Dicle University
Journal of Engineering

https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

A k-mer based metaheuristic approach for detecting COVID-19 variants

Hilal ARSLAN[1*]

[1] Ankara Yıldırım Beyazıt University, Software Engineering Department, hilalarslanceng@gmail.com, Orcid No: 0000-0002-6449-6952

ARTICLE INFO

Article history:

Received 27 October 2022
Received in revised form 31 January 2023
Accepted 3 February 2023
Available online 23 March 2023

Keywords:

COVID-19, SARS-CoV-2, Whale Optimization Algorithm, Classifiers, Feature Selection, Machine Learning

Doi: 10.24012/dumf.1195600

* Corresponding author

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) belongs to coronaviridae family and a change in the genetic sequence of SARS-CoV-2 is named as a mutation that causes to variants of SARS-CoV-2. In this paper, we propose a novel and efficient method to predict SARS-CoV-2 variants of concern from whole human genome sequences. In this method, we describe 16 dinucleotide and 64 trinucleotide features to differentiate SARS-CoV-2 variants of concern. The efficacy of the proposed features is proved by using four classifiers, k-nearest neighbor, support vector machines, multilayer perceptron, and random forest. The proposed method is evaluated on the dataset including 223,326 complete human genome sequences including recently designated variants of concern, Alpha, Beta, Gamma, Delta, and Omicron variants. Experimental results present that overall accuracy for detecting SARS-CoV-2 variants of concern remarkably increases when trinucleotide features rather than dinucleotide features are used. Furthermore, we use the whale optimization algorithm, which is a state-of-the-art method for reducing the number of features and choosing the most relevant features. We select 44 trinucleotide features out of 64 to differentiate SARS-CoV-2 variants with acceptable accuracy as a result of the whale optimization method. Experimental results indicate that the SVM classifier with selected features achieves about 99% accuracy, sensitivity, specificity, precision on average. The proposed method presents an admirable performance for detecting SARS-CoV-2 variants.

Introduction

SARS-CoV-2 detected in 2019 caused a disease called COVID-19 by spreading rapidly around the world. The spread of SARS-CoV-2 in many countries has led to multiple SARS-CoV-2 variants and accurate detection of SARS-CoV-2 variants is crucial to fight the COVID-19 pandemic. Early detection of SARS-CoV-2 is crucial to prevent infection. Several methods have been released to detect SARS-CoV-2. While some methods detect COVID-19 from images belonging to people, others detect the disease from genome sequences. Using genome sequencing including four nucleotides, A, G, C, and T in 30, 000 bps is preferred to monitor SARS-CoV-2 variants. Recent dominant variants of SARS-CoV-2 are B.1.1.7, B.1.351, P.1., B.1.617, and B.1.1.529. The Alpha variant, B.1.1.7 [1] was determined in the United Kingdom in the fall of 2020, and it spreads about 50% more quickly than the original SARS-CoV-2 [2]. Although current treatments against Alpha variant are effective, the Alpha variant may cause more severe COVID-19 disease. Beta variant, B.1.351 [3] is diagnosed in South Africa and Gamma variant, P.1 [4] first detected in Brazil at the end of 2020 spread less quickly than Alpha variant; however, current treatments against Beta and Gamma variants are less effective. Delta variant, B.1.617 [5] first identified in India may cause more severe disease when compared to the other variants. Furthermore, Delta variant spreads about 100% more quickly than SARS-CoV-2 [2]. It is not adequate information on whether it causes more severe COVID-19 disease, or not. Finally, Omicron variant, B.1.1.529 [6] was detected in South Africa in November 2021.

Although several types of studies are released to diagnose SARS-CoV-2 [7, 8, 9, 10, 11, 12, 13], there are a limited number of algorithms for determining SARS-CoV-2 variants. Ahmed et al. [14] clustered Omicron variant by analyzing mutations. Mohiuddin and Kasahara [15] investigated Omicron variant and suggest possible treatment strategies. Wang et al. [16] applied principal component analysis to diagnose COVID-19 by analyzing more than 20,000 RNA sequences. Khan et al. [17] applied deep learning techniques to detect Omicron variant from chest X-ray and computed tomography. Basu and Campbell [18] classified COVID-19 variants by applying deep learning models from genome sequences. They proposed k-mer based long short-term memory model that is an alignment-free method. Their method achieved an accuracy of 92.5%. Mann et al. [19] classified SARS-CoV-2 variants with mass spectrometry. They defined peptide signatures of unique mass to detect SARS-CoV-2 main variants of concern. Recently, Togrul and Arslan [20] proposed a deep learning method to detect SARS-CoV-2 variants and Arslan [21] published a paper to detect SARS-CoV-2 variants in Turkey.

Although there are many studies to detect SARS-CoV-2, a limited number of studies have been published to detect SARS-CoV-2 variants of concern. In this study, we introduce a method for determining SARS-CoV-2 variants from genome sequences. We list our contributions below:

- We proposed an accurate method to detect SARS-CoV-2 variants from SARS-CoV-2 nucleotide sequences
- We describe 16 dinucleotide and 64 trinucleotide features
- Whale optimization algorithm that is a state-of-the-art feature selection method is employed to select most representative features
- We evaluate the effectiveness of dinucleotide and trinucleotide features, separately by using four classifiers, k-nearest neighbor, multilayer perceptron, support vector machines, and random forest
- We construct a large dataset including 223,326 SARS-CoV-2 genome sequences. The dataset includes various types of SARS-CoV-2, Alpha, Beta, Delta, Gamma, and Omicron
- The proposed method accurately detects SARS-CoV-2 variants of concern

The remaining part of this study is organized as follows. The proposed method is introduced in Section 2. Experimental results are evaluated and compared in Section 3. Finally, Section 4 includes the conclusion.

