communication technology [18, 19], it is straightforward to collect the required big data. Among the solutions using machine learning, one of which is predictive modeling [20–22]. Furthermore, machine learning can uncover hidden patterns in big data, distinguish patterns better and more

Manuscript received June 14, 2022; revised July 19, 2022; accepted August 4, 2022; published February 13, 2023.

doi: 10.12720/jait.14.1.56-65

accurately [13], and provide high-accuracy prediction results [23]. For this reason (or why), this study's objective is to propose a machine learning system model for decision-making solutions (predictions) for the treatment status of COVID-19 patients, both inpatient and self-isolation, using data mining methods.

The implication is that the proposed machine learning system model can help and even replace the role of medical experts (specialist doctors) in making medical decisions for COVID-19 patients, whether hospitalization or self-isolation. Machine learning performs tasks like a medical specialist in deciding the results of the diagnosis of the nursing status of COVID-19 patients based on medical data of COVID-19 patients. Furthermore, machine learning can work tirelessly, time and place, and has intelligence like an expert, so it is not surprising that previous research confirms that intelligent machines can make superior decisions to experts because humans have a human error factor [24].

Machine learning can predict classification to predict class membership and regression to show numerical values [12]. While data mining is part of machine learning that can make system models have artificial intelligence. Artificial intelligence is a breakthrough in today's technology that has been widely used in prediction [25]. The embodiment of artificial intelligence in machine learning with data mining methods is an iterative process of training and repeated testing of data sets (big data) on the system model. In short, machine learning has an artificial intelligence role in predicting new data with high accuracy [23]. After all, predicting individuals with symptoms of being infected with COVID-19 mandates machine learning (application-based) and contributes to effectively isolating COVID-19 patients [26].

Big data demands large storage media [27]. However, big data is no longer traditionally processed [28]. Instead, today's big data processing relies on machines that can provide systematic results [29]. Big data storage is generally on a computer server with a large storage capacity. Still, some also make it happen by renting online cloud data storage services such as Amazon Simple Storage Service (Amazon S3) and Google Cloud [30]. Cloud facilitates cost-effective big data storage and analysis [30].

Big data processing techniques in data mining include several stages: target data, preprocessed data, data mining, and evaluation/analysis of mining results [30]. Target data and preprocessed data are the processes of extracting raw data from big data [30]. Target data is to select the required data (sample data) and classify data. The preprocessed data is to prepare data sets for data mining, including cleaning up incomplete data, duplicate data, and converting string data into numeric coded data. Finally, data mining and evaluation results extract hidden information by applying data mining methods suitable for the objectives and analyzing them.

Many Data Mining methods include K-Means, Naïve Bayes, KNN (k-Nearest Neighbor), ID3 (Iterative Dichotomiser 3), C4.5, Cart, RF, SVM, and others. There are two types or methods of machine learning, namely supervised machine learning and unsupervised machine learning. It is referred to as a supervised learning method when the subject's membership is known, and training is carried out to classify new data into its category. On the other hand, it is referred to as an unsupervised learning method when the subject's membership is unknown, and the closest distance search is to categorize the groups. The Data Mining methods used in this research are RF and SVM algorithms. RF and SVM are prevalent machine learning algorithms used in various scientific studies [31] and constitute data classification techniques with supervised learning methods. The RF machine learning algorithm has been widely applied for classification [32, 33], as well as SVM algorithm has a widely known technique used for classification [33]. It is why this study uses SVM and RF to classify treatment status. Given that the SVM and RF machine learning algorithms are both popularly used by many researchers, the SVM and RF machine learning algorithms are the most appropriate combination used in research, including research to classify and predict the treatment status of COVID-19 patients.

However, it is essential to know the accuracy of predicting the care status of COVID-19 patients from the system model proposed in this study and whether the patient should be hospitalized or self-isolated. Therefore, this study also further tested the percentage of machine learning efficacy or accuracy in predicting the treatment status of COVID-19 patients. The accuracy of predicting the treatment status of COVID-19 patients is tested on both RF and SVM machine learning methods.

The organization of the following writing of this manuscript is as follows: The second subsection discusses several of the related works of previous researchers and their relevance to the work in this research article. The third subsection describes Research Methodology, which discusses methods used in research in recommending patient care status. Meanwhile, the fourth subsection discusses the results of the study. Finally, it ends with a subsection that concludes the study's findings, the novelty of the research results, and advice for further research.

II. Related Works

This subsection provides an overview of some related works from the latest scientific articles compared with the work in this research article.

Kavzoglu, Bilucan, and Teke (2020) performed the classification of satellite remote sensing images using machine learning algorithms with RF, SVM, and Decision Tree classifier (DT) [31]. This previous research is different from the research in the article on the research objectives and the object under study. In the meantime, Iwendi *et al*. (2020) proposed the Random Forest model to predict the disease severity of COVID-19 patients [34]. The difference between previous research and the research in this article is that the previous research only used one method, namely Random Forest. In contrast, the research in this article used two methods, namely Random Forest and SVM. The difference also lies in the prediction criteria and class; previous research predicts the severity of the illness of COVID-19 patients, while the research in this article predicts the treatment status of COVID-19 patients.

Aroef, Rivan, and Rustam (2020) proposed a machine learning model to classify breast cancer by applying RF and SVM methods [35]. Previous research and the article in this research are both using RF and SVM methods. However, the previous research has research objectives that are not the same as the research in this article. The prior study classified breast cancer as patients with breast cancer. In contrast, the research in this article predicts the treatment status of COVID-19 patients.

Based on patient clinical data, using statistical methods, Hao *et al*. (2020) developed a model to predict pneumonia severity in COVID-19 patients using the Natural Language Processing tool [36]. However, this previous research differs in the research purpose and way compared to the research in this article. Meanwhile, Anggrawan *et al*. (2021) implemented machine learning to diagnose drug users and types of drug-using Forward Chaining and Certainty Factor methods [23]. Meanwhile, the research in this article develops machine learning to predict the patient's treatment status, whether inpatient or self-isolation, based on symptoms or patient medical data using RF and SVM.

