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.
Classifying COVID-19 based on amino acids encoding with machine learning algorithms

Walaa Alkady a,**, Khaled ElBahnasy b, Víctor Leiva c,*, Walaa Gad b

a Department of Bioinformatics, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
b Department of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
c School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

ARTICLE INFO

Keywords:
Amino acid composition
ANOVA
Artificial intelligence
Bagging ensemble and gradient boosting
Chi-square test
Deep learning
Feature extraction and selection
Information gain
LASSO
Molecular modeling
Protein sequence
SARS-CoV-2

ABSTRACT

COVID-19 disease causes serious respiratory illnesses. Therefore, accurate identification of the viral infection cycle plays a key role in designing appropriate vaccines. The risk of this disease depends on proteins that interact with human receptors. In this paper, we formulate a novel model for COVID-19 named “amino acid encoding based prediction” (AAPred). This model is accurate, classifies the various coronavirus types, and distinguishes SARS-CoV-2 from other coronaviruses. With the AAPred model, we reduce the number of features to enhance its performance by selecting the most important ones employing statistical criteria. The protein sequence of SARS-CoV-2 for understanding the viral infection cycle is analyzed. Six machine learning classifiers related to decision trees, k-nearest neighbors, random forest, support vector machine, bagging ensemble, and gradient boosting are used to evaluate the model in terms of accuracy, precision, sensitivity, and specificity. We implement the obtained results computationally and apply them to real data from the National Genomics Data Center. The experimental results report that the AAPred model reduces the features to seven of them. The average accuracy of the 10-fold cross-validation is 98.69%, precision is 98.72%, sensitivity is 96.81%, and specificity is 97.72%. The features are selected utilizing information gain and classified with random forest. The proposed model predicts the type of Coronavirus and reduces the number of extracted features. We identify that SARS-CoV-2 has similar physicochemical characteristics in some regions of SARS-CoV. Also, we report that SARS-CoV-2 has similar infection cycles and sequences in some regions of SARS CoV indicating the affectedness of vaccines on SARS-CoV-2. A comparison with deep learning shows similar results with our method.

1. Introduction, bibliographical review, notations, and objectives

1.1. Introduction and bibliographical review

The severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) infects humans with potentially serious pathogens and was discovered in 2019 [1]. This virus causes a pulmonary pathology called coronavirus disease 2019 (COVID-19). The SARS-CoV-2 or COVID-19 has infected an enormous number of people provoking thousands of deaths all around the world [2–4]. In March 2020, the World Health Organization (WHO) considered COVID-19 as a global pandemic [5].

The SARS-CoV-2 is classified within the Coronavirus family, which belongs to the Nidovirales order. This family [6,7] has four genera: alpha coronavirus, beta coronavirus, gamma coronavirus, and delta coronavirus. Human beings are infected by alpha and beta genera. Alpha has two human coronaviruses: HCoV-229E and HCoV-NL63. Beta has five human coronaviruses: HCoV-HKU1, HCoV-OC43, middle east respiratory syndrome coronavirus (MERS-CoV), SARS-CoV, and SARS-CoV-2.

The coronaviruses are positive single-stranded ribonucleic acid (RNA) viruses. They have uncommonly long and complex genomes in comparison to other RNA viruses. Two overlapped open reading frames (ORF1a and ORF1b) are occupying 66% of the corona-virus genome. These reading frames are translated to sixteen non-structural proteins (NSPs) encoded from NSP1 to NSP16. The remaining 34% of the genome is occupied by the coding regions of structural proteins. Coronavirus has four primary basic structural proteins [8]: the spike surface glycoprotein, envelope protein, membrane protein, and nucleocapsid protein. The

* Corresponding author.
** Corresponding author.
E-mail addresses: walaa.samir@cis.asu.edu.eg (W. Alkady), victor.leiva@pucv.cl, victorleivasanchez@gmail.com (V. Leiva).

https://doi.org/10.1016/j.chemolab.2022.104535
Received 24 October 2021; Received in revised form 27 February 2022; Accepted 5 March 2022
Available online 15 March 2022
0169-7439/© 2022 Elsevier B.V. All rights reserved.
spike proteins are responsible for interacting with the host receptors. In addition, the viral genome contains several auxiliary proteins.

According to phylogenetic studies, SARS-CoV, SARS-CoV-2, and other viruses in the beta coronavirus genus may be grouped into one cluster. Most human coronaviruses, such as SARS-CoV and MERS-CoV, are produced in bats and are transmitted to humans via intermediate hosts [9]. The similarity between other coronaviruses and SARS-CoV-2 is recognized in the receptor-binding domain (RBD) of the spike protein, which plays a significant role in the drug design process for this virus and determines the potential hosts of the infection [10,11].

Coronavirus crosses the species barrier and infects humans with very dangerous diseases [12,13]. Public health is challenged due to the novelty of the antigen for the human host. Due to the danger of COVID-19, many methods have been introduced to develop models as well as to diagnose and differentiate it from other viruses.

In [14], eighteen radiological semantic features and seventeen clinical features were selected from 67 features (41 radiological and 26 clinical features). Also, univariate analysis methods, as the least absolute shrinkage and selection operator (LASSO), have been used for feature selection and its classification performance has been evaluated based on accuracy, sensitivity, and specificity measures. Using clinical features, accuracy, sensitivity, and specificity were reached to 88.8%, 89.2%, and 88.4%, respectively. Employing radiological features, accuracy was 92.9%, sensitivity was 99%, and specificity was 85.1%. Utilizing both types of features, accuracy, sensitivity, and specificity were of 95.9%, 96.1%, and 95.7%, respectively.

In [15], the infection risk of non-human-origin coronavirus is classified considering the spike protein sequence of coronaviruses. A feature extraction method named amino acid composition (AAC) was used to reduce the 41 features of the protein sequence. The AAC counts the frequency of twenty amino acids in the sequence and then the features were reduced to 20. Protein sequences of 2666 coronaviruses were collected from the National Genomics Data Center (NGDC) database (ngdc.cnbc.ac.cn, accessed on October 24, 2021) [16]. The predictive model achieved an accuracy between 96.15% and 98.18% utilizing the random forest (RF) classifier and 10-cross validation. The results showed that SARS-CoV-2 is like SARS-CoV. It means both viruses have the same human receptor. Until now, the origin of SARS-CoV-2 is unknown and needs further analysis and study.