The Proposed Approach for Detecting SARS-CoV-2 Variants

In this section, we introduce the proposed approach for detecting SARS-CoV-2 variants. The fundamental steps of the proposed approach are presented in Figure 1, and detailed steps of the algorithm are also given in Algorithm 1. The algorithm receives complete genome sequences of SARS-CoV-2 as the input. First, features separating SARS-CoV-2 variants are extracted from complete human genome sequences. In this step, we use dinucleotide occurrences or trinucleotide occurrences as features separating SARS-CoV-2 variants. The information in the sequence is stored using four bases, which are adenine (A), thymine (T), cytosine (C), and guanine (G). Dinucleotide is a sequence of two nucleotides, and trinucleotide is a triplet of nucleotides. There are 16 dinucleotide and 64 trinucleotide patterns in total. Thus, we extract 16 dinucleotide features and 64 trinucleotide features for each sequence in the dataset. In this step, we propose to use trinucleotide features since they represent the genome sequence at a higher level. To determine the optimal trinucleotide subset, whale optimization algorithm, which is a state-of-the-art method for feature selection tasks is used. As a result of the whale optimization algorithm, the occurrences of the trinucleotides that are AAA, ACA, ATG, ACC, AGT, AGC, AGG, CTA, CCC, TAT, TTC, TTT, TTG, TCT, GTT, GTG, GCC, GCT, GGC, and GGT are excluded from the feature set, and the remaining 44 trinucleotide occurrences are used to differentiate SARS-CoV-2 variants. We apply four types of machine learning classifiers to evaluate the performance of the proposed features. Next we briefly explain feature selection and classifiers performed in this study.

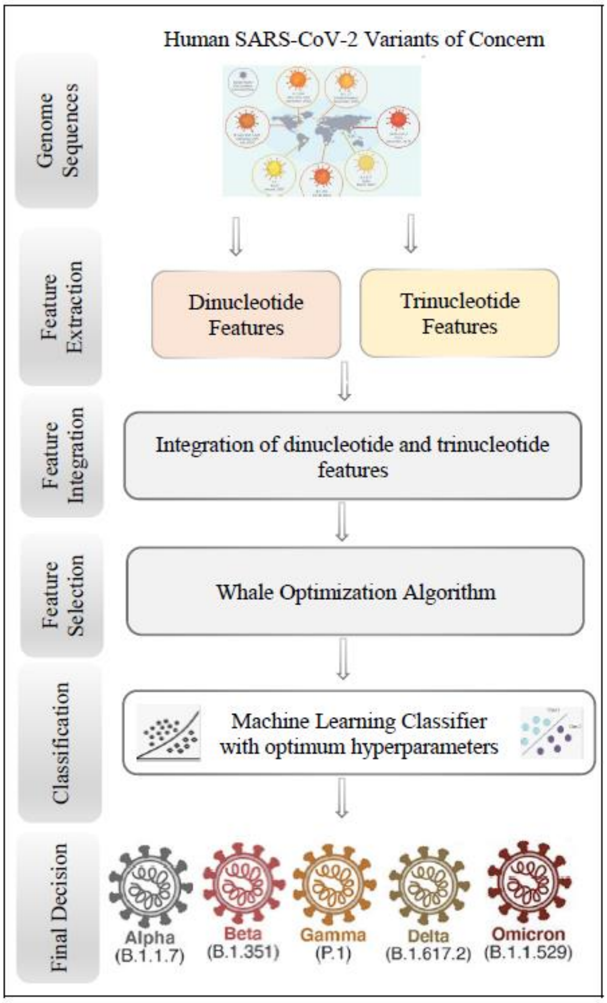


Figure 1. Main steps of proposed algorithm

Whale Optimization Algorithm (WOA) for Feature Selection

We perform WOA to reduce the dimensionality of the data with acceptable accuracy. WOA is a bioinspired algorithm focused on hunting behavior of humpback whales [22]. This method consists of three main steps. The first step is encircling prey, and in this step, the method produces k humpback whales which are randomly scattered in the search space. The best whales are determined by evaluating the position of each humpback whale. The second step is exploitation phase, and humpback whales initiate to attack using a bubble-net strategy in this step. Two strategies used in this step are shrinking encircling and spiral updating position for bubble-net attacking. Each whale proposes a subset of the features that are evaluated based on the accuracy of the classifier. The last step is exploration phase, and in this step, humpback whales look for prey for the position of each other randomly. The main steps and the pseudocode of the WOA can be found in [22].