Zhao *et al*. (2021) built a model to predict the number of cases of COVID-19 patients in the future using the Poisson and Gamma distribution [37]. Similarities between articles in this study and the previous one proposed a model with a machine learning approach. However, this previous research differs in the research purpose and method compared to the research in this article. In the meantime, Mahboub *et al*. (2021) developed a model to predict the length of hospital stay with the Decision Tree (DT) method [8]. This article's research differs from previous research; the difference lies in the research objectives and techniques used. If prior research predicts the length of stay for COVID-19 patients, the research in this article predicts whether COVID-19 patients should be hospitalized or self-isolated. This article's research does not use the DT method but uses the RF and SVM methods.

Guhathakurata *et al.* (2021) predicted whether a person is infected with COVID-19 or not using SVM [38]. However, this previous study differed in its objectives from this article's research. The previous research predicts patients suffering from COVID-19 or not utilizing the SVM data mining method. In contrast, the research in this article indicates the patient's care status using RF and SVM data mining methods. At the same time, Mehrotraa and Agarwal (2021) reviewed the usefulness of the Data Mining method for the COVID-19 pandemic [39]. This previous research is a literature review study that concludes that the Data Mining method plays an essential role in health care, diagnosing diseases, and recommending cures. However, it is different from this article’s research because it is an experimental study, not a literature review.

Guleria *et al*. (2022) proposed a machine learning model to predict the death rate of COVID-19 patients [14]. However, previous research has different objectives and data mining methods compared to this article's research. The difference is that previous studies examined the infection rate of COVID-19 patients to predict the cure/death rate of COVID-19 patients using the SVM, Decision Trees, and Naïve Bayes data mining methods. In contrast, the research in this article predicts the care status of COVID-19 patients using RF and SVM data mining methods [14].

Anggrawan *et al.* (2022) developed a machine learning model for scholarship recipients' recommendations by using Analytical Hierarchy Process (AHP) and the Multi-Objective Optimization Method by Ratio Analysis (Moora) methods [40]. However, the previous research differs in the purpose and way compared to this article's research. In contrast, Demichev *et al*. (2022) offered a model to optimize the treatment or intensive care of seriously ill COVID-19 patients with plasma proteomics [41]. This previous research is different from the research in the article on the research objectives, research method, and the object under study.

Table I compares some of the most recent previous related work with the work carried out in this study. By referring to the elaboration of the most recent last related work by some researchers, the research carried out in this article has novelties (from the prior research gap) that previous researchers have not studied. In essence, the gap in earlier research is that no one has researched machine learning models to predict the inpatient status or self-isolation of COVID-19 patients by involving RF and SVM algorithms. In addition, the 12 criteria used to indicate the treatment status of COVID-19 patients are entirely different from previous similar studies (as shown in Table I in the Criteria/Attributes column). So, the study's originality lies in proposing a machine learning model to predict the nursing status of COVID-19 patients, whether inpatient or self-isolation, which previous researchers have never done. Besides that, the novelty is also in the method used, not just one data mining method in predicting the treatment status of COVID-19 patients, but using two data mining methods. So this study can show differences in the accuracy of the RF and SVM methods in predicting the treatment status of COVID-19 patients.

III. Methodology

This study applies two data mining methods or machine learning algorithms: RF and SVM. The big data is on COVID-19 patients from a regional hospital in Mataram, Indonesia. The significant data source used in this study is primary data from patient medical records/documents. The attributes of the patient's disease symptoms and care status classes amount to thousands of patient medical record data. Patient datasets containing non-COVID-19 and duplicate and incomplete COVID-19 patient data are removed, so only data is left as a dataset for data mining processes. Medical record data of disease symptoms obtained from string data is then converted into numeric data. The development of the application program in this study uses the Python computer programming language.

TABLE I. COMPARISON OF THIS ARTICLE'S WORK WITH SOME PREVIOUS RELATED WORKS

Research by	**Research methods**			**Criteria/Attributes**		**Research Object**	**Accuracy Test**
	RF	SVM	ML	Number	Name		
Kavzoglu, Bilucan, and Teke (2020) [31]	Yes	Yes	Yes	10	Coastal Aerosol, Blue, Green, Red, Vegetation Red Edge, NIR, Narrow NIR, Water vapor, SWIR-Cirrus, SWIR	Satellite remote sensing images	Yes
Iwendi *et al*. (2020) [34]	Yes	No	Yes	6	Symptom1. symptom2, symptom3, symptom4, symptom5, symptom6	Illness severity of COVID-19 patients	No
Aroef, Rivan, and Rustam (2020) [35]	Yes	Yes	Yes	9	Age, Body Mass Index (BMI), Glucose, Insulin, Homa, Leptin, Adiponectin, Resistin, MCP 1	Breast cancer	Yes
Hao *et al*. (2020) [36]	No	No	No	10	Radiology Opacities, Respiratory Rate, Age, Fever Male, Albumin, Anion Gap, SpO2, LDH, Calcium	The severity of pneumonia in COVID-19 patients	Yes
Anggrawan *et al*. (2021) [23]	No	No	Yes	27	Out of breath, Anxious, Nausea, Diarrhea, Convulsions, Easily angry, Depression, Sleep patterns change, Sweating, Chills, Shaking, Insomnia, Fast heart rate, Blood pressure rises, Difficult to focus, Difficult to rest, Weight loss, Dry mouth, Blurred vision, Changed skin color, Constipation, Stomachache, Drowsiness, Itching, Difficulty urinating, Mood swings, Dizziness	Drug users and types of drug-using	Yes
Zhao *et al*. (2021) [37]	No	No	No	0	-	Number of cases of COVID-19 patients	No
Mahboub *et al*. (2021) [8]	No	No	Yes	5	Age, Gender, Nationality, Blood group, BMI	The treatment period for COVID-19 patient	Yes
Guhathakurata *et al*. (2021) [38]	No	Yes	Yes	8	Temp, Breathing rate, Hypertension, Heartbeat rate (HBR), Acute respiratory disease syndrome (ARDS), Chest pain, Heart disease, Cough with sputum (CWS)	Predicting whether patients are infected with COVID-19 or not	No
Mehrotraa and Agarwal (2021) [39]	No	No	No	0	-	Discussing the data mining method's role in the COVID-19 pandemic	No
Guleria *et al*. (2022) [14]	No	Yes	Yes	0	-	The death rate of COVID-19 patient	Yes
Anggrawan, *et al*. (2022) [40]	No	No	Yes	6	Achievement index, achievement points, recommendation, organizational activity, semester level, and completeness of documents	Scholarship recipient	Yes
Demichev *et al*. (2022) [41]	No	No	No	0	-	Optimization of treatment for COVID-19 patients	Yes
Our research	Yes	Yes	Yes	12	Pneumonia, ARDS, CHF, AKI, CAD, Dyspnea, NSTEMI, ADHF, HHD, Febris, Anosmia, Ageusia	Care status of COVID-19 patients, whether inpatient or self-isolation	Yes

