Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus

Hilal Arslan

PII: S0360-8352(21)00570-2
DOI: https://doi.org/10.1016/j.cie.2021.107666
Reference: CAIE 107666

To appear in: Computers & Industrial Engineering

Received Date: 16 March 2021
Accepted Date: 5 September 2021

Please cite this article as: Arslan, H., COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus, Computers & Industrial Engineering (2021), doi: https://doi.org/10.1016/j.cie.2021.107666

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Ltd.
COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus

Hilal Arslan
Ankara Yıldırım Beyazıt University, Ankara, Turkey
hilalarslanceng@gmail.com

Hilal Arslan is an Assistant Professor in Software Engineering department at Ankara Yıldırım Beyazıt University in Turkey. She received her B.S degree in Mathematics department from Ankara University and M.E. degree in Computer Engineering department from Middle East Technical University (METU). Her research interests are machine learning, bioinformatics, pattern recognition, parallel computing, scientific computing, numerical linear algebra, and graph algorithms.
The highlights of our contributions are:

- Similarity features are introduced by comparing human SARS-CoV-2 and bat SARS-CoV-like Coronavirus

- Integration of CpG based features and similarity features are used to create more reliable feature vectors

- Efficiency of proposed features are shown by using various machine learning techniques

- The best performance is achieved among the current COVID-19 detection studies from genome sequences

With best regards,

Hilal Arslan
All parts of the paper are prepared by Hilal Arslan.
COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus

Hilal Arslan

Received: date / Accepted: date

Abstract This paper proposes an efficient and accurate method to predict coronavirus disease 19 (COVID-19) based on the genome similarity of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and a bat SARS-CoV-like coronavirus. We introduce similarity features to distinguish COVID-19 from other human coronaviruses by comparing human coronaviruses with a bat SARS-CoV-like coronavirus. In the proposed method, each human coronavirus sequence is assigned to three similarity scores considering nucleotide similarities and mutations that lead to the strong absence of cytosine and guanine nucleotides. Next, the proposed features are integrated with CpG island features of the genome sequences to improve COVID-19 prediction. Thus, each genome sequence is represented by five real numbers. We exhibit the effectiveness of the proposed features using six machine learning classifiers on a dataset including the genome sequences of human coronaviruses similar to SARS-CoV-2. The performances of the machine learning classifiers are close to each other, and k-nearest neighbor classifier with similarity features achieves the best results with an accuracy of 99.2%. Moreover, k-nearest neighbor classifier with the integration of CpG based and similarity features has an admirable performance and achieves an accuracy of 99.8%. Experimental results demonstrate that similarity features remarkably decreases the number of false negatives and significantly improve the overall performance. The superiority of the proposed method is also highlighted by comparing with the state-of-the-art studies detecting COVID-19 from genome sequences.

Keywords Covid-19 · feature extraction · SARS-CoV-2 · machine learning methods · CpG islands · similarity

1 Introduction

Coronaviruses (CoVs) are crucial human and animal pathogens causing diseases ranging from the common flu to more severe respiratory diseases [Paules, Marston & Fauci] 2020. Currently, a novel coronavirus SARS-
CoV-2 detected in China in December 2019 has led to COVID-19 disease, which causes severe pneumonia and fatal human infections. COVID-19 seriously threatens the public health system since there is no scientifically proven treatment method. According to the report published by the World Health Organization (WHO), as of 16 February 2021, there have been 108,822,960 people infected by COVID-19 and 2,403,641 people have died from COVID-19. Well known symptoms of the COVID-19 reported by the WHO are cough, fever, chest pressure, pneumonia, diarrhea, and shortness of breath. Since the symptoms of the COVID-19 are similar to influenza, early diagnosis of COVID-19 is difficult. However, it is crucial to distinguish positive cases quickly as it spreads speedily and threatens the public health system as well as leads to severe outcomes.

One of the standard techniques for detecting COVID-19 is the reverse transmission polymerase chain reaction (RT-PCR) (Anika et al., 2020). In some studies, the sensitivity and accuracy of the RT-PCR test have been criticized and they reported that the RT-PCR test suffers from a great number of false negatives and false positives (Udugama et al., 2020; Holshue et al., 2020; Silva et al., 2020; F.-Jiang et al., 2020; D. Wang et al., 2020) due to the mutation of SARS-CoV-2 (Ai et al., 2020).

Many state-of-the-art machine learning algorithms have recently been published to detect COVID-19 (Zoabi, Deri-Rozov, & Shomron, 2021; Muhammad et al., 2021; Ardabili et al., 2020; Tayarani N., 2021; Kushwaha et al., 2020; De Felice & Polimeni, 2020). These methods can identify COVID-19 disease mostly by analyzing Computed Tomography (CT) scans, chest X-ray images, and whole genome sequences (Annarumma et al., 2019; Udugama et al., 2020; Arslan, 2021). Studies detecting COVID-19 from CT scans or X-ray images have some limitations. For instance, medical images may not distinguish COVID-19 from other viral pneumonia since they have several common features (Udugama et al., 2020; Zargari Khuzani, Heidari, & Shariati, 2021). The absence of any abnormalities on a chest X-ray or CT scan does not guarantee the absence of SARS-CoV-2. Furthermore, there is a lack of various annotated images that can be used in experiments in image-based analysis (Mohamadou, Halidou, & Kapen, 2020).

Molecular techniques have attracted attention recently as they can easily track specific pathogens and coronavirus genes. Hence, they have produced more satisfactory results compared to CT scans. Machine learning algorithms exhibit remarkable performance for analyzing genome sequences in a state-of-the-art manner. In these techniques, it is critical to identify features that distinguish COVID-19 from other coronaviruses. Many evolutionary and population based methods may be used for choosing the most relevant features discriminating COVID-19 (L. Abualigah, Yousri, et al., 2021; L. Abualigah & Diabat, 2021; L. Abualigah, Diabat, Mirjalili, Abd Elaziz, & Gandomi, 2021; L. M. Q. Abualigah, 2019; Too & Mirjalili, 2021) and modelling the potential effect of coronavirus (Salgotra, Gandomi, & Gandomi, 2020).

When the origin of SARS-CoV-2 virus is investigated, some studies point out that bats might be the original host of SARS-CoV-2 (Lu et al., 2020; Wu et
al. 2020). Zhou et al. (2020) showed that SARS-CoV-2 presents overall genome sequence similarity to a bat coronavirus and is 96% identical to a bat coronavirus. Taking advantages of this similarity, in this study, we propose an accurate and efficient COVID-19 detection method based on genome similarity between a bat SARS-like coronavirus and SARS-CoV-2. We define three similarity features by comparing human coronavirus sequences with the bat coronavirus sequence. Furthermore, to create more reliable feature vectors, we integrate similarity features with CpG based features that efficiently distinguish SAR-CoV-2 (Arслan and Arслan, 2021). Six machine learning classifiers are employed to present the effectiveness of the proposed features. To achieve the best results, hyperparameters of the machine learning classifiers are tuned using grid search. Experimental results exhibit that similarity features detect SARS-CoV-2 sequences from other coronavirus sequences genetically similar to SARS-CoV-2 in a short time and remarkably reduce the number of false negatives. Furthermore, using similarity features with CpG based features increases overall accuracy. Lastly, the proposed method is compared to the current state-of-the-art COVID-19 prediction methods (Lopez-Rincon et al. 2021; Arслan & Arслan 2021) and it exhibits an admirable performance.

The remaining parts of the study are organized as follows. In Section 2, we summarize studies that have diagnosed COVID-19. In Section 3, we explain the proposed method and summarize machine learning techniques used in this study. In Section 4, we evaluate experimental results and compare them with the-state-of-the-art methods. Finally, Section 5 presents important conclusions pointing out future directions.

2 Related Work

In this section, we briefly explain the studies diagnosing COVID-19 in three categories: studies using laboratory findings and common symptoms, studies using CT scans and chest X-ray images, and studies using whole genome sequences.