Algorithm 1 Proposed approach for detecting SARS-CoV-2 variants of concern

```
Inputs:

• Genome sequences of human SARS-CoV-2: genomicData

• Label of each SARS-CoV-2 sequence: Alpha, Beta, Gamma, Delta, and Omicron

• SARS-CoV-2 sequence for testing: unknownSeq

Output: Determine the variant of unknownSeq

Trinucleotide Features:
1: for each sequence seq in genomicData do
2:         Compute trinucleotide features
3: end for Feature Reduction
4: Apply whale optimization algorithm to reduce the number of features

Parameter Tuning
5: Apply grid search to obtain best performing hyperparameters of the classifier

Classification Step:
6: Compute trinucleotide features for unknownSeq
7: Perform the machine learning classifier (SVM is suggested)
8: Determine the variant of unknownSeq
```

Applied Machine Learning Techniques

K-nearest neighbor (KNN) [23, 24] is a non-parametric machine learning method. The method includes *k* hyperparameter that represents the number of neighbors. Data samples are classified with respect to *k* neighbors. The accuracy of the method depends on two hyperparameters, the selection of *k* and distance measures. We perform grid search approach through 5-fold cross validation to define optimum hyperparameters. In grid search approach, k value is chosen between 1 and 10, and the possible distance measures are manhattan, euclidean, and chebyshev.

Multilayer Perceptron (MLP) [25] is a type of artificial neural network. In this study, a MLP model with one hidden layer is used. For optimal determination of the number of neurons in the hidden layer and activation function, we perform 5-fold cross validation with grid search. In grid search, the number of neurons in the hidden layer is set to 50, 100, and 150, and the logistic sigmoid, hyperbolic tangent as well as the rectified linear unit are used as the activation functions.

Support Vector Machines (SVM) [26, 27, 28] is a machine learning method used for solving classification and regression problems. The goal is to construct a hyperplane that separates data samples for the classification problems. We use Radial Basis Function (RBF) for achieving non linearity [29, 30]. The selection of RBF kernel parameter (γ) and penalty parameter (*c*) related to SVM model are crucial. We determine these parameters by performing 5-fold cross validation with grid search. The possible values of *c* are $\{2^{-5}, 2^{-1}, 2^{9}\}$ and the possible values of γ are $\{2^{-9}, 2^{-5}, 2^{-1}, 2^{3}\}$.

Random Forest (RF) [31, 32] is an ensemble classifier that constructs multiple decision trees and subset of training samples are selected randomly. We perform grid search with 5-fold cross validation for achieving the best results.

Results and Discussion

In this section, we conduct several experimental studies to prove the efficacy of the proposed features discriminating SARS-CoV-2 variants. All experiments are implemented on a 64-bit Windows 10 Enterprise operating system running on Intel i7-6700HQ CPU CPU @2.50 GHz processor and 16GB RAM. All methods are implemented using Python language.

Dataset

Our dataset includes SARS-CoV-2 genome sequences of SARS-CoV-2 from the Global Initiative on Sharing All Influenza Data (GISAID) database [33]. All genome sequences in the dataset are complete and high coverage to minimize sequencing errors. WHO Label, scientific name, date of designation, and the number of sequences used in this study are presented in Table 1.

Table 1. Variants of SARS-CoV-2

WHO Label	Scientific Name	Date of Designation	Number of sequences
Alpha	B.1.1.7	October, 2020	54,467
Beta	B.1.351	December, 2020	25,455
Delta	B.1.617.2	October, 2020	46,221
Gamma	P.1	January, 2021	53,501
Omicron	B.1.1.529	November, 2021	43,682

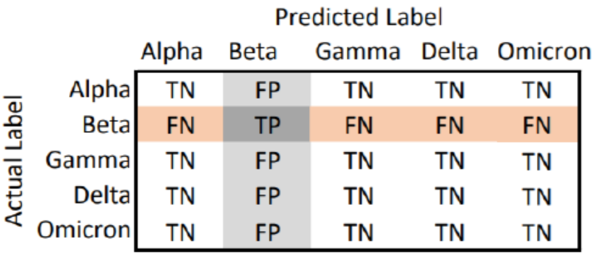
Performance Metrics

We perform multi-class classification since our dataset includes seven variants of SARS-CoV-2. We perform 5-fold cross validation technique to evaluate the performance of the methods. In this approach, the dataset is divided in 5 parts. While four parts are used for training, the other one part is used for resting. The method is continued until all parts are tested. The performances of the classifiers are measured using different metrics, which are precision, sensitivity, specificity, and accuracy. We use macro-averaging [34] to evaluate overall performance of a class as shown in Table 2.

Furthermore, we show the confusion matrices for each classifier separately. We illustrate the confusion matrix focusing on Beta class labelling the tiles accordingly in Figure 2.

Table 2. Performance measurements for evaluating classifiers

Performance Metric	Formula for each class c	Average Metric
Precision(Pre)	$\frac{TP(c)}{TP(c)+FP(c)}$	$\frac{1}{7}\sum_{i=1}^{7} Pre(i)$
Sensitivity(Sen)	$\frac{TP(c)}{TP(c)+FN(c)}$	$\frac{1}{7}\sum_{i=1}^{7} Sen(i)$
Specificity (Spe)	$\frac{TN(c)}{TN(c)+FP(c)}$	$\frac{1}{7}\sum_{i=1}^{7} Spe(i)$
Accuracy(Acc)	$\frac{TP(c)+TN(c)}{TP(c)+FN(c)+FP(c)++TN(c)}$	$\frac{1}{7}\sum_{i=1}^{7} Acc(i)$



Predicted Label

Actual Label	Alpha	Beta	Gamma	Delta	Omicron
Alpha	TN	FP	TN	TN	TN
Beta	FN	TP	FN	FN	FN
Gamma	TN	FP	TN	TN	TN
Delta	TN	FP	TN	TN	TN
Omicron	TN	FP	TN	TN	TN

Figure 2. Confusion matrix for the class Beta

Experimental Results

In this section, we evaluate and present results of the machine learning classifiers using dinucleotide *and* trinucleotide features, separately on the dataset including variants of the SARS-CoV-2.

