Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Identification and classification of coronavirus genomic signals based on linear predictive coding and machine learning methods

Amin Khodaei [a,*], Parvaneh Shams [b], Hadi Sharifi [a], Behzad Mozaffari-Tazehkand [a]

[a] *Faculty of Electrical & Computer Engineering of University of Tabriz, 29 Bahman Blvd, Tabriz, Iran*
[b] *Computer Engineering Department, Istanbul Aydin University, Turkey*

ARTICLE INFO

Keywords:
Linear predictive coding
Corona
DNA Sequence
Support vector machine
Signal processing
Machine learning

ABSTRACT

Corona disease has become one of the problems and challenges of humankind over the past two years. One of the problems that existed from the first days of this epidemic was clinical symptoms similar to other infectious viruses such as colds and influenza. Therefore, diagnosis of this disease and its coping and treatment approaches is also been difficult. In this study, Attempts has been made to investigate the origin of this disease and the genetic structure of the virus leading to it. For this purpose, signal processing and linear predictive coding approaches were used which are widely used in data compression. A pattern recognition model was presented for the detection and separation of covid samples from the influenza virus case studies. This model, which was based on support vector machine classifier, was tested successfully on several datasets collected from different countries. The obtained results performed on all datasets by more than 98% accuracy. The proposed model, in addition to its good performance accuracy, can be a step forward in quantifying and digitizing medical big data information.

1. Introduction

The coronavirus pandemic, which has plagued the world, has become more prevalent than other pervasive diseases in history. This is due to the high incidence of this disease and its high transmission rate. According to the World Health Organization statistics, in the last two years, tens of thousands of people die every day due to this disease [1]. Of course, this is not the world's first pandemic virus in a century. Previously, influenza, Ebola, Mers, and SARS, were among the viral epidemics that have engulfed human societies [2].

One of the main problems and challenges in the diagnosis of this disease is its seemingly dodgy primary symptoms. The initial clinical manifestations of this virus are similar to common infectious viruses such as colds and influenza diseases. There are several things that have been done about symptoms of patients with this disease, including [3]. In this study, the symptoms of patients in different countries have been considered and the effect of various characteristics and parameters has been investigated. Also, compared with pervasive diseases and previous versions of this virus such as MERS and SARS. In [4], the issue has also been investigated and compared with a series of well-known influenza viruses such as H1N1, H7N9, H5N1 types of this virus.

In medical science, diagnosis phase of the disease may be considered as the most important phase among the existing phases. Because in addition to the advantage of accelerating the patient's recovery with proper and timely diagnosis, the results of this phase can also be used in other phases. Diagnostic approaches of this disease can be classified into 3 general sections [5]. One of the important reasons for evaluating the diagnostic methods of this disease is the reporting of large amounts of the diagnostic tests wrong results [6]. Many claims and researches have been conducted regarding the origin of corona and the main cause of the emergence of this virus, including Y. Zhang's research, noted [7].

It is certain that the origin of corona disease is a virus that causes genetic changes in the nature of the cell by causing genetic changes. Analysis of coronavirus has not started two years ago. In articles such as [8910], the internal structure of this virus has been investigated from a chemical and biomolecular point of view. In another study in 2003, the spread of this disease has been noted in one of the French cities [11]. Other scientists have also investigated the virus. So that in 2010, P. Woo et al examined the genome of the virus from a bioinformatics perspective and in intelligent ways [12].

Genetic components analysis is one of the main tasks in the analysis of genetic disease. For example finding the host of the virus from other organisms or finding the similarity of the human virus with other living organisms can be valuable for future research, which was tested on

[*] Corresponding author.
E-mail addresses: amin.khodaei.13@gmail.com (A. Khodaei), Parvanehshams@aydin.edu.tr (P. Shams), mozaffary@tabrizu.ac.ir (B. Mozaffari-Tazehkand).

https://doi.org/10.1016/j.bspc.2022.104192
Received 19 May 2022; Received in revised form 12 August 2022; Accepted 10 September 2022
Available online 23 September 2022

about 2625 viruses and 429 corresponding hosts in [13]. In this research, the data have been compared from the statistical perspective, and there is no machine learning evaluation procedure. Research on the mutations of this virus has also been very important. For example, in [14], the severity of the virus based on mutations analyzed on 1594 viral genomes. They used an algorithm with about 94 % accuracy by using 22 mutations. Some research has focused on one or more specific genes. One of them is the Y. Cao and et al. research, that studied the activation of this gene on different parts of the world samples [15]. This article is not based on machine learning scheme and a comparative approach is used in this paper, too.

S. Yan and colleagues focused on the prediction of the position of nucleotide mutations on amino acid sequences [16]. The basis of the proposed research approach is the use of feed-forward backpropagation neural network. In addition to the covid virus, they used Influenza A virus data to aid in mutations in the learning phase. They tested their algorithm on influenza virus variants by more than 95 % correct rate accuracy and made comparisons based on them, but there is no comparison and guarantee for the success of the algorithm on coronavirus samples. Machine learning methods proposed in, K. Kuzmin et al. research, to analyze different variants of the virus, in the form of phylogenetic tree [17]. They compared several machine learning methods for this research and SVM and decision tree approaches classified the feature space with 95–99 % accuracy, on 1238 samples.

