Journal of Turkish

Operations Management

Classification of SARS-CoV-2 variants in Turkey

Hilal Arslan

Department of Software Engineering, Faculty of Engineering and Natural Sciences, Ankara Yıldırım Beyazıt University, Ankara
e-mail: hilalarslanceng@gmail.com, ORCID No: http://orcid.org/0000-0002-6449-6952

Article Info

Article History:
Received: 08.04.2022
Revised: 22.04.2022
Accepted: 25.04.2022

Keywords

SARS-CoV-2 variants,
COVID-19,
Coronavirus,
Classifiers

Abstract

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) causes the COVID-19 disease, which turns into a pandemic and threatens public health. Appearing of SARS-CoV-2 variants shows a significant challenge in determining the risk of infection, develop vaccines as well as antiviral agents, monitor the changes, and assess the evolution of SARS-CoV-2. In this study, we propose a method for identifying SARS-CoV-2 variants in Turkey. To achieve this goal, nucleotide occurrences are computed from the whole genome sequences that include four nucleotides, A, C, T, and G. Thus, 30 000 bps genome sequences are represented by only four integer numbers. After features are extracted, four classification methods, support vector machines, k-nearest neighbor, neural network, and decision tree are employed to identify SARS-CoV-2 variants. Experimental results are conducted on a dataset including 1403 genome sequences from Turkey and belonging to variants of SARS-CoV-2, B.1.1.7 (Alpha), B.1.351 (Beta), P.1. (Gamma), as well as B.1.617 (Delta). Experimental results present that the KNN classifier achieves an accuracy of 0.94, a precision of 0.81, a recall of 0.80, and an F-score of 0.80 on average.

1. Introduction

Similar to other RNA viruses, SARS-CoV-2 continually changes through mutation and new variants have emerged over time. While SARS-CoV-2 has been spreading, the genetic code of the virus is constantly changed through mutations. Khateeb and Zhang (2021) investigated mutations in variants of SARS-CoV-2. While most of the mutations in SARS-CoV-2 have a lower impact on the functioning of the SARS-CoV-2, mutations in the spike protein of SARS-CoV-2 have a considerable effect on spread of the virus (Lauring and Malani, 2021).

Various types of SARS-CoV-2 are defined and only a small proportion of SARS-CoV-2 variants threatens public health since they can be more transmissible and cause more severe disease. With respect to their crucial spread and increasing death rates, there are five variants of SARS-CoV-2 called as variants of concern, recently. They are B.1.1.7 (Alpha) (Volz et al., 2021), B.1.351 (Beta) (Tegally et al., 2021), P.1. (Gamma) (Sabino et al., 2021), B.1.617 (Delta) (Micochova et al., 2021), and B.1.1.529 (Omicron) (Padane et al., 2022). The Alpha variant is known to come from the UK at the end of 2020. The Beta variant originates from South Africa in late 2020 and recent studies present that Beta variant is less effective to vaccination when comparing to the other variants (Hatirnaz et al., 2021). Gamma variant determined in September 2020 in Brazil caused serious disease and has a high infection rate as well as mortality (Faria et al., 2021). The studies showed that the virus level is 3-4 times higher than earlier variants in Gamma variant (Tao et al., 2021). Delta variant identified in India in early 2021 showed increased transmissibility. Turkey has the highest number of cases for Delta variant in the GISAID (Shu and McCauley, 2017). A study published by Davies et al. (2021) indicated that the Alpha variant is more transmissible than the Beta as well as Gamma variants, and results in increasing severity of the illness with

infection. On the other hand, the Delta variant is over five times as contagious when compared with previous variants, and recent researches have indicated that it might be more possibly than the original SARS-CoV-2 to drop infected people in the hospital. The Omicron variant identified on 26 November 2021 in South Africa. Kandeel et al. (2021) investigated Omicron variant, and they concluded that Omicron variant includes the largest number of genomic mutations among all variants, and when the sequence identity is considered, Alpha variant is the closest variant of Omicron.

