

## Profiling DNA sequence of SARS-Cov-2 virus using machine learning algorithm

Lailil Muflikhah, Muh. Arif Rahman, Agus Wahyu Widodo

Department of Informatics Engineering, Faculty of Computer Sciences, Brawijaya University, Malang, Indonesia

### Article Info

#### Article history:

Received Dec 12, 2021

Revised Mar 2, 2022

Accepted Mar 17, 2022

#### Keywords:

Covid-19

DNA sequence

Feature extraction

k-mer

Random forest

### ABSTRACT

Corona virus disease-19 (COVID-19) is growing rapidly because it is an infectious disease. This disease is caused by a virus belonging to the type of DNA virus with very diverse genetics. This study proposes a feature extraction method using k-mer to obtain nucleotide frequencies in protein coding. In profiling viral DNA sequences, this study proposes to obtain similarity by country using hierarchical k-means, where the results are averaged by the hierarchical clustering method and then find the initial cluster center. The experimental results show that the silhouette, purity, and entropy are 0.867, 0.208, and 0.892, respectively. Then, we apply the Gini index feature selection to find the important components as characteristics in each country. The selected components are implemented using the ensemble method, Random Forest, to evaluate their performance. The experimental results showed high performance, including sensitivity, accuracy, specificity, and area under the curve (AUC).

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Lailil Muflikhah

Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University

Jl. Veteran No. 8, Malang, 65145, Indonesia

Email: lailil@ub.ac.id

## 1. INTRODUCTION

Coronavirus disease-19 (COVID-19) is caused by SARS Cov-2 virus infection. It spreads quickly around the world. First discovered in Wuhan, China, this virus is a sense-RNA-positive virus [1]. The virus infects the human body presumably by binding to the protein Angiotensin-converting accelerator a pair of (ACE2) [2] found within the lower tract. After entering the cell, this virus hijacks the cell system to multiply, eventually destroying the host cell and infecting other surrounding cells. This adversely affects lung function and even leads to acute respiratory distress syndrome (ARDS), leading to death [3].

Various research institutions and health institutions have made many efforts to get the right vaccine to overcome the virus mutation after infecting the human body. The application of information technology (IT) in molecular biology, known as bioinformatics, is proliferating. The main goal of bioinformatics is to improve our understanding of biological processes [4]. One of them is to analyze the development of the virus through the characteristics of its DNA sequence. The high rate of mutations affected many variations of DNA sequences and encouraged many researchers to investigate using a computational approach. Kandpal and Davuluri [4] identified SARS CoV-2 missense mutations in specific regions using the Random Forest (RF) and feature selection methods. In their research, DNA sequences were taken at specific mutation sites in building a classifier model for identification. In this study, we used a whole-genome DNA sequence for profiling SARS Cov-2 using machine learning methods to identify the demographic origin. This paper is presented in four sections. The first section describes the background of the research. As for the second topic, it discusses the research method, then presents and discusses the results. The final section is the conclusion.

## 2. METHOD AND MATERIAL

In general, the stages of our research include crawling DNA sequences data, extracting features to compose the protein-coding, and profiling the sequence using machine learning algorithms. The feature extraction is a transformation by quantifying the certain patterns in the sequences. Then, profiling the sequence is proposed using clustering, feature selection, and classification method. The details are illustrated in Figure 1.

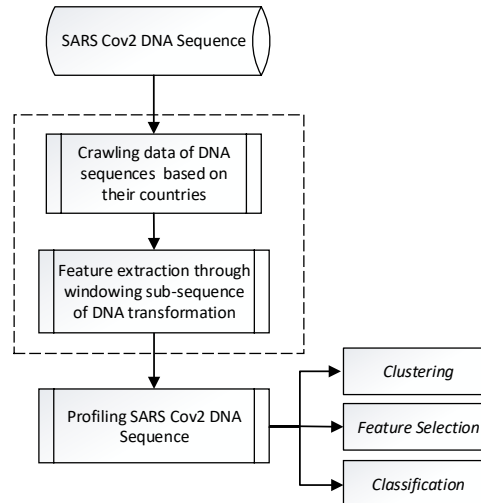


Figure 1. General block diagram of profiling SARS CoV-2 DNA sequence

### 2.1. Severe acute respiratory syndrome coronavirus (SARS CoV-2)

SARS Coronavirus (SARS CoV-2) is a type of coronavirus and has a sense-RNA-positive virus [5].  $\alpha$ - and  $\beta$ -CoV can infect mammals, meanwhile  $\gamma$  and  $\delta$ -CoV tend to infect birds. Previously, six coronaviruses have been identified as human susceptible viruses, including  $\alpha$ -CoVs HCoV-229E and HCoV-NL63, and low pathogenic  $\beta$ -CoVs HCoV-HKU1 and HCoV-OC43, which cause a fever similar to ordinary fever. Of mild respiratory symptoms. There are two other known  $\beta$ -CoV that cause severe and potentially fatal respiratory tract infections [6], SARS-CoV and MERS-CoV. The SARS-CoV-2 genome was 96.2% identical to the genome of the bat CoV RaTG13.

