Received: May 12, 2020. Revised: June 5, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Optimization of Machine Learning Algorithms for Predicting Infected COVID-19 in Isolated DNA

Berlian Al Kindhi[1]*

[1]*Institut Teknologi Sepuluh Nopember, Indonesia*
* Corresponding author's Email: berlian@its.ac.id

Abstract: The stipulation of the COVID-19 (Corona Virus Disease 2019) as a global pandemic by the WHO (World Health Organization) made a number of countries lockdown. Countries like Italy, Denmark, China, and Ireland have taken lockdown steps to prevent this disease from spreading and taking many lives. COVID-19, SARS (Severe Acute Respiratory Syndrome), and MERS (Middle-East Respiratory Syndrome) are viral infections in the respiratory tract that can be fatal. SARS first became an epidemic in China in 2002, while MERS first appeared in the Middle East in 2012. At the end of 2019, a new disease appeared in China called COVID-19. These three viruses are still in the same family so they have very similar nucleotide sequences. The tested COVID-19 primer was able to adhere well with a similarity level of more than 70% in all DNA SARS and MERS isolates tested. To distinguish DNA samples between MERS, SARS, and COVID-19 using the basic local alignment sequence nucleotide approach alone is not enough. We propose an optimization of machine learning methods to predict the COVID-19, the optimization method depends on the method we improved. In Discriminant Analysis, we use Wilks Lamda's approach and change Linear into Diagonal Discriminant Matrix. In the Decision Tree method, we make optimization by making gain formulation to minimize the entropy value to get more information on the result. We optimized K-NN with add weighted distance optimization, and in SVM we try several kernels and optimize the hyperplane with SRM (Structural Risk Minimization) approach to looking for the best result. Besides that, in preparation for input features, we use Edit Levenshtein Method with the calculation of the optimum similarity from each DNA sequence. The results of our test, optimization of the Decision Tree method produces an accuracy of 98.3%, optimization of Discriminant Analysis 98.3%, and optimization of SVM and KNN 100% respectively. We also find a fact in the DNA Alignment process, when the primer being compared is 'R', the nucleotides in the COVID-19 sample data are always 'A' and this approach from the bioinformatic side can be used as analytical material in the medical world.

Keywords: COVID-19, Discriminant analysis, K-NN, Decision Tree, SVM, DNA.

1. Introduction

Since the discovery of a new type of Coronavirus at the end of 2019 which call COVID-19, the number of infected patients has increased significantly by March 2020. US reports the largest number of deaths worldwide, followed by Italy. This study conducts trials and analysis of the proximity of MERS, COVID-19, and SARS in terms of DNA nucleotide patterns that can be used as decision support in biomedical research. The incubation period is the time needed by germs to multiply in a person's body to cause complaints. In other words, the incubation period is the time span between the occurrence of infection and the appearance of symptoms [1]. Although the viruses COVID-19, SARS, and MERS are from the same family of viruses, namely coronavirus, these three diseases have different incubation periods, for SARS disease is 1–14 days (average 4-5 days). The incubation period for MERS disease is 2–14 days (average 5 days), while the incubation period for COVID-19 is 1–14 days, with an average of 5 days.

These three diseases can cause fever, cough, sore throat, nasal congestion, weakness, headaches, and muscle aches. If it gets worse, the symptoms of the

three can resemble pneumonia. The big difference between these three diseases is that COVID-19 is rarely accompanied by colds and digestive complaints, such as bowel movements, nausea, and vomiting. The spread of coronavirus from animals to humans is actually very rare, but this is what happened to COVID-19, SARS, and MERS. Humans can get the coronavirus through direct contact with animals infected with this virus. This method of transmission is called zoonotic transmission [2].

SARS is known to be transmitted from mongoose to humans and MERS is transmitted from humped camels. While in COVID-19, there are allegations that the animal that first transmitted the disease to humans was a bat. A person can become infected with the Coronavirus if he inhales a splash of saliva released by a COVID-19 sufferer when sneezing or coughing. Not only that, but transmission can also occur if someone holds an object that has been contaminated with COVID-19 saliva splashes and then holds the nose or mouth without washing hands first. SARS and COVID-19 are known to spread more easily from human to human than MERS [3]. And when compared with SARS, the transmission of COVID-19 from human to human is easier and faster. So far, the death rate from COVID-19 is not higher than SARS and MERS. The SARS death rate reaches 10%, while MERS reaches 37%. However, the transmission of COVID-19 which is faster than SARS and MERS cause the number of sufferers of this disease to increase sharply in a short time. So far, there is no proven drug that is effective in dealing with COVID-19 [4]. Several antiviral drugs, such as oseltamivir, cloroquine, lopinavir, and ritonavir, have been tried to be given to COVID-19 patients while continuing to be studied. Whereas in SARS and MERS, administration of lopinavir, ritonavir, and the latest broad-spectrum antiviral drug called Remdesivir has been proven effective as a treatment. In patients with Coronavirus infection with severe symptoms, in addition to antiviral drugs, they also need to get fluid therapy (infusion), oxygen, antibiotics, and other medicines according to symptoms that appear. Patients with COVID-19 also need to be treated in the hospital so that their condition can be monitored and not transmit the infection to others [5].

In this study, we compared the similarity patterns of the SARS and MERS nucleotide structures with COVID-19 to determine the similarity of the nucleotides with the bioinformatic approach. The data we used consisted of 20 COVID-19 DNA samples, 20 SARS DNA samples, 20 MERS DNA samples, and primers from COVID-19. The three types of DNA samples tested have a short enough distance or in other words have a high enough similarity value when compared to the Primary COVID-19. So if we detect the presence of a coronavirus simply by matching a DNA sample with a COVID-19 primer, then all DNA samples, both SARS and MERS, will be detected as COVID-19. Apart from biomedical, if it is discussed from the perspective of bioinformatics, the process of string similarity alone or the basic sequence alignment is not enough to prove that the DNA sample includes Covid-19 because SARS and MERS still have close kinship values.