Jiang et al. (2020) applied machine learning techniques, support vector machines (SVM), k-nearest neighbor (KNN), decision tree (DT), and random forest (RF) for diagnosing COVID-19 using the combination of clinical features of COVID-19. They reported that they achieved the best result with an accuracy of 80% when the SVM classifier is used. In another study, Batista et al. (2020) performed the machine learning methods for predicting COVID-19 using 18 clinical findings. They applied SVM, RF, logistic regression, neural networks, and gradient boosted trees methods. Their method achieved the best results with an AUC score of 0.847 when the SVM or RF classifier was used. Zoabi et al. (2021) proposed a COVID-19 prediction method based on COVID-19 symptoms. They used eight binary features including sex, age, and contact with a person having COVID-19 disease. When they applied gradient boosting with DT, their proposed method reached an AUC-ROC of 0.90. These studies pointed out that diagnosis of COVID-19 using clinical data or symptoms achieves a constrained diagnosis efficiency.

Mohamadou et al. (2020) and Shi et al. (2021) reviewed artificial intelligence methods for predicting and
managing COVID-19. In these techniques, X-ray and CT scans have been mostly used. Akram et al. (2021) introduced a framework for rapid diagnosis of COVID-19 by extracting textual and statistical features from raw CT images including 2400 normal chest CT scans and 3500 COVID-19 pneumonia. They performed the genetic algorithm for selecting meaningful features and Naive Bayes for classification. Their proposed method reached an accuracy of 92.6%. Ucar et al. (2021) proposed a deep learning based method from X-ray images to quickly identify COVID-19. They analyzed 1125 images in total and their method reached an accuracy of 92.4%. Dansana et al. (2020) proposed a method for early diagnosing of COVID-19 from X-ray and CT images which includes 360 images. They used convolutional neural networks and their method achieved an accuracy of 91%. Khuzani et al. (2021) developed a machine learning model to separate chest X-ray images of COVID-19 cases from other types of pneumonia. They used 420 images and they created an optimum set of features of X-ray images by performing dimensionality reduction method. In the classification step, they performed the multilayer neural network and they achieved satisfactory results.

On the other hand, DNA sequencing techniques for identifying SARS-CoV-2 are highly useful for monitoring coronavirus genes that change frequently as the disease passes from one person to another as well as understand the behaviour of the virus (Nawaz et al., 2021; Annarumma et al., 2019; Udugama et al., 2020; Arslan, 2021). The DNA sequencing methods can be investigated in two categories, which are alignment-based and alignment-free methods. Alignment based methods are preferred when the size of the data is small, and they are expensive if the size of the data is large. Machine learning algorithms are effectively used for analyzing large numbers of genome sequences. Nawaz et al. (2021) investigated sequential pattern mining and mutation analysis techniques. They reported that these techniques demonstrated interesting knowledge and patterns to explore the progress and variations in COVID-19 strains.

Randhawa et al. (2020) proposed a method for rapid classification of COVID-19 cases that combines machine learning techniques with genomic signatures. They downloaded various types of virus sequences from the National Center for Biotechnology Information (NCBI) database. In their method, first they converted the sequences into discrete numeric values using Chaos Game Representation (CGR) (Jeffrey, 1990), which was based on two dimensional k-mer. In their study, the value of k was selected as 7. After they performed discrete Fourier transform (DFT) to the sequences, the Pearson correlation coefficient (PCC) was used to compute the pairwise distance. They used six supervised machine learning methods, linear discriminant, linear support vector machine (SVM), quadratic SVM, fine KNN, subspace discriminant, and subspace KNN, which were evaluated using 10-fold cross validation. When SARS-CoV-2 sequences were classified under four genera, gammacoronavirus, deltacoronavirus, betacoronavirus, and alphacoronavirus, their results verified that SARS-CoV-2 sequences belonged to sarbecoronavirus within betacoronavirus family.
Naeem et al. (2020) proposed another method detecting SARS-CoV-2 sequences among human coronaviruses. Their dataset included 76 sequences for the type of SARS-CoV, MERS-CoV, and SARS-CoV-2 recorded in the NCBI. They extracted DFT and discrete cosine transform (DCT), and the seven moment invariants (MI) features from genome sequences. They performed k-nearest neighbor (KNN) method and their method reached an accuracy of 100%. Arslan and Arslan (2021) proposed a recent method diagnosing SARS-CoV-2 from various types of human coronaviruses. They used 1000 SARS-CoV-2 sequences and 592 other types of human coronaviruses recorded in the 2019 novel coronavirus resource (2019nCoVR). They extracted two CpG based features from the genome sequences. Thus, they represented whole genome sequences with two real numbers. They used the KNN algorithm for classification and they showed that their method achieved 98.4% accuracy when any $L_1$ type distance measure in the KNN was used. Recently, Lopez-Rincon et al. (2021) proposed classification and specific primer design for detecting SARS-CoV-2 from genome sequences. In their approach, they extracted max pooling features that were the 21-bps sequences. They detected 12 of 3827 max pooling features found in SARS-CoV-2 using a convolutional neural network (CNN) classifier. They reported that their method separated different virus genomes from the coronavirus family with 98.73% accuracy.

3 Proposed COVID-19 Prediction Method

Although evolution has led to the specification of all living organisms, most of the genomes of the living organisms exhibit a high amount of similarities because of the conservation of the genetic data from one generation to another. We develop an efficient and accurate COVID-19 prediction method based on genome similarity of human SARS-CoV-2 and a bat SARS-CoV-like coronavirus. The main steps of the proposed method are presented in Figure 1. Furthermore, the pseudocode of the proposed method is given in Algorithm 1. The method takes complete sequences of human coronaviruses including alphacoronavirus and betacoronavirus as well as a bat coronavirus, RaTG13 as the input data. Then, features discriminating SARS-CoV-2 sequences from the other coronaviruses are extracted. We introduce similarity features to differentiate SARS-CoV-2 sequences from the other human coronaviruses.

The bat coronavirus RaTG13 was observed in Rhinolophus affinis from Yunnan, and presented the highest complete genome sequence similarity to SARS-CoV-2 (Zhou et al., 2020; C. Li, Yang, & Ren, 2020). Xing-guang et al. (2020) showed that the sequence similarity between bat coronavirus RaTG13 and SARS-CoV-2 is about 96%. Based on this similarity due to the conservation of the genetic data, we describe three similarity features by comparing each human SARS-CoV-2 sequence and the bat coronavirus RaTG13 sequence. The first similarity feature between genome sequences of human and bat coronavirus called $sim_1$ is determined by computing the total number of nucleotides that are identical and located at the same location, which is shown in line 2 in Algorithm 1. Thus, it checks whether sequence of human coronavirus resembles to the sequence of the bat coronavirus. The second and third similarity features
are based on low CG content in human SARS-CoV-2 genome sequence. The main reason for this is that CG is mutated into CA or CT and also CG is mutated into AG and TG (Y. Wang et al., 2020). To extract second similarity feature called $sim_2$ demonstrated in line 3 in Algorithm 1, we count the number mutated C of CG by comparing human SARS-CoV-2 with bat RaTG13 SARS-CoV nucleotides at the same positions. Similarly, the third similarity feature called $sim_3$ presented in line

![Fig. 1 Main steps of the proposed COVID-19 prediction method](image-url)
4 in Algorithm 1 is extracted by counting the number mutated G of CG.

On the other hand, CpG dinucleotides are rarely found in the genome of SARS-CoV-2, and Arslan and Arslan (2021) used this property to discriminate SARS-CoV-2 sequences. Computation of CpG based features is shown in lines 5-9 in Algorithm 1. The first CpG based feature, $CpG_1$, is computed by adding $\text{ratio}_C$ and $\text{ratio}_G$ where the ratio of the nucleotide is calculated by dividing frequency of occurrences of the nucleotide to the sequence length. Similarly, $CpG_2$ is computed by dividing $\text{ratio}_{CG}$ to $\text{ratio}_C \times \text{ratio}_G$. To improve COVID-19 prediction, we propose to use an integration of CpG based and similarity features. Thus, in the feature extraction step, each genome sequence is represented by five real numbers.

After the feature extraction step, any machine learning classifier may be used to determine SARS-CoV-2 sequences. In the classification step, we recommend using KNN with an $L_1$ type metric (Arslan & Arslan, 2021) since it is a simple non-parametric algorithm, and does not require any specific training phase. It also requires fewer hyperparameters that are tuned. In addition to KNN, support vector machines, decision tree, random forest, AdaBoost, and multilayer perceptron methods are performed to evaluate efficiency of the proposed features.