Based on the results of the sequencing of the bats' genome and evolutionary analysis, the initial hosts of the virus have been suspected as the bats, and it is possible that the virus was transmitted from bats to humans through an unknown intermediate host. The evolutionary development of SARS-CoV-2 appears to be very fast, following the host according to the region of the country. Data from Nextstrain containing simulations of the virus sequence from various countries shows a mutation rate of 23,697 subs per year as shown in Figure 2. These changes allow the variation of the target gene to change [7].

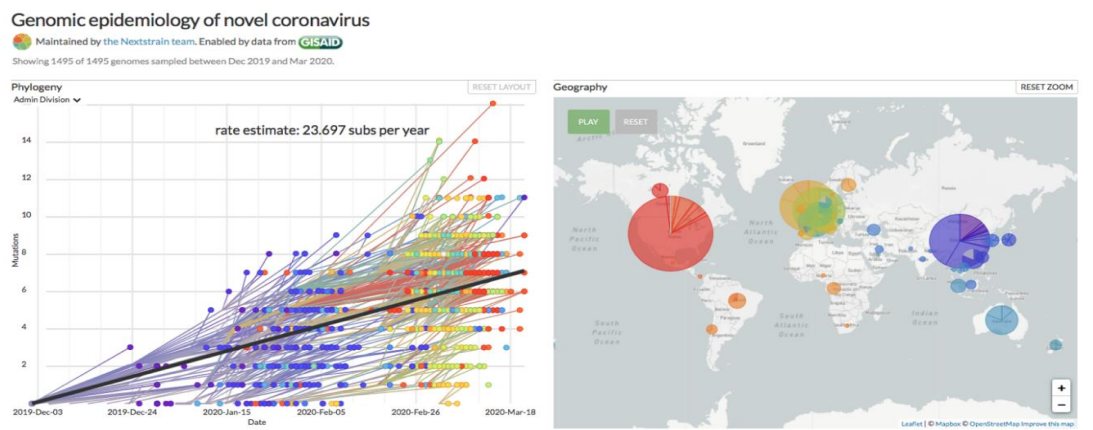


Figure 2. The evolutionary pattern of SARS-CoV-2 that follows the location or race of the host

## 2.2. Feature extraction of SARS-CoV-2 DNA sequence

In protein coding, codons are composed of three DNA nucleotides from four nucleotides, including Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) or RNA in one sequence to form 20 kinds of amino acids. These 20 amino acids are found in protein and some other foods. They are compounds that contain an amine and a carboxyl functional group. This data set contains amino acid sequences that can be translated into proteins. The protein coding reference standard genetic code is a set of three letters that tells which instructions the protein will be coded as shown in Table 1. This data set contains sequences of amino acids that can be translated into proteins. The letters of the genetic code represent different amino acids.

Table 1. Standard genetic code

1 <sup>st</sup> Nucleotide	2 <sup>nd</sup> Nucleotide				3 <sup>rd</sup> Nucleotide
	T	C	A	G	
T	Phe/F	Ser/S	Tyr/Y	Cys/C	T
	Leu/L		Stop	Stop	C
		Pro/P	Stop	Trp/W	A
			His/H	Arg/R	G
C			Gln/Q		T
					C
					A
					G
A	Ile/I	Thr/T	Asn/N	Ser/S	T
			Lys/K	Arg/R	C
					A
					G
G	Met/M	Ala/A	Asp/D	Gly/G	T
	Val/V				C
			Glu/E		A
					G

In this study, the virus DNA sequences were extracted to protein coding construction using sliding window in three of four nucleotides from the sequences. The extraction describes a set of analytical techniques to process metagenome of sequence data. The raw sequence information is transformed into reusable data structures using feature selection techniques and machine learning algorithms. The analysis and transformation from raw sequences into reusable structures is performed using  $k$ -length DNA substrings, known as the  $k$ -mer method. This study is applied the algorithm with  $k=3$  (three-dimensional) to represent a codon in a protein coding construct. This is based on previous research conducted by Cho *et al.* [8] that sequences can be used as a rapid genomic screening to find out the evolution of protein variants. Furthermore, the number of  $k=3$  refers to the table of genetic code as in Table 1 [9].