Therefore, it is necessary to add a machine learning method to study the distance pattern of each DNA sample so that it can be known and predicted where the DNA infected with COVID-19 really is. We optimize the four machine learning methods, namely Decision Tree, Discriminant Analysis, K-NN, and SVM. The optimization process of each machine learning method varies according to the need to get the best prediction results. Good input features will provide predictive analysis of machine learning with good results. For the DNA Alignment process we use the Edit Levenshtein algorithm with the addition of a DNA sequence normalization filter that meets the positive minimum limit and has the greatest similarity to the primers being compared as an input feature. We describe the optimization process in each method in Chapter 3, while we present the analysis of the results and the discussion in Chapter 4, and Chapter 5 contains conclusions from the results of our research. The results of the study show that the optimization of machine learning method is very helpful in predicting DNA samples by producing accuracy values above 98% for all methods that have been optimized, that were not able to be done in the previous string similarity process.

2. Literature study

DNA alignment is a method for analyzing the sequence of a DNA sample by aligning the sequence with another sequence. In bioinformatics, the nucleotide alignment method can also be said with the character comparison method. In one isolated file DNA can consist of tens of thousands of nucleotide sequences. In large numbers, the process of finding patterns in a sample will require significant time, therefore the speed of an algorithm in determining patterns is an important factor. Research before comparing the performance of the Brute Force, Knuth-Morris-Pratt, and Boyer Moore algorithms to find patterns in isolated DNA [6]. In the process of finding DNA patterns, there are millions of sequences that are compared, so the speed and accuracy of an

algorithm in finding these patterns is a major factor. In addition, the length of the primary characters that are not always the same can also provide different distance measurement results, one of the solution problems is by adding the normalization method to the Hamming algorithm so that the comparison process between primers can be balanced [6].

Decision Tree method is often used to determine a problem with multilevel consideration factors [7]. A condition can be chosen based on the selection of previous conditions and continues to flow until the final decision. This method can help provide a decision on the number of hospital costs to be paid by a patient by looking at the background factors of the patient [8]. Decision tree is one of the strong data mining that can be used to understand the factors that influence health condition decisions. Decision trees can be used to design factors in an urban environment that can affect health outcomes [10]. Previous research used a decision tree learning algorithm called classification and regression tree (CART) for CAD diagnosis as an alternative to the currently available diagnostic methods [10]. In machine learning, sometimes a problem occurs because of an unbalanced data set, this can be overcome by applying ensemble learning. Decision Tree method can be used as an initial classification in the ensemble learning method [11].

Beside decision tree method, machine learning methods that are also often compared are discriminant analysis and SVM [13]. Discriminant Analysis can be applied as a kernel for discrete cross-models to reduce the loss in some cases on quantization [13]. Linear Discriminant Analysis (LDA) can be used to classify patterns, this technique is often used to detect illness early in the data set being tested [14]. However, LDA sometimes cannot provide a good classification if it meets data that are matrices covariant and unseparated linear [15]. Problems in this LDA model can be overcome with a new model approach called Lp- and Ls-Norm Distance Based Robust Linear Discriminant Analysis (FLDA-Lsp) [16]. Linear Discriminant Analysis is also able to classify the bent of a cell based on bispectral invariant features and the results of this classification can be analyzed in more detail by combining the SVM method [17]. For speaker recognition, Discriminant Analysis can be used by make optimization in Kernel Discriminant Analysis (KDA) in higher dimension [18]. In addition to a linear approach, to solve unstructured Covariance matrices is by applying Vanishing Non-Linear Discriminant Analysis (VNDA), this method is able to solve the ratio of trace problems on limited polynomials data [19].

KNN is one of the supervised machine learning methods that are able to solve various problems flexibly [21]. KNN can also be easily combined with other machine learning methods such as SVM, string distance, and neural network [21]. The results of the KNN classification process can increase significantly if at the time of comparison the pattern is given two paired criteria [23]. To determine the node on the KNN sometimes use the average value of the data, the disadvantage of this method is that it cannot determine the really good variable [23]. One solution to this problem is to choose sparse group features as candidates for relevant classes [24]. KNN algorithm is also able to recognize patterns in high-resolution images by calculating the similarity distance around the pixels being compared [25].

Support Vector Machine (SVM) is a supervised machine learning algorithm that is able to solve both classification and regression problems [26]. The way SVM works are to maximize the Hyperplane limit (maximum Hyperplane margin) [27]. There are a number of possible hyperplane choices for a data set, to get the best results from SVM is to determine the maximum Hyperplane [28]. Hyperplane with maximum margins will give better generalization to the classification method [30]. Hyperplane in SVM is not always linear, this model can be in the form of a quadratic curve, or Gaussian in accordance with the kernel that is applied to the data classification process [30].

3. Optimization of machine learning algorithms

3.1 DNA alignment

Sample data from this study totaled 60 isolated DNA consisting of isolated positive DNA infected with COVID-19, MERS, and SARS each of them is 20 samples. All data is taken from the world gene bank [31]. For comparison, we use published Primary COVID-19 data [32, 33]. In one isolated DNA complete gene COVID-19, MERS, and SARS, consisting of 20,000 to 30,000 nucleotide sequences, this number is far more than the other isolated DNA in our previous study [6]. All samples will be compared with each primer, with a total of about 18,000,000 nucleotide comparison processes.