M. El-dosuky et al research [18] main idea is the deep learning approach and CNN method to categorize the data into two categories. This research, which has been tested on about 500 genomic samples, and its performance accuracy was 99 %. One of the main challenges of this research is the use of algorithms with high time complexity. M. S. Nawaz et al, analyzed genomic data [19], by sequential pattern mining, which leads to the revelation of a series of important features. They focused on frequent pattern of genomic sequences and the mutations of the mentioned sequences, by machine learning methods. S. Ma and his colleagues' research is another similar study that has been done in the field of genomics on two coronaviruses and influenza [20]. A large number of attributes were compared in this research, but the number of records was not sufficient. They succeeded in separating these two viruses with about 88 % accuracy.

In [21], the effect of a specific gene has been studied on influenza data. A. Tsonis et al, focuses on the corona and influenza viruses comparison [22]. The genomic sequences are mapped in the form of signals, and wavelet transform is used in the feature extraction phase. Finally, a graph is drawn to measure the similarity of the results of different viruses, which are based on peak point values. One of the applications of deep learning is in the analysis of genomic data, which has also been used for corona disease. In the study of A. Lopez-Rincon et al., Deep learning approaches have been used to diagnose the sequences associated with this disease [23].

In another study, A. Lopez-Rincon et al. used in-depth learning approaches to diagnose disease-related sequences [24]. The use of deep

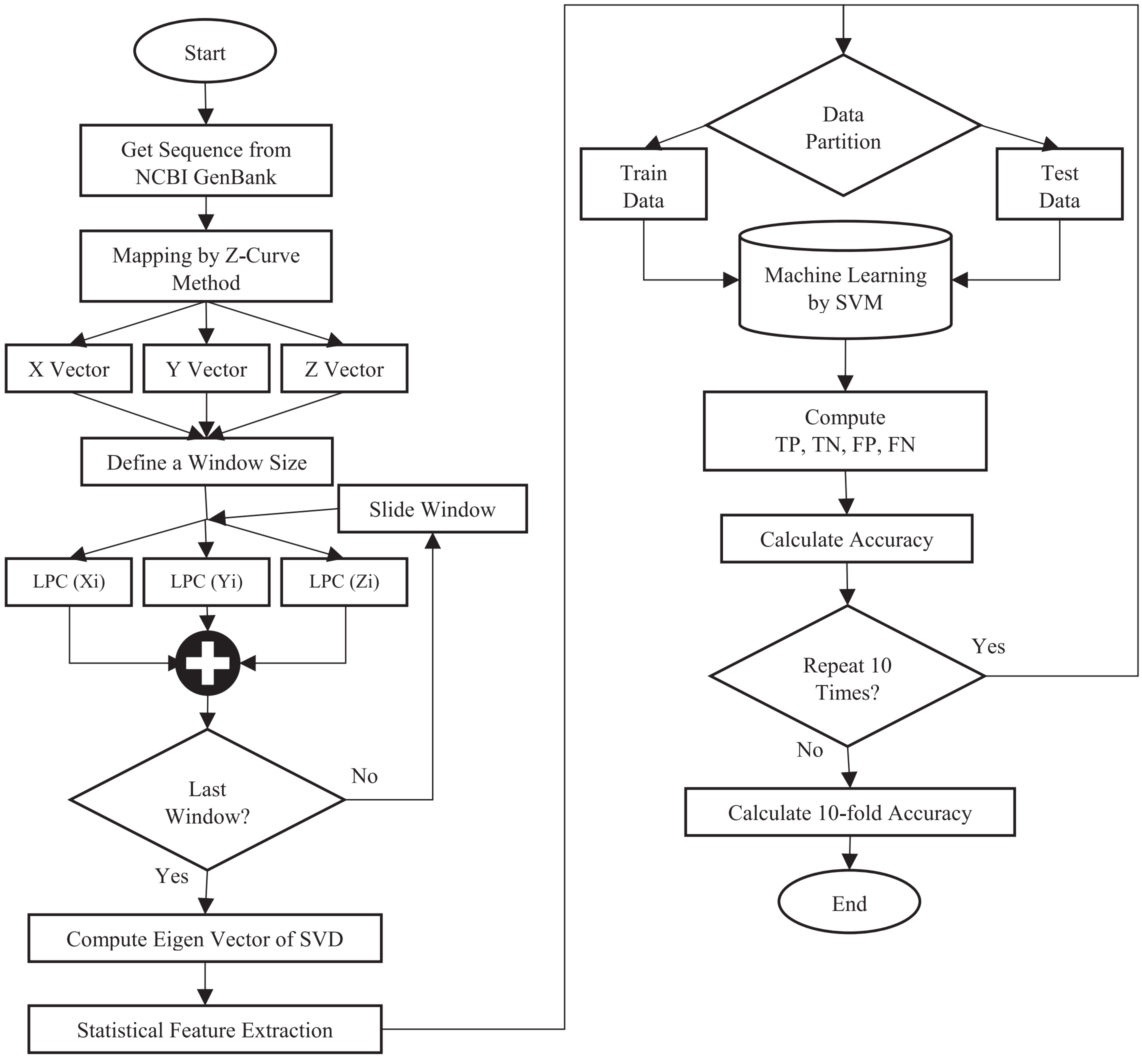


Fig. 1. The proposed approach flowchart.

learning approaches on genomic data has been used in other studies such as [25]. In this research, a new method has been used to display and encode genomic sequences. Another application of deep learning is in the R. Khan et al. research, who proposed an algorithm for the prediction of genetic mutations in the form of nucleotides and amino acids [26]. The selected data in the form of complete genome is selected from different countries.

In another study by H. Arslan et al., the cgp-island genetic component was used as a distinguishing feature of several groups of genomic sequences [27]. A database of Chinese Wuhan genome data has been used as case studies. S. Naeem et al., used signal processing approaches to classify the coronavirus genomic data [28]. In this study, the genomic form of SARS and MERS viruses compared with the new covid-19 virus. Another signal processing-based study to identify the type of corona was presented by T. Paul et al. [29]. The basis of this research is using the wavelet transform and the use of a series of music signals features. The proposed approach is tested on the amino acid sequences.