Results of the machine learning classifiers on dinucleotide features

In this part, we evaluate results of machine learning classifiers to prove the effectiveness of the dinucleotide features to predict SARS-CoV-2 variants. The hyperparameters of the classifiers are determined by grid search with 5-fold cross validation. In the KNN classifier, k is set to 3, and manhattan distance is used. In the MLP method, the hyperbolic tangent function is used as an activation function, and the number of neurons in the hidden layer is 100. In the SVM method, c is 512 and γ is 2^{-9}. Finally, in RF, criterion is *gini*, maximum depth of tree is 25, the minimum number of the sample leaf is 1, and the minimum number of sample split is 3. The results of the classifiers are obtained with respect to these parameters.

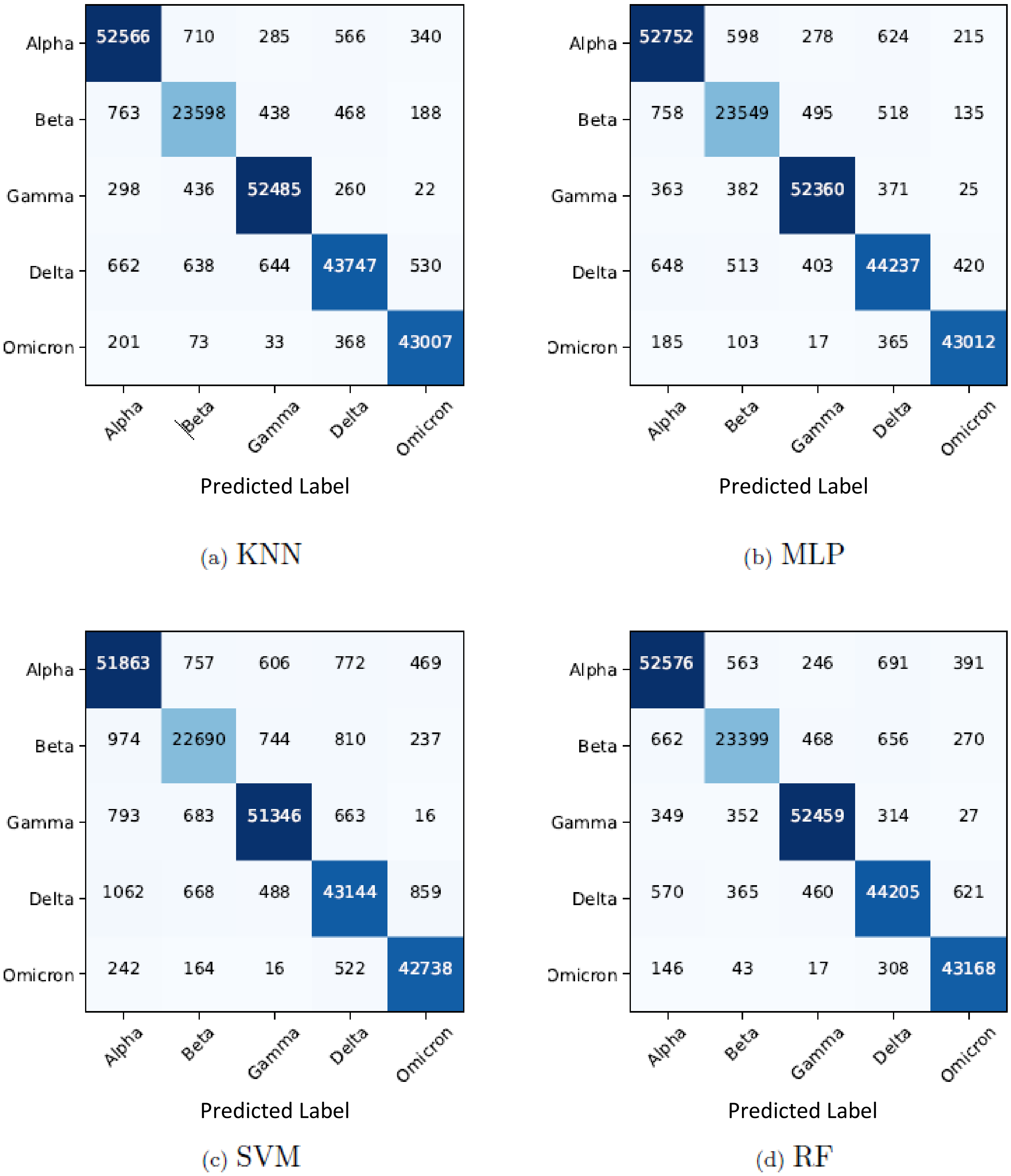
Figure 3 shows confusion matrices of machine learning classifiers on the dinucleotide features extracted from the genome sequences. The results of four classifiers are close to each other. The MLP classifier achieves better performance, and it correctly labels 52,752 of 54,467 genome sequences of *Alpha* variant, 23,549 of 25,455 genome sequences of *Beta* variant, 52,360 of 53,501 genome sequences of *Gamma* variant, 44,237 of 46,221 genome sequences of *Delta* variant, and 43,012 of 43,682 genome sequences of *Omicron* variant. Table 3 presents both variant-based and average results of the machine learning classifiers. As seen in Table 3, average results of the machine learning classifiers are close to each other. The average accuracy values of the classifiers vary between 0.98 and 0.99. Similarly, average specificity values are about 0.99. On the other hand, average sensitivity and precision values are lower when compared to average accuracy and specificity values. Average sensitivity and precision values vary between 0.94 and 0.96.