The process of comparing DNA alignment with COVID-19 primers using the Levenshtein distance Edit method. Each isolated DNA will be cut into pieces as long as the number of primary characters and then compared to the primer, calculated the distance of its proximity then shifts again to the next nucleotide. An isolated DNA is said to be positive for

a primary virus or bacterium if the similarity level of the nucleotide fragment reaches greater than 70% [34]. In this process, all isolated DNA tested at least one sequence has a similarity greater than 70% in the forward primer, so it can be said that all of the samples are Covid-19. SARS and MERS are indeed still in one group with Covid-19, which is a Coronavirus group, so it has a similar pattern. Therefore, a further predictive analysis process needs to be carried out.

$$dist_{a,b}(i,j) = \begin{cases} \max(i,j), \ if \ \min(i,j) = 0 \\ \min \begin{cases} dist_{a,b}(i,j) + 1 \\ dist_{a,b}(i,j-1) + 1 \\ dist_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} \end{cases} \quad (1)$$

$$Sim_{a,b} = \frac{dist_{a,b}(i,j)}{n} \times 100\% \quad (2)$$

$$Var_{(x,b)} = \begin{cases} Sim_{a,b}, \ if \ \max(Sim_{a,b}) \geq 70 \\ 0 \quad , if \ \max(Sim_{a,b}) \leq 70 \end{cases} \quad (3)$$

Eq. (1) is an algorithm to calculate the distance between the sequence of DNA slice (*a*) to the Primer (*b*), while *i* is the character index of *a* and *j* is the character index of *b*. Then from the results of distance calculation, the similarity percentage will be calculated as in Eq (2). The variable *n* is the amount or length of the DNA slice being compared, so the percentage of similarity is calculated by dividing the resulting distance value by the number of characters multiplied by 100%. Eq. (3) is an explanation of how we fill the value of the variable independence in the matrix that we build as input features machine learnings.

We conducted various simulations to change the data from the comparison results so that it could be used as an appropriate input feature for machine learning. From some simulation results, the right simulation model in our opinion is to use primers as each input feature. Then every sequence comparison that produces higher similarity from 70% will be entered into the application database. From all data stored in the database, one comparison result that has the highest similarity value on each isolated DNA (has the shortest distance) to a primary will be selected as an input feature. If there is one isolated DNA that does not have a similarity level greater than 70% in a particular primer, then the dataset will be written 0. The amount of training data is the amount of isolated DNA compared to 60 and the number of input features is eight (the number of primers compared), for the target output, there are three classes namely 0 for COVID-19, 1 for SARS, and 2 for MERS.

3.2 Decision tree optimization

The first Machine Learning algorithm that we tried is the Decision Tree. Decision trees use a hierarchical structure for supervised learning. The process of the decision tree starts from the root node to the leaf node which is done recursively. Where each branching states a condition that must be met and at each end of the tree states the class of data.

We use the Entropy concept which is used to measure "how informative" a node (which is usually called how good it is). Entropy *(S) = 0*, if all the examples in *S* are in the same class. *Entropy (S) = 1*, if the number of examples positive and the number of negative examples in *S* is the same. *0 <Entropy (S) <1*, if the number of positive and negative examples in *S* is not the same. *S* is the case dataset and *k* is the number of *S* partitions, while p_j is the probability obtained from Sum (Yes / values more than 70%) divided by Total Cases. *k* is the number of input features being selected, and *P* is the condition of the input feature. The Entropy algorithm can be analyze in Eq. (4-5). After getting the entropy value, the attribute selection is done with the largest information gain value.

$$Entropy\ (S) = -\sum_{j=1}^{k} p_j log_2\, p_j \quad (4)$$

which can be applied to this case study:

$$Entropy\ (S) = -(P_{cov19} log_2 p_{cov19} + P_{sar} log_2 p_{sar} + P_{mer} log_2 p_{mer} \quad (5)$$

So the *Gain (A)* value in this case study can be calculated with:

$$Gain\ (A) = Entropy\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \quad (6)$$

In Eq. (6), *S* is the sample data space used for training. Variable *A* is the number of attributes, |*Si*| is the number of samples for values *V* and |*S*| is the sum of all sample data, both of which have absolute values. Whereas *Entropy (Si)* is entropy for samples that have a value of *i*. From the application of the formula (6), it can be concluded that the greater the information gain we get, the greater the entropy value that we delete. Because the main purpose of applying this

gain is to get an entropy value close to 0 or equal to 0.

3.3 Discriminant analysis optimization

The next method that we tested was a discriminant analysis. In our case study, the dataset tested will be divided into three classes, so it cannot use the linear discriminant analysis method. Then we do the discriminant analysis optimization process by forming an optimal discriminant function with several assumptions about the data used. These assumptions include the data on our independent variables, the multivariate normal distribution and the similarity of variance-covariance matrices between groups. In the preparation of discriminant functions, there are two methods that can be used, namely simultaneous estimation and stepwise estimation. The general model of discriminant analysis is a linear combination of data that can be observed in Eq. (7). $\vec{w}$ and $\vec{x}$ are two vectors whose distances are measured using the diagonal discriminant method. To find out the independent variables that can discriminate against a group we use Wilks Lambda method as in Eq. (8).

$$S_{jk} = a + \vec{w}_j \cdot \vec{x}_{ik} + \cdots + \vec{w}_n \cdot \vec{x}_{nk} \tag{7}$$

To find out which independent variables can be discriminated against:

$$\lambda = \frac{\det(A)}{\det(A+B)} = \frac{\left|\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)\prime\right|}{\left|\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})\prime\right|} \tag{8}$$

In this case study, because there are three groups, so the linear model is converted into a diagonal model. With a diagonal matrix $Dicr = diag(a_1, \dots, a_2)$ and a vector for this dataset become $vec = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. So vector operations can be observed in Eq. (9).

$$Dicr_{vec} = diag(y_1, \dots, y_n)\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_n \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 x_1 \\ \vdots \\ y_n x_n \end{bmatrix} \tag{9}$$

In the process of optimization, we tested several kernel analysis including linear, multiple, and diagonal. The test results show that the diagonal discriminant analysis gives the best results compared to other kernels in this case study.