Note: ML = Machine Learning

This research uses a confusion matrix and k-fold cross-validation to measure the classification performance of RF and SVM methods. The data mining process in this research uses CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM is a standard data mining process. The process in CRISP-DM comprises a six-stage [42], as shown in Fig. 1 [43]. Fig. 2 shows the process carried out at each stage of CRISP-DM.

In Figure 2, business understanding is the stage of sorting out thousands of hospital patient medical data to collect the required patient data. The next stage is understanding the data collected as representative data for COVID-19 patients. This COVID-19 patient data classifies the patient's signs and symptoms and treatment status, which needs further processing at the next stage. The next stage is the data preparation stage, which essentially determines the attributes of the names of signs and symptoms of COVID-19 patients. The embodiment of the dataset containing knowledge according to treatment status refers to the signs and symptoms that the patient has (marked with Yes) or does not have (marked with No). The next thing to do is preprocess the dataset, changing the category value of the symptom attribute and the class attribute with the number 1 and the number −1. The process of extracting raw data obtained in the previous process stage is data that is further processed by data mining methods or used as learning machine learning data at the modeling stage. So that machine learning can predict. The next stage is the evaluation stage, namely, knowing the predictive reliability of the data mining or machine learning method. Then the last stage is the deployment stage to disseminate research results so that they are helpful for implementation by various parties, especially hospitals and other professionals, in the form of developing application programs and scientific articles.

This research uses a confusion matrix and k-fold cross-validation to measure the classification performance of RF and SVM methods.

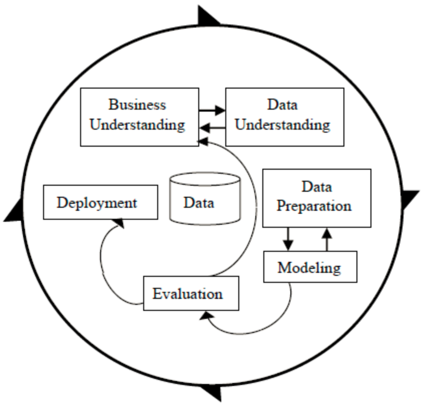

Figure 1. The CRISP-DM process.

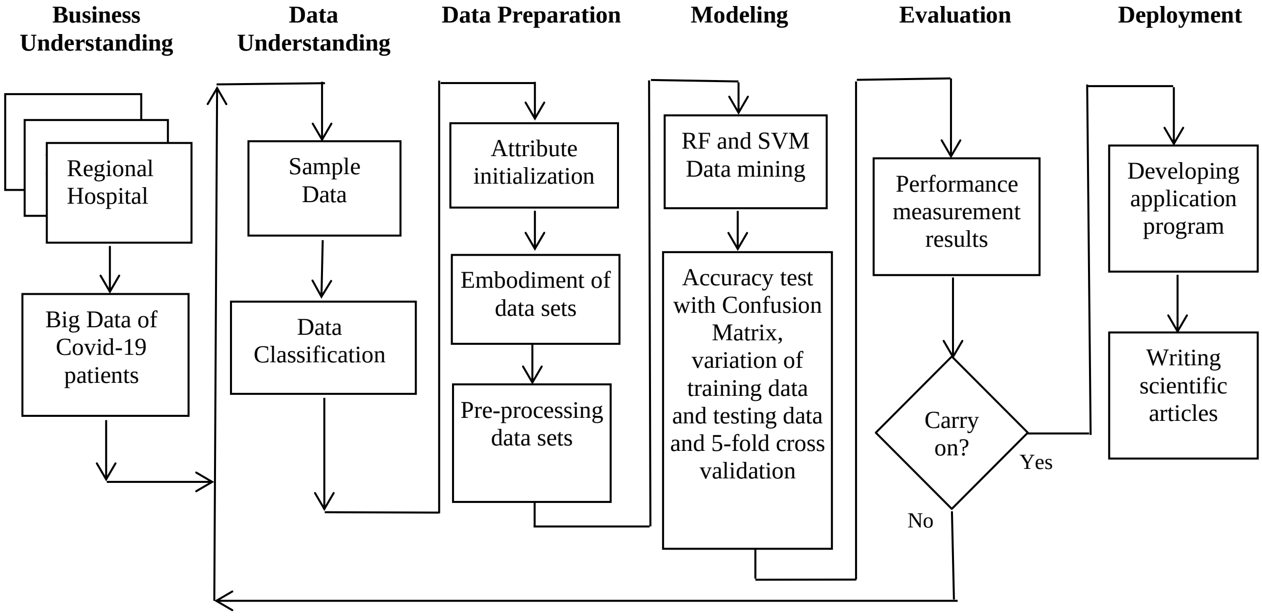

Figure 2. Data mining process of COVID-19 patient big data.