Machine learning classifiers may require some parameters that need to be optimized to achieve the best prediction accuracy. One of the most common techniques for tuning parameters is grid search (Hamida, Ganhour, Cherradi, Ouajji, & Rahami, 2020; Syarif, Prugel-Bennett, & Wills, 2016; Das, Mishra, & Saraswathy Gopalan, 2020), and we employ the grid search to determine the optimum value of the hyperparameters. In the following subsection, we briefly explain these classifiers and their hyperparameters that are tuned.

Algorithm 1 Proposed COVID-19 Prediction Method

Input:

- Various coronavirus types of human complete genome sequences $SeqData$
- Bat coronavirus RaTG13 complete genome sequence $batSeq$
- Class label of each genome sequences: COVID-19 or non-COVID
- A test complete genome sequence $testSeq$

Output: Determine whether $testSeq$ is COVID-19 or non-COVID

Feature Extraction Step:

1: for each sequence $S$ in $SeqData$ do
   // Extraction of Similarity Features comparing $S$ to $batSeq$
2: $\text{sim1} \leftarrow$ compute total number of nucleotides that are identical and located at the same location
3: $\text{sim2} \leftarrow$ the number of CG mutations into AG and TG by comparing $S$ and $batSeq$
4: $\text{sim3} \leftarrow$ the number of CG mutations into CA and CT by comparing $S$ and $batSeq$
   // Extraction of CpG island features (Arslan & Arslan, 2021)
5: $\text{ratio}_C \leftarrow$ compute ratio of C nucleotide in $S$
6: $\text{ratio}_G \leftarrow$ compute ratio of G nucleotide in $S$
7: $\text{ratio}_{CG} \leftarrow$ compute ratio of CG nucleotide in $S$
8: $CpG_1 = \text{ratio}_C + \text{ratio}_G$
9: $CpG_2 = \text{ratio}_{CG} / (\text{ratio}_C \times \text{ratio}_G)$
10: end for

Parameter Tuning

11: Determine hyperparameters to be tuned for the classifier and perform grid search

Classification Step:

12: Extract $CpG_1$, $CpG_2$, $\text{sim1}$, $\text{sim2}$, and $\text{sim3}$ features for $testSeq$
13: Apply the machine learning classifier (KNN is suggested) with its optimum hyperparameters
14: Determine whether $testSeq$ is SARS-CoV-2 or not
3.1 Machine Learning Classifiers

We perform several types of classifiers to predict COVID-19 disease, and in this section, we briefly summarize these classifiers.

3.1.1 Support Vector Machine (SVM)

SVM (Burges, 1998; Vapnik, 1995) separates the data samples by finding a hyperplane maximizing the margin, which is the smallest distance between two classes. The hyperplane can be determined as a solution to the following problem:

$$\min \ C \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2$$  \hspace{1cm} (1)

where $C$ called the regularization parameter controls the trade-off between controlling complexity of the model and minimizing training errors. If the data is not divided linearly, non linear kernel functions are used to transform the data into a high dimensional space. The polynomial, sigmoid, and radial basis kernel function (RBF) (Liao, Fang, & L.W. Nuttle, 2004) are commonly used kernel functions. The choice of the kernel function and parameter $C$ remarkably affects the performance of the SVM. The RBF is the most commonly used kernel function to classify multi-dimensional data. Moreover, the linear kernel is a specific version of the RBF (Keerthi & Lin 2003). Furthermore, the RBF requires fewer parameters to set than polynomial kernel, and performance of the RBF has similar to the other kernel functions (Lin, Ying, Chen, & Lee 2008). For these reasons, in this study, we used the RBF kernel function given in Equation (2).

$$K(t_i, t_j) = \exp(-\gamma ||t_i - t_j||^2), \quad \gamma > 0$$  \hspace{1cm} (2)

Grid search is used to determine optimum parameters and we follow the similar strategy with (Duarte & Wainer 2017) to determine possible values of $C = 2^{-5}, 2^{-1}, 2^3, 2^9, 2^{15}$ and $\gamma = 2^{-15}, 2^{-9}, 2^{-5}, 2^{-1}, 2^{3}$.

3.1.2 k-Nearest Neighbor (KNN)

KNN (Brown & Mues, 2012; Deng, Zhu, Cheng, Zong, & Zhang 2016; Abu Alfeilat et al., 2019; Bishop, 2006) is a non parametric and lazy learning classification method. Thus, it does not require a training phase and it memorizes the training dataset instead of learning the training dataset. In the KNN method, to classify a new data sample, the $k$ nearest samples from the training dataset are identified, and the new data sample is labelled to the class based on its majority votes. The type of the distance measure and the $k$ parameter remarkably affect the performance of the KNN. Arslan and Arslan (2021) used 19 distance metrics analyzed in five groups. They concluded that $L_1$ type metrics such as Manhattan and Chebyshev metrics have the highest accuracy. In this study, we perform the grid search to determine the optimum $k$ value and the distance measure. The possible metric types are set to Manhattan and Chebyshev (Arslan & Arslan 2021) and $k$ parameter is changed from 1 to 30 in the grid search.

3.1.3 Decision Tree (DT)

The basic idea behind DT (Safavian & Landgrebe, 1991; Aha, Kibler, & Albert, 1991; Bishop, 2006) is to divide
a complex decision into a union of several simple decisions to get the final decision. The performance of the DT classifier significantly depends on how well the tree is constructed. There exist some splitting methods for decision trees, such as information gain and Gini index. In this study, the Gini index \cite{Menze2009,Ghiasi2020} is employed. Using grid search, we also perform the hyperparameter tuning, and we follow the similar strategy with \cite{Das2020} for determining possible values of hyperparameters (maximum tree depth: 5, 10, 15, 20, 25, minimum samples per leaf: 1, 2, 3, and minimum samples per split: 3, 4, 5, 6, 7).

### 3.1.4 Random Forest (RF)

RF \cite{Breiman2001} is an ensemble method combining multiple decision trees and can be widely applied to solve classification problems. The final decision is reached by combining the results from a sequence of decision trees. Each tree is independently created and relies on a random vector sampled from the dataset so that all of the trees have the same distribution. Random feature selection and bootstrap aggregation are performed to average the predictions from all trees \cite{Ho1998,Breiman1996}. We note that the hyperparameter tuning is the same as the DT method.

### 3.1.5 Adaptive Boosting (AdaBoost)

AdaBoost \cite{Freund1997} is a widely known boosting technique for combining multiple simple weak classifiers to create a strong classifier. Furthermore, AdaBoost is an iterative algorithm since it sequentially trains the base classifiers. The weights of the training samples are set to the same value (for instance 1.0) in the first iteration. The weights of the training samples are sequentially improved with respect to the error rates. If the training samples have higher error rates, they will have higher weights. Hyperparameter tuning is a significant process of boosting algorithms to cope with overfitting since it reduces bias error and constructs robust predictive models. In the Adaboost method, two hyperparameters that are learning rate whose possible values are 1.0, 0.15, 0.1, 0.05, 0.01, 0.005, 0.001 and a maximum number of estimators whose possible parameters are 20, 40, 60, 80, 100, 250, 1000 are tuned by using the grid search.

### 3.1.6 Multilayer Perceptron (MLP)

MLP, a kind of a feed forward artificial neural network, includes three layers, which are an input layer, multiple hidden layers, and an output layer \cite{Hornik1989}. The input layer contains the neurons in the same number as the number of features, and the output layer contains the single neuron presenting the output of the MLP \cite{Bishop2006}. MLP models the data by using activation functions. The logistic sigmoid, the rectified linear unit (ReLU), and the hyperbolic tangent (tanh) functions are widely known activation functions. In this study, we use a basic MLP model that consists of one hidden layer. The grid search is applied to specify the number of neurons in the hidden layer and activation function. The number of neurons in the hidden layer is set to 5, 10, and 15, and sigmoid, tanh,
as well as ReLU activation functions are used in the grid search.

4 Experiments

4.1 Experimental setup

All experimental results have been conducted on a computer with a 2.6 GHZ Intel Core i7 CPU and 16 GB RAM under Ubuntu 18.04.03 LTS operating system. Scikit-learn library (Pedregosa et al., 2011), an open-source software for machine learning library in Python programming language is used to implement the classifiers, the grid search, and the cross validation.