In [17,18], a diagnostic model based on X-ray images was proposed. In this model, related features to image texture and shapes were extracted employing the Zernike-Haralick method. The Kyoto Encyclopedia of Genes and Genomes (KEGG) dataset (www.genome.jp/kegg, accessed on October 24, 2021) was considered to test this model. It contains 464 COVID-19 images, 1490 viral images, 2783 bacterial images, and 1583 healthy images. The proposed model employed the support vector machine (SVM) classifier and got an accuracy of 89.78%, precision of 89.85%, sensitivity of 90.9%, and specificity of 99.63%.

Note that the selection of human monoclonal antibodies may identify immunodominant antigenic sites associated with neutralization and provide reagents for stabilizing and solving the structure of viral surface proteins [19]. Understanding the structural basis of the SARS-CoV-2 can guide the selection of vaccine targets. Studying the similarity between coronaviruses and the sequences analysis of these viruses results in knowing whether the other viruses’ vaccines may affect SARS-CoV-2 or not. COVID-19 has an infection cycle and a similar sequence in some regions of the proteins that indicate the efficacy of vaccines against SARS-CoV-2.

Machine-learning methods and artificial intelligence [20,21,45] can be used to analyze, predict, and diagnose infection rates. In addition, machine learning helps in the drug discovery process for a vaccine. To the best of our knowledge, there are no studies that accurately differentiate between SARS-CoV-2 and other coronavirus types with a small number of features. Therefore, machine learning models are needed to make this differentiation.

As mentioned in [44], differentiating between coronavirus types is helpful for designing an effective vaccine. Moreover, the mutation pattern of the COVID-19 can be partially predicted after applying more analysis on the difference of the protein sequences of different strains.

### 1.2. Abbreviations, acronyms, notations, and symbols

In this section, all abbreviations and symbols employed in this paper are defined in Table 1 in alphabetical order. One of the used abbreviations is the standard one-letter abbreviation of the twenty amino acids found in [22].

| Acronym/ Symbol | Meaning | Acronym/ Symbol | Meaning |
|-----------------|---------|----------------|---------|
| A, G, M, ... | The 20 amino acids 1-letter abbreviations [19] | MI | Mutual information |
| AAC | Amino acid composition | MMS | Mean of sums of squares in the between groups |
| AAPred | Amino acid encoding based prediction | MMS | Mean of sums of squares in the same group |
| Acc | Accuracy | NCBI | National Center for Biotechnology Information |
| ANOVA | Analysis of variance | NGDC | National Genomics Data Center |
| BE | Bagging ensemble | NSPs | Non-structural proteins |
| c, d, f | Number of classes, dataset, feature | Num | Number of occurrences of class |
| COVID-19 | Coronavirus disease 2019 | ORF | Open reading frame |
| CSV | Comma-separated values | Pr | Probability of selecting a sample of a class |
| DNA | Desoxyribonucleic acid | PK | Polynomial kernel |
| DT | Decision tree | Prec | Precision |
| Ent | Entropy | RBD | Receptor-binding domain |
| FASTA | Format of text for nucleotide or amino acids | RBF | Radial basis function |
| FN | False-negative | RF | Random forest |
| FP | False-positive | RNA | Ribonucleic acid |
| Freq | Frequency of each amino acid class | SARS-CoV | Severe acute respiratory syndrome coronavirus |
| GB | Gradient boosting | Sens | Sensitivity |
| IG | Information gain | SK | Sigmoid kernel |
| KEGG | Kyoto Encyclopedia of Genes and Genomes | Spec | Specificity |
| KNN | k-nearest neighbor | Spyder | Scientific python environment |
| LASSO | Least absolute shrinkage and selection operator | SVM | Support vector machine |
| Len | Length of the protein sequence | TN | True-negative |
| LK | Linear kernel | TP | True-positive |
| MD | Manhattan distance | WHO | World Health Organization |
| MERS-CoV | Middle East respiratory syndrome coronavirus | $\chi^2$ | Chi-square |

### 1.3. Objectives and outline

The objectives of this paper are as follows:

- To formulate a novel accurate model for COVID-19 that classifies the various coronavirus types and differentiates SARS-CoV-2 from other coronaviruses.
- To use machine learning techniques for evaluating the model in terms of accuracy, precision, sensitivity, and specificity.
- To reduce the number of features of the novel model to enhance its performance.
- To analyze the protein sequence of SARS-CoV-2 for understanding the viral infection cycle.

Table 1

| Acronym/ Symbol | Meaning | Acronym/ Symbol | Meaning |
|-----------------|---------|----------------|---------|
| A, G, M, ... | The 20 amino acids 1-letter abbreviations [19] | MI | Mutual information |
| AAC | Amino acid composition | MMS | Mean of sums of squares in the between groups |
| AAPred | Amino acid encoding based prediction | MMS | Mean of sums of squares in the same group |
| Acc | Accuracy | NCBI | National Center for Biotechnology Information |
| ANOVA | Analysis of variance | NGDC | National Genomics Data Center |
| BE | Bagging ensemble | NSPs | Non-structural proteins |
| c, d, f | Number of classes, dataset, feature | Num | Number of occurrences of class |
| COVID-19 | Coronavirus disease 2019 | ORF | Open reading frame |
| CSV | Comma-separated values | Pr | Probability of selecting a sample of a class |
| DNA | Desoxyribonucleic acid | PK | Polynomial kernel |
| DT | Decision tree | Prec | Precision |
| Ent | Entropy | RBD | Receptor-binding domain |
| FASTA | Format of text for nucleotide or amino acids | RBF | Radial basis function |
| FN | False-negative | RF | Random forest |
| FP | False-positive | RNA | Ribonucleic acid |
| Freq | Frequency of each amino acid class | SARS-CoV | Severe acute respiratory syndrome coronavirus |
| GB | Gradient boosting | Sens | Sensitivity |
| IG | Information gain | SK | Sigmoid kernel |
| KEGG | Kyoto Encyclopedia of Genes and Genomes | Spec | Specificity |
| KNN | k-nearest neighbor | Spyder | Scientific python environment |
| LASSO | Least absolute shrinkage and selection operator | SVM | Support vector machine |
| Len | Length of the protein sequence | TN | True-negative |
| LK | Linear kernel | TP | True-positive |
| MD | Manhattan distance | WHO | World Health Organization |
| MERS-CoV | Middle East respiratory syndrome coronavirus | $\chi^2$ | Chi-square |
To implement the obtained results computationally.
To apply the results to real-world data, in our case using the NGDC database.