3.4 K-NN optimization

The K-Nearest Neighbor algorithm uses Neighborhood Classification as the predicted value of the new instance value. In this case, the variables we use are independent variables (variables that are not related to each other) so it can be said that these variables are input features. To calculate the distance between nodes and surrounding neighbors we use the Euclidean distance algorithm, we add weighted distance optimization between one node and another [35]. The kNN optimization algorithm can be observed in Eqs. (10)-(12), where L is the data set to be grouped.

$$L = \{(y_i, x_i), i = 1, \dots, n_L \tag{10}$$

$$d\left(x, x_{(1)}\right) = min_i(dist_{a,b}(x, x_i))$$

with distance:

$$dist_{a,b} = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2} = \left(\sum_{a=1}^{b}(x_{ia} - x_{ja})^2\right)^{\frac{1}{2}}$$

node turn into the class by weighted

$$\hat{y} = max_r\left(\sum_{i=1}^{k} w_{(I)}\, I(y_{(i)} = r)\right) \tag{11}$$

$\hat{y}$ is the max value of a node to the neighbor value compared whether the node has a similarity to the neighbor. rom our test results analysis, the amount of K that we determined also determines the results of the classification. The number of output classes produced can be influenced by the number of distance neighbors or the specified number of K. It can be observed a pattern that by using an odd K, our test results produce a more precise predictive value, the K we use in this study is 1.

3.5 SVM optimization

Support Vector Machine (SVM) is a learning system that uses hypothetical spaces in the form of linear functions in a high-dimensional feature space, trained with learning algorithms based on optimization theory by implementing learning bias derived from statistical learning theory. To classify data that cannot be separated linearly the SVM formula must be modified because no solution will be found. Therefore, the two bounding fields must be changed so that they are more flexible (for certain conditions) by adding the variable S_i ($S_i \geq 0, \forall_i: S_i = 0$ if x_i is classified correctly) to be $x_i w + b \geq 1 - S_i$ for class 1 and $x_i w + b \leq -1 + S_i$ for

class 2. Finding the best separator field by adding the variable S_i is often also called the soft margin hyperplane. In this study we use a Gaussian kernel that can be optimized as in Eqs. (12)-(13).

$$k(x_i, x_j) = \exp(-\gamma \|x_i, x_j\|)^2 \quad (12)$$

Can be applied for $\gamma = 0$, if the parameter is different then $\gamma = \frac{1}{(2\sigma^2)}$ and the hyperplane optimization become:

$$y_i(\vec{w} \cdot \overrightarrow{x_1} - b) \geq 1, for\ i = 1, \dots, n$$

$$\left[\frac{1}{2}\sum_{i=1}^{n} \max(0.1 - y_i(\vec{w} \cdot \overrightarrow{x_1} - b))\right] + \gamma\|\vec{w}\|^2$$

$$min\frac{1}{2}|w|^2 + C\ (\sum_{i=1}^{n} S_i) \quad or$$

$$s.t.\, y_i(w.x_i + b) \geq 1 - S_i \quad or$$

$$S_i \geq 0 \quad (13)$$

C is the parameter that determines the large selection and the data value is determined by the user. This optimization process follows the rules of Structural Risk Minimization (SRM). SRM principle is finding a subset of space. The hypothesis is chosen so that the upper limit is the actual risk by using that subset minimized. SRM aims to minimize actual risk by minimizing error in training data. In this study, minimizing $\frac{1}{2}|w|^2$ are equivalent to minimizing VC dimension and minimize $C(\sum_{i=1}^{n} S_i)$ means minimizing error in training data [36].

4. Result and discussion

In the string similarity process, the results of matching the character of each primer to each isolated DNA tested give varying degrees of similarity. What's interesting about this study is that all isolated DNA tested both SARS, MERS, and COVID-19 all yield a similarity percentage of higher than 69% at least in one of the COVID-19 primers compared, so it can be said that the sample is positive for COVID-19. Table 1 shows the number of sequences that have a higher similarity percentage of 69% in each primer. It can be observed that the sequence tends to have a high similarity value in the forward primary, but some also have a high similarity value in the primary refers.

Below is a piece of positive DNA COVID-19 accession code LR757996.1 on index 15850, MERS accession code MG923468.1 on index 1858, SARS accession code NC_004718 on index 15798. On the MERS DNA, there is one insert command that is adding T nucleotides blue characters) to get the shortest distance.

```
Primer      : GTGARATGGTCATGTGTGGCGG
COVID-19    : GTGAAATGGTCATGTGTGGCGG
MERS        : GTGACATTGTCAGGTGTGGGGG
SARS        : GTGAGATGGTCATGTGTGGCGG
```