4.2 Human Genome Sequences Dataset

In this study, the experiments were carried out on genome sequences retrieved from the 2019 Novel Coronavirus Resource (2019nCoVR) by the China National Center for Bioinformation (Song et al., 2020; Gong et al., 2020). 2019nCoVR is an open source database and shares genomic sequences of all influenza data by linking with relevant databases. Coronaviruses can be classified into four genera. Alphacoronavirus and betacoronavirus infect mammals; on the other hand, gammacoronavirus and deltacoronavirus are commonly detected in avian species (Perlman & Netland, 2009). There have been identified seven coronaviruses infecting human beings so far. HCoV-NL63 and HCoV-229E are types of alphacoronaviruses; on the other hand, HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV, and SARS-COV-2 are types of betacoronavirus (Priyanka, Choudhary, & Singh, 2021). A bat coronavirus RaTG13 sequence, which is the type of betacoronavirus, is also used to extract similarity features. All sequences used in this study are longer than 27,000 bps and high quality. The number of sequences and the types of the coronavirus are presented in Table 1. The dataset consists of 1615 human genome sequences labelled as either COV19+ or COV19-. Genome sequences labelled as COV19+ are diagnosed with COVID-19 disease, and those labelled as COV19- are not diagnosed with COVID-19 disease. We note that the bat coronavirus (RaTG13) sequence is not used for training and testing, and it is only used to extract similarity features. For this reason, the sequence of the bat coronavirus is labelled as neither COV19+ nor COV19-.

4.3 Performance Evaluation

Precision, recall, F1-score, and accuracy (Goutte & Gaussier, 2005) are used to evaluate performance of the machine learning methods. In addition to these metrics, we also used the Matthews correlation coefficient (MCC) (Chicco, Tötsch, & Jurman, 2021) and the area under the receiver operating characteristic (ROC) curve (AUC) metrics. MCC is an appropriate measure for evaluating binary classifiers and summarizes the confusion matrix elements into a single value. The ROC presents the relationship between true positive rate and false positive rate. These metrics are described in Table 2 where true positive (TP) is the number of accurately predicted genome sequences in COV19+ class and false positive (FP) is the number of inaccurately predicted genome sequences in COV19- class. True negative (TN) is the number of accurately predicted genome sequences
Table 1  Types of human coronavirus genomes

| Species        | Genus            | Number of entries | Target class |
|----------------|------------------|-------------------|--------------|
| HCoV-NL63      | Alphacoronavirus | 64                | COV19-       |
| HCoV-229E      |                  | 28                | COV19-       |
| HCoV-HKU1      |                  | 27                | COV19-       |
| HCoV-OC43      |                  | 145               | COV19-       |
| MERS-CoV       | Betacoronavirus  | 339               | COV19-       |
| SARS-CoV       |                  | 12                | COV19-       |
| SARS-CoV-2     |                  | 1000              | COV19+       |
| RaTG13         | Betacoronavirus  | 1                 |              |

Table 2  Performance Measurements

| Measure         | Formula |
|-----------------|---------|
| Precision (Pre) | \( \frac{TP}{TP+FP} \) |
| Recall (Re)     | \( \frac{TP}{TP+FN} \) |
| F1-score        | \( \frac{2 \times \text{Pre} \times \text{Re}}{\text{Pre}+\text{Re}} \) |
| Accuracy (Acc)  | \( \frac{TP+TN}{FN+TP+TN+FP} \) |
| MCC             | \( \frac{TP \times TN - FP \times FN}{\sqrt{(TN+FP)(TP+FN)(TN+FN)(TP+FP)}} \) |

in COV19- class and false negative (FN) is the number of inaccurately predicted genome sequences in COV19+ class.

4.4 Results

In this section, we discuss and compare the efficiency of CpG based features, similarity features, and the integration of them for predicting COVID-19 disease. Analyzing the performance of the classifier is important to discuss the efficiency of the proposed features. To achieve the best performance, hyperparameters of the machine learning classifiers are tuned by using the grid search. In the grid search approach, possible values of each hyperparameter shown in Section 3.1 are determined, and the grid search tries all possible combinations of the parameters. Next, k-fold cross-validation technique is employed to evaluate the machine learning model considering \( k = 10 \) (Govindarajan & Chandrasekaran, 2010) because of its relatively low bias and variance (Sulistiana & Muslim, 2020; Lin et al., 2008; Merghadi et al., 2020; Aggarwal, 2015). Thus, the dataset is separated into ten equal size random groups for training and testing. Nine groups are used for training and the remaining one group is used for testing. This process is repeated ten times until each group is tested. Average measures are computed to achieve the effectiveness of the model.

Table 3 presents the optimum hyperparameters obtained by the grid search function for each machine learning classifier when CpG based, similarity, and combination of them are separately used. In the following subsections, we evaluate the performance of MLP, AdaBoost, RF, KNN, DT, and SVM methods for CpG
based features, the similarity features, and integration of them, separately. All results are obtained by using these hyperparameters.

### 4.4.1 Evaluation of the effectiveness of CpG-based features for COVID-19 prediction

In this part, we discuss the performance of the machine learning methods when CpG based features are only used. Table 4 summarizes the performance of the machine learning classifiers with precision, recall, F1-score, accuracy, MCC, and AUC scores. Results of the machine learning methods are close to each other. Considering the precision values varying between 0.986 and 0.999, the SVM achieves the best result. Recall values ranged from 0.957 to 0.979 and the best value is obtained when the KNN is used. F1-score varies between 0.976 and 0.983, the best value is achieved by the KNN classifier. Accuracy, MCC, and AUC values range from 0.971 to 0.98, from 0.971 to 0.973 to 0.98, respectively, and the best results are obtained when the KNN is used. Furthermore, confusion matrices for each machine learning classifier are presented in Figure 2 when CpG based features are only used. Considering all performance measures, the performance of the KNN is remarkable. As can be also seen in this figure, the KNN only misclassifies 21 out of 1000 COVID-19 genome sequences, and it misclassifies 12 out of 615 non COVID-19 cases.

### 4.4.2 Evaluation of the effectiveness of similarity features for COVID-19 prediction

In this part we discuss the performance of the machine learning methods when the similarity features, which are defined in this study, are used. Table 5 summarizes the performance of the machine learning methods with precision, recall, F-measure, accuracy, MCC, and AUC values. As can be seen in this table, precision values range from 0.984 to 0.992, and the KNN achieves the best precision value. Recall values range from 0.983 to 0.997, and the SVM and MLP achieve the best results. F1-score varies between 0.985 and 0.994, and the best value is achieved by the SVM classifier. On the other hand, accuracy, MCC, and AUC values range from 0.982 to 0.992, from 0.962 to 0.983, and from 0.982 to 0.991, respectively, and the KNN and SVM achieve better results. Furthermore, confusion matrices for each machine learning classifier are presented in Figure 3. The results of the KNN and SVM methods are close to each other, and they classify the genome sequences more successfully compared to the other classifiers. They predict 1602 genome sequences correctly among 1615 sequences. Furthermore, similarity features remarkably reduce the number of false negatives compared to the machine learning methods using CpG based features.

### 4.4.3 Evaluation of the effectiveness of integration of CpG based and similarity features for COVID-19 prediction

In this part, we discuss the results of the machine learning methods using the integration of the CpG based
Table 3 The best performing hyperparameters of the machine learning methods with CpG-based, similarity, and integration of them, separately.

| Method | Hyperparameter | CpG based features | Similarity features | Integration of them |
|--------|----------------|--------------------|---------------------|--------------------|
| SVM    | C              | $2^{-5}$           | $2^9$               | $2^3$              |
|        | $\gamma$      | $2^{-1}$           | $2^{-5}$            | $2^{-1}$           |
| k-NN   | k              | 3                  | 3                   | 2                  |
|        | metric         | chebyshev          | chebyshev           | manhattan          |
| DT     | max depth      | 15                 | 20                  | 10                 |
|        | min samples leaf | 2              | 3                   | 1                  |
|        | min samples split | 3             | 6                   | 3                  |
| RF     | max depth      | 5                  | 20                  | 25                 |
|        | min samples leaf | 3              | 3                   | 1                  |
|        | min samples split | 3             | 7                   | 3                  |
|        | num of estimators | 20            | 60                  | 60                 |
| Adaboost| learning rate | 0.1               | 0.1                 | 1.0                |
|        | num of estimators | 20           | 1000                | 40                 |
| MLP    | num of neurons | 5                 | 5                   | 5                  |
|        | activation     | logistic           | tanh                | relu               |