## 2.3. Machine learning method

In principle, machine learning methods are divided into two, namely supervised learning and unsupervised learning. Most machine learning methods use supervised learning, which has an input variable (X) and an output variable (Y), and an algorithm to train the matching function from input to output,  $Y=f(X)$ . The goal is to have a good estimate of the mapping function when given input data (X) can predict the output variable (Y) for that data. As an example, Figure 3 shows the difference between a machine learning algorithm and traditional programming.

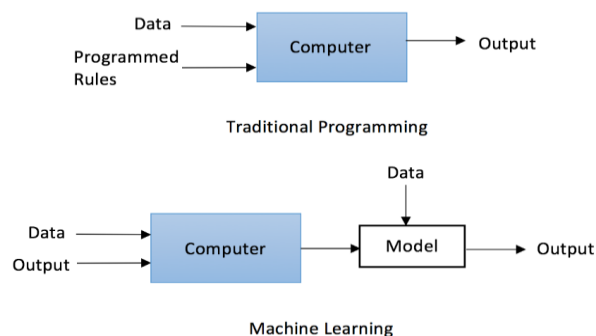


Figure 3. Illustration of machine learning algorithm

Two kinds of machine learning method including unsupervised learning algorithm and supervised learning algorithm. In the unsupervised learning method, it only has input data (X) and no output. The purpose of the unsupervised learning method is to model the underlying structure or distribution in the data for further study. Several algorithms of unsupervised learning method are clustering, association rule, and sequential pattern mining. Another hand, supervised learning methods are need the training data for labeling such as: classification and regression methods This study is applied both kinds of methods such us clustering and classification algorithm.

### 2.3.1. Clustering algorithm

Clustering is a way to group objects based on their characteristics. K-Means is an unsupervised machine learning method for clustering with high computing speed and good accuracy [10]. However, this method is depend on the initial centroid [11]–[13]. It was created for the purpose of automatically identifying wrist fractures and it was 80% accurate [14]. To improve performance, there are various ways to find the best initial centroid, including using hierarchical cluster method. The hierarchical method is a bottom-up clustering method with initial stage, each object represents a cluster of other objects [15]. The cluster tree is constructed based on objects with or without relationships. The end result is a collection of all objects, called as a root. However, this algorithm required high computational time to complete hierarchy of cluster.

Hierarchical k-means is a hybrid clustering method of hierarchical and k-means to overcome their limitations [16]. This method is like K-Means with a defined initial centroid [17], [18]. Therefore, this study aims to profile Sars CoV-2 DNA sequences in various ASEAN countries using the Hierarchical k-means algorithm based on the similarity of the country of origin.

### 2.3.2. Classification algorithm

Classification is a supervised method of using training data to build a classifier model. Usually used for prediction, detection, and classification or recommendation purposes. Several supervised learning algorithms were used for representative classification methods including decision tree, naive Bayes [19], [20] and SVM [21] as well as ensemble method (RF) [22]. In this study, we used RF to classify SARS-Cov-2 DNA sequences by country.

RF algorithm is an algorithm that is suitable for classifying large data without pruning variables in the decision tree. The formation of the RF tree is built by conducting training sample data. The selected variables are taken to be separated randomly. Classification is run after all trees are formed and is taken based on the most votes from each tree. This method is an enhanced CART algorithm by applying the bootstrap aggregation method and random feature selection. Predictions are made by averaging the predictions of all individual models. Each tree is constructed based on several N samples from the original data in turn. For various input variables (M), where m has a much smaller number than N, and the variable m is chosen independently of M. Therefore, the best split in m is used for simplification of nodes, no pruning is performed [23].

### 2.3.3. Gini index feature selection

In RF method, the construction of decision tree required candidate nodes selection based on the Gini index measurement. It means that the node is considered as the important information to compose in the tree. In this study, we proposed to apply this measure value to select the potential features. The Gini index is calculated by subtracting the sum of the squared probabilities of each organism's class. The Gini value of the segmented data is greater than the value of information acquisition. The attribute for the smallest Gini values is proposed to be a candidate node in the construction of the RF decision tree. It indicated that it is an important feature in building a decision tree. The important value is calculated using a Gini index measurement as given in (1).

$$Gini\ Index = 1 - \sum (P(x = k))^2 \quad (1)$$

Where,  $P(x=k)$  is probability of each class (k) from dataset (x)