In this study, we focus on four variants of concerns detected in Turkey including Alpha, Beta, Gamma, and Delta. Omicron variant was not considered in this study since this variant was not yet seen in Turkey. We propose an efficient method for identifying these variants in Turkey using complete genome sequences. The genome sequences include four nucleotides, A, C, T, and G and each genome sequence used in this study includes approximately 30 000 nucleotides. Based on the nucleotide occurrence, four discriminative features are extracted by computing the frequency of nucleotide. Four machine learning methods, support vector machine (SVM) (Noble, 2006), k-nearest neighbor (KNN) (Guo et al., 2003), neural network (NN) (Charytoniuk et al., 2000), and decision tree (DT) (Galar et al., 2011) methods are employed to achieve classification. The other parts of this study is organized as follows. In Section 2, related works are given. Section 3 includes the proposed method. We present and analyze the results in Section 4. Finally, Section 5 includes conclusion.

2. Related Works

It is crucial to identify variants of SARS-CoV-2 since determining specific variants aids in to understanding their underlying patterns, propose effective strategies, determine the capability of known vaccines as well as fight future outbreaks. There is a recent study for identifying variants of the human SARS-CoV-2. Ali et al. (2021) used spike sequences to classify variants and used order information of the amino acids. Harvey et al. (2021) discussed the mutations of SARS-CoV-2 spike proteins by focusing on their effects on antigenicity and also discussed mutation frequencies. Simon-Loriere and Schwartz (2022) explained the concept of SARS-CoV-2 serotype that is a variation in a microbial species distinguished by humoral immune response. Burioni and Topol (2021) discussed the human immune response to variants of SARS-CoV-2. Arora, Kumar, and Panigrahi (2020) predicted COVID-19 cases in India using deep learning methods. Arslan and Arslan (2021) used machine learning techniques with CpG based features and similarity features to detect COVID-19 positive cases (Arslan, 2021). Garcia-Beltran et al. (2021) proposed a method for predicting COVID-19 disease severity. They considered 113 patients with COVID-19 and detected severe cases resulting in death and incubation. Han and Ye (2021) published a review on main variants of SARS-CoV-2 and its effects on vaccines. In another study, Arslan and Aygun (2021) detected COVID-19 cases from main symptoms as well as basic information about the patient such as age, gender, and contact with a person with COVID-19. Mann et al. (2021) detected SARS-CoV-2 variants with mass spectrometry. They determined peptide signatures of unique mass to identify variants of SARS-CoV-2, Alpha, Beta, Gamma, and Delta. Davi et al. (2021) characterized SARS-CoV-2 genome sequences using sequence alignment techniques.

3. Proposed Method

The main focus of this study is to present a stable and efficient method that successfully separates SARS-CoV-2 variants. To achieve this goal, whole genome sequences belonging to the Alpha, Beta, Gamma, and Delta variants of the SARS-CoV-2 are used. The pseudocode of the proposed method is given in Algorithm 1 and the main steps of the proposed method is also shown in Figure 1. To extract the features discriminating SARS-CoV-2 variants, we compute the frequency of each nucleotide in the genome sequences. The frequency is the number of occurrences of each nucleotide. Since genome sequences include four nucleotides, A, C, T, and G, we compute the frequency of A, C, T, and G. Thus, each SARS-CoV-2 sequence is represented by four integer values. After the feature extraction step, any machine learning method may be performed to classify SARS-CoV-2 variants and in this study, we use Support Vector Machines (SVM), k-Nearest Neighbor (KNN), Neural Network (NN), and Decision Tree (DT) classifiers. In the following, we briefly explained these machine learning methods.

SVM is a classification algorithm that is frequently used for solving binary classification or multi-class classification problems (Arslan, 2021). It is based on statistical learning theory, decision planes, as well as risk minimization. In the SVM method, each sample in the training data is treated as a point in an n-dimensional feature space, and the hyperplane separates these points with respect to different classes. The main purpose of the algorithm is to ensure that the hyperplane constructed has the largest margin.

In the KNN (Hamed et al., 2020) algorithm, the k which is a user defined constant value, is first determined, and then k closest samples are chosen. In the selection of the k samples, any distance measurements can be used such as Euclidean, Manhattan, and Minkowski. The newest data is classified with respect to majority voting. Since the KNN is a simple and easy to implement, it is among popular algorithms for solving classification problems.