Specifically, we propose a novel model named amino acid encoding-based prediction (AAPred) to carry out our investigation. The AAPred model classifies and distinguishes between COVID-19 and other coronaviruses adopting the amino acid encoding method [23] for features extraction and to improve the model performance. The amino acid encoding method is modified by adding a new amino acid class called “zero”, which contains unknown amino acids. Also, ambiguous amino acids are added to amino acids classes. In addition, the feature selection phase is applied to remove irrelevant features. The proposed AAPred model employs three different methods [24] for feature selection which are detailed in Subsection 2.1. This model optimizes the performance when predicting the virus type and utilizes mathematical and statistical methods to reduce the number of features.

The paper is organized as follows. Section 2 introduces the methodology and the AAPred model. Section 3 presents the experimental results employing the NGDC database. Finally, Section 4 provides a discussion as well as the conclusions, limitations, and ideas for future research.

2. Methodology

2.1. The proposed AAPred model

The AAPred model uses selection algorithms to extract features and then reduces the number of them eliminating those that are irrelevant. Three different methods of feature selection are utilized:

- Information gain (IG) [25],
- Analysis of variance (ANOVA) test [26], and
- Chi-square ($\chi^2$) statistic test [27].

The number of extracted features is reduced by removing irrelevant features. Specifically, protein sequence analysis requires feature extraction to convert sequence characters to a numerical form employing amino acid encoding. Lastly, the classification is performed predicting the coronavirus type using six different machine learning algorithms:

- Bagging ensemble (BE),
- Decision trees (DT),
- Gradient boosting (GB),
- k-nearest neighbors (KNN),
- RF, and
- SVM.

In summary, as shown in Fig. 1, the AAPred model consists of three phases:

- Feature extraction,
- Feature reduction, and
- Classification.

2.2. Feature extraction phase

In the feature extraction phase, the amino acid encoding method is used to obtain features from the viral protein sequences. The amino acid encoding method employs two physicochemical properties of the amino acids: volume and dipole [30]. The volume and dipole values are calculated utilizing molecular modeling and density-functional methods [31]. The calculated volumes and dipole values of amino acids divide the twenty amino acids into seven classes as shown in Table 2.

Table 2

| Class number | Dipole scale | Volume scale | Amino acids |
|--------------|--------------|--------------|-------------|
| 1            | –            | –            | A, G, V     |
| 2            | –            | +            | I, L, F, P  |
| 3            | +            | +            | Y, M, T, S  |
| 4            | ++           | +            | H, N, Q, W  |
| 5            | ++ ++        | +            | R, K        |
| 6            | ++ ++ ++     | +            | D, E        |
| 7            | +            | +            | C           |

In summary, as shown in Fig. 1, the AAPred model consists of three phases:

- Feature extraction,
- Feature reduction, and
- Classification.

Note that the volume scale “–” means that the volume is less than 50; the “+” means that volume is greater than 50; and cysteine (C) amino acid is moved from Class 3 to Class 7 because it can form disulfide bonds.

Due to the ambiguity of protein sequencing, there are four ambiguous amino acid characters added to the twenty amino acids. An eighth class labeled with zero is added to the seven classes of amino acid encoding. The newly added class contains three ambiguous amino acids (X, Z, and B), where X is an unknown amino acid, Z is a glutamic acid (E) or

![Fig. 1. Phases of the AAPred model and their connection.](image-url)
glutamine (Q), and B is an aspartic acid (D) or asparagine (N). The fourth ambiguous amino acid (J) is leucine (L) or isoleucine (I), which is added to Class 2 because I and L amino acids belong to this class. The ambiguity of amino acid codes comes from the error percentage of the sequencer of amino acids. During the sequencing process, the machine can be unable to detect the coming amino acid. For example, is it (L) or (I)? Thus, the sequencer puts (J), which means this amino acid may be (I) or (L).

Table 3 shows the amino acid classification. The amino acid encoding method replaces each amino acid character with its corresponding class number. This method converts the amino acids to numbers according to physicochemical properties.

### 2.3. Feature reduction phase

In this phase, optimal features are selected employing the mentioned feature selection techniques: information gain (recall it is denoted as IG in short), ANOVA, and $\chi^2$.

The IG is a filter-based method that calculates the information amount for each feature item related to the target class. It measures features' importance for the classification phase. The IG of a dataset (d) concerning one feature is calculated as

$$IG(d, f) = \text{Ent}(d) - \text{Ent}(d|f),$$  
(2)

where $\text{Ent}(d)$ stated in (2) is the entropy of the dataset d and $\text{Ent}(d|f)$ is the conditional entropy of d given the feature f. The entropy of d with the classes is defined as

$$\text{Ent}(d) = - \sum_{i=1}^c (p_i \log_2(p_i)),$$

(3)

where $p_i$ given in (3) is the probability of selecting a sample of Class i, and c is the number of classes. The conditional entropy is obtained by dividing the dataset into several groups, with each group consisting of two features. Then, we compute the quantitative relation of samples in each group out of the complete dataset multiplied by the entropy of each group based on the expression stated as

$$\text{Ent}(d|f) = - \sum_{d \in D, f \in F} \left( \text{Ent}(d) \log \left( \frac{p(f)}{p(d)} \right) \right),$$

(4)

where $p(f)$ given in (4) is the probability of selecting a feature and $D, F$ are the universe of datasets and features, respectively. Mutual information (MI) is calculated between two features with MI measuring the IG for one feature given a known value of the other feature. Thus, the MI between two features $f_1$ and $f_2$ is formulated as

$$MI(f_1, f_2) = IG(d, f_1) - IG(f_1|f_2),$$

(5)

where $MI(f_1, f_2)$ stated in (5) is the mutual information for $f_1$ and $f_2$. $IG(d, f_1)$ is the information gain for $f_1$, and $IG(f_1|f_2)$ is the conditional information gain for $f_1$ given $f_2$. Therefore, we have that the expression formulated as

$$IG(f_1|f_2) = - \sum_{d \in D, f_1 \in F} \left( IG(f_1) \log \left( \frac{p(f_1|f_2)}{p(f_1)} \right) \right),$$

(6)

which defines the conditional information gain for $f_1$ given $f_2$, where $IG(f)$ stated in (6) is the IG of a feature f with respect to the dataset d and $p(f_1|f_2)$ is the probability of selecting the two features. In the proposed model, IG is calculated according to the information of each amino acid class.