Table 3. Performances of the machine learning classifiers using dinucleotide features

Method	**SARS-CoV-2 Variant**	**Variant based results**				**Average results**			
		Acc	**Sen**	**Spe**	**Pre**	**Acc**	**Sen**	**Spe**	**Pre**
KNN	Alpha Beta Gamma Delta Omicron	0.98 0.98 0.99 0.98 0.99	0.97 0.93 0.98 0.95 0.98	0.99 0.99 0.99 0.99 0.99	0.96 0.93 0.97 0.96 0.98	0.99	0.96	0.99	0.96
MLP	Alpha Beta Gamma Delta Omicron	0.98 0.98 0.99 0.98 0.99	0.97 0.93 0.98 0.96 0.98	0.99 0.99 0.99 0.99 1	0.96 0.94 0.98 0.96 0.98	0.99	0.96	0.99	0.96
SVM	Alpha Beta Gamma Delta Omicron	0.97 0.98 0.98 0.97 0.99	0.95 0.89 0.96 0.93 0.98	0.98 0.99 0.99 0.98 0.99	0.94 0.91 0.97 0.94 0.96	0.98	0.94	0.99	0.94
RF	Alpha Beta Gamma Delta Omicron	0.98 0.98 0.99 0.98 0.99	0.97 0.92 0.98 0.96 0.99	0.99 0.99 0.99 0.99 0.99	0.97 0.95 0.98 0.96 0.97	0.99	0.96	0.99	0.96

Results of the machine learning classifiers on trinucleotide features

In this part, we evaluate results of machine learning classifiers to prove the effectiveness of the trinucleotide features to predict SARS-CoV-2 variants. The hyperparameters of the classifiers are determined by grid search. In the KNN classifier, k is set to 1, and manhattan distance is used. In the MLP method, the hyperbolic tangent function is used as an activation function, and number of



(a) KNN

(b) MLP

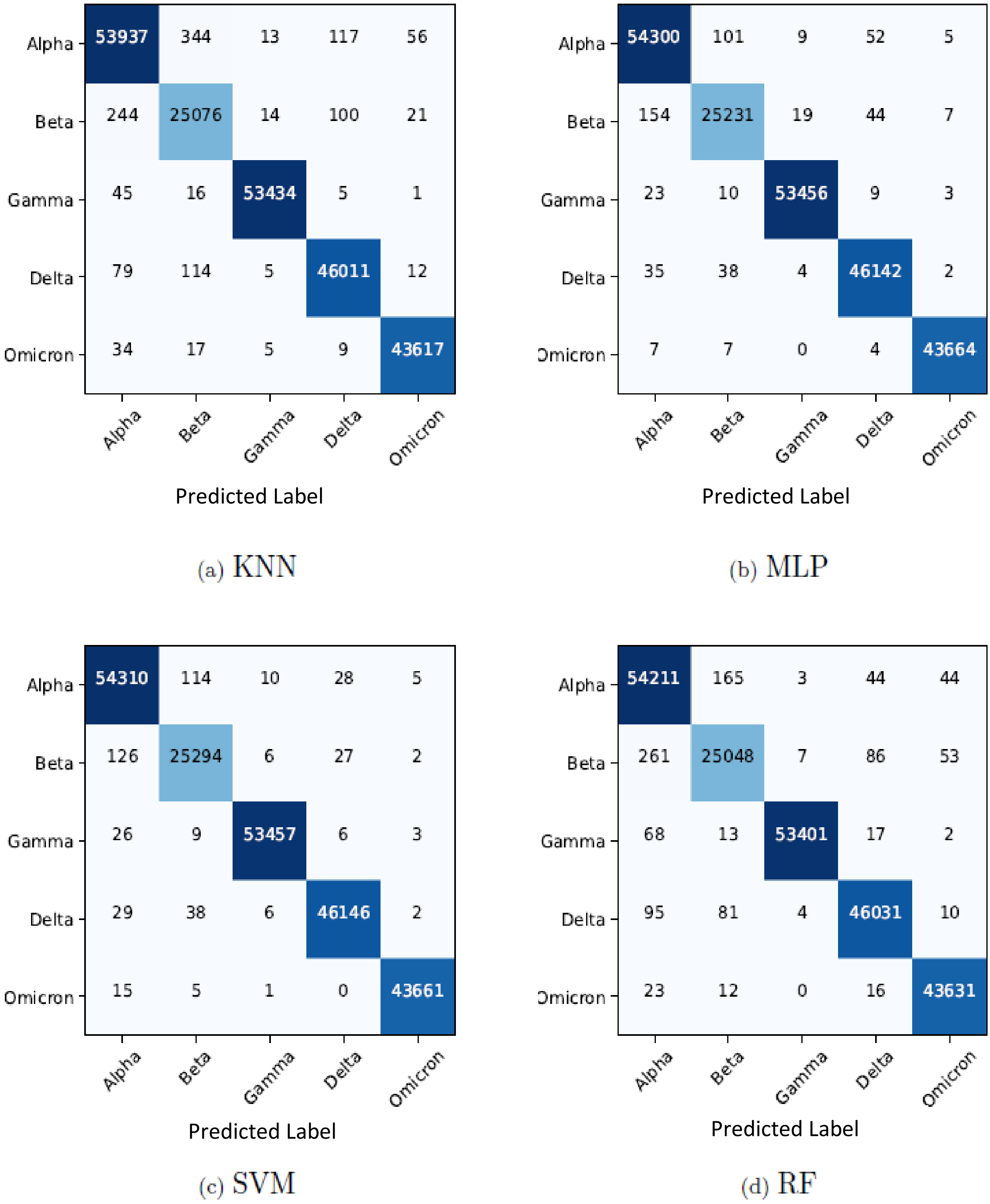
(c) SVM

(d) RF

Figure 3 Confusion matrices of machine learning classifiers on the dinucleotide features