Algorithm 1 Classification of the SARS-CoV-2 variants

Require: SARS-CoV-2 genome sequences
Ensure: Determine the variant of given SARS-CoV-2 sequences
Feature Extraction
for each genome sequences **do**
Compute the frequency of C nucleotide
Compute the frequency of G nucleotide
Compute the frequency of A nucleotide
Compute the frequency of T nucleotide
end for
Classification
Use any machine learning method to determine the variant of the SARS-CoV-2 sequences

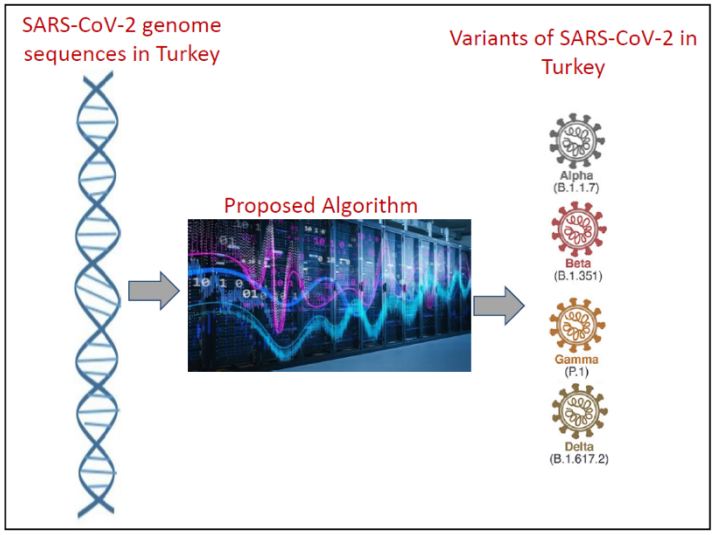

Figure 1. Proposed method

NN (Hasan, 2020) classifier is inspired by the biological nervous system of the human brain. It consists of an input layer, multiple hidden layers, and an output layer. Each node that is also called artificial neuron connects another node with respect to the determined threshold as well as associated weight. The node is activated if the output value of the node is above this threshold and the node sends its data to the next layer. Otherwise, data of the node is not passed to the next layer.

DT (Yoo et al., 2020) is a classifier that identifies the class labels of the samples by creating a tree-shaped model from the training data sets. The decision tree consists of a root node that does not receive any input and internal nodes that receive inputs. The output of one node is passed as input to another node. Inputs and outputs are called branches in the tree concept. If the output of a node in the tree is not transmitted as an input to another node, that node is called a leaf node. The concepts of node, branch and leaf in the tree structure symbolize the attribute, attribute values and class labels, respectively, in the dataset. Decision trees are frequently used in many classification studies because they are easy to construct and interpret algorithms. Another advantage of decision trees is that they do not take any external parameters and can be used to estimate both categorical and numerical values. The most important feature that distinguishes decision trees from other classifiers is that when estimating the class label of a sample, which attribute is more decisive.

4. Results

In this part, we evaluate the results of the DT, KNN, SVM, and NN methods when the nucleotide occurrences are used to discriminate SARS-CoV-2 variants. In DT method, Gini's diversity index is used, and the maximum number of splits is set to 20. In KNN classifier, the number of neighbors is set to 10 and the distance metric is chosen as cosine. In SVM classifier, Gaussian kernel is used. In NN method, two fully connected layers are used. The first and second layer sizes are set to 10, and ReLU activation function is used.

4.1 Dataset

In this study, the experiments were carried out on the SARS-CoV-2 genome sequences which were extracted from the Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017) and the location is chosen as Turkey. The properties of the human complete genome sequences are shown in Table 1. Four variants of the SARS-CoV-2 sequences, which are alpha (B.1.1.7), beta (B.1.351), gamma (P.1), and delta (B.1.617), are emerged in Turkey. We download all available sequences of alpha, beta, and gamma variants in Turkey and repeated sequences are removed. We used 436 genome sequences of Alpha variant, 357 genome sequences of Beta variant, 110 genome sequences of Gamma variant, and 500 genome sequences of Delta variant. We used 1403 sequences in total, and all genome sequences are complete as well as high quality. We note there are many sequences from Delta variant in Turkey and we used 500 Delta variant of the SARS-CoV-2 sequences to have a balanced dataset.