The ANOVA is a statistical test that studies the variance of means of squares in different groups. The ANOVA is constructed from the expressions stated as

$$MMS_{\text{within}} = \frac{\sum_{i=1}^G \sum_{j \in F_i} (f_j - \bar{F}_i)^2}{(G - 1)},$$

(7)

where $MMS_{\text{within}}$ is the mean sum of squares in the same group, f is the feature in this group, $\bar{F}_i$ is the mean of the features in the group, G is the number of groups, $(G - 1)$ are the degrees of freedom in the group, and

$$MMS_{\text{between}} = \frac{\sum_{i=1}^G (\bar{F}_i - \bar{F})^2}{(S - G)},$$

(8)

where $MMS_{\text{between}}$ defined in (8) is the mean sum of squares between groups, with $\bar{F}_i$ and G being defined as in (5), $\bar{F}$ is the mean of all features, S is the total number of samples, and $(S - G)$ corresponds to the degrees of freedom between groups. Therefore, the statistic used to compare groups is given by

$$\text{ANOVA} = \frac{MMS_{\text{between}}}{MMS_{\text{within}}},$$

(9)

Note that the ANOVA establishes the variation between the frequency of each amino acid class with respect to the variation of all virus sequences.

The $\chi^2$ test is applied to the groups of features to evaluate the association between them using the corresponding statistic defined as

$$\chi^2 = \sum_{i=1}^G \frac{(O_i - E_i)^2}{E_i},$$

(10)

where $O_i$ and $E_i$ are observed and predicted or expected frequencies in Class i, respectively, with G, as mentioned in (7), being the number of groups.

All these feature selection methods, that is, IG, ANOVA, $\chi^2$ test, are efficient in selecting features on protein sequence datasets because of the relationship between the amino acid classes. The proposed AAPred model uses the IG established in (2), ANOVA defined in (9), and $\chi^2$ statistic stated in (10) to assign scores to the extracted features. Then, the features are sorted, and the highest scores are considered in the classification phase. This step is very important to select the most significant features and remove unwanted features.

### 2.4. Classification phase

In the classification phase, the significant features are used to classify the coronaviruses to COVID-19 and non-COVID-19 employing the BE, DT, GB, KNN, RF, and SVM algorithms. Prediction of coronavirus types is
considered a binary classification problem.

The BE algorithm [42] utilizes a basis classifier for a given number of iterations. The BE classifier applies weight to the training set to help the next iteration of the classifier to achieve better performance.

The DT classifier splits the dataset based on the entropy by means of the formula given by

$$\text{Ent}(d) = -\sum_{x \in X} \left[ \text{Pr}(x) \log_2 \text{Pr}(x) \right].$$  \hspace{1cm} (11)

where Ent(d) given in (11) is the entropy of the dataset d, x is the set of classes (COVID-19 class or non_CoVID-19 class), and Pr(x) is the proportion (probability) between samples in Class x and the total number of samples in d. After calculating the entropy values for each sample, the DT method splits the dataset using the IG defined as

$$\text{IG}(d, a) = \text{Ent}(d) - \sum_{b \in B} \left( \text{Pr}(b) \text{Ent}(b) \right).$$ \hspace{1cm} (12)

with Ent(d) being defined in (12), B is the universe of subsets of the samples that are divided using the random variable a, Pr(b) is the proportion between samples in subset b \(\in B\) and the total number of the dataset, and Ent(b) is the entropy of subset b.

The GB algorithm [43] aims to maximize the results of the classifier correlated with the negative gradient of the loss function.

The KNN classifier is applied using the Manhattan distance (MD), which is calculated as the sum of the absolute values of the differences of the samples and then defined as

$$\text{MD} (x, y) = \sum_{i=1}^{k} |x_i - y_i|.$$ \hspace{1cm} (13)

where k stated in (13) is the number of features (dimensions), and x, y are two samples (vectors).

The RF algorithm is an ensemble classifier that combines several classifiers. The RF method consists of multiple DTs, each of which works as a classifier to predict the class label and then the majority voting of these trees’ outputs is utilized to predict the class label.

The SVM algorithm is applied using different kernel functions: Linear, hyperbolic tangent or sigmoid, polynomial, and radial basis function (RBF) [34]. The simplest kernel is the linear function, which is computed by the inner product of the feature vector and the class label vector, being it expressed as

$$\text{LK}(f, y) = \langle f, y \rangle + g.$$ \hspace{1cm} (14)

where LK(f, y) is the linear kernel, f is the feature vector, y is the class label vector, “\(\langle f, y \rangle\)” is the corresponding inner product, and g is a free general parameter (constant). The polynomial kernel function is calculated as

$$\text{PK}(f, y) = \langle f, y \rangle^q + g.$$ \hspace{1cm} (15)

where PK(f, y) given in (15) is the polynomial kernel, f, y, g are defined as in (14), and q is the polynomial degree. The sigmoid kernel utilizes the bipolar sigmoid function formulated as

$$\text{SK}(f, y) = \tanh(\alpha \langle f, y \rangle + g).$$ \hspace{1cm} (16)

where SK(f, y) stated in (16) is the sigmoid kernel, f, y, g are defined as in (14)-(15), and \(\alpha\) is a value of the slope in the linear function. The RBF is an exponential function established as

$$\text{RK}(f, y) = \exp \left( \gamma |f - y| \right).$$ \hspace{1cm} (17)

where RK(f, y) given in (17) is the radial basis kernel, f, y, g are defined as in (14)-(16), and \(\gamma\) is an adjustable parameter.