neurons in the hidden layer is 100. In the SVM method, c is 512, γ is 2^{-9}, and radial basis function is used. Finally, in RF, criterion is *gini*, maximum depth of tree is 25, the minimum number of the sample leaf is 1, and the minimum number of sample split is 3. The results of the classifiers are obtained with respect to these parameters.

Figure 4 shows confusion matrices of machine learning classifiers on the trinucleotide features extracted from the genome sequences. The SVM classifier achieves the best results and it correctly labels 54,310 out of 54,467 genome sequences of *Alpha* variant, 25,294 out of 25,455 genome sequences of *Beta* variant, 53,457 out of 53,501 genome sequences of *Gamma* variant, 46,146 out of 46,221 genome sequences of *Delta* variant, and 43,661 out of 43,682 genome sequences of *Omicron* variant.



(a) KNN

(b) MLP

(c) SVM

(d) RF

Figure 4. Confusion matrices of machine learning classifiers on the trinucleotide features

Table 4 presents variant-based and average results of the machine learning classifiers when trinucleotide features are used. As seen in Table 4, the machine learning classifiers with trinucleotide features have an admirable performance. The average accuracy and specificity values of the classifiers are close to 1.0. Average sensitivity and precision values vary between 0.99 and 1.0. When compared to dinucleotide features, sensitivity and precision values are significantly improved with trinucleotide features.

Table 4. Performances of the machine learning classifiers using trinucleotide features

Method	SARS-CoV-2 Variant	Variant based results				Average results			
		Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
KNN	Alpha	1	0.99	1	0.99	1	0.99	1	0.99
	Beta	1	0.99	1	0.98				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				
MLP	Alpha	1	1	1	1	1	1	1	1
	Beta	1	0.99	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				
SVM	Alpha	1	1	1	1	1	1	1	1
	Beta	1	0.99	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				
RF	Alpha	1	1	1	0.99	1	0.99	1	0.99
	Beta	1	0.98	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				

Feature Selection using Whale Optimization Algorithm

The trinucleotide features identify SARS-CoV-2 variants more accurately than dinucleotide features. In order to choose the most relevant trinucleotide features and reduce dimensionality of the dataset, we use the WOA. As a result of the WOA, the trinucleotides occurrences that are AAA, ACA, ATG, ACC, AGT, AGC, AGG, CTA, CCC, TAT, TTC, TTT, TTG, TCT, GTT, GTG, GCC, GCT, GGC, and GGT are excluded from the feature set, and the remaining trinucleotide occurrences are used to detect SARS-CoV-2 variants. Thus, the initial set of 64 features is reduced to 44. Table 5 presents variant-based and average results of the machine learning classifiers when 44 trinucleotide features are used. When the SVM classsifier with 44 trinucleotide features is used, an average accuracy, precision, sensitivity, and specificity is ∼ 1.0. Furthermore, Table 6 presents average results of classifiers using trinucleotide features for full set of features (64 features in total) and reduced set of features (44 features in total). The results of full set and the reduced set of features are close to each other.

We present the total number of incorrectly classified instances for each classifier when 16 dinucleotide, 64 trinucleotide, and 44 selected features are separately used in Figure 5. As can be seen in Figure 5a, the number of genome sequences that are incorrectly classified remarkably decreases when the trinucleotide features are used. For instance, the SVM classifier with trinucleotide features misclassifies 157 out of 54,467 genome sequences of *Alpha* variant, 161 out of 25,455 genome sequences of *Beta* variant, 44 out of 53,501 genome sequences of *Gamma* variant, 75 out of 46,221 genome sequences of *Delta* variant, and 21 out of 43,682 genome sequences of *Omicron* variant. In total, it misclassifies 458 out of 223,326 genome sequences. Furthermore, the results of 64 trinucleotide features and 44 selected features are close as shown in Figure 5b.

Table 5 Performances of the machine learning classifiers using 44 trinucleotide features

Method	SARS-CoV-2 Variant	Variant based results				Average results			
		Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
KNN	Alpha	1	0.99	1	0.99	1	0.99	1	0.99
	Beta	1	0.98	1	0.98				
	Gamma	1	1	1	1				
	Delta	1	0.99	1	1				
	Omicron	1	1	1	1				
MLP	Alpha	1	1	1	1	1	1	1	1
	Beta	1	0.99	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				
SVM	Alpha	1	1	1	1	1	1	1	1
	Beta	1	0.99	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				
RF	Alpha	1	0.99	1	0.99	1	0.99	1	0.99
	Beta	1	0.98	1	0.99				
	Gamma	1	1	1	1				
	Delta	1	1	1	1				
	Omicron	1	1	1	1				