IV. Result and Discussion

A. Business Understanding

The significant data acquisition of COVID-19 patients needed for research is obtained from the hospital. The data collected from medical document data from all COVID-19 patients registered at the hospital includes the patient's name, disease symptoms, and the treatment status specified. There are thousands of data on COVID-19 patients. The patient dataset containing incomplete data and non-COVID-19 was omitted or ignored. The critical information extracted at this stage is first to find the attributes or criteria of the class of treatment status (hospitalization or self-isolation); second, to find the category of each feature of the treatment status class. The existing attributes and categories represent disease symptoms and other medical data from COVID-19 patients. Based on COVID-19 patient data adopted from the hospital, there are 12 symptom criteria or patient medical data that are used as references by expert doctors in determining the status of patient care, whether to be hospitalized or self-isolated. Furthermore, big data containing several symptom criteria or patient medical data is used in training and testing the prediction model proposed in this study. Therefore the offered machine learning model has artificial intelligence in predicting.

TABLE II. Data Set of the Signs and Symptoms and the Treatment Status of COVID-19 Patients

No	Disease Sign and Symptom	Treatment
1	Pneumonia, Dyspnea	Inpatient
2	Pneumonia, ARDS, AKI, Febris	Inpatient
3	Pneumonia, CHF, CAD, Dyspnea	Inpatient
4	Pneumonia, AKI	Inpatient
5	Pneumonia, CAD	Inpatient
6	Pneumonia, Dyspnea, Anosmia, Ageusia	Inpatient
7	CHF, NSTEMI	Inpatient
..	…..	….
..	…..	….
114	Febris, Anosmia, Ageusia	Self-isolation
115	Pneumonia, Anosmia	Inpatient
116	Pneumonia, Anosmia, Ageusia	Inpatient
117	HHD	Inpatient

B. *Data Understanding*

The Data Understanding stage is preparing the data set from the research. The dataset from this study is a data representation of the COVID-19 patient sample, which contains sign and symptom data and treatment status. Table II shows the association between signs and symptoms of disease and treatment status in the study data set.

C. *Data Preparation*

Each patient confirmed positive for COVID-19 has a different diagnosis from others, and some patients have similar diagnoses. There were 117 patients with COVID-19 who had a different diagnosis from the others. In this study, the number of signs and symptoms or the number of research criteria is 12 signs and symptoms, or the number of research criteria is 12 signs and symptoms or 12 criteria (see Table III).

TABLE III. DATA SET RELATED TO RESEARCH ATTRIBUTES AND DISEASE SIGNS AND SYMPTOMS

Attribute	Sign and Symptoms	Word extension
G01	Pneumonia	Pneumonia
G02	ARDS	Acute Respiratory Distress Syndrome
G03	CHF	Congestive Heart Failure
G04	AKI	Acute Kidney Injury
G05	CAD	Coronary Artery Disease
G06	Dyspnea	Dyspnea
G07	NSTEMI	Non-ST-Segment Elevation Myocardial Infarction
G08	ADHF	Acute Decompensated Heart Failure
G09	HHD	Hypertensive Heart Disease
G10	Febris	Febris
G11	Anosmia	Anosmia
G12	Ageusia	Ageusia

The signs and symptoms of each COVID-19 patient (G01, G02, ..., G12 or Gi where *i* = 1, 2, 3 ..., 12) are not all the same from one patient to another. For this reason, the attributes of each patient's data are different, and some are the same between one patient and another, as shown in Table IV. If the sign or symptom attribute is No, the patient does not have these signs or symptoms. On the other hand, if the sign or symptom attribute is Yes, the patient has these signs or symptoms.

TABLE IV. DATA SET OF KNOWLEDGE BASED ON TREATMENT STATUS REFERRING TO THE SIGNS AND SYMPTOMS

No	G01	G02	G03	G04	…	…	G11	G12	Class
1	Yes	No	No	No	…	…	No	No	Inpatient
2	Yes	Yes	No	Yes	…	…	No	No	Inpatient
3	Yes	No	Yes	No	…	…	No	No	Inpatient
4	Yes	No	No	No	…	…	No	No	Inpatient
5	Yes	No	No	No	…	…	No	No	Inpatient
6	Yes	No	No	No	…	…	Yes	Yes	Inpatient
7	No	No	Yes	No	…	…	No	No	Inpatient
..	…	…	…	…	…	…	…	…	…
..	…	…	…	…	…	…	…	…	…
114	No	No	No	No	…	…	No	Yes	Self-isolation
115	Yes	No	No	No	…	…	Yes	No	Inpatient
116	Yes	No	No	No	…	…	Yes	Yes	Inpatient
117	No	No	No	No	…	…	No	No	Inpatient

Furthermore, the preprocessing of the data set is done by changing the Gi with xi and the Gi Yes attribute value with the number 1 while the Gi No attribute with the number −1. In addition, dataset preprocessing is also carried out on class attributes, namely changing the independent isolation class attribute category with the number 1 and the inpatient class attribute category with the number −1, as shown in Table V.

TABLE V. PREPROCESSING OF THE DATA SET RESULT

No	G01	G02	G03	G04	…	…	G11	G12	Class
1	1	−1	−1	−1	…	…	−1	−1	−1
2	1	1	−1	1	…	…	−1	−1	−1
3	1	−1	1	−1	…	…	−1	−1	−1
4	1	−1	−1	−1	…	…	−1	−1	−1
5	1	−1	−1	−1	…	…	−1	−1	−1
6	1	−1	−1	−1	…	…	1	1	−1
7	−1	−1	1	−1	…	…	−1	−1	−1
..	…	…	…	…	…	…	…	…	…
..	…	…	…	…	…	…	…	…	…
114	−1	−1	−1	−1	…	…	−1	1	1
115	1	−1	−1	−1	…	…	1	−1	−1
116	1	−1	−1	−1	…	…	1	1	−1
117	−1	−1	−1	−1	…	…	−1	−1	−1

D. *Modeling*

The proposed machine learning model to predict COVID-19 treatment status in this study applies the RF and SVM data mining classification methods. In addition, known various programming language [44], which has their respective advantages in building application programs [45, 46]. The application program built in this research uses the Python programming language to facilitate patient care status prediction (as shown in Table VI).