3. Application

3.1. Data description

As mentioned, the proposed model is evaluated using the NGDC dataset that is described by two types of files: comma-separated values (CSV) and FASTA, a format that is based on the text for representing either nucleotide or amino acid sequences, in which nucleotides or amino acids are denoted by single-letter codes. This dataset was accessed on July 2020 and has 113,927 samples of protein sequences for COVID-19 and other coronavirus types. There are different types of coronaviruses in this dataset such as alpha coronaviruses, bat coronaviruses, MERS-CoV, SARS-CoV, and SARS-CoV-2. The NGDC dataset contains 60,539 protein sequences of COVID-19 and the remaining 53,388 sequences are non-COVID-19 viruses. Therefore, the dataset is balanced between the two coronaviruses types the COVID-19 and non-COVID-19 by selecting only 53,388 from the COVID-19 protein sequences randomly. The remaining COVID-19 sequences were considered in the model training. Also, the NGDS dataset was accessed on November 2020 and the proposed model is evaluated with the new uploaded sequences totaling 520,789 protein sequences. Moreover, the AAPred model is evaluated employing another dataset that contains only the spike protein of coronaviruses from the National Center for Biotechnology Information (NCBI) coronavirus dataset [35].

For COVID-19, the minimum protein sequence length is 21 amino acids, and its maximum value is 7097 amino acids. For non-COVID-19, the minimum protein sequence length is 26 amino acids, and its maximum value is 7247 amino acids. The CSV file has data about viruses’ protein sequences, such as accession numbers, collection date of protein sequences, species, genus, family, protein sequence length, isolation source, host, and geographical location as shown in Table 4.

The FASTA file contains the protein sequences with the accessions number and virus type as a header for each protein sequence as shown in Fig. 2.

3.2. Model evaluation criteria

As mentioned, the AAPred model is evaluated using four performance measures: accuracy, precision, sensitivity, and specificity defined next. The accuracy is stated as

$$\text{Acc} = \left( \frac{\text{TP} + \text{TN}}{\text{S}} \right) \times 100,$$ \hspace{1cm} (18)

where TP given in (18) is the total number of true positives, TN is the total number of true-negatives, and S is the total number of samples in the dataset.

The precision criterion calculates the positive prediction or the correctness of the model when the classifier predicts the protein sequence as COVID-19. The precision formula is defined by

$$\text{Prec} = \left( \frac{\text{TP}}{\text{TP} + \text{FP}} \right) \times 100,$$ \hspace{1cm} (19)

where TP is given in (18) and FP in (19) is the total number of false-positives. The summation of TP and FP represents the total number of protein sequences predicted as COVID-19 (predicted as presence of the virus).

Table 4

| Accession | Species        | Length | Host        |
|-----------|----------------|--------|-------------|
| AVP78037  | SARS-CoV-2     | 121    | Homo Sapiens|
| AVP78039  | SARS-CoV-2     | 97     | Homo Sapiens|
| AVP78040  | SARS-CoV       | 70     | Homo Sapiens|
| BBE15202  | Alpha          | 237    | Felis Catus |
| QP71705   | Avian          | 125    | Gallus Gallus|
| AXM42849  | Porcine        | 161    | Sus Scrofa  |
| ATG48898  | MERS           | 4391   | Homo Sapiens|
The sensitivity states the true-positive rate of the proposed model when the classifier predicts the protein sequence as COVID-19 and it is really a COVID-19. The sensitivity is also called recall and formulated as
\[
\text{Sens} = \frac{TP}{TP + FN} \times 100,
\]
where FN stated in (20) is the total number of false negatives. The summation of TP and FN represents the number of COVID-19 protein sequences (when the presence of the virus is true).

The specificity determines the true-negative rate of the model when the classifier predicts the protein sequence as non-COVID-19 and it is really a non-COVID-19. The specificity is defined as
\[
\text{Spec} = \frac{TN}{TN + FP} \times 100,
\]
where TN is given in (18) and FP stated in (21) is the total number of false-positives. The summation of TN and FP represents the number of non-COVID-19 protein sequences (when the absence of the virus is true).

3.3. Results

In the dataset, protein sequences are categorized into two classes: COVID-19 and non-COVID-19. The 106,776 protein sequences are divided into 80% for training and 20% for testing. Moreover, all protein sequences were evaluated using 10-fold cross-validation. Fig. 3 shows the frequencies of the eight amino acid classes for the samples in the dataset. In Fig. 3, the x-axis represents the eight classes of amino acids, and the y-axis is the frequency of each amino acid class in each sample. As shown in Fig. 3, the COVID-19 samples contain high amounts of Class 2 amino acids. This class dipole value is less than 1.0 and includes the amino acids: I, L, Phenylalanine (F), and Proline (P). Concerning non-COVID-19 samples, it contains high amounts of amino acids of Classes 3 and 6. These classes include the amino acids: D, E, Methionine (M), Serine (S), Threonine (T), and Tyrosine (Y). Moreover, the third and sixth classes of amino acids dipole values range between 1.0 and 3.0.

Therefore, SARS-CoV-2 has similar physicochemical characteristics to some regions of SARS-CoV. In addition, the polarizabilities of most amino acids found in non-COVID-19 protein sequences are higher than the polarizabilities of amino acids found in COVID-19 protein sequences. Therefore, COVID-19 is potentially nonpolar and tends to form chemical bonds with very small dipole values.

The proposed model was fitted on the scientific python development environment (Spyder) using the Scikit-learn software package for python programming language. The model was applied on a computer with Intel(R) Core(TM) i7-9750H CPU and 16 GB RAM.

In the AAPred model, amino acid encoding extracts the features and, as mentioned, three different selection methods are utilized to reduce the number of features and optimize the performance. Then, the performance measures are employed for evaluation, that is, accuracy, precision, sensitivity, and specificity.

Table 5 reports the performance of the six different classifiers based on the NGDC dataset using 80% training and 20% testing. As reported in this table, the highest performance is reached using the $\chi^2$ method for feature selection in the KNN classifier. The $\chi^2$ method reduces the number of features to seven of them. Accuracy, precision, sensitivity, and specificity are reached 99.69%, 99.72%, 99.65%, and 99.72%, respectively.