Table 1. The properties of SARS-CoV-2 genome sequences

Variant	The number of genome sequences
Alpha (B.1.1.7)	436
Beta (B.1.351)	357
Gamma (P.1)	110
Delta (B.1.617)	500

4.2 Training and Testing

The k-fold cross validation, which is a more acceptable method to reserve the dataset as training and test data, is employed to evaluate the performance of the classifiers (Refaelzadeh et al., 2009). In the k-fold cross validation, the dataset is separated into k parts and while four sets are used as the training, the remaining one subset is used to test dataset. Thus, it minimizes the bias and variance to overcome the overfitting problem. In this study, *k* is set to 5 and computed the average performance measures to evaluate the effectiveness of the proposed method.

4.3 Performance Parameters

The performances of the classifiers are evaluated and compared by utilizing the precision, recall, accuracy, and F-score metrics that are given in Equations 1-4. Respectively. For each class c,

$$Precision(c) = \frac{TP(c)}{TP(c) + FP(c)} \tag{1}$$

$$Recall(c) = \frac{TP(c)}{TP(c) + FN(c)} \tag{2}$$

$$Accuracy(c) = \frac{TP(c) * TN(c)}{TP(c) + TN(c) + FN(c) + FP(c)} \tag{3}$$

$$F - score(c) = \frac{2 * Precision(c) * Recall(c)}{Precision(c) + Recall(c)} \tag{4}$$

where *TP(c)* is the number of sequences correctly classified for class *c*, whereas *FN(c)* and *FP(c)* are incorrectly classified sequences genome sequences on the row and column of class *c*, respectively. *TN(c)* are all the other tiles. An example of confusion matrix for class *c* is shown in Figure 2.

Actual Label	Predicted Label: Alpha	Beta	Gamma	Delta
Alpha	TN	FP	TN	TN
Beta	FN	TP	FN	FN
Gamma	TN	FP	TN	TN
Delta	TN	FP	TN	TN

Figure 2. Confusion matrix for class Beta.

4.4 Experiments and results

After four nucleotide features are extracted from SAR-CoV-2 sequences, DT, KNN, NN, and SVM classifiers are applied. To evaluate and compare the performance of each classifier, 5-fold cross validation technique is employed. Figure 3 presents confusion matrices for DT, KNN, NN, and SVM methods, separately, and Table 2 summarizes the performance of the classifiers with respect to average performance measurements. First, we evaluate results of DT classifier. It correctly classifies 387 out of 436 genome sequences of Alpha variant as can be seen in Figure 3, and achieves 0.89 precision, recall, and F-measure values, and 0.93 accuracy value as in shown in Table 2. It correctly labels 317 out of 357 genome sequences of Beta variant, and achieves 0.84, 0.89, 0.93, and 0.86 precision, recall, accuracy, and F-measure values, respectively. When we analyze Gamma variant, DT classifier correctly classifies 42 out of 110 genome sequences of Gamma variant, and achieves 0.61, 0.38, 0.93, and 0.47 precision, recall, accuracy, and F-measure values, respectively. When we compare to the other variants, the performance results related to the Gamma variant has lower than the other variants. The main reason for this is that the dataset includes fewer genome sequences for the Gamma variant, and we have less information about the Gamma variant. Finally, the DT classifier successfully classifies 459 out of 500 genome sequences of Delta variant, and achieves 0.88, 0.92, 0.93, and 0.90 precision, recall, accuracy, and F-measure values, respectively. On average, the DT classifier achieves 0.8 precision, 0.77 recall, 0.93 accuracy, and 0.78 F-measure.

Second, we evaluate the results of the KNN classifier based on variants. The KNN achieves better performance, and it correctly labels 403 out of 436 genome sequences of Alpha variant, and achieves 0.89, 0.92, 0.94, and 0.91 precision, recall, accuracy, and F-measure values, respectively. When we look at the results of Beta variant, it correctly classifies 319 out of 357 genome sequences, and achieves 0.86, 0.89, 0.93, and 0.87 precision, recall, accuracy, and F-measure values, respectively. For Gamma variant, it correctly labels 49 of 110 genome sequences, and achieves 0.58, 0.45, 0.93, and 0.50 precision, recall, accuracy, and F-measure values, respectively. Finally, it correctly classifies 459 out of 500 genome sequences of Delta variant, and achieves 0.93, 0.92, 0.95, and 0.92 precision, recall, accuracy, and F-measure values, respectively. On average, the KNN classifier achieves 0.81 precision, 0.80 recall, 0.94 accuracy, and 0.80 F-measure.