## 3. RESULTS AND DISCUSSION

### 3.1. Data sets

The data used in this study were taken from the official GISAID website with the URL address: <https://www.gisaid.org/>. The acquisition of whole-genome data on DNA sequences of the SARS Cov2 virus was carried out by selecting several countries in ASEAN, including Indonesia, Malaysia, Singapore, the Philippines, and Myanmar using control China (Wuhan) as the first spreader of the virus. pandemic. On January 10, 2020, the first viral genome and related data were shared publicly on the GISAID website. As a pandemic progresses, scientists worldwide are investigating viruses and their genome sequences to ensure optimal viral

```
> ti$sequence
DNA data for 6917 sequences
> TACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTAAACG... + 29799 bases
> CTCRAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCT... + 29860 bases
> ATTAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCT... + 29863 bases
> TAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGT... + 29841 bases
> ACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTAAACGA... + 29830 bases
> ATTAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCT... + 29843 bases
> ATTAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCT... + 29843 bases
> TTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTA... + 29835 bases
> GTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCT... + 29836 bases
> TACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTAAACG... + 29831 bases
> TACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTAAACG... + 29831 bases
> TATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTTGTAGATCTGTTCTCTAAA... + 29833 bases
... with 6905 more sequences.> |
```

Table 2. The total DNA sequence in each country

Country	The number of DNA sequences
Indonesia	13
Malaysia	21
Vietnam	23
Thailand	28
Myanmar	1
Control (Wuhan)	2

After obtaining the feature extraction as a predictor, we applied a clustering algorithm to determine the similarity of DNA sequences in protein-coding characteristics. The method constructed a link in a high similarity of characters to make a join. The result shows that Indonesia and Vietnam have high similarities of sequence to Wuhan (as control), as described in Figure 5. To know the performance result, we use three evaluation measurements, such as silhouette width, purity, and entropy. Silhouette width ( $S_i$ ) analysis is a measure of the average distance between clusters as illustrated in Figure 6 to be defined formulation (2). The range is -1 and 1 ( $-1 \leq s(i) \leq 1$ ). This cluster is well-clustered if the value is close to 1. In contrast, if the value is close to -1, then it is placed in the wrong cluster [24].

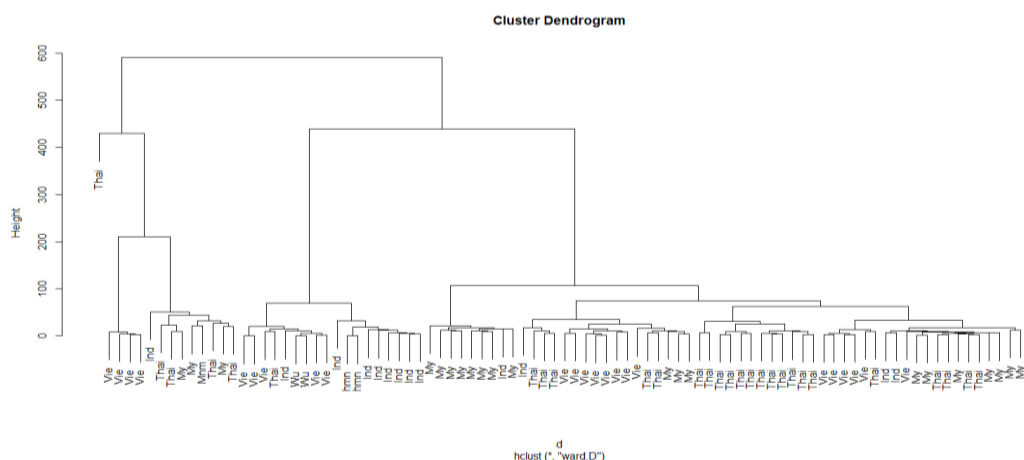


Figure 5. Clustering result using hierarchical k-Means algorithm

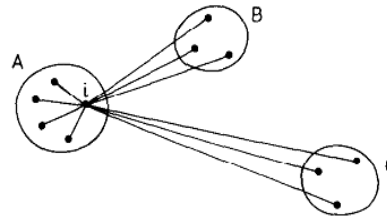


Figure 6. Illustration of silhouette clustering measurement

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

Remarks:

$s(i)$  = silhouette with between clusters

$a(i)$  = average dissimilarity of  $i$  to all other objects of  $A$

$b(i)$  = minimum cluster distance in neighbor of object  $i$

Then, purity and entropy values are evaluation measurements to determine the ability of clustering techniques to recover the appropriate class. Suppose the clustering technique produces  $k$  clusters, while given  $l$  categories. The purity of clustering for known categories in  $n$  dataset is given by (3) [25].

$$Purity = \frac{1}{n} \sum_{q=1}^k \max_{1 \leq j \leq l} n_q^j \quad (3)$$

The purity of the object dataset is defined as the number of object data in the group that belongs to a certain class. This number ranges from 0 to 1. The more efficient the clustering process, the purer the result. Another evaluation is the entropy for grouping by category as in the formulation given in (4) [25].

$$Entropy = - \frac{1}{n \log_2 l} \sum_{q=1}^k \sum_{j=1}^l n_q^j \log_2 \frac{n_q^j}{n_q} \quad (4)$$