TABLE VI. USE OF HYPERPARAMETER ON SVM AND RF METHOD

Classifier Method	Hyperparameter	Value
SVM	C	1
	Kernel	Rbf
	Degree	3
	Gamma	Scale
	Coef	0
	Tol	0,0001
	Max_iter	−1
RF	n_estimators	100
	Criterion	Gini
	Max_depth	None
	Min_samples_split	2
	Min_samples_leaf	1

1) *SVM data mining method*

The process of realizing the classification using the SVM data mining method is as follows: 1) Forming a linear equation from the training data that has gone through the preprocessing stage; 2) Finding the values of w and b by means of elimination and substitution of linear equations; 3) Finding the value of the classification decision with the function.

In SVM, there are two implementation models: mathematical programming techniques and kernel functions. This study applies kernel functions and focuses on classifying two categories of class attributes. The class attribute is a treatment for $yi = +1, -1$. The formula of the SVM data mining method is: 1) to form a linear equation from the training data; 2) find the value of w and b, and 3) the value of the classification decision is as follows.

$$S = ((x1, y1), \dots, (xl, yl)) \quad (1)$$

$$yi\ ((w.xi) + b) \geq 1, i = 1 \quad (2)$$

$$(x) = w.x + b \quad (3)$$

Description:

S= set; *x* = attribute; *y* = class; *w* = weight; *b* = bias

2) RF data mining method

The process of realizing the classification using the RF data mining method is as follows: 1) Generating a random subset of data; 2) Creating a decision tree (Root tree, branch tree & leaves tree) from each attribute and class; and 3) Testing each decision tree with data testing and calculating the accuracy of each decision tree.

RF uses bootstrap samples from training data to create a tree from a randomly selected subset. The chosen predictor is a candidate for splitting the decision tree. The results of the category predictions from the treatment class based on the results of the highest voting were chosen as the final prediction results. The formula for the RF data mining method is the Gini criterion and the Entropy criterion:

$$Gini = 1 - \sum_{i=1}^{c} pi^2 \quad (4)$$

$$Entropy(S) = \sum_{i=1}^{c} -pi \text{ x } log_2(pi) \quad (5)$$

Description:

S = Set of cases

pi = the proportion of case *i* to the Set of cases

3) Confusion matrix

This research uses a confusion matrix to measure the performance of the classification method. The confusion matrix is a method that can be used to measure the performance of a classification method. In essence, the confusion matrix can produce information by comparing the system's classification results with the classification results that should be.

In measuring performance using the Confusion Matrix, four terms represent the results of the classification process, namely: True Positive (*TP*) or positive data detected correctly; False Positive (*FP*) or negative data detected is positive; True Negative (*TN*) or negative data detected correctly, and False Negative (*FN*) or positive data detected is negative. Meanwhile, the calculation of accuracy, prediction, and Recall in the confusion matrix can use the following equation:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall= \frac{TP}{TP+FN} \quad (8)$$

Accuracy states the closeness of the measurement results to the actual value, while Precision shows how close the difference in the measurement results is on repeated measurements. On the other hand, recall states the level of success in retrieving information. Precision and Recall are necessary because Precision denotes a measure of quality, and Recall denotes a measure of quantity.

Measurement of accuracy is based on the ratio between the correct predictions (positive and negative) with the overall data. In contrast, precision measurements are based on the percentage of true positive predictions compared to overall positive predicted outcomes. Meanwhile, the recall measurement is based on the ratio of true positive predictions compared to the general actual positive data.

The format of the confusion matrix table is as shown in Table VII. The results of the predictions of the SVM and RF methods are shown in Tables VIII and IX.

TABLE VII. CONFUSION MATRIX

Class	Classified Positive	Classified Negative
Positive	True Positive	False Negative
Negative	False Negative	True Positive

TABLE VIII. CONFUSION MATRIX OF SVM

		Prediction	
Actual	Class	Self-isolation	Inpatient
	Self-isolation	4	0
	Inpatient	2	12

TABLE IX. CONFUSION MATRIX OF RF

		Prediction	
Actual	Class	Self-isolation	Inpatient
	Self-isolation	4	0
	Inpatient	1	13

4) K-fold cross-validation

This study used K-fold cross-validation to measure the performance of the classification method. K-fold cross-validation helps assess the performance of data mining methods by dividing the data sample randomly and grouping the data as much as the k-fold value. In the performance testing of this study with k-fold cross-validation, the dataset is partitioned into five subsets (k = 5). It allows each subgroup to have the same number and fold, which refers to the number of resulting subsets. Dataset partitioning is done by taking random samples from the dataset. However, data that has been taken previously will not be retrieved.

In the first fold, the first subset serves as the validate set (Dval), and the remaining four subsets serve as the training set (Dtrain). In the second fold, the second subset is the validate set, the remaining subset is the training set, and so on until the 5th fold.

E. Evaluation

The evaluation of the proposed model in this study is to measure the performance of the resulting prediction system model. The model's performance evaluation is based on the prediction system model generated by the RF and SVM methods.

1) Evaluation of prediction model with confusion matrix

Evaluation of the prediction results of the proposed system model uses the confusion matrix technique. The evaluation result using the confusion matrix is shown in Table X and Fig. 3. The accuracy in predicting with 85% of training data and 15% of test data shows that the RF machine learning method is more accurate and precise than the SVM machine learning method.