Table 4 Results of machine learning methods when CpG based features are used

| Feature | Method | Pre | Re  | F1-score | Acc  | MCC  | AUC  |
|---------|--------|-----|-----|----------|------|------|------|
| CpG     | SVM    | 0.999 | 0.957 | 0.978   | 0.973 | 0.944 | 0.978 |
|         | KNN    | 0.988 | 0.979 | 0.983   | 0.980 | 0.957 | 0.980 |
|         | DT     | 0.986 | 0.976 | 0.981   | 0.976 | 0.950 | 0.977 |
|         | RF     | 0.991 | 0.969 | 0.980   | 0.975 | 0.948 | 0.977 |
|         | AdaBoost | 0.986 | 0.968 | 0.977   | 0.972 | 0.940 | 0.973 |
|         | MLP    | 0.986 | 0.967 | 0.976   | 0.971 | 0.938 | 0.973 |

Table 5 Performance comparison of the machine learning methods when similarity features are used

| Feature                  | Method | Pre | Re  | F1-score | Acc  | MCC  | AUC  |
|--------------------------|--------|-----|-----|----------|------|------|------|
| Similarity features      | SVM    | 0.990 | 0.997 | 0.994   | 0.992 | 0.983 | 0.990 |
|                         | KNN    | 0.992 | 0.995 | 0.993   | 0.992 | 0.983 | 0.991 |
|                         | DT     | 0.988 | 0.983 | 0.985   | 0.982 | 0.962 | 0.982 |
|                         | RF     | 0.984 | 0.992 | 0.988   | 0.985 | 0.968 | 0.983 |
|                         | AdaBoost | 0.989 | 0.989 | 0.989   | 0.986 | 0.971 | 0.986 |
|                         | MLP    | 0.988 | 0.997 | 0.993   | 0.991 | 0.980 | 0.989 |
Table 6 summarizes the performance of the machine learning methods with precision, recall, F-measure, accuracy, MCC, as well as AUC values. Precision values range from 0.995 to 0.997, and recall values range from 0.989 to 0.999. F1-score varies between 0.992 and 0.998. On the other hand, accuracy, MCC, and AUC values range from 0.991 to 0.998, from 0.98 to 0.995, and from 0.991 to 0.997, respectively. It is seen that the results of the machine learning methods are close to each other, and the proposed features accurately separate SARS-CoV-2 sequences no matter what type of classifier is used.

Fig. 2 Confusion matrices for each classifier when CpG based features are used

Fig. 3 Confusion matrices for each classifier when similarity features are used

and similarity features, which is the proposed approach.
Table 6 Performance comparison of the machine learning methods using CpG based + similarity features

| Feature                  | Method  | Pre  | Re   | F1-score | Acc  | MCC  | AUC  |
|--------------------------|---------|------|------|----------|------|------|------|
| Proposed Method:         | SVM     | 0.997| 0.998| 0.998    | 0.997| 0.993| 0.997|
| CpG based + Similarity   | KNN     | **0.997** | **0.999** | **0.998** | **0.998** | **0.995** | **0.997** |
| features                 | DT      | 0.996| 0.989| 0.992    | 0.991| 0.980| 0.991|
|                          | RF      | 0.997| 0.996| 0.996    | 0.996| 0.991| 0.996|
|                          | AdaBoost| 0.997| 0.996| 0.996    | 0.996| 0.991| 0.995|
|                          | MLP     | 0.995| 0.998| 0.997    | 0.996| 0.991| 0.995|

The confusion matrices for each machine learning classifier are presented in Figure 4. As can be seen in this figure, the KNN exhibits admirable performance and while it misclassifies 1 out of 1000 COVID-19 genome sequences, it misclassifies 3 out of 615 non COVID-19 cases. Moreover, Figure 5 shows that the total number of incorrectly classified instances for each classifier when CpG based features, similarity features, and integration of them are used, separately. Using integration of the similarity and CpG based features remarkably reduces the total number of incorrectly classified samples.

4.5 Comparison with the-state-of-the-art COVID-19 detection methods

In this part, we compare the proposed method with existing studies detecting COVID-19 disease from genome sequences. Table 7 gives comparative results of COVID-19 detection methods. Naeem et al. (2020) used genomic signal processing to detect COVID-19. They converted the nucleotide bases in the genome sequences into numbers to extract features (Ghosh & Barman, 2013). Next, they extracted nine features using DFT, DCT, and MI methods requiring costly mathematical operations such as integral, determinant. Thus, their feature extraction methods are expensive. They performed two types of classifiers, KNN and trainable cascade-forward back propagation neural network. The KNN was applied with default parameters where k was set to 1 and the metric was set to Euclidean without employing any hyperparameter tuning. They used 46 genomes for training and 30 genomes for testing. Their method achieved 100% accuracy when the KNN was used. The amount of the training data has a crucial role in making decisions and using the small number of samples in the dataset tends to overfit (Ghasemian, Hosseinmardi, & Clauset, 2019). Thus, when the number of genome sequences in the dataset is increased, the overall accuracy of their method may remarkably decrease. Furthermore, selection of k parameter in the KNN method significantly affects the performance of the KNN. If the number of sequences becomes larger, a small value of k may not give the expected output, and the hyperparameter tuning may be required (Sahu, Kumar, & Singh, 2018) to find the optimum k value.

Randhawa et al. (2020) used machine learning methods for taxonomic classification of COVID-19 genomes. They used 20 genome sequences of alphacoronavirus, betacoronavirus, and deltacoronavirus for training. Moreover, they used 29 SARS-COV-2 sequences for testing. When SARS-CoV-2 sequences were classified under four genera, gammacoronavirus, deltacoronavirus, betacoron-
avivirus, and alphacoronavirus, their results verified that SARS-CoV-2 sequences belonged to sarbecoronavirus within betacoronarious family. They perform genus-level classification and their method may not distinguish SARS-COV-2 sequences among the sequences belonging to betacoronavirus family such as SARS-COV or MERS-CoV.

A recent efficient method has been proposed by Lopez-Rincon et al. (2021). They classified SARS-CoV-2 sequences from other human coronavirus sequences by employing a deep learning classifier, convolutional neural network (CNN). They built a model including one convolution layer containing 12 different filters, a fully connected layer with 196 rectified units, a final

**Fig. 4** Confusion matrices for each classifier when the integration of CpG based and similarity features is used

**Fig. 5** The total number of incorrectly classified instances (i.e. sum of FP and FN) for each classifier when CpG based, similarity features, and combination of them are used
Table 7 Overview of the state-of-the-art COVID-19 detection methods

| Study                     | Method       | Features                  | Sequence Dataset          | Accuracy (%) |
|---------------------------|--------------|----------------------------|----------------------------|--------------|
| Naeem et al. (2020)       | KNN          | 7 MI 1 DFT 1 DCT          | 76 SARS-CoV-2 76 SARS-CoV 76 MERS-CoV | 100          |
| Randhawa et al. (2020)    | LD KNN SVM   | DFT PCC                   | 20 betaCoV 20 alphaCoV 20 deltaCoV | 100          |
| Lopez-Rincon et al. (2021)| CNN          | 3827 max pooling features | 66 SARS-CoV-2 487 others | 98.73        |
| Arslan and Arslan (2021)  | KNN with L1 metrics | 2 CpG based features   | 1000 SARS-CoV-2 592 others | 98.4         |
| Proposed Method           | MLP AdaBoost RF KNN DR SVM | 2 CpG based and 3 similarity features | 1000 SARS-CoV-2 92 AlphaCoV 523 BetaCoV 1 RaTG13 | 99.8         |

softmax layer containing five units. The convolution layer identified the subsequences whose lengths were 21 base pairs (bps). They extracted 3827 features (i.e. 21 bps sequences) using 553 human coronavirus sequences. Next, they applied a state-of-the-art feature selection algorithm to decrease the sequences required to determine different virus strains to the bare minimum. Finally, they listed 12 SARS-CoV-2 specific 21 bps subsequences to describe SARS-CoV-2 sequences. They performed 10-fold cross validation and their method could detect SARS-CoV-2 from any other virus with >99% accuracy. Furthermore, they showed that their methods separated the genome sequences obtaining from different coronavirus families with 98.73% accuracy. When we compare this method with the proposed method, their feature extraction method is expensive and the learning time of their method is long. We recall that the proposed method uses only five effective features. Furthermore, their dataset is imbalanced and includes few SARS-CoV-2 sequences. Therefore, there is relatively less information about the class of SARS-CoV-2 sequences and their proposed method may not separate rare SARS-CoV-2
sequences from majority \cite{sun2009,japkowicz2002}. Lastly, Lopez-Rincon et al. \cite{lopez2021} released their dataset and when we run the proposed method on their dataset, the proposed method achieves 100% accuracy.