The amount of object data in cluster  $q$  is the amount of object data in cluster  $q$  multiplied by the number of objects in cluster  $q$ . The number of object data in cluster 1 is the number of object data in cluster 1 multiplied by the number of objects in cluster 1. The purity and entropy are measurements that help you know how the cluster method recovers known classes, even if the number of clusters is different from the number of known classes. [24]. Based on the experiment result, the purity, entropy, and silhouette value are achieved of 0.867, 0.208, and 0.892 in respectively. On the other hand, we applied a feature selection method based on the Gini index value to obtain significant or potential protein coding as a predictor of DNA sequences from a particular country. The descending order of the Gini index values results in the top ten characteristics based on the Gini index as a measure of significant value as shown in Figure 7 and Table 3.

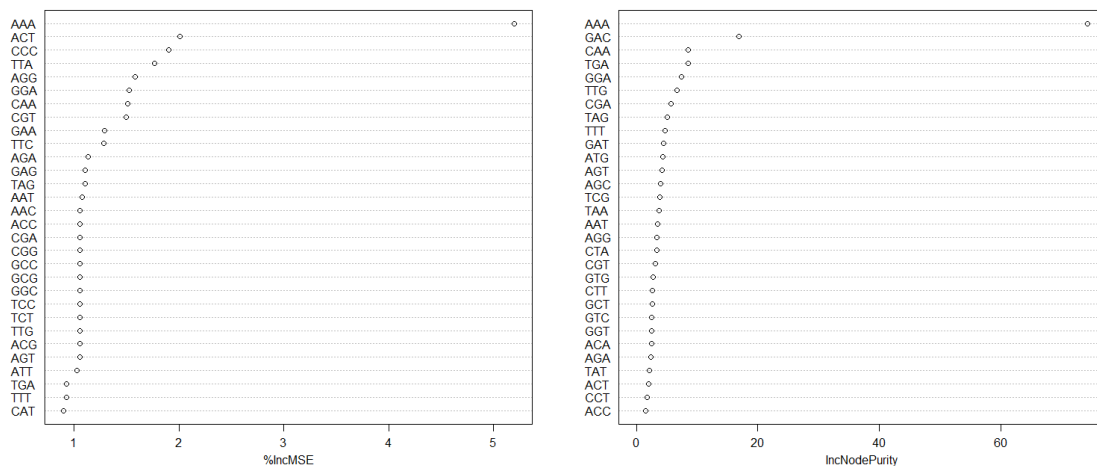


Figure 7. The importance values of extracted features (codon)

Table 3. The ten highest important values of the extracted features

Codon	Amino acid	Important value
AAA	Lys (K)	80.45
GAC	Asp (D)	23.05
ATG	Met (M)	10.90
CGA	Arg (R)	9.20
AGT	Ser (S)	5.28
TGG	Trp (W)	4.65
CAA	Gln (Q)	4.16
GCG	Ala (A)	4.96
GTT	Val (V)	4.04
ATT	Ile (I)	3.72
TGG	Trp (W)	4.65

As an example, we will figure out the correlation of the selected features to the target class (dy). These correlations are displayed in a range of colors. The intensity of the color and the size of the circle are inversely proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors. The features selected have a significant relationship with one another as shown in Figure 8(a). Then the selected features are constructed a classifier model of decision-tree in RF as shown in Figure 8(b). The nodes are representative of the decision class target with a p-value of less than 0.001.