TABLE X. SYSTEM MODEL PERFORMANCE TESTING WITH 85% OF TRAINING DATA AND 15% OF TESTING DATA

Method	**Accuracy**	**Precision**	**Recall**
SVM	89%	83%	93%
RF	94%	90%	96%

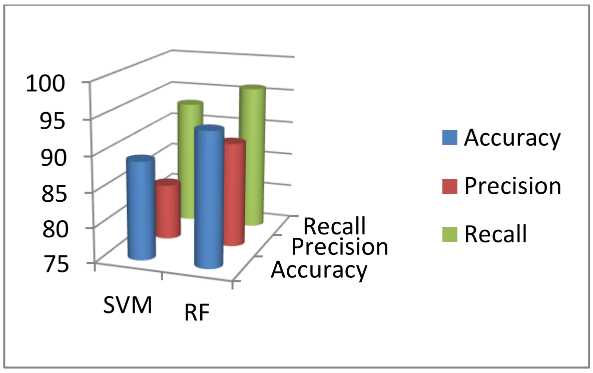

Figure 3. System model performance testing with 85% of training data and 15% of testing data.

Further comparison of the accuracy and precision of the prediction system model with 50% training data and 50% testing data, 60% training data and 40% testing data, 70% training data and 30% testing data, 80% training data and 20% data 90% testing and training data and 10% testing data are as shown in Table XI.

TABLE XI. PREDICTION SYSTEM MODEL PERFORMANCE TESTING WITH VARIOUS TEST DATA AND TRAINING DATA VARIATIONS

Data (in %)		**Accuracy (in %)**		**Precision (in %)**		**Recall (in %)**	
Training	Testing	RF	SVM	RF	SVM	RF	SVM
50	50	95	97	97	96	91	96
60	40	96	91	97	90	93	90
70	30	94	92	96	91	92	90
80	20	96	92	94	89	97	94
90	10	92	83	75	67	95	91
Average		95	91	92	87	94	92

Predicting with various test data and training data variations shows that the RF machine learning method is more accurate and precise than the SVM machine learning method. In other words, the prediction system model proposed to predict the treatment status of COVID-19 patients using the RF method is better (more accurate and precise) than the SVM machine learning method based on performance tests with a confusion matrix.

2) Evaluation of prediction model with k-fold cross-validation

The performance of the model proposed in this study uses a 5-fold cross-validation on both RF and SVM prediction models presented in Table XII and Fig. 4.

TABLE XII. PREDICTION PERFORMANCE TESTING WITH K-FOLD CROSS-VALIDATION

RF (in %)	**SVM (in %)**
98.290	97.436

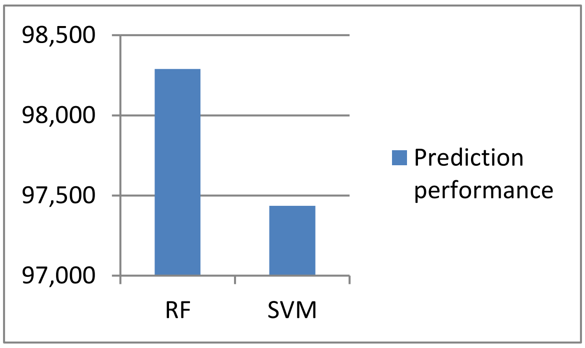

Figure 4. Predictive performance testing using K-fold cross-validation.

F. Deployment

One of the deployments in this research is making scientific articles on machine learning system models that are produced to be published in reputable scientific papers. Thus, the results obtained can be developed and become the knowledge of many parties as a responsibility for the correctness of the effects of research carried out as professional researchers. Another form of deployment is to make reports to cooperative hospital partners where data on COVID-19 patients is obtained.

V. CONCLUSION

This study found that the prediction system model for the treatment status of COVID-19 patients using the RF machine learning method had better predictive performance than the SVM machine learning method. The test of accuracy and precision in predicting the treatment status of COVID-19 patients using the confusion matrix showed that the RF machine learning method has a prediction accuracy of 94% and a precision of 90%; In comparison, the SVM machine learning method has a prediction accuracy of 89% and a precision of 83%. Further testing of the accuracy of the system model in predicting the treatment status of COVID-19 patients using k-fold cross-validation showed that the RF machine learning method had a prediction accuracy of 98.290% and the SVM machine learning method had a prediction accuracy of 97.436%. The research result implication is that RF machine learning can help or replace the role of medical personnel in predicting the treatment status of COVID-19 patients, whether inpatient or self-isolation, with high accuracy.

The novelty of this study is to propose a system model for predicting the treatment status of COVID-19 patients, whether inpatient or self-isolation, which researchers have never studied before using two machine learning methods of RF and SVM.

Further research needs to develop a machine learning system model to predict the death or recovery status of each COVID-19 patient. Another suggestion for future study is: to conduct further research using other data mining methods to predict patient care status and the status of death or recovery from COVID-19 patients and various other diseases, to build a system that not only predicts but also performs clustering, association, and estimates of various other fields of science, including patients' care status, with a combination of machine learning and the Internet of Things.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors undertake work assignments to complete the research and writing of this article jointly. Anthony Anggrawan wrote the entire manuscript including writing research background, research analysis and research related works, research methodology and conclusions including supporting references. Mayadi has collected data and completed the necessary calculations, including calculating the accuracy of the results of the research method. Christofer Satria has worked on edited and visualized graphs and tables, including helping the Mayadi’s work. Bambang Krismono Triwijoyo checked to provide additional technical input needed in the manuscript including checking the correctness of the sentence structure and grammar in the writing in the manuscript, while Ria Rismayati rechecked the contents of the manuscript. All authors had approved the final version.

References

[1] C. Rothe, M. Schunk, P. Sothmann, *et al.*, “Transmission of 2019-nCoV infection from an asymptomatic contact in germany,” *N. Engl. J. Med.*, vol. 382, no. 10, pp. 970-971, 2020.

[2] Q. Li, X. Guan, P. Wu, *et al.*, “Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia,” *N. Engl. J. Med.*, vol. 382, no. 13, pp. 1199-1207, 2020.