Signal processing algorithms on genomic data have already been used in other coronavirus researches. For example, M. Cohen-Mcfarlane et al have created a new database of Crohn's coughs [30]. Other work has been done on the analysis of cough or voice signals in people such as [31] and [32]. It should be noted that, signal processing and machine learning approaches has already successful results on other genetic diseases genomic data. For example S. S. Roy research [33], which have been tested on relatively large number of cancerous sequences. Among the newly published articles in this scope, we can mention the [34–40] researches. These studies are not limited to cancer disease and in the research of D. Dalwadi et al, these methods have been used on AIDS disease [41]. In addition to signal processing approaches, computational, statistical and electronic approaches have also been tested in the classification of DNA and RNA biological sequences [42–46].

This review has shown that in this limited time, this virus has been studied from different perspectives. The high length, volume and variety of this virus case studies make it difficult to analysis of this type of data. For example, in some of the presented methods, a small number of samples have been tested. Therefore, the classification phase or evaluation and comparison of the results of these researches have not been done appropriately. Some of the presented methods also have a high complexity, which makes it challenging to generalize them to more data. Some of these researches have been done without mapping to a signal format, and a few articles have been presented based on signal processing methods. The Proposed model have been used on influenzas and coronavirus diseases' genomic data in the form of nucleotide sequences by DSP-based techniques. The following is a brief description of the proposed research contributions:

- data-independent (in terms of length and type of genomic sequences) performance of the utilized methods, facilitate the development of the model.
- Feature extraction results of the corresponding signals has led to separation of the coronavirus from the influenza virus with high accuracy.
- The proposed approach has been tested on a large number of coronavirus genomics datasets from different geographical areas and variants (not specific country, area or dataset).
- Using a new combinatorial compression approach on the genomic sequences decreases the time complexity.
- Dealing with the randomness of the results by adopting k-fold machine learning evaluation technique.
- Successful result (with more than 0.98 accuracy) of the proposed method on a much larger number of data (107000), compare with previous researches.
- The low complexity and memory consumption of applied methods is important in genomic sequences analysis research. This case is also considered in this study and there is no simulation or electrical modeling approach.

In the following parts, the proposed approach will be introduced step by step.

2. Material and methods

The proposed approach is presented in the form of a pattern recognition. The proposed approach aims to identify and differentiate samples of covid-19 virus from other viruses with similar clinical symptoms. Influenza virus samples have been selected for this purpose. The proposed approach flowchart presented in Fig. 1, in two general phases. The first phase on the left side is the feature extraction and feature selection stages and the right side of the flowchart is related to the machine learning and evaluation of the proposed model.

As shown in Fig. 1, the research has been proposed on DNA nucleotide sequences. DNA sequences are known as the principal source of genetic information within the cell. These sequences are converted to RNA and protein sequences respectively during organized steps. Any change at the nucleotide level, which can take many forms, are known as genetic mutations. These genetic mutations can also change the genetic structure of other cells [47]. The purpose of this study is to identify and determine a series of distinctive features that distinguish between different samples, that will be described in following sections.

2.1. Linear predictive coding (LPC)

The first step in analyzing genomic information using DSP approaches is converting the sequences into a signal format. For this purpose, several methods have been presented. A. Yang et al have divided the existing methods into several general sections [47]. In this research, the Z-Curve method has been used for nucleotide to signal mapping. The basis of this method is the conversion of nucleotide sequences into three signal vectors based on four vectors specific to each type of nucleotide. In other words, in this method, four vectors are formed for each type of nucleotide, and then based on Equation (1), the corresponding final vector is calculated [48].

$$\begin{aligned} x_i &= (A_i + G_i) - (C_i + T_i) \\ y_i &= (A_i + C_i) - (G_i + T_i) \\ z_i &= (A_i + T_i) - (C_i + G_i) \end{aligned} \tag{1}$$

In equation (1), there are three relations from top to bottom, which show how to calculate the three vectors x, y, z. To calculate these vectors, it is first necessary to compute 4 separate vectors for each type of nucleotide. Then, by using the four mentioned vectors (consists of A, C, G, T), the x, y and z vectors are calculated. According to these relations, the mentioned vectors are calculated for each sample.

After this step, linear predictive coding model performed on signal data to feature extraction. The idea of this method is based on the principle that the current speech sample can be made as a linear combination of the previous samples. In other words, LPC method model the future signals according to the previous samples. The main relation of this method is based on Eq. (2). Based on this formula, the goal is finding the prediction coefficients [49].

$$S[n] = \sum_{k=1}^{P} a_k S[n-k] = a_1 S[n-1] + a_2 S[n-2] + \cdots + a_p S[n-p] \tag{2}$$

In Eq. (2), p refers to the degree of prediction. This equation can be considered as an all-pole filter, which can predict the instantaneous instance of signal S [n] with the linear composition p of the previous s examples. Prediction coefficients is also specified as a_i, which is known as linear prediction coefficients. Finally, from this equation, each sample is estimated from the coefficients and previous samples. There are several methods for computing linear prediction coefficients, including autocorrelation and covariance methods [49,50].