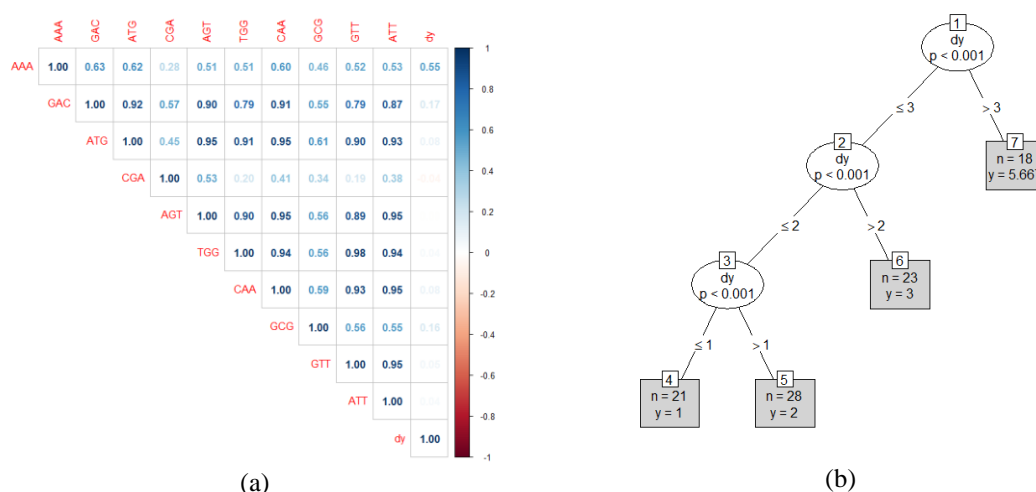


Figure 8. Correlation (a) and tree construction (b) of the selected features

### 3.3. K-fold cross-validation

One of the evaluation methods in data composition requires k-fold cross-validation. This is intended to avoid data limitations and data imbalances in each class. Test data is used as m-fold and training data is used as k-fold. In this study, the number of folds will be divided into two, one as training data and one as test data. The data is divided by k parts, and then the data is repeated, for k iterations, in different folds [26].

### 3.4. Performance result evaluation

The proposed method for feature selection is evaluated by applying a classification algorithm, as a method for identifying the features associated with the target. By using a confusion matrix, the data test gets the correct value for the actual data. In this matrix, true positive (TP) is the only unit that is correctly classified for our class, while false positive (FP) and false negative (FN) are items that are misclassified in the column and row of the class, respectively. Negative true (TN) all other tiles, as shown in the confusion matrix[27].

The top ten codons extracted from the genome sequence are evaluated using a confusion matrix. The results of the performance of RF showed that the accuracy, sensitivity, specificity, and area under the curve (AUC) were high, very close to 1 as shown in Figure 9. Furthermore, in the RF classifier method, many trees were constructed. The total trees of a recursive filter have an influence an error rate as illustrated in Figure 10. In this study, we applied 30 trees in the R-squared algorithm.



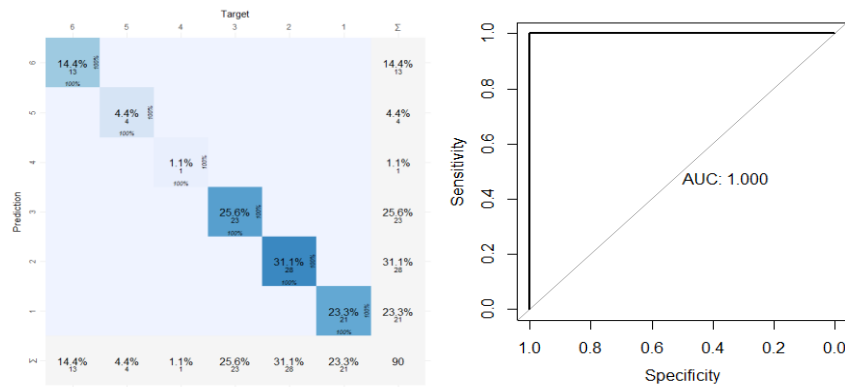


Figure 9. Confusion matrix of the result using (a) RF algorithm and (b) the AUC

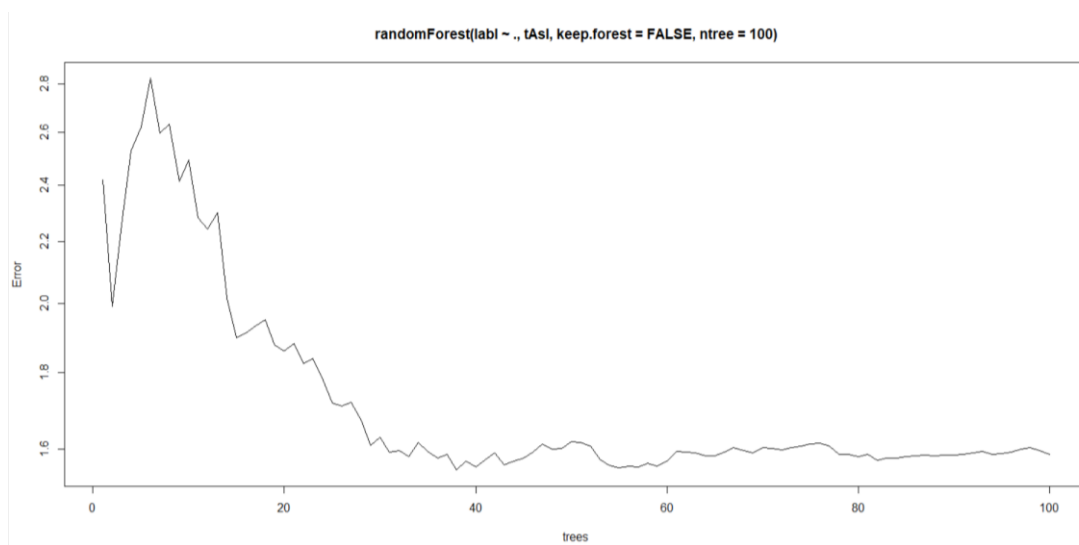


Figure 10. Error rate of number of trees in RF

Finally, the selected features include ten (10FS), five (5FS), and three (3FS) to deploy using a representative machine learning algorithm, such as a RF, supporting vector machines with a radial basis function kernel (SVM RBF), the K-nearest neighbor (KNN) method, and the C5.0 decision tree. Performance results such as accuracy, sensitivity, and specificity are shown in Table 4. The RF algorithm is more dominant than other algorithms, including the AUC performance value as shown in Table 5.