[3] L. Zou, F. Ruan, M. Huang, *et al.*, “SARS-CoV-2 Viral load in upper respiratory specimens of infected patients,” *N. Engl. J. Med.*, vol. 382, no. 12, pp. 1171-1179, 2020.

[4] M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, *et al.*, “Identification of high-risk COVID-19 patients using machine learning,” *PLoS One*, vol. 16, no. 9, pp. 1-21, September 2021.

[5] C. Wang, Z. Wang, G. Wang, *et al.*, “COVID-19 in early 2021: Current status and looking forward,” *Signal Transduct. Target. Ther.*, vol. 6, no. 1, 2021.

[6] J. Nayak and B. Naik, “Intelligent system for COVID-19 prognosis: A state-of-the-art survey,” *Appl. Intell.*, vol. 51, no. 5, pp. 2908-2938, 2021.

[7] C. Huang, Y. Wang, X. Li, *et al.*, “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” *Lancet*, vol. 395, no. 10223, pp. 497-506, 2020.

[8] B. Mahboub, M. T. A. Bataineh, H. Alshraideh, *et al.*, “Prediction of COVID-19 hospital length of stay and risk of death using artificial intelligence-based modeling,” *Front. Med.*, vol. 8, pp. 1-9, 2021.

[9] T. Singhal, “A review of coronavirus disease-2019 (COVID-19),” *Indian J. Pediatr.*, vol. 87, pp. 281-286, April 2020.

[10] P. H. D. Santos, S. M. Neves, D. O. Sant'Anna, *et al.*, “The analytic hierarchy process supporting decision making for sustainable development: An overview of applications,” *J. Clean. Prod.*, vol. 212, pp. 119-138, 2019.

[11] D. Laureiro-Martínez and S. Brusoni, “Cognitive flexibility and adaptive decision-making: Evidence from a laboratory study of expert decision-makers,” *Strateg. Manag. J.*, vol. 39, no. 4, pp. 1031-1058, 2018.

[12] D. Prasad, S. K. Goyal, A. Sharma, *et al.*, “System model for prediction analytics using k-nearest neighbors algorithm,” *J. Comput. Theor. Nanosci.*, vol. 16, no. 10, pp. 4425-4430, 2019.

[13] Z. Zhang, “Introduction to machine learning: K-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, pp. 1-7, 2016.

[14] P. Guleria, S. Ahmed, A. Alhumam, *et al.*, “Empirical study on classifiers for earlier prediction of COVID-19 infection cure and death rate in the Indian states,” *Healthc.*, vol. 10, no. 1, 2022.

[15] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, “Data mining and machine learning models for predicting drug likeness and their disease or organ category,” *Front. Chem.*, vol. 6, pp. 1-11, May 2018.

[16] J. Bullock, A. Luccioni, K. H. Pham, *et al.*, “Mapping the landscape of artificial intelligence applications against COVID-19,” *J. Artif. Intell. Res.*, vol. 69, pp. 807-845, 2020.

[17] L. Wynants, B. V. Calster, G. S. Collins, *et al.*, “Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal,” *BMJ*, vol. 369, no. m1328, pp. 1-16, 2020.

[18] A. Anggrawan, A. H. Yassi, C. Satria, *et al.*, “Comparison of online learning versus face to face learning in english grammar learning,” in *Proc. the 5th International Conference on Computing Engineering and Design*, 2018, pp. 1-4.

[19] A. Anggrawan, “Interaction between learning preferences and methods in face-to-face and online learning,” *ICIC Express Lett.*, vol. 15, no. 4, pp. 319-326, 2021.

[20] H. Zhang, L. L. Wang, Y. Y. Chen, *et al.*, “A tool to early predict severe corona virus disease 2019 (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China,” *Cancer*, vol. 46, pp. 1-17, May 2020.

[21] X. Jiang, M. Coffee, A. Bari, *et al.*, “Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity,” *Comput. Mater. Contin.*, vol. 63, no. 1, pp. 537-551, 2020.

[22] J. Liu, Y. Liu, P. Xiang, *et al.*, “Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage,” *J. Transl. Med.*, vol. 18, no. 1, pp. 1-12, 2020.

[23] A. Anggrawan, C. Satria, C. K. Nuraini, *et al.*, “Machine learning for diagnosing drug users and types of drugs used,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 111-118, 2021.

[24] A. Agrawal, J. S. Gans, and A. Goldfarb, “Exploring the impact of artificial Intelligence: Prediction versus judgment,” *Inf. Econ. Policy*, vol. 47, pp. 1-6, 2019.

[25] K. Shankar, A. R. W. Sait, D. Gupta, *et al.*, “Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model,” *Pattern Recognit. Lett.*, vol. 133, pp. 210-216, 2020.

[26] C. W. Yancy, “From mitigation to containment of the COVID-19 pandemic: Putting the SARS-CoV-2 genie back in the bottle,” *JAMA - J. Am. Med. Assoc.*, vol. 323, no. 19, pp. 1891-1892, 2020.

[27] V. C. Storey and I. Y. Song, “Big data technologies and Management: What conceptual modeling can do,” *Data Knowl. Eng.*, vol. 108, pp. 50-67, February 2017.

[28] A. Ribeiro, A. Silva, and A. R. D. Silva, “Data modeling and data analytics: A survey from a big data perspective,” *J. Softw. Eng. Appl.*, vol. 8, no. 12, pp. 617-634, 2015.

[29] B. Suvarnamukhi and M. Seshashayee, “Big data concepts and techniques in data processing,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 712-714, 2018.

[30] Y. Li, L. Guo, and Y. Guo, “An efficient and performance-aware big data storage system,” *Communications in Computer and Information Science*, vol. 367, pp. 1-17, April 2018.

[31] T. Kavzoglu, F. Bilucan, and A. Teke, "Comparison of support vector machines, random forest and decision tree methods for classification of sentinel - 2A image using different band combinations," in *Proc. 41st Asian Conf. Remote Sens.*, November 2020, pp. 1-9.