Table 6. Average results of classifiers using trinucleotide features for full set of features and reduced set of features

Method	Full set of features				Reduced set of features			
	Average(%)				Average(%)			
	Acc	Sen	Spe	Pre	Acc	Sen	Spe	Pre
KNN	1.0	0.99	1.0	0.99	1.0	0.99	1.0	0.99
MLP	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SVM	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
RF	1.0	0.99	1.0	0.99	1.0	0.99	1.0	0.99

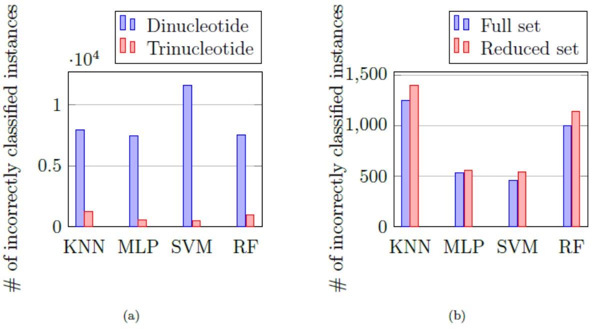


Figure 5. Incorrectly classified instances for each classifier

Comparison with Existing Methods Detecting SARS-CoV-2 Variants

Table 7 analyzes the performances of the methods detecting SARS-CoV-2 variants in the literature. Jamil and Rahman [17] proposed a deep learning approach to detect SARS-

CoV-2 variants, Alpha, Beta, Gamma, and Delta from CT scans and X-ray images. They used five convolution units with a rectified unit as an activation function. They reported accuracy results for each variant. Their prediction accuracies are 99.7%, 99.6%, 99.6%, 98.6% for detecting Alpha, Beta, Gamma, and Delta variants, respectively on X-ray images. Ali et al. [16] proposed to use k-mer based features to detect SARS-CoV-2 variants. Then they applied lasso regression and ridge regression methods, which are feature selection methods to reduce the dimension of the dataset. As a result of lasso regression, they used 964 features out of 4977 to predict SARS-CoV-2 variants. K-means clustering with lasso regression achieved F1-scores of 99.87%, 27.05%, 99.91%, 99.98%, and 97.04% for identifying Alpha, Beta, Delta, Gamma, and Epsilon variants, respectively. Their method failed to predict Beta variants. Togrul and Arslan [20] performed CNN to detect features that discriminate variants of SARS-CoV-2. After feature extraction, they used various types of ML algorithms including SVM, KNN, RF, and MLP. Their experimental results achieved about 100% accuracy on the dataset including 1000 sequences of each variants of concern when 1563 features are used. Main disadvantage of their method was the use a large number of features, which required a lot of time. Arslan [21] used nucleotide frequencies to diagnose SARS-CoV-2 variants. Their method reached a relatively low accuracy (94%) on average using a dataset including fewer sequences from Turkey when four features are used.

Table 7. Comparison of the methods identifying SARS-CoV-2 variants

Study	Method	Fetaures	Image Dataset	Acc(%)
Jamil and Rahman [17]	CNN	Vocabulary of features	1345 Alpha 10,192 Beta 6,012 Gamma 3,616 Delta	99.7 99.6 99.6 98.6
Study	**Method**	**# of Fetaures**	**# of Amino Acid Sequence**	**Acc(%)**
Ali et al. [16]	K-means with Lasso Regression	964	13,966 Alpha 1,727 Beta 7,551 Delta 26,629 Gamma 12,784 Epsilon	99.87 27.05 99.91 99.98 97.04
Togrul and Arslan [20]	CNN,KNN,MLP, SVM, RF	1563	1000 Alpha 1000 Beta 1000 Gamma 1000 Delta 1000 Omicron	100 100 100 100 100
Arslan [21]	KNN	4	436 Alpha 357 Beta 110 Gamma 500 Delta	94 93 93 95
Study	**Method**	**# of Fetaures**	**# of Amino Acid Sequence**	**Acc(%)**
Proposed Method	SVM with WOA	44 trinucleotide	54,467 Alpha 25,455 Beta 53,501 Gamma 46,221 Delta 43,682 Omicron	100 100 100 100 100

When we compare the proposed method with these methods, the dataset used in this study is larger and includes current SARS-CoV-2 variants of concern. Moreover, most of the methods shown in Table 7 are more expensive than the proposed method since our method predicts SARS-CoV-2 variants by using fewer number of features. The proposed method can accurately predict current SARS-CoV-2 variants of concern, and achieves an accuracy of 100%.

Conclusion

Emerging variants of SARS-CoV-2 causes a devastating effect on human health. Determining variants of SARS-CoV-2 is crucial to follow correct treatment strategy and taking under control to contagious of the virus. In this study, we introduce a method to determine SARS-CoV-2 variants. We determine 16 dinucleotide and 64 trinucleotide features representing the whole genome sequences. The WOA is applied to select the most relevant features and reduce the dimensionality. The proposed method reaches full accuracy for detecting current SARS-CoV-2 variants of concern when the SVM classifier with 44 trinucleotide features are employed. In future, we will investigate effect on SARS-CoV-2 variants on patients with any types of cancer to decrease the date ratio of COVID-19.

Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared.

There is no conflict of interest with any person / institution in the article prepared.

Authors' Contributions

All parts of the paper are prepared and implemented by Hilal Arslan.