Table 4. Comparison of performance result for the representative machine learning algorithms

Algorithm	Accuracy				Sensitivity				Specificity			
	NonFS	10FS	5FS	3FS	NonFS	10FS	5FS	3FS	NonFS	10FS	5FS	3FS
KNN	0.811	0.833	0.789	0.867	0.686	0.874	0.662	0.87	0.958	0.963	0.953	0.971
SVM	1	0.956	0.856	0.744	1	0.897	0.828	0.748	1	0.989	0.966	0.94
C5.0	0.944	0.778	0.822	0.689	0.951	0.535	0.613	0.533	0.988	0.95	0.96	0.927
RF	1	0.989	0.922	0.789	1	0.993	0.868	0.784	1	0.998	0.982	0.951

Table 5. Comparison of area under the curve (AUC) for the representative machine learning algorithms

Algorithm	AUC			
	NonFS	10 Features	5 Features	3 Features
KNN	0.82	0.92	0.81	0.92
SVM	1	0.94	0.90	0.84
C5.0	0.97	0.74	0.80	0.73
RF	1	1.00	0.93	0.87



#### 4. CONCLUSION

Research related to DNA sequence profiling of the SARS Cov2 virus has been carried out regionally in several ASEAN countries. Profiling was done by clustering after DNA sequence alignment and feature extraction with three grams of four basal codons used as protein-forming codons. The silhouette, entropy, and purity values achieved are close to one, with the number of clusters of 6 countries as the total countries in this study. Then based on the information gain value, 10 features (codons) were selected as predictors in classifying DNA sequences with six countries in ASEAN. Experimental results achieved high performance using RF including accuracy, sensitivity, and specificity. Furthermore, the AUC also reaches 1. It means that the classifier model is perfect.

#### ACKNOWLEDGEMENTS

This research is supported by Faculty of Computer Science, Brawijaya University through the granted program of DIPA 2021 under contract number: 2639/UNI0.F15/PN/2021.