Table 6 reports the average performance using a 10-fold cross-validation. The average performance reaches, employing the $\chi^2$ method for feature selection with the KNN classifier, an accuracy, precision, sensitivity, and specificity of 90.69%, 91.72%, 89.65%, and 88.72%, respectively. The highest performance employing the 10-cross-validation is reached with the RF classifier and the IG method as shown in Figs. 4–6.

The accuracy of the SVM classifier employing the $\chi^2$ method is 94.15%. In addition, the DT classifier reaches an accuracy of 94.39%, a precision of 90.48%, a sensitivity of 96.31%, and a specificity of 90.48% using an ANOVA. The accuracy of the RF classifier with IG is 98.69% and this is the maximum accuracy reached.

Note that the three feature selection methods reduce the number of features to seven. The selected seven features from each method are different, but there are two common features among the three feature selection methods: the frequency of Classes 2 and 6.

The proposed model performance using 10-cross-validation based on IG, ANOVA, and $\chi^2$ test is shown in Figs. 4–6. These figures reveal that the RF classifier outperforms the other classifiers, and the IG method is the best one concerning accuracy.

The proposed model was applied to the spike protein dataset to predict the infection risk. The results reached 99.01%, 96.41%, 98.56%, and 97.02% for accuracy, precision, sensitivity, and specificity, respectively. Fig. 7 shows the performance of the RF classifier based on the spike protein dataset using the three selection methods: IG, ANOVA, and $\chi^2$ test.

The proposed model is evaluated using six different classifiers of machine learning algorithms [20,28,29]: BE, DT, GB, KNN, RF, and SVM. The proposed model reached an accuracy of 99.69%, a precision of
99.58%, a sensitivity of 99.80%, and a specificity of 99.58% based on 80% of the data for training and 20% for testing. The average performance measures using 10-fold cross-validation are 98.69%, 98.72%, 96.81%, and 97.72% for accuracy, sensitivity, specificity, and precision, respectively. The AAPred model outperforms the existing models in terms of accuracy and computation time. Moreover, it reduces the

Table 5
AAPred model performance for the indicated method and classifier with the NGDC dataset.

| Classifier | IG | Sens | Spec | Prec | ANOVA | IG | Sens | Spec | Prec | χ² | Acc | Sens | Spec | Prec |
|------------|----|------|------|------|-------|----|------|------|------|----|------|------|------|------|
| BE         | 98.53 | 98.21 | 98.43 | 98.35 | 96.91 | 96.32 | 96.44 | 96.83 | 98.89 | 98.87 | 98.33 | 98.33 |
| DT         | 99.23 | 99.19 | 99.37 | 99.37 | 99.39 | 99.31 | 99.48 | 99.48 | 99.28 | 99.17 | 99.40 | 99.40 |
| GB         | 97.61 | 96.80 | 97.85 | 97.51 | 95.62 | 95.34 | 95.64 | 95.64 | 97.81 | 97.56 | 97.74 | 97.73 |
| KNN        | 99.66 | 99.65 | 99.67 | 99.67 | 99.63 | 99.55 | 99.71 | 99.71 | 99.69 | 99.65 | 99.72 | 99.72 |
| RF         | 99.69 | 99.80 | 99.58 | 99.58 | 99.69 | 99.81 | 99.56 | 99.56 | 99.68 | 99.78 | 99.57 | 99.57 |
| SVM        | 95.13 | 95.40 | 94.86 | 94.89 | 95.14 | 95.39 | 94.89 | 94.91 | 95.15 | 95.40 | 94.90 | 94.92 |

Table 6
AAPred model performance for the indicated method and classifier with the NGDC dataset using 10-fold cross-validation.

| Classifier | IG | Sens | Spec | Prec | ANOVA | IG | Sens | Spec | Prec | χ² | Acc | Sens | Spec | Prec |
|------------|----|------|------|------|-------|----|------|------|------|----|------|------|------|------|
| BE         | 98.53 | 96.21 | 97.43 | 98.35 | 96.51 | 96.32 | 94.44 | 90.83 | 93.91 | 90.44 | 90.83 |
| DT         | 89.23 | 89.19 | 89.37 | 89.37 | 94.39 | 96.31 | 90.48 | 90.48 | 89.28 | 89.17 | 89.4   | 89.4   |
| GB         | 97.61 | 96.8  | 97.85 | 97.51 | 95.62 | 95.34 | 95.64 | 90.64 | 92.62 | 89.34 | 88.64 | 88.64 |
| KNN        | 89.66 | 89.65 | 89.67 | 89.67 | 89.63 | 89.55 | 89.71 | 89.71 | 90.69 | 89.65 | 88.72 | 89.17 |
| RF         | 98.69 | 96.81 | 97.72 | 98.72 | 96.69 | 95.81 | 95.56 | 91.56 | 94.68 | 90.78 | 91.57 | 89.57 |
| SVM        | 85.13 | 85.4  | 84.86 | 84.89 | 85.14 | 85.39 | 84.89 | 84.91 | 94.15 | 85.4  | 84.9   | 84.9   |

Fig. 3. Bar plot with frequencies of eight amino acids classes in: (a) COVID-19 and (b) non-COVID-19 samples for the NGDC dataset.
number of features to only seven of them. The performance of the AAPred model changed when applying different training and testing methods. As reported in Table 5, the KNN method outperforms the other classifiers using 80% of the data for training and 20% for testing. On the contrary, when using a 10-fold cross-validation, the RF method outperforms the other classifiers. This indicates that splitting the dataset into training and testing affects the prediction of the viruses, especially to the KNN classifier that depends on the neighbors. Thus, applying the RF classifier using a 10-fold cross-validation makes the prediction results more reliable.

4. Discussion and conclusions, limitations, and future research

4.1. Discussion

To verify the performance of the proposed AAPred model, we compare this model with the performance of two different models: LASSO [14] and AAC [15] for feature selection. With LASSO, the selected number of features is 35 (eighteen radiological semantic features and seventeen clinical features). Using both types of features, the performance measures are 95.9%, 96.1%, and 95.7%, in terms of accuracy, sensitivity, and specificity, respectively. When the AAC is utilized, the 41 features of the protein sequence are reduced to twenty important features. We employ the RF classifier such as in [15] to evaluate the models. The achieved accuracy is 98.18% using 10-cross validation.