References

[1] Volz, E., Mishra, S., Chand, M., Barrett, J. C., & al., R. J. et. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, *593*(7858), 266–269. doi:10.1038/s41586-021-03470-x

[2] Lauring, A. S., & Malani, P. N. (09 2021). Variants of SARS-CoV-2. *JAMA*, *326*(9), 880–880. doi:10.1001/jama.2021.14181

[3] Tegally, H., Wilkinson, E., Giovanetti, M., & al., A. I. et. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, *592*(7854), 438–443. doi:10.1038/s41586-021-03402-9

[4] Sabino, E. C., Buss, L. F., Carvalho, M. P. S., & al., E. (2021). Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet*, *397*(10273), 452–455. doi:10.1016/s0140-6736(21)00183-5

[5] Mlcochova, P., Kemp, S. A., Dhar, M. S., & al., G. P. et. (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*, *599*(7883), 114–119. doi:10.1038/s41586-021-03944-y

[6] Sahoo, J. P., & Samal, K. C. (2021). World on alert: WHO designated south African new COVID strain (Omicron/B.1.1.529) as a variant of concern. *Biotica Research Today*, *3*(11), 1086–1088.

[7] Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., … Huang, Y. (2020). Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. *Computers, Materials $\&$ Continua*, *62*(3), 537–551. doi:10.32604/cmc.2020.010691

[8] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4*(1), 3. doi:10.1038/s41746-020-00372-6

[9] Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, *2*(1), 11. doi:10.1007/s42979-020-00394-7

[10] Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., … Shen, D. (2021). Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, *14*, 4–15. doi:10.1109/RBME.2020.2987975

[11] Mohamadou, Y., Halidou, A., & Kapen, P. T. (2020). A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. *Applied Intelligence*, *50*(11), 3913–3925. doi:10.1007/s10489-020-01770-9

[12] Arslan, H., & Arslan, H. (2021). A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Engineering Science and Technology, an International Journal*. doi:10.1016/j.jestch.2020.12.026

[13] Arslan, H. (2021a). COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. *Computers $\&$ Industrial Engineering*, *161*, 107666. doi:10.1016/j.cie.2021.107666

[14] Ahmed, W., (2022). Detection of the Omicron (B.1.1.529) variant of SARS-CoV-2 in aircraft wastewater. In *Science of The Total Environment*, 820, p. 153171. doi:10.1016/j.scitotenv.2022.153171

[15] Mohiuddin, M., & Kasahara, K. (2022). Investigating the aggressiveness of the COVID-19 Omicron variant and suggestions for possible treatment options. In *Respiratory Medicine*, vol. 191, p. 106716. doi: 10.1016/j.rmed.2021.106716

[16] Wang, B., & Jiang, L. (2021). Principal Component Analysis Applications in COVID-19 Genome Sequence Studies. In *Cognitive Computation*. doi:10.1007/s12559-020-09790-w

[17] Khan, A., Khan, S. H., Saif, M., Batool, A., Sohail, A., & Waleed Khan, M. (2023). A Survey of Deep Learning Techniques for the Analysis of COVID-19 and their usability for Detecting Omicron. In *Journal of Experimental & Theoretical Artificial Intelligence* pp. 1–43. doi: 10.1080/0952813x.2023.2165724

[18] Basu, S., & Campbell, R. H. (2022). Classifying COVID-19 Variants Based on Genetic Sequences Using Deep Learning Models. In *Springer Series in Reliability Engineering* pp. 347–360. doi: 10.1007/978-3-031-02063-6_19

[19] Mann, C., Griffin, J. H., & Downard, K. M. (2021). Detection and evolution of SARS-CoV-2 coronavirus variants of concern with mass spectrometry. *Analytical and Bioanalytical Chemistry*, *413*(29), 7241–7249. doi:10.1007/s00216-021-03649-1

[20] M. Togrul and H. Arslan. (2022). Detection of SARS-CoV-2 Main Variants of Concerns using Deep Learning. *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Antalya, Turkey, 2022, pp. 1-5, doi: 10.1109/ASYU56188.2022.9925559.

[21] Arslan, H. (2022). Classification of SARS-CoV-2 Variants in Turkey. *Journal of Turkish Operations Management*, 6(1), 1092–1101.

[22] Mafarja, M., & Mirjalili, S. (2018). Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, *62*, 441–453. doi:10.1016/j.asoc.2017.11.006

[23] Abu Alfeilat, H., Hassanat, A., Lasassmeh, O., Tarawneh, A., Alhasanat, M., Eyal-Salman, H., & Prasath, S. (08 2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, *7*. doi:10.1089/big.2018.0175

[24] Bishop, C. M. (2006). *Pattern recognition and Machine Learning*. Springer.

[25] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. doi:10.1016/0893-6080(89)90020-8

[26] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. doi:10.1023/A:1009715923555

[27] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. doi:10.1007/978-1-4757-2440-0

[28] Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. doi:10.1109/72.991427

[29] Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, *28*(4), 603–614. doi:10.1016/j.eswa.2004.12.008

[30] Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. *Neural Computation*, *15*(7), 1667–1689. doi:10.1162/089976603321891855

[31] Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

[32] Breiman, L. (2001b). *Machine Learning*, *45*(1), 5–32. doi:10.1023/a:1010933404324

[33] Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, *22*(13). doi:10.2807/1560-7917.ES.2017.22.13.30494

[34] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing $\&$ Management*, *45*(4), 427–437. doi:10.1016/j.ipm.2009.03.002