#### REFERENCES

- [1] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *The New England journal of medicine*, vol. 382, pp. 727-733, doi: 10.1056/NEJMoa2001017 2020.
- [2] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *nature*, vol. 579, no. 7798, pp. 270-273, 2020, doi: 10.1038/s41586-020-2012-7.
- [3] Jing Gao *et al.*, "Risk factors of hepatocellular carcinoma-current status and perspectives," *Asian Pacific Journal of Cancer Prevention*, vol. 13, no. 3, pp. 743-752, 2012, doi: 10.7314/APJCP.2012.13.3.743.
- [4] M. Kandpal and R. V. Davuluri, "Identification of geographic specific SARS-Cov-2 mutations by random forest classification and variable selection methods," *Statistics and applications*, vol. 18, no. 1, p. 253, 2020.
- [5] N. Zhu *et al.*, "A Novel Coronavirus from Patients with Pneumonia in China, 2019," *N Engl J Med*, vol. 382, no. 8, pp. 727-733, Feb. 2020, doi: 10.1056/NEJMoa2001017.
- [6] Y. Yin and R. G. Wunderink, "MERS, SARS and other coronaviruses as causes of pneumonia," *Respirology*, vol. 23, no. 2, pp. 130-137, Feb. 2018, doi: 10.1111/resp.13196.
- [7] J. Hadfield *et al.*, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121-4123, 2018, doi: 10.1093/bioinformatics/bty407.
- [8] N. Cho *et al.*, "De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries," *Nat Commun.*, vol. 6, p. 8351, Sep. 2015, doi: 10.1038/ncomms9351.
- [9] "Table of Genetic Code.pdf - Table of Genetic Code a | Course Hero," Accessed on: Feb. 19, 2022. [Online]. Available: <https://www.coursehero.com/file/130275051/Table-of-Genetic-Codepdf>
- [10] R. Salman, V. Kecman, Qi Li, R. Strack, and E. Test, "Fast K-Means Algorithm Clustering," *International Journal of Computer Networks & Communications*, vol. 3, no. 4, pp. 17-31, Jul. 2011, doi: 10.5121/ijcnc.2011.3402.
- [11] M. D. R. Wahyudi, "Evaluation of TF-IDF Algorithm Weighting Scheme in The Qur'an Translation Clustering with K-Means Algorithm," *Journal of Information Technology and Computer Science*, vol. 6, no. 2, pp. 117-129, 2021, doi: 10.25126/jitecs.202162295.
- [12] Y. A. Auliya, W. F. Mahmudy, and Sudarto, "Land Clustering for Potato Plants Using Hybrid Particle Swarm Optimization and K-Means Improved by Random Injection," *Journal of Information Technology and Computer Science*, vol. 4, no. 1, pp. 42-56, 2019, doi: 10.25126/jitecs.20194183.
- [13] A. H. Khaleel and I. Q. Abduljaleel, "A novel technique for speech encryption based on k-means clustering and quantum chaotic map," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 160-170, 2021, doi: 10.11591/eei.v10i1.2405.
- [14] K. B. Kim, D. H. Song, and S.-S. Yun, "Automatic segmentation of wrist bone fracture area by K-means pixel clustering from X-ray image," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 5205, 2019, doi: 10.11591/ijece.v9i6.pp5205-5210.
- [15] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies: 1. Hierarchical systems," *The computer journal*, vol. 9, no. 4, pp. 373-380, 1967, doi: 10.1093/comjnl/9.4.373.
- [16] L. Muflikhah, Widodo, W. F. Mahmudy and Solimun, "DNA Sequence of Hepatitis B Virus Clustering Using Hierarchical k-Means Algorithm," *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2019, pp. 1-4, doi: 10.1109/ICETAS48360.2019.9117565.
- [17] Tung-Shou Chen *et al.*, "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray," *2005 International Symposium on Intelligent Signal Processing and Communication Systems*, 2005, pp. 405-408, doi: 10.1109/ISPACS.2005.1595432.
- [18] J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang, "An effective and efficient hierarchical K-means clustering algorithm," *International Journal of Distributed Sensor Networks*, vol. 13, no. 8, p. 1550147717728627, Aug. 2017, doi: 10.1177/1550147717728627.
- [19] Y. I. Kurniawan, F. Razi, N. Nofiyati, B. Wijayanto, and M. L. Hidayat, "Naive Bayes modification for intrusion detection system classification with zero probability," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, p. 2751, Jan. 2021, doi: 10.11591/eei.v10i5.2833.
- [20] N. P. Aprilia, D. Pratiwi, and A. B. Ariwibowo, "Sentiment Visualization of Covid-19 Vaccine Based On Naive Bayes Analysis," *Journal of Information Technology and Computer Science*, vol. 6, pp. 195-208, Oct. 2021.
- [21] E. A. Mahareek, A. S. Desuky, and H. A. El-Zhni, "Simulated annealing for SVM parameters optimization in student's performance prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, Art. no. 3, Jun. 2021, doi: 10.11591/eei.v10i3.2855.
- [22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [23] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272-278, 2012.
- [24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.




- [25] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007, doi: 10.1093/bioinformatics/btm134.
- [26] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*, Hoboken, NJ: Wiley-IEEE press 2013.
- [27] S. Loukas, "Multi-class Classification: Extracting Performance Metrics From The Confusion Matrix," Jun. 19, 2020. Accessed on: Jan. 6, 2022. [Online]. Available: <https://towardsdatascience.com/multi-class-classification-extracting-performance-metrics-from-the-confusion-matrix-b379b427a872>

## BIOGRAPHIES OF AUTHORS






**Lailil Muflikhah**    received B.Sc. in Informatics Engineering from Sepuluh Nopember Institute of Technology (ITS), M.Sc in IT from Universiti Teknologi PETRONAS, Malaysia, and Doctor in Bioinformatics from Brawijaya University. She is an Associate Professor at Faculty of Computer Science, Brawijaya University. Her research interests include artificial intelligent, bioinformatics, data mining, and machine learning. She can be contacted at email: [lailil@ub.ac.id](mailto:lailil@ub.ac.id).



**Muh. Arif Rahman**    received B.Sc. in Mathematics from Sepuluh Nopember Institute of Technology (ITS), M.Sc. in Computer Science from Universitas Indonesia. His research interests include artificial intelligent, computer vision, and image processing. He can be contacted at email: [m\\_arif@ub.ac.id](mailto:m_arif@ub.ac.id).



**Agus Wahyu Widodo**    received B.Sc. in Engineering from Brawijaya University, M.Sc. in Computer Science from Gadjah Mada University. His research interests include image processing, data mining, artificial intelligent. He can be contacted at email: [a\\_wahyu\\_w@ub.ac](mailto:a_wahyu_w@ub.ac).