Table 7 summarizes the performance of the AAPred model compared to the method presented in [14]. The proposed AAPred model outperforms Chen et al. [14]'s method. It is evaluated employing six classifiers and four performance measures, reaching a higher accuracy of 98.69% in 6.43 seconds compared to 95.9% with Chen et al. [14]'s method in 5.62 seconds. In addition, our model reduces the number of features to seven of them compared to the 35 features presented in [14]. The second comparison is conducted with the spike protein dataset. The AAPred model was compared with Qiang et al. [15]'s model that has an
accuracy of 98%. Our AAPred model outperforms Qiang et al. [15]'s model, with an accuracy of 99.01% and seven features in 3.58 seconds compared to Qiang et al. [15] that has twenty features as shown in Table 8. The results report that SARS-CoV-2 has similar infection cycles and sequences in some regions of the SARS-CoV proteins that indicate the affectedness of vaccines on SARS-CoV-2.

4.2. Conclusions, limitations, and future research

In this paper, a novel model, named AAPred in short, was proposed to predict coronavirus types. The AAPred model extracts features from the protein sequences by replacing the amino acid characters in each sequence with the normalized frequencies of the eight amino acid classes. We used the information gain, analysis of variance, and chi-square methods to reduce the number of features and select the most significant ones. The model reduces the features to only seven of them. The AAPred model was evaluated employing six classifiers: Bagging ensemble, decision trees, gradient boosting, k-nearest neighbors, random forest, and support vector machine. Experimental results showed that the AAPred model can differentiate between COVID-19 and non-COVID-19 viruses’ sequences. This differentiation is supported by the average accuracy of the model of 98.69%, a precision of 98.72%, a sensitivity of 96.81%, and a specificity of 97.72%. The maximum accuracy was reached utilizing the random forest classifier and the information gain method.

Table 7
AAPred model performance compared to the method proposed in Ref. [14]. Source: The authors.

| Method         | Number of features | Acc   | Sens  | Spec  | Prec  | Computing time (in seconds) |
|----------------|--------------------|-------|-------|-------|-------|-----------------------------|
| IG             | 7                  | 96.89 | 96.81 | 97.72 | 98.72 | 6.43                        |
| ANOVA          | 7                  | 96.69 | 95.81 | 95.56 | 91.56 | 1.17                        |
| $\chi^2$       | 7                  | 94.68 | 90.78 | 91.57 | 89.57 | 1.12                        |
| Chen et al. [14]| 35                 | 95.90 | 96.10 | 95.70 | 98.60 | 5.62                        |

Fig. 6. Proposed model performance based on the $\chi^2$ method for the NGDC dataset using a 10-fold cross-validation.

Fig. 7. Proposed model performance using the spike protein dataset.
In addition, the 10-Fold cross-validation method was used. Moreover, after studying the polarity and dipole values of the protein sequence of the coronavirus, we conclude that the COVID-19 protein sequence has a high level of amino acids of the second class that contains nonpolar amino acids, compared to the non-COVID-19 protein sequence.

We choose machine learning tools rather than deep learning methods because machine learning techniques can reach excellent results in our dataset with less computational burden and they are simpler to use. Indeed, it has only seven features and is not complicated as an image classification, for example. In any case, we applied a generative adversarial network and resulted in an accuracy of around 98% as the random forest classifier.

A limitation of the proposed model is that it relies only on protein sequence analysis and does not consider other aspects such as the protein structure of the virus or its DNA sequence. As further research directions, we aim to apply the same model using different feature extraction methods according to the sequence and the structure of the proteins to obtain more detailed biological information about the virus behavior and its infection cycle. Other classification methods will also be explored in future studies such as principal components analysis and its new derivations, including supervised and unsupervised approaches, as well as functional data analysis, partial least squares structures, and other recent methodologies [36–41,46–49].

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at https://ngdc.cncb.ac.cn/gwh/browse/virus/coronavirus.

Author statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript which resulted in this improved version. The research of V. Leiva was partially funded by FONDICYT, project grant number 1200525 from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science and Technology, Knowledge, and Innovation.