Recently, Arslan and Arslan \cite{arslan2021} used the KNN classifier with 2 effective CpG based features to detect SARS-CoV-2 from complete genome sequences whose lengths are 30 000bps. Their method achieved an accuracy of 98.4% on 1000 SARS-CoV-2 sequences and 592 sequences including other types of human coronaviruses. In this study, we improve their approach integrating CpG based features with the similarity features. Experimental results show that using integration of CpG based and similarity features remarkably reduces the total number of false negative as well as false positive, and the proposed method achieves an accuracy of 99.8% when the KNN is employed. Furthermore, the proposed method is conducted on the dataset used by Arslan and Arslan \cite{arslan2021} and it achieves 100% accuracy. This underlines the efficiency of the integration of the CpG based and similarity features.

5 Conclusion
COVID-19 caused by a novel SARS-CoV-2 virus rapidly spread worldwide and threatens the public health system as well as leads to severe outcomes. It is essential to quickly identify SARS-CoV-2 to prevent the spread of COVID-19 among people. In this study, we propose a COVID-19 detection method based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. By considering nucleotide similarities, we define three similarity features by comparing sequences of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. The efficiency of the proposed features is evaluated by conducting various machine learning classifiers on 1000 genome sequences of SARS-CoV-2 and 612 genome sequences of the other types of human coronaviruses. Results are also shown on the datasets released by the current state of the art studies predicting SARS-CoV-2 from genome sequences. Experimental results show that using the integration of CpG based and similarity features achieves an accuracy of 99.8% on our dataset and 100% accuracy on the dataset released by the current state of the art studies. The proposed method remarkably achieves better results than the state-of-the-art COVID-19 prediction methods by using only five effective features. This underlines the efficiency of the proposed method.

In the future, various types of viruses causing zoonotic diseases can appear, and it may possible to apply the same strategy to detect the disease. We will also investigate whether CpG motifs with similarity features may be used to develop vaccines in future studies.

References
Abu Alfeilat, H., Hassanat, A., Lasassmeh, O., Tarawneh, A., Alhasanat, M., Eyal-Salman, H., & Prasath, S. (2019, 08). Effects of distance measure choice on k-nearest neighbor classifier performance: A review. Big Data, 7. doi: 10.1089/big.2018.0175
Abualigah, L., & Diabat, A. (2021, April). Advances in Sine Cosine Algorithm: A compre-
COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus

Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H. (2021). The arithmetic optimization algorithm. Computer Methods in Applied Mechanics and Engineering, 376, 113609. Retrieved from https://www.sciencedirect.com/science/article/pii/S0045782520307945 doi: https://doi.org/10.1016/j.cma.2020.113609

Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A. A., Al-qaness, M. A., & Gandomi, A. H. (2021). Aquila optimizer: A novel meta-heuristic optimization algorithm. Computers & Industrial Engineering, 157, 107250. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360835221001546 doi: https://doi.org/10.1016/j.cie.2021.107250

Abualigah, L. M. Q. (2019). Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering (Vol. 816). Springer International Publishing. Retrieved 2021-06-29, from http://link.springer.com/10.1007/978-3-030-10674-4 doi: 10.1007/978-3-030-10674-4

Aggarwal, C. C. (2015). Data mining: The textbook. Springer Publishing Company, Incorporated.

Aha, D. W., Kibler, D., & Albert, M. K. (1991, January). Instance-based learning algorithms. Machine Learning, 6(1), 37–66. Retrieved 2021-06-25, from http://link.springer.com/10.1007/BF00153759 doi: 10.1007/BF00153759

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., ... Xia, L. (2020, August). Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in china: A report of 1014 cases. Radiology, 296(2), E32–E40. Retrieved from https://doi.org/10.1148/radiol.2020200642 doi: 10.1148/radiol.2020200642

Akram, T., Attique, M., Gul, S., Shahzad, A., Altaf, M., Naqvi, S. S. R., ... Maskeliunas, R. (2021, January). A novel framework for rapid diagnosis of COVID-19 on computed tomography scans. Pattern Analysis and Applications. Retrieved 2021-06-29, from http://link.springer.com/10.1007/s10044-020-00950-0 doi: 10.1007/s10044-020-00950-0

Anika, S., Monika, P., Andre, C., Jamie, L. B., Vanessa, S., Joanna, E., ... Robin, G. (2020). Duration of infectiousness and correlation with rt-pcr cycle threshold values in cases of covid-19. Euro Surveill., 25(32). doi: 10.2807/1560-7917.ES.2020.25.32.2001483

Annarumma, M., Withey, S. J., Bakewell, R. J., Pesce, E., Goh, V., & Montana, G. (2019). Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology, 291(1). doi: 10.1148/radiol.2018180921

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... Atkinson, P. M. (2020, October). COVID-19 Outbreak Prediction with Machine Learning. Algorithms, 13(10), 249. Retrieved 2021-02-17, from https://

Artificial Intelligence Review, 54(4), 2567–2608. Retrieved 2021-06-29, from http://link.springer.com/10.1007/s10462-020-09909-3 doi: 10.1007/s10462-020-09909-3

Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H. (2021). The arithmetic optimization algorithm. Computer Methods in Applied Mechanics and Engineering, 376, 113609. Retrieved from https://www.sciencedirect.com/science/article/pii/S0045782520307945 doi: https://doi.org/10.1016/j.cma.2020.113609

Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A. A., Al-qaness, M. A., & Gandomi, A. H. (2021). Aquila optimizer: A novel meta-heuristic optimization algorithm. Computers & Industrial Engineering, 157, 107250. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360835221001546 doi: https://doi.org/10.1016/j.cie.2021.107250

Abualigah, L. M. Q. (2019). Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering (Vol. 816). Springer International Publishing. Retrieved 2021-06-29, from http://link.springer.com/10.1007/978-3-030-10674-4 doi: 10.1007/978-3-030-10674-4

Aggarwal, C. C. (2015). Data mining: The textbook. Springer Publishing Company, Incorporated.

Aha, D. W., Kibler, D., & Albert, M. K. (1991, January). Instance-based learning algorithms. Machine Learning, 6(1), 37–66. Retrieved 2021-06-25, from http://link.springer.com/10.1007/BF00153759 doi: 10.1007/BF00153759

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., ... Xia, L. (2020, August). Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in china: A report of 1014 cases. Radiology, 296(2), E32–E40. Retrieved from https://doi.org/10.1148/radiol.2020200642 doi: 10.1148/radiol.2020200642

Akram, T., Attique, M., Gul, S., Shahzad, A., Altaf, M., Naqvi, S. S. R., ... Maskeliunas, R. (2021, January). A novel framework for rapid diagnosis of COVID-19 on computed tomography scans. Pattern Analysis and Applications. Retrieved 2021-06-29, from http://link.springer.com/10.1007/s10044-020-00950-0 doi: 10.1007/s10044-020-00950-0

Anika, S., Monika, P., Andre, C., Jamie, L. B., Vanessa, S., Joanna, E., ... Robin, G. (2020). Duration of infectiousness and correlation with rt-pcr cycle threshold values in cases of covid-19. Euro Surveill., 25(32). doi: 10.2807/1560-7917.ES.2020.25.32.2001483

Annarumma, M., Withey, S. J., Bakewell, R. J., Pesce, E., Goh, V., & Montana, G. (2019). Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology, 291(1). doi: 10.1148/radiol.2018180921

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... Atkinson, P. M. (2020, October). COVID-19 Outbreak Prediction with Machine Learning. Algorithms, 13(10), 249. Retrieved 2021-02-17, from https://
Arslan, H. (2021). Machine learning methods for covid-19 prediction using human genomic data. *Proceedings, 74*(1). Retrieved from https://www.mdpi.com/2504-3900/74/1/20 doi: 10.3390/proceedings2021074020

Arslan, H., & Arslan, H. (2021). A new covid-19 detection method from human genome sequences using cpg island features and kmn classifier. *Engineering Science and Technology, an International Journal*. Retrieved from https://www.sciencedirect.com/science/article/pii/S2215098620342877 doi: https://doi.org/10.1016/j.jestch.2020.12.026