Through DNA alignment above, it can be observed that the distance difference lies in the nucleotide R, where R is one component of RNA that can be transformed into nucleotides A, T, G, C. From the observations above, that the changes are not always specific to certain nucleotides. But from our deeper observation, from 20 COVID-19 samples, all of them turned into nucleotide A (Adenine). It can be concluded that the pattern of COVID-19 tends to be A, as in some of the alignment examples below:

```
Primer      : GTGARATGGTCATGTGTGGCGG
LC528232.1  : GTGAAATGGTCATGTGTGGCGG
LC528233.1  : GTGAAATGGTCATGTGTGGCGG
LR757995.1  : GTGAAATGGTCATGTGTGGCGG
LR757996.1  : GTGAAATGGTCATGTGTGGCGG
LR757997.1  : GTGAAATGGTCATGTGTGGCGG
MN908947.3  : GTGAAATGGTCATGTGTGGCGG
MN996531.1  : GTGAAATGGTCATGTGTGGCGG
MN994468.1  : GTGAAATGGTCATGTGTGGCGG
```

Table 1 describes the number of sequences that have a percentage similarity of ≥ 70% with respect to each primer. This amount is cumulative of all isolated DNA grouped according to the type of virus that infected it. Indeed, the COVID-19 DNA sample has the greatest number of similar sequences because what is tested is the COVID-19 primer. However, SARS also has a sequence of similarity above 70% in some primers, and MERS, although only two types of primers, can still be said to have a high degree of similarity to primers COVID-19.

In the Decision Tree algorithm, to decide on an isolated DNA including the type of which virus are quite difficult because the percentage of similarity in COVID-19 and SARS is almost the same, therefore this Decision Tree algorithm needs to add an entropy approach to measure how informative the value is given from the measurement results similarity distance in the previous process, this process also decide the value of gini index. The optimization process of Decision Tree algorithm can be observed in Fig. 1, the result of this system show that maximum split is tree using maximum deviance reduction method.

Table 1. The number of DNA Sequence having similarity level ≥ 70% in each of the COVID-19 primer tested

Primer	COVID-19	MERS	SARS
5'-TGGGGYTTTACRGGTAACCT-3'(Forward)	80	*0*	98
5'-AACRCGCTTAACAAAGCACTC-3'(Reverse)	26	*0*	*0*
5'-TAATCAGACAAGGAACTGATTA-3'(Forward)	141	*0*	151
5'-CGAAGGTGTGACTTCCATG-3'(Reverse)	14	28	*0*
5′-GTGARATGGTCATGTGTGGCGG-3' (Forward)	90	2	86
5'-CARATGTTAAASACACTATTAGCATA-'3 (Reverse)	*0*	*0*	*0*
5'-ACAGGTACGTTAATAGTTAATAGCGT-3' (Forward)	122	*0*	126
5'-ATATTGCAGCAGTACGCACACA-3' (Reverse)	41	*0*	1

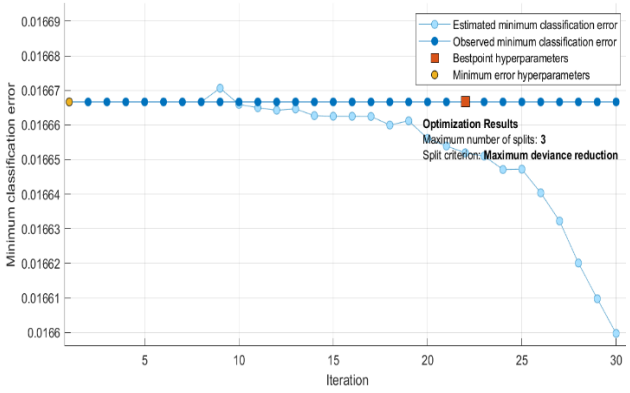

Figure. 1 Optimization process of decision Tree algorithm

Discriminant analysis algorithms usually use a linear approach to determine which node belongs in which class. However, because the class needed in this study amounted to three, so the linear discriminant analysis is less precise in solving problems. We add the Wilks Lambda method to determine the independence variable used as input features of discriminant analysis. The test results using the Optimize Discriminant Analysis algorithm produce an accuracy rate of more than 98%. The optimization process of Discriminant Analysis algorithm can be observed in Fig. 2.