Third, we evaluate results of the SVM classifier. It correctly classifies 401 out of 436 genome sequences of Alpha variant as can be seen in Figure 3, and achieves 0.88, 0.92, 0.94, and 0.90 precision, recall, accuracy, and F-measure values, respectively as in shown in Table 2. It correctly labels 328 out of 357 genome sequences of Beta variant, and achieves 0.84, 0.92, 0.94, and 0.88 precision, recall, accuracy, and F-measure values, respectively. When we analyze Gamma variant, SVM classifier correctly classifies 25 out of 110 genome sequences of Gamma variant, and achieves 0.57, 0.23, 0.93, and 0.32 precision, recall, accuracy, and F-measure values, respectively. Finally, the SVM classifier successfully classifies 458 out of 500 genome sequences of Delta variant, and achieves 0.89, 0.92, 0.93, and 0.90 precision, recall, accuracy, and F-measure values, respectively. On average, the SVM classifier achieves 0.80 precision, 0.75 recall, 0.93 accuracy, and 0.75 F-measure.

Finally, we evaluate the results of the NN classifier based on variants. It correctly labels 403 out of 436 genome sequences of Alpha variant, and achieves 0.92 precision and recall values, 0.95 accuracy, and 0.92 F-measure. When we look at the results of Beta variant, it correctly classifies 320 out of 357 genome sequences, and achieves 0.85, 0.90, 0.93, and 0.87 precision, recall, accuracy, and F-measure values, respectively. For Gamma variant, it correctly labels 45 of 110 genome sequences, and achieves 0.60, 0.41, 0.93, and 0.49 precision, recall, accuracy, and F-measure values, respectively. Finally, it correctly classifies 464 out of 500 genome sequences of Delta variant, and achieves 0.90, 0.93, 0.94, and 0.91 precision, recall, accuracy, and F-measure values, respectively. On average, the NN classifier achieves 0.82 precision, 0.79 recall, 0.94 accuracy, and 0.80 F-measure.