In this research, the LPC model was applied on the corresponding signal vectors of the nucleotide sequences. In other words, the LPC

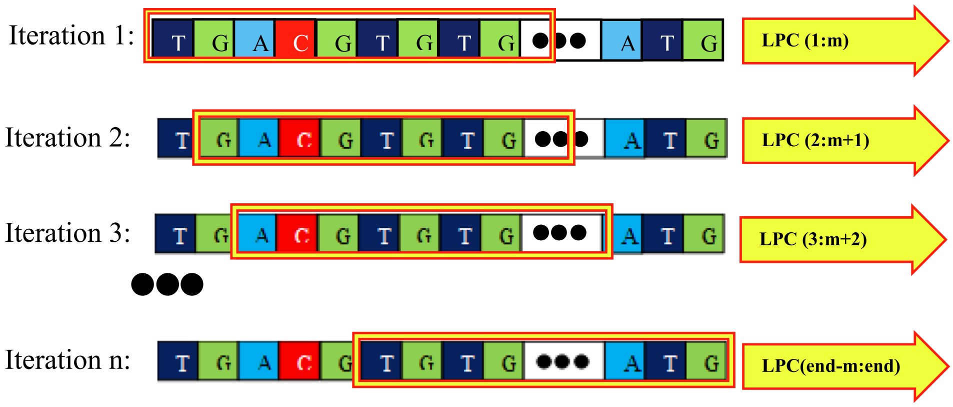

Iteration 1:
T G A C G T G T G ●●● A T G
LPC (1:m)
Iteration 2:
T G A C G T G T G ●●● A T G
LPC (2:m+1)
Iteration 3:
T G A C G T G T G ●●● A T G
LPC (3:m+2)
●●●
Iteration n:
T G A C G T G T G ●●● A T G
LPC(end-m:end)

Fig. 2. Sliding window technique procedure on the corresponding signal vectors of nucleotide sequences.

model is applied to the three-dimensional vectors obtained from the Z-Curve mapping method. Due to the relatively long length of the obtained vector signals, they are analyzed at specific intervals (equal to the specified window size). So that from the beginning of the signal vector of this model is applied on a part of the signal vector of nucleotides and the obtained results are stored. This process is repeated until it finally reaches the last window and its output is calculated and the results are stored in the form of a matrix. Fig. 2 shows an example of windowing technique on the corresponding vectors of a nucleotide sequence.

As shown in Fig. 2, a part of the sequence is selected, which is marked with a rectangular box. First, the linear prediction method is applied only to the corresponding signals of this window. The output is saved as a vector and then the window is slid. The window has the same size as before, but it refers to a different range of the signal. It should be noted that, the overlap between two consecutive windows is taken into account to prevent the loss of border information between two adjacent windows. Again, the result is stored in the form of a vector and the same process continues until the end of the signal. If the length of each output vector is assumed to be m and this process is repeated n times, finally the output is an n*m matrix. Considering the length of the signals) which is caused by the long genomic sequences), the use of the windowing technique reveals more details and prevents the randomness of the final result.

2.2. Singular value decomposition (SVD)

Applying the sliding window LPC model on the corresponding vectors of nucleotide sequences yields several output vectors. The covariance matrix is used to manage the feature space of the created matrix. The covariance matrix, which will be a square matrix, is calculated from the Eq. (3) formulas [51].

$$\begin{gathered}X=[x_1,x_2,x_3,\cdots,x_m],\,C_{xx}=E\{X.X^T\}\\ D=diag[\lambda_1,\lambda_2,\cdots,\lambda_m]\,,\,V=[v_1,v_2,\cdots,v_m]\end{gathered} \tag{3}$$

In the above-mentioned equation, “D” is a square matrix, whose diameter is λ_i the eigen-values of the matrix, and V is the eigen-vector of the matrix. In this research, two statistical features extracted as the final distinguishing criteria. The maximum, standard deviation of the mentioned vectors values, which is calculated based on the Eq. (4):

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}which\mu=\sum_{i=1}^{n}x_i \tag{4}$$

In Eq. (4), x_i and μ denoted the mean and standard deviation quantities, respectively.

Table 1
The common kernel functions of the SVM learning model.

Kernel Name	Equation
Linear	$x_i^T.x_j$
Polynomial	$(x^Tx_i+1)^P$
Sigmoid	$tanh(x^Tx_i+1)$
RBF	$e^{(-\frac{1}{2\gamma^2}\|x-x_i\|^2)}$
MLP	$\tanh(\beta_0x^Tx_i+\beta_1)$

2.3. Support vector machine (SVM) for classification

The aim of this study is to differentiate and classify the samples into two classes. One of the efficient methods in the field of machine learning for the classification is the support vector machine (SVM). The main goal of the SVM approach is to find a hyperplane that creates the maximum separation margin between the two existing categories. In the simplest possible way, if the feature space is assumed as {(x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n)}, the goal is to find a line or hyperplane that separates yi = 1 from yi = -1. This hyperplane can be shown with Eq. (5) as follows [52].

$$W.x+b=0 \tag{5}$$

In this regard, W is the vector of weights and b is a number for adjusting the weight in different conditions. One of the advantages of this method is the ability to generalize to multidimensional space at a relatively good speed. One of the capabilities of this model that improves the classification performance, is kernel function concept. Kernel functions model the non-linear feature spaces as well as linear dimensions. Table 1 lists the known kernel functions with their mathematical relations.

In these formulas, x is the feature vector of the existing problem, which is mapped to a multidimensional space with the mentioned functions.