In K-NN algorithm, determining the value of K,

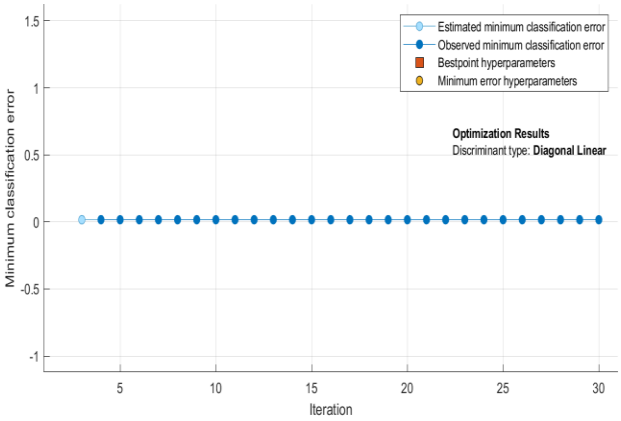

Figure. 2 Optimization process of discriminant analysis

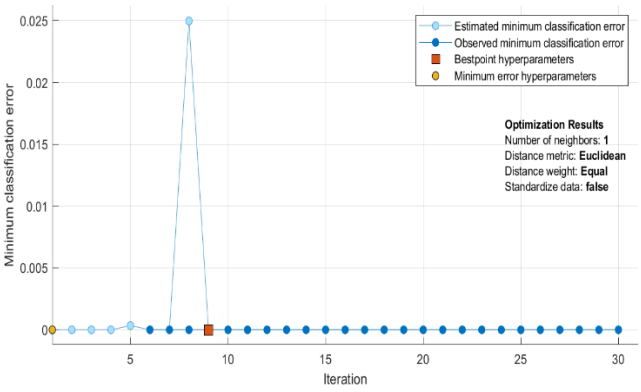

Figure. 3 Optimization process of K-NN algorithm

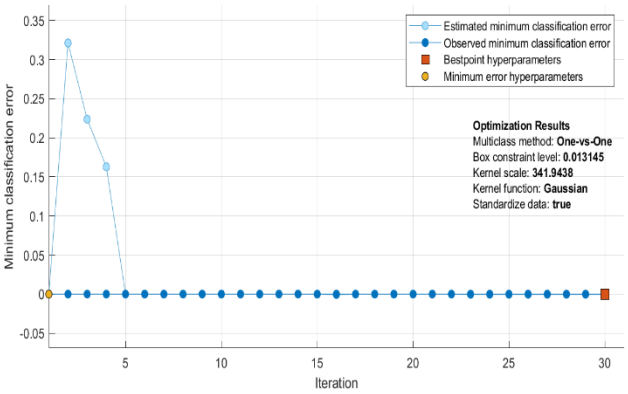

Figure. 4 Optimization process of SVM algorithm

if the sum of our classifications is even then we better use even K values, and vice versa if our total classifications are odd then we better use even K values because if it is not so, there is a possibility that we will not be optimal results from testing. In this study, we use K = 1, which is choosing 1 neighbor who have high proximity values with the node that we are comparing. The training result of K-NN can be observe in Fig. 3.

Similar to the Discriminant Analysis Algorithm, the SVM algorithm also provides results with low accuracy in linear SVM. The advantage of SVM, this method has a kernel that can be adjusted to the needs-

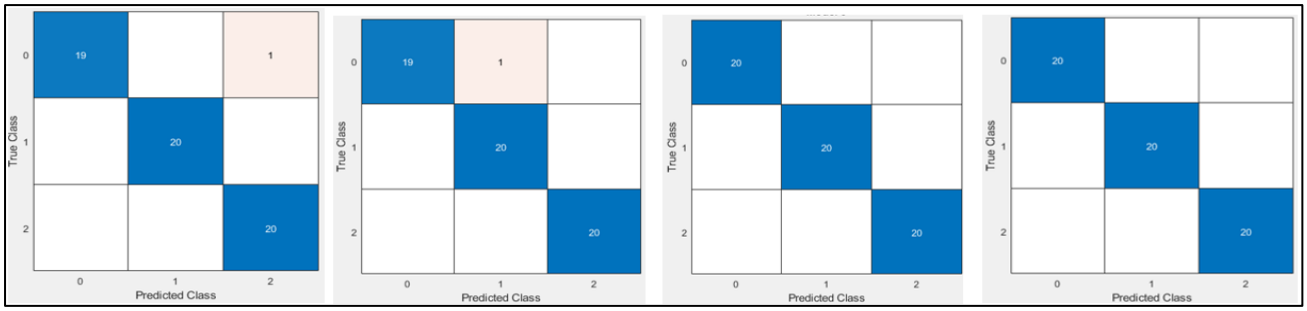

Figure. 5 Confusion matrix results of (the order of images from left to right) Opt. decision tree, Opt. discriminant analysis, Opt. K-NN, and Opt. SVM

Table 2. Sensitivity, specificity, precision (Positive Predictive Value), and negative predictive value (NPV) values for each class

Algorithm	Class	Sensitivity	Specificity	Precision	NPV
Opt. Decision Tree	COVID-19	0.950	1.000	1.000	0.976
	SARS	1.000	1.000	1.000	1.000
	MERS	1.000	0.975	0.952	1.000
Opt. Discriminant Analysis	COVID-19	0.950	1.000	1.000	0.976
	SARS	1.000	0.975	0.952	1.000
	MERS	1.000	1.000	1.000	1.000
Opt. K-Nearest Neighbors	COVID-19	1.000	1.000	1.000	1.000
	SARS	1.000	1.000	1.000	1.000
	MERS	1.000	1.000	1.000	1.000
Opt. Support Vector Machine	COVID-19	1.000	1.000	1.000	1.000
	SARS	1.000	1.000	1.000	1.000
	MERS	1.000	1.000	1.000	1.000

based on input data or the number of output classes desired. In the SVM optimization process, we tested several kernels to produce the best hyperplane. The optimization process to get the best hyperplane uses the SRM principle and considers the actual risk factor. Kernel testing can be observed in Fig. 4.

The validation process in this study uses the Cross-Validation approach with K as many as 10. In each of our tested optimization methods, we divided the data into two groups, 90% for training data and 10% for test data. Then our application will form the composition of the data randomly 10 times to test its accuracy. The test results showed that the optimization of the Decision Tree algorithm and Discriminant analysis each resulted in 1 data error prediction. In the Decision Tree, one data that should be DNA infected with COVID-19 is predicted to be DNA infected with MERS. Whereas in Discriminant Analysis, one data which should be DNA infected with COVID-19 is predicted to be DNA infected with SARS. Comparison of these data uses the COVID-19 primer, but instead, the error data is found in COVID-19, while MERS and SARS can be predicted well, this shows that the pattern on COVID-19 is still changing more.

In the SVM and K-NN algorithms, each data can be predicted well and produces 0 prediction errors. The confusion matrix of the four optimization algorithms can be observed in Fig. 5. From the confusion matrix in Fig. 5, the sensitivity, specificity, precision/Positive Prediction Value (PPV), and Negative Prediction Value (NPV) values can be calculated. Sensitivity values are obtained from correctly predicted data values divided by the amount of correct data in real conditions.