References

[1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, The species severe acute respiratory syndrome-related coronavirus classifying 2019-CoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (2020) 536–544.
[2] P. Zhou, X. Yang, X. Wang, B. Hu, L. Zhang, W. Zhang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.
[3] N. Jerez-Lillo, B. Lagos-Alvarez, J. Muñoz-Gutierrez, J.J. Figureeroa-Zúñiga, V. Leiva, A statistical analysis for the epidemiological surveillance of COVID-19 in Chile, Signa Vitae 18 (2022) 19–30, https://doi.org/10.22514/sv.2021.130, in press.
[4] C. Martín-Barreiro, J.A. Ramírez-Figueroa, X. Cabezas, V. Leiva, M.P. Galindo-Villarría, Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data, Sensors 21 (2021) 4094.
[5] World Health Organization, WHO Announces COVID-19 Outbreak a Pandemic, 09 March 2020. Available from: http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.
[6] A.A. Agranovsky, Structure and expression of large (+)RNA genomes of viruses of higher eukaryotes, Biochemistry 86 (2021) 248–261.
[7] International Committee on Taxonomy of Viruses. Available from: http:// ICTVonline.org/virusTaxonomy.asp (accessed on 24 October 2021).
[8] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (2016) 237–261.
[9] L.S.K. J., T. Kk, C. Ve, W. Py, C. Kk, Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease, Clin. Microbiol. Rev. 28 (2015) 465–522.
[10] Y. Wu, B. Peng, X. Huang, X. Ding, P. Niu Wang, Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China, Cell Host Microbe 27 (2020) 325–328.
[11] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (2016) 237–261.
[12] M.J. Adams, E.J. Letko, A.M. King, B. Harrach, R.L. Harrison, N.J. Knowles, et al., Ratification vote on taxonomic proposals to the international committee on taxonomy of viruses (2016), Arch. Virol. 161 (2016) 2921–2949.
[13] V. Menachery, B. Youn, K. Debbink, S. Agnihotri, L. Gralinski, J. Plante, A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat. Med. 21 (2015) 1508–1513.
[14] X. Chen, Y. Tang, Y. Mo, S. Li, D. Lin, Z. Yang, et al., A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study, Eur. Radiol. 30 (2020) 4893–4902.
[15] X. Qiang, P. Xu, G. Fang, W. Liu, Z. Kou, Using the spike protein feature to predict infection risk and monitor the evolutionary dynamics of coronavirus, Infect. Dis. Poverty 9 (2020) 35.
[16] W.M. Zhao, S.H. Song, M.L. Chen, et al., The 2019 novel coronavirus species, Ciuchan 42 (2020) 212–221.
[17] J. Gomes, V. Barbosa, M. Santana, J. Bandeira, IONOS: an intelligent tool to support diagnosis of COVID-19 by texture analysis of x-ray images, medRxiv (29 May 2020).
[18] N. Bustos, M. Tello, G. Droppelmann, N. García, F. Feijoo, V. Leiva, Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases, Signa Vitae 18 (2022) 23–33.
[19] P. Vokovs, A. Krtatzel, S. Steiner, H. Stalder, V. Thiel, Coronavirus biology and replication: implications for SARS-CoV-2, Nat. Rev. Microbiol. 19 (2020) 155–170.
[20] Braga-Neto, Fundamentals of Pattern Recognition and Machine Learning, Springer, New York, 2020.
[21] C.A. Palacios, J.A. Reyes-Suarez, L.A. Beazzotti, V. Leiva, C. Marchant, Knowledge discovery for higher education student retention based on data mining: machine learning algorithms and case study in Chile, Entropy 23 (2021) 485.
[22] The Ddbj/ENA/GenBank Feature Table Definition. International Nucleotide Sequence Database Collaboration. Available from: https://www.insdc.org/docs ments/feature-table (accessed on 24 October 2021).
[23] M. Zhang, Q. Su, Y. Lu, M. Zhao, B. Niu, Application of machine learning approaches for protein-protein interactions prediction, Med. Chem. 13 (2017) 506–514.
[24] S. Asim, A. Shah, H. Shahbib, S. Rehman, M. Waqas, A comparative study of feature selection approaches: 2016–2020, Int. J. Sci. Eng. Res. 11 (2020) 469.
[25] S. Lekovits, L. Lekovits, Gaborf feature selection based on information gain, Process Eng. 181 (2017) 892–898.
[26] F. Ardalan, Case study using analysis of variance to determine groups’ variations, MATEC Web Conf. 126 (2017), 04008.
[27] E. Benhamou, V. Melot, Seven proofs of the Pearson chi-squared independence test and its graphical interpretation, SSRN (2010), http://doi.org/10.2139/ssrn.3239829.
[28] A. Torsello, L. Ronsi, M. Peilolo, B. Biggio, A. Robles-Kelly, Structural, Syntactic, and Statistical Pattern Recognition, Springer, New York, USA, 2021.
[29] W. Alkady, W. Gad, K. Bahnawy, Swarm intelligence optimization for feature selection of biomolecules, in: 2019 14th International Conference on Computer Engineering and Systems (ICCES), 2019, pp. 380–385, http://doi.org/10.1109/ ICCES48960.2019.9068178.
[30] D. Xiuquan, L. Xinrui, H. Zhang, Y. Zhang, Prediction of protein-protein interaction by metasample-based sparse representation, Math. Probl. Eng. (2015) 852526.
[31] J. Phillip, R. Keith, I.J. Probert, R. Jonathan, J. Stewart, J. Chris, Density functional theory in the solid-state, Phil. Trans. R. Soc 372 (2014) 20130270.
[32] N. Xiao, D.S. Cao, M.F. Zhu, Q.S. Xu, protor/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, Bioinformatics 31 (2015) 1857–1859.
X. Wang, Y. Wu, R. Wang, Y. Wei, Y. Gui, A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences, PLoS ONE 14 (2019) e0217312.

M.A. Cano Lengua, E.A. Papa Quiroz, A systematic literature review on support vector machines applied to Classification, IEEE Eng. Int. Res. Conferen. (EIRCON) (2020) 1–4, https://doi.org/10.1109/EIRCON51178.2020.9254028.

NCBI coronavirus datasets. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType=Protein (accessed on 24 October 2021).

J.A. Ramirez-Figueroa, C. Martin-Barreiro, A.B. Nieto, V. Leiva, M.P. Galindo-Villardón, A new principal component analysis by particle swarm optimization with an environmental application for data science, Stoch. Environ. Res. Risk Assess. 35 (2021) 1969–1984.

J.L. Martinez, V. Leiva, Estimating the covariance matrix of the coefficient estimator in multivariate partial least squares regression with chemical applications, Chemometr. Intell. Lab. Syst. 214 (2021) 104328.

T.L. Campos, P.K. Korhonen, N.D. Young, Gradient boosting machines: a tutorial, Front. Neurorob. 7 (2013) 21.

M. Cascella, M. Rajnik, A. Aleem, et al. Features, evaluation, and treatment of Coronavirus (COVID-19) [Updated 2021 Sep 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 January. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554776/.

A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses, Sensors 21, 8020, https://doi.org/10.3390/s21238020.

A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Abnormality detection and failure prediction using explainable bayesian deep learning: methodology and case study with industrial data, Mathematics 10 (2022) 554.

M. Huerta, V. Leiva, S. Liu, M. Rodriguez, D. Villegas, On a partial least squares regression model for asymmetric data with a chemical application in mining, Chemometr. Intell. Lab. Syst. 190 (2019) 55–68.

L. Ma, Y. Zhang, V. Leiva, S. Liu, T. Ma, A new clustering algorithm based on a radar scanning strategy with applications to machine learning data, Expert Syst. Appl. 191 (2022) 116143.

E. Mahdi, V. Leiva, S. Mara’Beh, C. Martin-Barreiro, A new approach to predicting cryptocurrency returns based on the gold prices with support vector machines during the COVID-19 pandemic using sensor-related data, Sensors 21 (2021) 6319.