3. Results

In this section, the experiments are presented and described with the obtained results. In this research, 107,000 human nucleotides sequences in FASTA format extracted from the Genbank database of the NCBI website [53]. 47,200 of these numbers are coronavirus data, and the rest related to the influenza virus sequences. The sequences length maximum size is about 30,000 base pairs, and had no ambiguous nucleotides. Due to the genetic differences of various geographical regions, experiments were performed on several datasets from 20 different countries separately. The mentioned datasets have been collected from some countries with higher mortality rates (such as USA, India, Brazil, Russia, e.g.,). It was also tested on countries, where different mutations and variants of

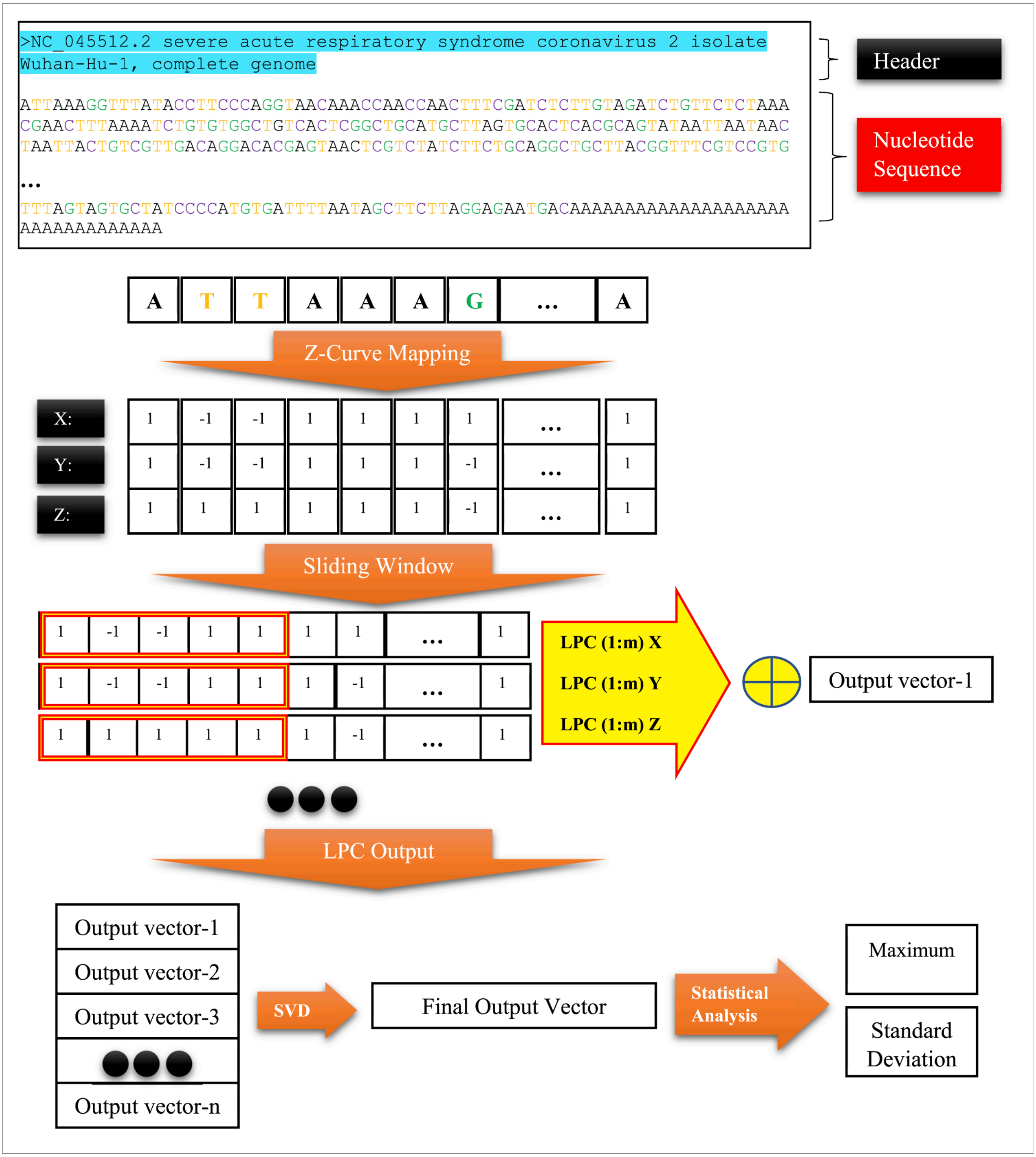


Fig. 3. Procedure of a genomic sequence feature extraction steps.

the virus were observed during this period (such as China, India, United Kingdom, e.g.,).

Fig. 3 shows an example of the data format and the steps of the proposed approach are depicted on it. Fig. 3 first shows an example of genomic raw data in the Fasta format. Each sample has two parts: "Header" and "Nucleotide sequence". The "Header" section gives information about the mentioned sample, which does not affect the analysis. The "Nucleotide sequence" section contains genomic data information in the form of a sequence of A, C, G, and T nucleotides. The proposed method is applied on this part of the data according to the flowchart of Fig. 1 step by step on each data.

As shown at the bottom of the raw sample, each type of the nucleotide is mapped to the corresponding three-dimensional vector using the Z-curve method. The sliding window LPC method is applied to all three vectors. The sum of these three vectors is stored in the form of a vector as the result of one repetition of the algorithm. The sliding window algorithm of the LPC continues and repeats until the end of the signal. Finally, due to the repetition of this process, the output of the LPC method can be considered as a matrix. The SVD technique converted it as a vector and statistical features are extracted from that. This procedure is repeated on all the data and the mentioned two features are extracted from each data.