The specificity value is obtained from dividing correctly predicted data not the class divided by real data that is not the class. Then Precision is obtained from all data that is in the class and correctly predicted divided by the amount of true data that is predicted correctly and incorrectly. Calculating the value of sensitivity, specificity, precision, and NPV on a multi-class matrix is different from the calculation of a two-class matrix basically. In this case, when calculating the sensitivity value for the COVID-19 class, the MERS and SARS data will be considered True Negative (TN) data, as well as the calculations on MERS, and SARS. In Table 2, the sensitivity value for the COVID-19 class using the Decision Tree optimization algorithm is 0.95. A COVID-19 data is predicted to be wrong into MERS data, which causes the specificity and precision values in the MERS class to be imperfect. Whereas

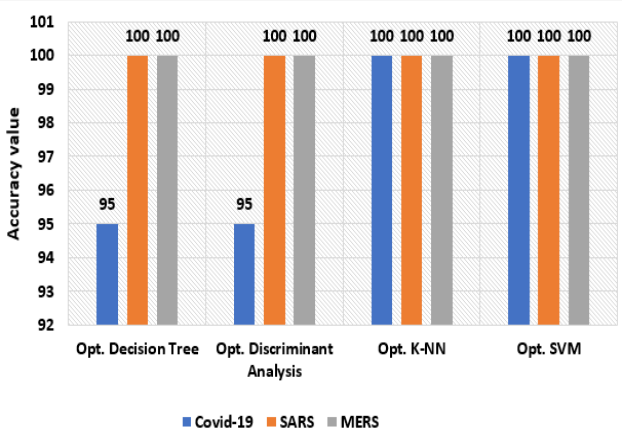

Fig. 6. The accuracy value of each class uses decision tree, discriminant analysis, K-NN, and SVM with optimization methods

in the Discriminant Analysis Optimization method, one member of the COVID-19 class is predicted to be wrong in the SARS class and also results in an imperfect precision and specificity value. For the K-NN and SVM Optimization methods, they can correctly predict data into each class. Fig. 6. is the accuracy value of each class for the tested methods.

5. Conclusion

The similarity in DNA structure between COVID-19, MERS, and SARS is one of the obstacles in predicting samples that are actually infected with COVID-19. The DNA alignment method with Primer produces a positive value of COVID-19 in all MERS and SARS samples. Machine learning methods can help the prediction process by observing changes in the pattern of DNA alignment that are included as input features. The results of predictions show Optimization of SVM and KNN are able to predict 100% correctly, while optimization of Discriminant Analysis and Decision Tree produces an accuracy of 98.3%. The prediction error is precisely in the COVID-19 sample data, even though the Primer tested was the COVID-19 primer. This shows that the composition of DNA in COVID-19 samples is still diverse and there is a possibility that mutations will continue to occur. In the process of DNA alignment between COVID-19 Primer and isolated DNA samples, we analyzed that when tested with certain primers containing RNA 'R', the sequence in isolated DNA infected COVID-19 always becomes 'A'

Conflicts of Interest

The authors declare no conflict of interest

Author Contributions

Berlian Al Kindhi in this study contributed to the entire process of machine learning and data set processing and writing paper.

Acknowledgments

This research is partially funded by the Institut Teknologi Sepuluh Nopember for research grants No. 853/PKS/ ITS/2020.

References

[1] T. Lupia, S. Scabini, S. M. Pinna, G. D. Perri, F. G. Rosa and S. Corcione, “2019 novel coronavirus (2019-nCoV) outbreak: A new challenge”, *Journal of Global Antimicrobial Resistance*, Vol. 21, pp. 22-27, 2020.

[2] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir and R. Siddique, “COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses”, *Journal of Advanced Research*, Vol. 24, pp. 91-98, 2020.

[3] J. A. Al-Tawfiq and P. Gautret, “Asymptomatic Middle East Respiratory Syndrome Coronavirus (MERS-CoV) infection: Extent and implications for infection control: A systematic review”, *Travel Medicine and Infectious Disease*, Vol. 27, pp. 27-32, 2019.

[4] P. B. Tim Smith, P. Jennifer Bushek and P. Tony Prosser, “COVID-19 Drug Therapy – Potential Options, Clinical Drug Information”, *Clinical Drug Information, Clinical Solutions*, 2020.

[5] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis and R. Agha, “World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)”, *International Journal of Surgery*, Vol. 76, pp. 71-76, 2020.

[6] B. A. Kindhi and T. A. Sardjono, “Pattern Matching Performance Comparisons as Big Data Analysis Recommendations for Hepatitis C Virus (HCV) Sequence DNA”, In: *Proc. of the 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, Kota Kinabalu, Malaysia, 2015.

[7] B. A. Kindhi, T. A. Sardjono, M. H. Purnomo and G. J. Verkerke, “Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis”, *Expert Systems with Applications*, Vol. 121, pp. 373-381, 2019.

[8] N. E. I. Karabadji, I. Khelf, H. Seridi, S. Aridhi, D. Remond, and W. Dhifli, “A data sampling and attribute selection strategy for improving