[32] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review," *Front. Aging Neurosci.*, vol. 9, pp. 1-12, 2017.

[33] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, *et al.*, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 6308-6325, 2020.

[34] C. Iwendi, A. K. Bashir, A. Peshkar, *et al.*, "COVID-19 patient health prediction using boosted random forest algorithm," *Front. Public Heal.*, vol. 8, pp. 1-9, 2020.

[35] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *Telecommun. Comput. Electron. Control*, vol. 18, no. 2, pp. 815-821, 2020.

[36] B. Hao, S. Sotudian, T. Wang, *et al.*, "Early prediction of level-of-care requirements in patients with COVID-19," *Elife*, vol. 9, pp. 1-23, 2020.

[37] H. Zhao, N. N. Merchant, A. McNulty, *et al.*, "COVID-19: Short term prediction model using daily incidence data," *PLoS One*, vol. 16, no. 4, pp. 1-14, April 2021.

[38] S. Guhathakurata, S. Kundu, A. Chakraborty, *et al.*, "A novel approach to predict COVID-19 using support vector machine," *Data Sci. COVID-19*, pp. 351-364, 2020.

[39] A. Mehrotra and R. Agarwal, "A review of use of data mining during COVID-19 pandemic," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 6, pp. 4547-4552, 2021.

[40] A. Anggrawan, Mayadi, C. Satria, *et al.*, "Scholarship recipients recommendation system using AHP and moora methods," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 260-275, 2022.

[41] V. Demichev, P. Tober-Lau, T. Nazarenko, *et al.*, "A proteomic survival predictor for COVID-19 patients in intensive care," *PLOS Digit. Heal.*, vol. 1, no. 1, pp. 1-17, 2022.

[42] O. Marban, G. Mariscal, and J. Segovia, "A data mining & knowledge discovery process model," in *Austria: Data Mining and Knowledge Discovery in Real Life Applications*, IntechOpen, 2009, p. 438.

[43] P. Chapman, J. Clinton, R. Kerber, *et al.*, *Crisp-Dm 1.0: Step-by-Step Data Mining Guide*, SPSS Inc, January 2000.

[44] A. Anggrawan, C. Satria, Mayadi, *et al.*, "Reciprocity effect between cognitive style and mixed learning method on computer programming skill," *J. Comput. Sci.*, vol. 17, no. 9, pp. 814-824, 2021.

[45] A. Anggrawan, C. K. Nuraini, Mayadi, *et al.*, "Interplay between cognitive styles and gender of two hybrid learning to learning achievements," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 10, pp. 2404-2413, 2021.

[46] A. Anggrawan, S. Hadi, and C. Satria, "IoT-Based garbage container system using NodeMCU ESP32 microcontroller," *J. Adv. Inf. Technol.*, vol. 13, 2022.

Anthony Anggrawan currently works as an associate professor in the Department of Computer Science as a lecturer, university Rector, and State Civil Apparatus at Bumigora University, Indonesia. In addition, he is a reviewer of articles in several reputable international scientific journals. He received his Master in Computer Science Information Technology (M.T) from the 10 November Institute of Technology, Surabaya, Indonesia. After that, he earned his first Doctoral degree (Ph.D.) in Accounting Information Systems from Universiti Utara Malaysia. Then, he received his second Doctoral degree (Dr.) from Hasanuddin University, Makassar, Indonesia, in linguistics. Finally, he earned his third Doctorate in Educational Technology from the State University of Jakarta. His research interests include Information Technology, Data Mining, Machine Learning, Online Learning, and the Internet of Things. From 2016 until now, he has been the chairman of the Association for Higher Education in Informatics and Computers (APTIKOM) for the West Nusa Tenggara region, Indonesia.

Mayadi obtained a bachelor's degree (S.Kom) in Computer Science from Bumigora University, Indonesia, and a master's degree (M.Kom) in Informatics Engineering from Amikom University, Indonesia. Currently a lecturer in the Computer Science Study Program at Bumigora University, Indonesia, and as Vice Chancellor III. His research interests are Artificial intelligence, Machine Learning, Data Mining, Big Data, Deep learning, the Internet of Things, and Data Science.

Christofer Satria received a bachelor's degree (S.Sn) in Visual Communication Design from Petra Christian University, Surabaya, Indonesia, and a master's degree (M.Sn) in Visual Communication Design from the Indonesian Art Institute (ISI) Denpasar, Bali, Indonesia. He is currently a lecturer in the Visual Communication Design Study Program at Bumigora University, Indonesia, and the head of the laboratory in photography, animation, and video. His research interests include animation learning media, video learning media, education method, Data Mining, and experimental Design. He is currently pursuing a doctorate in the same area as his expertise.

Bambang Krismono Triwijoyo currently works as an assistant professor in the Department of Computer Science as a lecturer and head of the quality assurance institute at Bumigora University, Indonesia. In addition, he is a reviewer of articles in several reputable international scientific journals. He completed his post-graduation in Computer Science from the Sepuluh November Institute of Technology, Surabaya, Indonesia, in 2003. Doctoral degree in Computer Science from Bina Nusantara University, Jakarta, Indonesia, in 2021. His research interests are Computer Vision, Image Processing, and Machine Learning. He is also active in several professional organizations, including IEEE Computer Science, the Indonesian Informatics Experts Association (IAII), and the Indonesian Pattern Recognition Association (INAPR).

Ria Rismayati obtained a bachelor's degree (S.Kom) in Informatics Engineering from the College of Management and Computer Informatics, Mataram, Indonesia, and a master's degree (M.Kom) from AMIKOM University, Yogyakarta, Indonesia. Currently, she is a lecturer in the Computer Science study program at Bumigora University, Indonesia. Her research interests include Data Mining, Artificial Intelligence, Enterprise Architecture, and Information Systems.