Batista, A. F. M., Miraglia, J. L., Rizzi Donato, T. H., & Porto Chiavegatto Filho, A. D. (2020, April). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach (preprint). Epidemiology. Retrieved 2021-06-30, from http://medrxiv.org/lookup/doi/10.1101/2020.04.04.20052092 doi: 10.1101/2020.04.04.20052092

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. Retrieved 2021-06-25, from http://link.springer.com/10.1023/A:1018054314350 doi: 10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. Retrieved 2021-06-25, from http://link.springer.com/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Brown, I., & Mues, C. (2012, February). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453. Retrieved from https://eprints.soton.ac.uk/204741/

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167. Retrieved 2021-06-23, from http://link.springer.com/10.1023/A:1009715923555 doi: 10.1023/A:1009715923555

Chicco, D., Tötsch, N., & Jurman, G. (2021, December). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining, 14*(1), 13. Retrieved 2021-06-11, from https://biodatamining.biomedcentral.com/articles/10.1186/s13040-021-00244-z doi: 10.1186/s13040-021-00244-z

Dansana, D., Kumar, R., Bhattacharjee, A., Hemanth, D. J., Gupta, D., Khanna, A., & Castillo, O. (2020, August). Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm. *Soft Computing*. Retrieved 2021-06-29, from http://link.springer.com/10.1007/s00500-020-05275-y doi: 10.1007/s00500-020-05275-y
Das, A. K., Mishra, S., & Saraswathy Gopalan, S. (2020, September). Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ*, 8, e10083. Retrieved 2021-06-10, from https://peerj.com/articles/10083. doi: 10.7717/peerj.10083

De Felice, F., & Polimeni, A. (2020). Coronavirus Disease (COVID-19): A Machine Learning Bibliometric Analysis. *In Vivo*, 34(3 suppl), 1613–1617. Retrieved 2021-02-17, from http://iv.iiarjournals.org/lookup/doi/10.21873/invivo.11951. doi: 10.21873/invivo.11951

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016, June). Efficient knn classification algorithm for big data. *Neurocomput.*, 195(C), 143–148. Retrieved from https://doi.org/10.1016/j.neucom.2015.08.112. doi: 10.1016/j.neucom.2015.08.112

Duarte, E., & Wainer, J. (2017). Empirical comparison of cross-validation and internal metrics for tuning svm hyperparameters. *Pattern Recognition Letters*, 88, 6-11. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167865517300077. doi: https://doi.org/10.1016/j.patrec.2017.01.007

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139. Retrieved from https://www.sciencedirect.com/science/article/pii/S002200009791504X. doi: https://doi.org/10.1006/jcss.1997.1504

Ghasemian, A., Hosseinmardi, H., & Clauset, A. (2019). Evaluating Overfit and Underfit in Models of Network Community Structure. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. Retrieved 2021-07-02, from https://ieeexplore.ieee.org/document/8692626/. doi: 10.1109/TKDE.2019.2911585

Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: Cart model. *Computer Methods and Programs in Biomedicine*, 192, 105400. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169260719308971. doi: https://doi.org/10.1016/j.cmpb.2020.105400

Ghosh, A., & Barman, S. (2013). Prediction of Prostate Cancer Cells based on Principal Component Analysis Technique. *Procedia Technology*, 10, 37–44. Retrieved 2021-07-02, from https://linkinghub.elsevier.com/retrieve/pii/S221201731300488X. doi: 10.1016/j.protcy.2013.12.334

Gong, Z., Zhu, J.-W., Li, C.-P., Jiang, S., Ma, L.-N., Tang, B.-X., ..., Zhao, W.-M. (2020). An online coronavirus analysis platform from the national genomics data center. *Zoological Research*, 41(6), 705–708. Retrieved from https://doi.org/10.24272/j.issn.2095-8137.2020.065. doi: 10.24272/j.issn.2095-8137.2020.065

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score,
with Implication for Evaluation. In D. Hutchison et al. (Eds.), Advances in Information Retrieval (Vol. 3408, pp. 345–359). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved 2021-06-11, from http://link.springer.com/10.1007/978-3-540-31865-1_25 doi: 10.1007/978-3-540-31865-1_25

Govindarajan, M., & Chandrasekaran, R. (2010). Evaluation of k-nearest neighbor classifier performance for direct marketing. Expert Systems with Applications, 37(1), 253-258. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417409004230 doi: https://doi.org/10.1016/j.eswa.2009.04.055

Hamida, S., Gannour, O. E., Cherradi, B., Ouajji, H., & Raihani, A. (2020). Optimization of machine learning algorithms hyper-parameters for improving the prediction of patients infected with covid-19. In 2020 ieee 2nd international conference on electronics, control, optimization and computer science (icecocs) (p. 1-6). doi: 10.1109/ICECOCS50124.2020.9314373

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. Retrieved from http://ieeexplore.ieee.org/document/709601/ doi: 10.1109/34.709601

Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., . . . Pillai, S. K. (2020). First case of 2019 novel coronavirus in the united states. New England Journal of Medicine, 382(10), 929-936. Retrieved from https://doi.org/10.1056/NEJMoa2001191 (PMID: 32004427) doi: 10.1056/NEJMoa2001191

Hornik, K., Stinchcombe, M., & White, H. (1989, January). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. Retrieved from https://doi.org/10.1016/0893-6080(89)90020-8

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 429–449.

Jeffrey, H. (1990, 04). Chaos game representation of gene structure. Nucleic Acids Research, 18(8), 2163-2170. Retrieved from https://doi.org/10.1093/nar/18.8.2163 doi: 10.1093/nar/18.8.2163

Jiang, F., Deng, L., Zhang, L., Cai, Y., Cheung, C. W., & Xia, Z. (2020, March). Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). Journal of General Internal Medicine, 35(5), 1545–1549. Retrieved from https://doi.org/10.1007/s11606-020-05762-w doi: 10.1007/s11606-020-05762-w

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. Retrieved 2021-06-30, from https://www.techscience.com/cmc/v63n1/38464 doi: 10.32604/cmc.2020.010691
Keerthi, S. S., & Lin, C.-J. (2003, July). Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. *Neural Computation, 15*(7), 1667–1689. Retrieved 2021-06-03, from https://direct.mit.edu/neco/article/15/7/1667-1689/6751 doi: 10.1162/089976603321891855

Kushwaha, S., Bahl, S., Bagha, A., Parmar, K., Javaid, M., Haleem, A., & Singh, R. (2020, December 1). Significant applications of machine learning for covid-19 pandemic. *Journal of Industrial Integration and Management, 5*(4). doi: 10.1142/S2424862220500268

Li, C., Yang, Y., & Ren, L. (2020, August). Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. *Infection, Genetics and Evolution, 82*, 104285. Retrieved 2021-07-01, from https://linkinghub.elsevier.com/retrieve/pii/S1567134820301167 doi: 10.1016/j.meegid.2020.104285

Li, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B. T., & Chaillon, A. (2020, June). Evolutionary history, potential intermediate animal host, and cross-species analyses of sars-cov-2. *Journal of Medical Virology, 92*(6), 602–611. Retrieved 2021-02-20, from https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25731 doi: 10.1002/jmv.25731

Liao, Y., Fang, S.-C., & L.W. Nuttle, H. (2004). A neural network model with bounded-weights for pattern classification. *Computers & Operations Research, 31*(9), 1411-1426. Retrieved from https://www.sciencedirect.com/science/article/pii/S0305054803000972 doi: https://doi.org/10.1016/S0305-0548(03)00097-2

Lin, S.-W., Ying, K.-C., Chen, S.-C., & Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications, 35*(4), 1817-1824. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417407003752 doi: https://doi.org/10.1016/j.eswa.2007.08.088

Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Mulders, D. G. J. C., Molenkamp, R., Perez-Romero, C. A., . . . Kraneveld, A. D. (2021, December). Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports, 11*(1), 947. Retrieved 2021-02-15, from http://www.nature.com/articles/s41598-020-80363-5 doi: 10.1038/s41598-020-80363-5

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., . . . Tan, W. (2020, February). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet, 395*(10224), 565–574. Retrieved from https://doi.org/10.1016/s0140-6736(20)30251-8 doi: 10.1016/s0140-6736(20)30251-8

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classifica-
tion of spectral data. BMC Bioinformatics, 10(1), 213. Retrieved 2021-06-25, from https://www.biomedcentral.com/1471-2105/10/213 doi: 10.1186/1471-2105-10-213