Table 2
Comparison of obtained results by changing various parameters.

no	Order	Window	Accuracy
1	40	60	0.951
2	40	90	0.952
3	40	150	0.943
4	20	40	0.961
5	20	90	0.964
6	20	150	**0.977**
7	20	200	0.961
8	10	40	0.943
9	10	90	0.950
10	10	150	0.961
11	10	200	0.961

3.1. *Evaluation methods*

The predicted labels and the real labels can be used as criteria for evaluating the supervised learning model. In this case, 4 states occur which are known as true positive (TP), true negative (TN), false positive (FP), false negative (FN). Some other criteria have been defined in the field of machine learning based on the four mentioned components. One of the most famous criteria is Accuracy, which is computed based on the correct prediction value of the mentioned labels. The Eq. (6) indicates how to calculate this index in terms of the above four criteria [51].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

Also, the 10-fold method has been used to validate the results.

3.2. *Experimental results*

Experimental tests were performed in the MATLAB 2018 software. Several experiments were performed to analyze the parameters of the used LPC model. For this purpose, the experiment (with the mentioned parameters) was performed on the China dataset samples with 1050 samples, of which 344 were coronavirus samples and the rest were non-

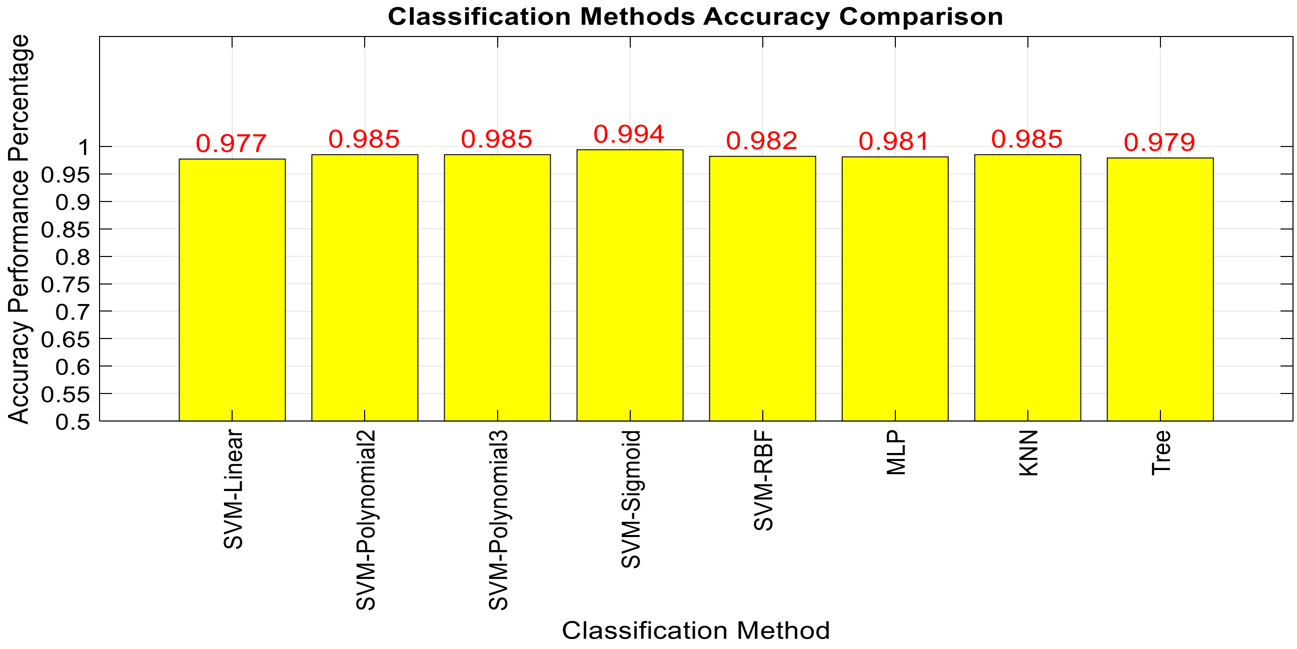


Fig. 4. 10-fold accuracy of different machine learning models comparison.