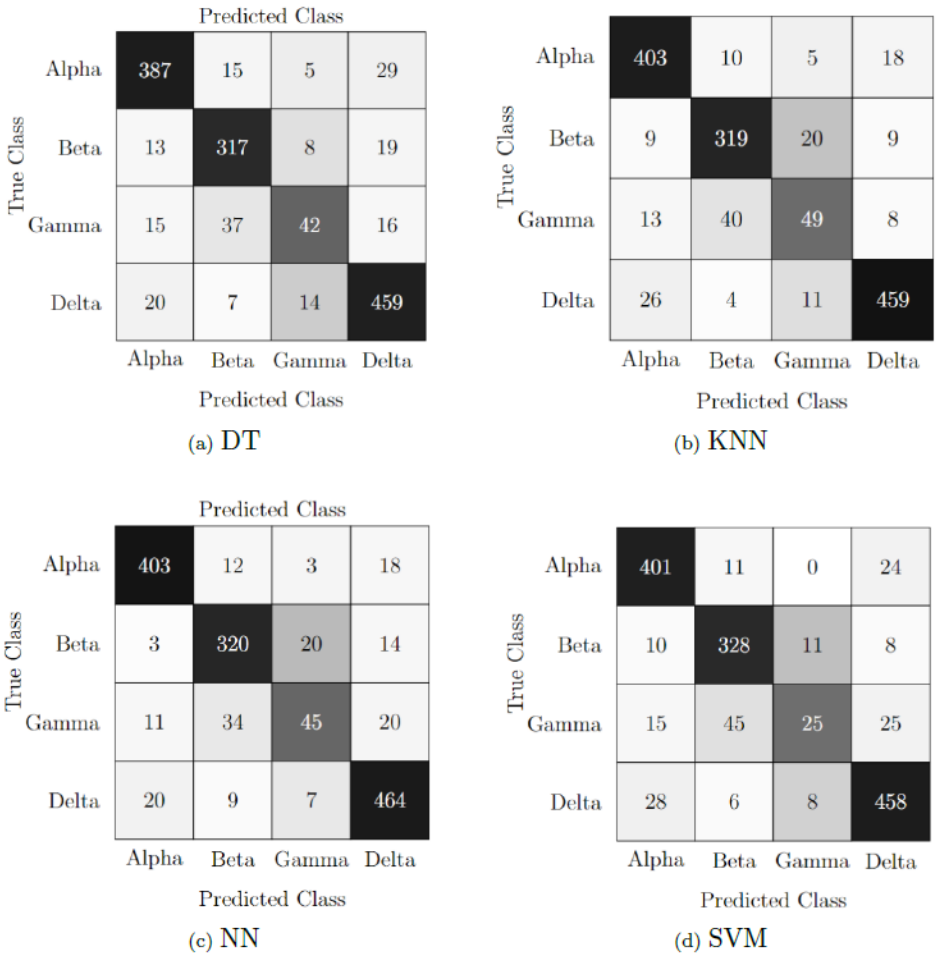

Figure 3. Confusion matrices for DT, KNN, NN, and SVM methods

Table 2. Variant based and average results of the machine learning classifiers

Method	SARS-CoV-2 Variant	Variant based results Pre	Re	Acc	F-score	Average results Pre	Re	Acc	F-score
DT	Alpha	0.89	0.89	0.93	0.89	0.8	0.77	0.93	0.78
	Beta	0.84	0.89	0.93	0.86				
	Gamma	0.61	0.38	0.93	0.47				
	Delta	0.88	0.92	0.93	0.90				
KNN	Alpha	0.89	0.92	0.94	0.91	0.81	0.8	0.94	0.8
	Beta	0.86	0.89	0.93	0.87				
	Gamma	0.58	0.45	0.93	0.50				
	Delta	0.93	0.92	0.95	0.92				
SVM	Alpha	0.88	0.92	0.94	0.90	0.8	0.75	0.93	0.75
	Beta	0.84	0.92	0.94	0.88				
	Gamma	0.57	0.23	0.93	0.32				
	Delta	0.89	0.92	0.93	0.90				
NN	Alpha	0.92	0.92	0.95	0.92	0.82	0.79	0.94	0.8
	Beta	0.85	0.90	0.93	0.87				
	Gamma	0.60	0.41	0.93	0.49				
	Delta	0.90	0.93	0.94	0.91				

5. Conclusion

Since the beginning of the pandemic, variants of SARS-CoV-2 have been identified. SARS-CoV-2 variants may be characterized by their transmissibility and disease severity. Recently, there have been identified four common variants of SARS-CoV-2 in Turkey. Alpha and Delta variants generally present higher infection rates and significant disease severity when compared to Beta and Gamma variants. Developing a method discriminating SARS-CoV-2 variants is significant to track mutations, monitor the changes, measuring the efficiency of the current vaccines, assess the evolution of SARS-CoV-2 as well as prevent its spread. In this study, we proposed a method separating variants of SARS-CoV-2 from genome sequences from Turkey. We determine four features representing the whole genome sequences. Next, we applied four machine learning methods to present effectiveness of the proposed features. The proposed method achieves the best accuracy with 94% on the dataset, including four variants of SARS-CoV-2 from Turkey when the KNN is employed. In future studies, we will analyze the sequences from all over the world and novel feature vectors will be described to increase overall accuracy. Furthermore, we will introduce new methods for developing PCR test kits by analyzing variants of SARS-CoV-2 genome sequences.

Contribution of Researchers

All parts of the paper are developed by Hilal Arslan.

Conflict of Interest

The authors declared that there is no conflict of interest.

References

Ali, S., Sahoo, B., Ullah, N., Zelikovskiy, A., Patterson, M., & Khan, I. (2021). A k-mer Based Approach for SARS-CoV-2 Variant Identification. In Bioinformatics Research and Applications (pp. 153–164). Springer International Publishing. https://doi.org/10.1007/978-3-030-91415-8_14

Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. In Chaos, Solitons & Fractals (Vol. 139, p. 110017). Elsevier BV. https://doi.org/10.1016/j.chaos.2020.110017

Arslan, H. (2021). COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. In Computers & Industrial Engineering (Vol. 161, p. 107666). Elsevier BV. https://doi.org/10.1016/j.cie.2021.107666

Arslan, H. (2021). Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. In The 7th International Management Information Systems Conference. International Management Information Systems Conference. MDPI. https://doi.org/10.3390/proceedings2021074020

Arslan, H., & Aygun, B. (2021). Performance Analysis of Machine Learning Algorithms in Detection of COVID-19 from Common Symptoms. In 2021 29th Signal Processing and Communications Applications Conference (SIU). 2021 29th Signal Processing and Communications Applications Conference (SIU). IEEE. https://doi.org/10.1109/siu53274.2021.9477809