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., … Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews, 207, 103225. Retrieved from https://www.sciencedirect.com/science/article/pii/S0012825220302713 doi: https://doi.org/10.1016/j.earscirev.2020.103225

Mohamadou, Y., Halidou, A., & Kapen, P. T. (2020, November). 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. Retrieved 2021-06-30, from https://link.springer.com/10.1007/s10489-020-01770-9 doi: 10.1007/s10489-020-01770-9

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakrabarty, C., & Mohammed, I. A. (2021, February). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. SN Computer Science, 2(1), 11. Retrieved 2021-02-17, from https://link.springer.com/10.1007/s42979-020-00394-7 doi: 10.1007/s42979-020-00394-7

Naeem, S. M., Mabrouk, M. S., Marzouk, S. Y., & Eldosoky, M. A. (2020, 08). A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19. Briefings in Bioinformatics. Retrieved from https://doi.org/10.1093/bib/bbaa170 (bbaa170) doi: 10.1093/bib/bbaa170

Nawaz, M. S., Fournier-Viger, P., Shojaee, A., & Fujita, H. (2021, May). Using artificial intelligence techniques for COVID-19 genome analysis. Applied Intelligence, 51(5), 3086–3103. Retrieved 2021-06-30, from https://link.springer.com/10.1007/s10489-021-02193-w doi: 10.1007/s10489-021-02193-w

Paules, C. I., Marston, H. D., & Fauci, A. S. (2020, February). Coronavirus Infections more Than Just the Common Cold. JAMA, 323(8), 707. Retrieved 2021-02-17, from https://jamanetwork.com/journals/jama/fullarticle/2759815 doi: 10.1001/jama.2020.0757

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011, November). Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12(null), 2825–2830.

Perlman, S., & Netland, J. (2009, June). Coronaviruses post-SARS: update on replication and pathogenesis. Nature Reviews Microbiology, 7(6), 439–450. Retrieved 2021-02-17, from https://www.nature.com/articles/nrmicro2147 doi: 10.1038/nrmicro2147

Priyanka, Choudhary, O. P., & Singh, I. (2021, January). Protective immunity against COVID-19: Unravelling the evidences for humoral vs. cel-
lular components. *Travel Medicine and Infectious Disease, 39*, 101911. Retrieved 2021-02-17, from https://linkinghub.elsevier.com/retrieve/pii/S1477893920304075 doi: 10.1016/j.tmaid.2020.101911

Randhawa, G. S., Soltysiak, M. P. M., El Roz, H., de Souza, C. P. E., Hill, K. A., & Kari, L. (2020, 04). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. In (Vol. 15, p. 1-24). Public Library of Science. Retrieved from https://doi.org/10.1371/journal.pone.0232391

Safavian, S., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660-674. doi: 10.1109/21.97458

Sahu, S. K., Kumar, P., & Singh, A. P. (2018, March). Modified K-NN algorithm for classification problems with improved accuracy. *International Journal of Information Technology, 10*(1), 65–70. Retrieved 2021-07-02, from http://link.springer.com/10.1007/s41870-017-0058-z doi: 10.1007/s41870-017-0058-z

Salgotra, R., Gandomi, M., & Gandomi, A. H. (2020). Evolutionary modelling of the covid-19 pandemic in fifteen most affected countries. *Chaos, Solitons & Fractals, 140*, 110118. Retrieved from https://www.sciencedirect.com/science/article/pii/S0960079920305154 doi: https://doi.org/10.1016/j.chaos.2020.110118

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

Silva, P., Luz, E., Silva, G., Moreira, G., Silva, R., Lucio, D., & Menotti, D. (2020). COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. *Informatics in Medicine Unlocked, 20*, 100427. Retrieved from https://doi.org/10.1016/j.imu.2020.100427

Song, S., Ma, L., Zou, D., Tian, D., Li, C., Zhu, J., ... Bao, Y. (2020, December). The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019ncovr. *Genomics, Proteomics & Bioinformatics*. Retrieved from https://doi.org/10.1016/j.gpb.2020.09.001 doi: 10.1016/j.gpb.2020.09.001

Sulistiana, S., & Muslim, M. A. (2020, 9). Support vector machine (svm) optimization using grid search and unigram to improve e-commerce review accuracy. *Journal of Soft Computing Exploration, 1*(1), 8–15.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009, June). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687–719. Retrieved 2021-07-04, from https://www.worldscientific.com/doi/abs/10.1142/S0218001409007326 doi: 10.1142/S0218001409007326
Syrif, I., Prugel-Bennett, A., & Wills, G. (2016, December). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4), 1502. Retrieved 2021-06-03, from http://journal.uad.ac.id/index.php/TELKOMNIKA/article/view/3956 doi: 10.12928/telkomnika.v14i4.3956

Tayarani N., M.-H. (2021, January). Applications of artificial intelligence in battling against covid-19: A literature review. *Chaos, Solitons & Fractals*, 142, 110338. Retrieved 2021-02-17, from https://linkinghub.elsevier.com/retrieve/pii/S0960077920307335 doi: 10.1016/j.chaos.2020.110338

Too, J., & Mirjalili, S. (2021). A hyper learning binary dragonfly algorithm for feature selection: A covid-19 case study. *Knowledge-Based Systems*, 212, 106553. Retrieved from https://www.sciencedirect.com/science/article/pii/S0950705120306821 doi: https://doi.org/10.1016/j.knosys.2020.106553

Ucar, E., Atila, U., Ucar, M., & Akyol, K. (2021). Automated detection of covid-19 disease using deep fused features from chest radiography images. *Biomedical Signal Processing and Control*, 69, 102862. Retrieved from https://www.sciencedirect.com/science/article/pii/S1746809421004596 doi: https://doi.org/10.1016/j.bspc.2021.102862

Udugama, B., Kadhiresan, P., Kozlowski, H. N., Malekjahani, A., Osborne, M., Li, V. Y. C., ... Chan, W. C. W. (2020, April). Diagnosing COVID-19: The Disease and Tools for Detection. *ACS Nano*, 14(4), 3822–3835. Retrieved 2021-06-29, from https://pubs.acs.org/doi/10.1021/acsnano.0c02624 doi: 10.1021/acsnano.0c02624

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. Retrieved 2021-06-03, from http://link.springer.com/10.1007/978-1-4757-2440-0 doi: 10.1007/978-1-4757-2440-0

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., ... Peng, Z. (2020, March). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *JAMA*, 323(11), 1061. Retrieved from https://doi.org/10.1001/jama.2020.1585 doi: 10.1001/jama.2020.1585

Wang, Y., Mao, J.-M., Wang, G.-D., Luo, Z.-P., Yang, L., Yao, Q., & Chen, K.-P. (2020, Jul 23). Human sars-cov-2 has evolved to reduce cg dinucleotide in its open reading frames. *Scientific Reports*, 10(1), 12331. Retrieved from https://doi.org/10.1038/s41598-020-69342-y doi: 10.1038/s41598-020-69342-y

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., ... Zhang, Y.-Z. (2020, March). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265–269. Retrieved 2021-02-17, from http://www.nature.com/articles/s41586-020-2008-3 doi: 10.1038/s41586-020-2008-3
Xingguang, L., Wei, W., Xiaofang, Z., Junjie, Z., Qiang, Z., Yi, L., & Antoine, C. (2020). Transmission dynamics and evolutionary history of 2019-ncov. *J. Med. Virol.*, 92, 501-511. Retrieved from [https://doi.org/10.1002/jmv.25701](https://doi.org/10.1002/jmv.25701) doi: 10.1002/jmv.25701

Zargari Khuzani, A., Heidari, M., & Shariati, S. A. (2021). COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. *Scientific Reports*, 11(1), 9887. Retrieved 2021-06-29, from [http://www.nature.com/articles/s41598-021-88807-2](http://www.nature.com/articles/s41598-021-88807-2) doi: 10.1038/s41598-021-88807-2

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., ... Shi, Z.-L. (2020, March). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. Retrieved 2021-02-17, from [http://www.nature.com/articles/s41586-020-2012-7](http://www.nature.com/articles/s41586-020-2012-7) doi: 10.1038/s41586-020-2012-7

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021, December). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, 4(1), 3. Retrieved 2021-02-17, from [http://www.nature.com/articles/s41746-020-00372-6](http://www.nature.com/articles/s41746-020-00372-6) doi: 10.1038/s41746-020-00372-6