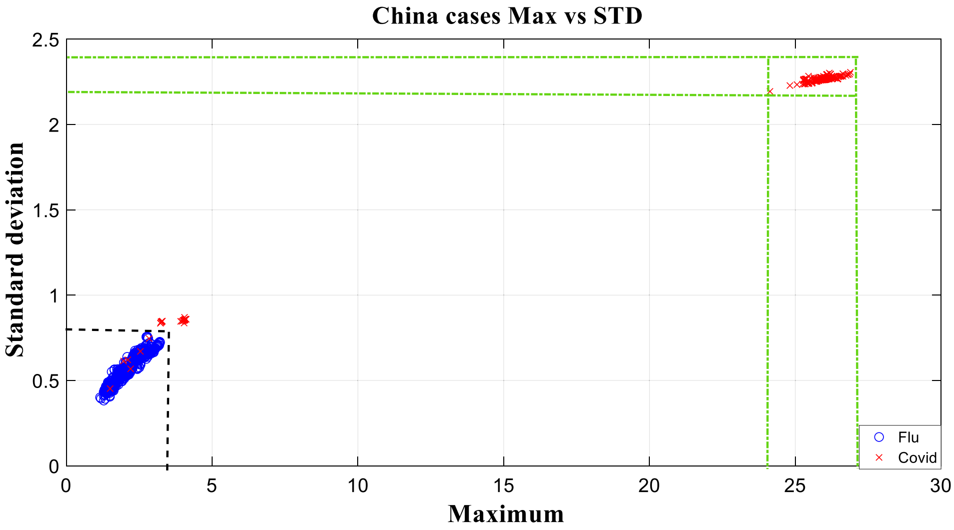
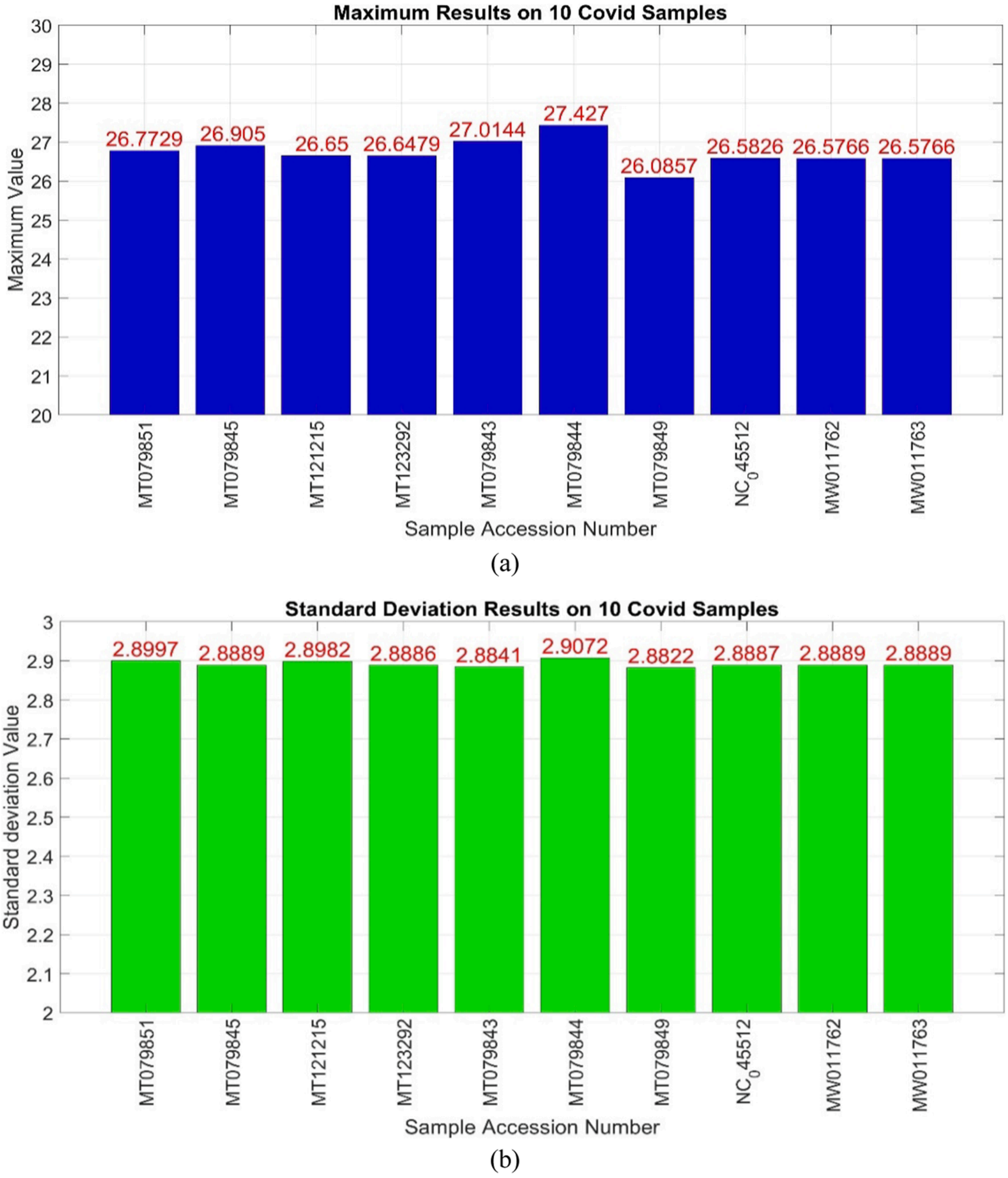


Fig. 5. Comparison of the maximum vs standard deviation values on China dataset.



(a)

(b)

Fig. 6. The result of two features on some covid samples (a) Maximum feature (b) Standard deviation feature.

corona cases. Table 2 lists the results some experiments with different coefficient order and different window length values. The last column from the left side indicates the obtained accuracy. It should be noted that, a common classification method (linear SVM) was performed.

The results of Table 2 show that the LPC model order 20 and 150 window lengths had better performance results. One of the important steps in pattern recognition studies is choosing a proper machine learning method. In this research, several machine learning approaches were tested, and the results are depicted in Fig. 4. In this picture, the horizontal axis represents the name of machine learning methods. The vertical axis also indicates the obtained accuracy of the 10-fold method in evaluating the classification accuracy.

Fig. 4 shows that the SVM model with nonlinear kernel functions have better performance than other models. It is obvious that the experiments were successful on about 99 % of data, by using Sigmoid kernel. The values of two mentioned distinguishing peak and standard deviation features of this dataset, are illustrated in Fig. 5. The horizontal and vertical axis of this diagram represents the peak and standard deviation of vector values, respectively. The covid virus samples are specified by red and cross symbols and flu samples are distinguished by blue and circle symbols.

In Fig. 5, it is clear that for the non-covid virus class, the first feature (maximum) is limited in the range of 1–4 and this feature is limited in the range of 24–28 for the coronavirus class. In the case of the second feature (standard deviation), the non-covid virus class ranged in 0.4–0.8 and the coronavirus class results is limited in the range of 2.1–2.4. It is necessary to mention this point again that due to the relatively large number of data, the results of some samples have been mixed., and this makes the data less visible.