decision tree construction", *Expert Systems with Applications*, Vol. 129, pp. 84-96, 2019.
[9] G. A. Kundakçi, M. Yılmaz, and M. KaanSözmen, "Determination of the costs of falls in the older people according to the decision tree model", *Archives of Gerontology and Geriatrics*, Vol. 87, p. 104007, 2020.
[10] A. C. Hillar, L. C. Donna, B. Charles, and M. M. Heather, "Using decision trees to understand the influence of individual- and neighborhood-level factors on urban diabetes and asthma", *Health & Place*, Vol. 58, p. 102119, 2019.
[11] M. M.Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model", *Computer Methods and Programs in Biomedicine*, Vol. 192, p. 105400, 2020.
[12] J. Obregon, A. Kim, and J.-Y. Jung, "RuleCOSI: Combination and simplification of production rules from boosted decision trees for imbalanced classification", *Expert Systems with Applications*, Vol. 126, pp. 64-82, 2019.
[13] C. L. M. Morais, K. M. G. Lima, and F. L. Martin, "Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines", *Analytica Chimica Acta*, Vol. 1063, pp. 40-46, 2019.
[14] Y. R. Y. Fang, "Supervised discrete cross-modal hashing based on kernel discriminant analysis", *Pattern Recognition*, Vol. 98, No. 1, p. 107062, 2020.
[15] S. Yang, J. Bian, Z. Sun, L. Wang, H. Zhu, H. Xiong, and Y. Li, "Early Detection of Disease Using Electronic Health Records and Fisher's Wishart Discriminant Analysis", *Procedia Computer Science*, Vol. 140, No. 1, pp. 393-402, 2018.
[16] Z. Jing, G. Wang, S. Zhang, and C. Qiu, "Building Tianjin driving cycle based on linear discriminant analysis", *Transportation Research Part D: Transport and Environment*, Vol. 53, pp. 78-87, 2017.
[17] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, "Lp- and Ls-Norm Distance Based Robust Linear Discriminant Analysis", Neural Networks, Vol. 105, No. 1, pp. 393-404, 2018.
[18] V. C. K. Al-Dulaimi, K. Nguyen, J. Banks, and I. Tomeo-Reyes, "Benchmarking HEp-2 specimen cells classification using linear discriminant analysis on higher order spectra features of cell shape", *Pattern Recognition Letters*, Vol. 1251, pp. 534-541, 2019.
[19] R. K. Das, A. B. Manam, and S. R. M. Prasanna, "Exploring kernel discriminant analysis for speaker verification with limited test data", *Pattern Recognition Letters*, Vol. 98, pp. 26-31, 2017.
[20] Y. Shao, G. Gao, and C. Wang, "Nonlinear discriminant analysis based on vanishing component analysis", *Neurocomputing*, Vol. 218, pp. 172-184, 2016.
[21] S. B. Chen, Y. L. Xu, C. H. Q. Ding, and B. Luo, "A Nonnegative Locally Linear KNN model for image recognition", *Pattern Recognition*, Vol. 83, pp. 78-90, 2018.
[22] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection", *Physica A: Statistical Mechanics and its Applications*, Vol. 517, pp. 29-35, 2019.
[23] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, "Granger Causality Driven AHP for Feature Weighted kNN", *Pattern Recognition*, Vol. 66, p. 4250436, 2017.
[24] J. N. Myhre, K. Ø. Mikalsen, S. Løkse, and R. Jenssen, "Robust clustering using a kNN mode seeking ensemble", *Pattern Recognition*, Vol. 76, pp. 491-505, 2018.
[25] C. D. S. Zheng, "A group lasso based sparse KNN classifier", *Pattern Recognition Letters*, Vol. 131, pp. 227-233, 2020.
[26] N. Liu, X. Xu, Y. Li, and A. Zhu, "Sparse representation based image super-resolution on the KNN based dictionaries", *Optics & Laser Technology*, Vol. 110, pp. 135-144, 2019.
[27] B. Lin, X. Wei, and Z. Junjie, "Automatic recognition and classification of multi-channel microseismic waveform based on DCNN and SVM", *Computers & Geosciences*, Vol. 123, pp. 111-120, 2019.
[28] T. I. Dhamecha, A. Noore, R. Singh, and M. Vatsa, "Between-subclass piece-wise linear solutions in large scale kernel SVM learning", *Pattern Recognition*, Vol. 95, pp. 173-190, 2019.
[29] D. Zhang, L. Jiao, X. Bai, S. Wang, and B. Hou, "A robust semi-supervised SVM via ensemble learning", *Applied Soft Computing*, Vol. 65, pp. 632-643, 2018.
[30] R. Sundar and M. Punniyamoorthy, "Performance enhanced Boosted SVM for Imbalanced datasets", *Applied Soft Computing*, Vol. 83, p. 105601, 2019.
[31] U. Khan, L. Schmidt-Thieme, and A. Nanopoulos, "Collaborative SVM classification in scale-free peer-to-peer networks", *Expert Systems with Applications*, Vol. 691, pp. 74-86, 2017.
[32] Gene Bank, 10 2 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/.

[33] J.-M. Kim, Y.-S. Chung, H. J. Jo, N.-J. Lee, M. S. Kim, S. H. Woo, S. Park, J. W. Kim, H. M. Kim, and M.-G. Han, "Identification of Coronavirus Isolated from a Patient in Korea with COVID-19", *Osong Public Health and Research Perspectives*, Vol. 11, No. 1, pp. 3-7, 2020.

[34] LKS Faculty of Medicine, School of Public Health, Hongkong University, "Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR", https://www.who.int/docs/default-source/coronaviruse/peiris-protocol-16-1-20.pdf?sfvrsn=af1aac73_4, Hongkong, 2020.

[35] S. Carson and D. Robertson, *Manipulation and Expression of Recombinant DNA, chapter: III Expression, Detection, and Purification of Recombinant Proteins from Bacteria*, Elsevier Academic Press, California, pp. 130–168, 2006.

[36] K. Hechenbichler and K. Schliep, "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification", *Sonderforschungsbereich*, Vol. 386, p. 399, 2004.

[37] E. E. Osuna, R. Freund, and F. Girosi, "Support vector machines; training and applications", *A. I. Memos No. 1602, CBCL Memos No. 144, Artificial Intelligence laboratory, Massachusetts Institute of Technology*, 1997.