Arslan, H., & Arslan, H. (2021). A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. In Engineering Science and Technology, an International Journal (Vol. 24, Issue 4, pp. 839–847). Elsevier BV. https://doi.org/10.1016/j.jestch.2020.12.026

Burioni, R., & Topol, E. J. (2021). Assessing the human immune response to SARS-CoV-2 variants. In Nature Medicine (Vol. 27, Issue 4, pp. 571–572). Springer Science and Business Media LLC. https://doi.org/10.1038/s41591-021-01290-0

Charytoniuk, W., & Chen, M. S. (n.d.). Neural network design for short-term load forecasting. In DRPT2000. International Conference on Electric Utility Deregulation and Restructuring and Power Technologies. Proceedings

(Cat. No.00EX382). International Conference on Electric Utility Deregulation and Restructuring, and Power Technologies 2000. IEEE. https://doi.org/10.1109/drpt.2000.855725

Davi, M. J. P., Jeronimo, S. M. B., Lima, J. P. M. S., & Lanza, D. C. F. (2021). Design and in silico validation of polymerase chain reaction primers to detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In Scientific Reports (Vol. 11, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41598-021-91817-9

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., Diaz-Ordaz, K., Keogh, R., … Eggo, R. M. (2020). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.12.24.20248822

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C. S., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., … Sabino, E. C. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. In Science (Vol. 372, Issue 6544, pp. 815–821). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.abh2644

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. In IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Vol. 42, Issue 4, pp. 463–484). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tsmcc.2011.2161285

Garcia-Beltran, W. F., Lam, E. C., Astudillo, M. G., Yang, D., Miller, T. E., Feldman, J., Hauser, B. M., Caradonna, T. M., Clayton, K. L., Nitido, A. D., Murali, M. R., Alter, G., Charles, R. C., Dighe, A., Branda, J. A., Lennerz, J. K., Lingwood, D., Schmidt, A. G., Iafrate, A. J., & Balazs, A. B. (2021). COVID-19-neutralizing antibodies predict disease severity and survival. In Cell (Vol. 184, Issue 2, pp. 476-488.e11). Elsevier BV. https://doi.org/10.1016/j.cell.2020.12.015

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (pp. 986–996). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62

Hamed, A., Sobhy, A., & Nassar, H. (2021). Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm. In Arabian Journal for Science and Engineering (Vol. 46, Issue 9, pp. 8261–8272). Springer Science and Business Media LLC. https://doi.org/10.1007/s13369-020-05212-z

Han, X., & Ye, Q. (2021). The variants of SARS-CoV-2 and the challenges of vaccines. In Journal of Medical Virology (Vol. 94, Issue 4, pp. 1366–1372). Wiley. https://doi.org/10.1002/jmv.27513

Harvey, W.T., Carabelli, A.M., Jackson, B. et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol 19, 409–424 (2021). https://doi.org/10.1038/s41579-021-00573-0

Hasan, N. (2020). A Methodological Approach for Predicting COVID-19 Epidemic Using EEMD-ANN Hybrid Model. In Internet of Things (Vol. 11, p. 100228). Elsevier BV. https://doi.org/10.1016/j.iot.2020.100228

Hatirnaz Ng, O., Akyoney, S., Sahin, I., Soykam, H. O., Bayram Akcapinar, G., Ozdemir, O., Kancagi, D. D., Sir Karakus, G., Yurtsever, B., Kocagoz, A. S., Ovali, E., & Ozbek, U. (2021). Mutational landscape of SARS-CoV-2 genome in Turkey and impact of mutations on spike protein structure. In M. Adnan (Ed.), PLOS ONE (Vol. 16, Issue 12, p. e0260438). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0260438

Kandeel, M., Mohamed, M. E. M., Abd El-Lateef, H. M., Venugopala, K. N., & El-Beltagi, H. S. (2021). Omicron variant genome evolution and phylogenetics. In Journal of Medical Virology (Vol. 94, Issue 4, pp. 1627–1632). Wiley. https://doi.org/10.1002/jmv.27515

Khateeb, J., Li, Y., & Zhang, H. (2021). Emerging SARS-CoV-2 variants of concern and potential intervention approaches. In *Critical Care* (Vol. 25, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s13054-021-03662-x

Lauring, A. S., & Malani, P. N. (2021). Variants of SARS-CoV-2. In *JAMA* (Vol. 326, Issue 9, p. 880). American Medical Association (AMA). https://doi.org/10.1001/jama.2021.14181

Mann, C., Griffin, J. H., & Downard, K. M. (2021). Detection and evolution of SARS-CoV-2 coronavirus variants of concern with mass spectrometry. In Analytical and Bioanalytical Chemistry (Vol. 413, Issue 29, pp. 7241–7249). Springer Science and Business Media LLC. https://doi.org/10.1007/s00216-021-03649-1

Mlcochova, P., Kemp, S. A., Dhar, M. S., Papa, G., Meng, B., Ferreira, I. A. T. M., Datir, R., Collier, D. A., Albecka, A., Singh, S., Pandey, R., Brown, J., Zhou, J., Goonawardane, N., Mishra, S., Whittaker, C., Mellan, T., Marwal, R., … Datta, M. (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. In *Nature* (Vol. 599, Issue 7883, pp. 114–119). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-021-03944-y

Noble, W. S. (2006). What is a support vector machine? In Nature Biotechnology (Vol. 24, Issue 12, pp. 1565–1567). Springer Science and Business Media LLC. https://doi.org/10.1038/nbt1206-1565

Padane, A., Mbow, M., Mboup, A., Diedhiou, C. K., Gueye, K., Lo, C. I., Ndiour, S., Leye, N., Ndoye, A. S., Selbé Ndiaye, A. J., Diagne, N. D., Ndiaye, S., Beye, M., Sarr, M., Lo, G., Wade, D., Ahouidi, A., Diaw, P. A., Camara, M., … Mboup, S. (2022). Rapidly rising cases with Omicron in Senegal. In *New Microbes and New Infections* (Vol. 45, p. 100959). Elsevier BV. https://doi.org/10.1016/j.nmni.2022.100959

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In Encyclopedia of Database Systems (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Sabino, E. C., Buss, L. F., Carvalho, M. P. S., Prete, C. A., Jr, Crispim, M. A. E., Fraiji, N. A., Pereira, R. H. M., Parag, K. V., da Silva Peixoto, P., Kraemer, M. U. G., Oikawa, M. K., Salomon, T., Cucunuba, Z. M., Castro, M. C., de Souza Santos, A. A., Nascimento, V. H., Pereira, H. S., Ferguson, N. M., Pybus, O. G., … Faria, N. R. (2021). Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. In *The Lancet* (Vol. 397, Issue 10273, pp. 452–455). Elsevier BV. https://doi.org/10.1016/s0140-6736(21)00183-5

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. In Eurosurveillance (Vol. 22, Issue 13). European Centre for Disease Control and Prevention (ECDC). https://doi.org/10.2807/1560-7917.es.2017.22.13.30494

Simon-Loriere, E., & Schwartz, O. (2022). Towards SARS-CoV-2 serotypes? In Nature Reviews Microbiology (Vol. 20, Issue 4, pp. 187–188). Springer Science and Business Media LLC. https://doi.org/10.1038/s41579-022-00708-x

Tao, K., Tzou, P. L., Nouhin, J., Gupta, R. K., de Oliveira, T., Kosakovsky Pond, S. L., Fera, D., & Shafer, R. W. (2021). The biological and clinical significance of emerging SARS-CoV-2 variants. In Nature Reviews Genetics (Vol. 22, Issue 12, pp. 757–773). Springer Science and Business Media LLC. https://doi.org/10.1038/s41576-021-00408-x

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Van Zyl, G., … de Oliveira, T. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. In *Nature* (Vol. 592, Issue 7854, pp. 438–443). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-021-03402-9

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., … Ferguson, N. M. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. In *Nature* (Vol. 593, Issue 7858, pp. 266–269). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-021-03470-x

Yoo, S. H., Geng, H., Chiu, T. L., Yu, S. K., Cho, D. C., Heo, J., Choi, M. S., Choi, I. H., Cung Van, C., Nhung, N. V., Min, B. J., & Lee, H. (2020). Deep Learning-Based Decision-Tree Classifier for COVID-19 Diagnosis From Chest X-ray Imaging. In Frontiers in Medicine (Vol. 7). Frontiers Media SA. https://doi.org/10.3389/fmed.2020.00427