Two mentioned extracted features result on 10 samples from this

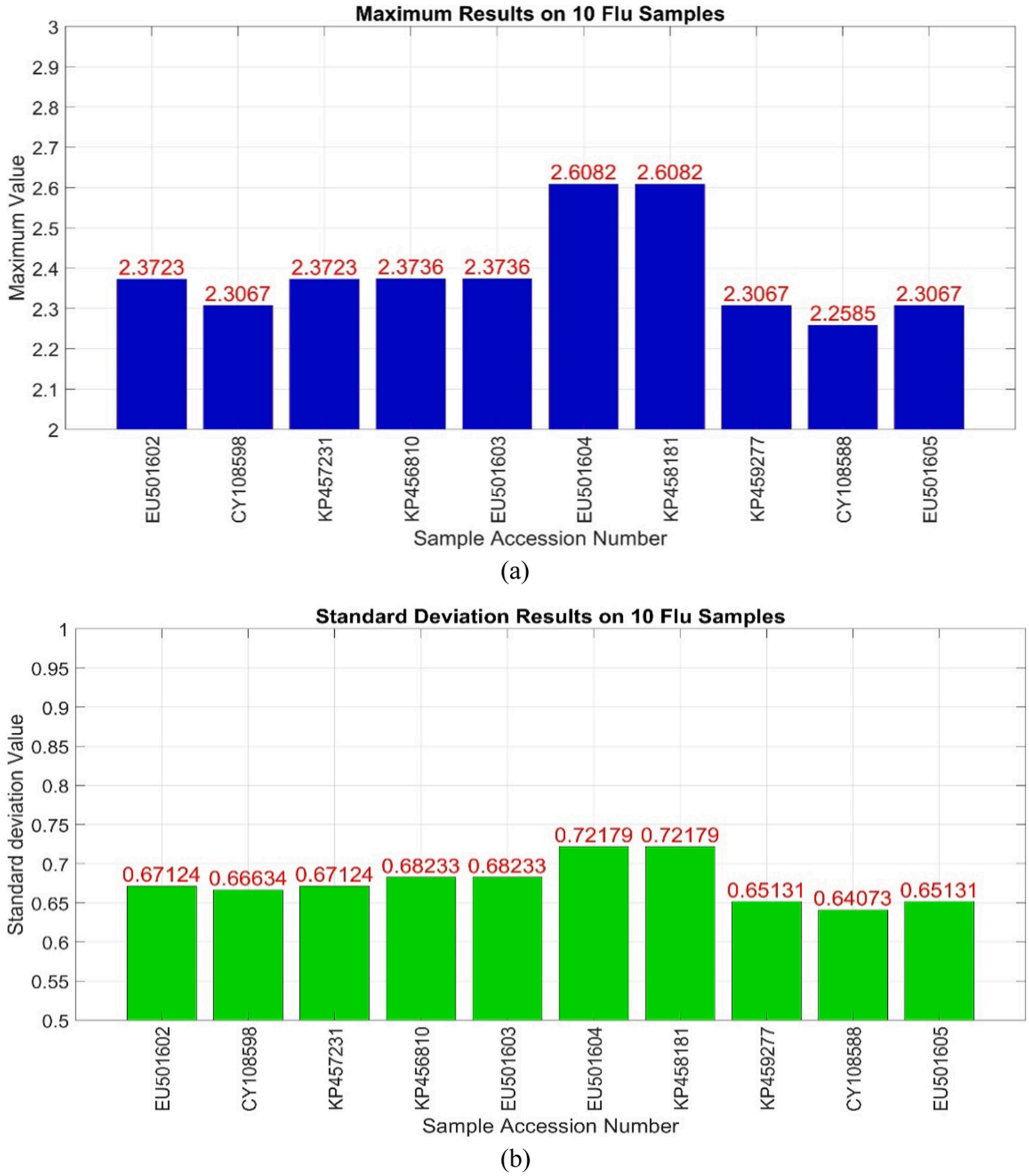


Fig. 7. The result of two features on some non-covid samples (a) Maximum feature (b) Standard deviation feature.

dataset are shown in Fig. 6, by their NCBI accession numbers. In this figure, the first feature (maximum) and the second feature (standard deviation) of the results are specified in two sections (a) and (b), respectively. The horizontal axis of both charts, represents the NCBI Genbank accession number of the genomic sequence at, and the vertical axis shows the obtained results.

As shown in Fig. 6, the results are similar, which is true for both features. These results have a similar interpretation for the non-corona class. Two mentioned features results were successful on other datasets as well. The same experiment was tested on 10 non-covid (flu virus) sequences, which are shown in Fig. 7.

Two types of virus difference are obvious according to the previous figures. This difference is evident in both features, which have also been tested on more several data. The proposed approach tested on the United Kingdom dataset with 4500 samples, 1735 of which belong to the covid class, also yielded successful results. Similarly, Fig. 8 illustrates the outputs in the form of two-dimensional diagrams in which the horizontal and vertical axes represent the two features, respectively.

In Fig. 8, it is also clear that the threshold values can be defined on both features of the two categories, similarly. Similarly, this experiment and diagrams can also be performed on the other countries datasets. Some of them is illustrated in Fig. 9, in the form of several parts by similar diagrams.

In Fig. 9, the a, b, c and d sub-sections are related to the USA, Australia, France, India countries datasets respectively. In these experiments, 53400, 23600, 420 and 2890 samples tested, respectively. The obtained diagrams well represent the ability to classify and differentiate the two categories. Similar to the previous figure diagrams, some rules can be defined in terms of numerical values for each feature per class of samples. The results of the maximum value in covid samples limited in the range of 25–30 and in flu cases, were in the range of 1–6. In Covid samples, the second feature values were dispersed in the range of 2–2.5

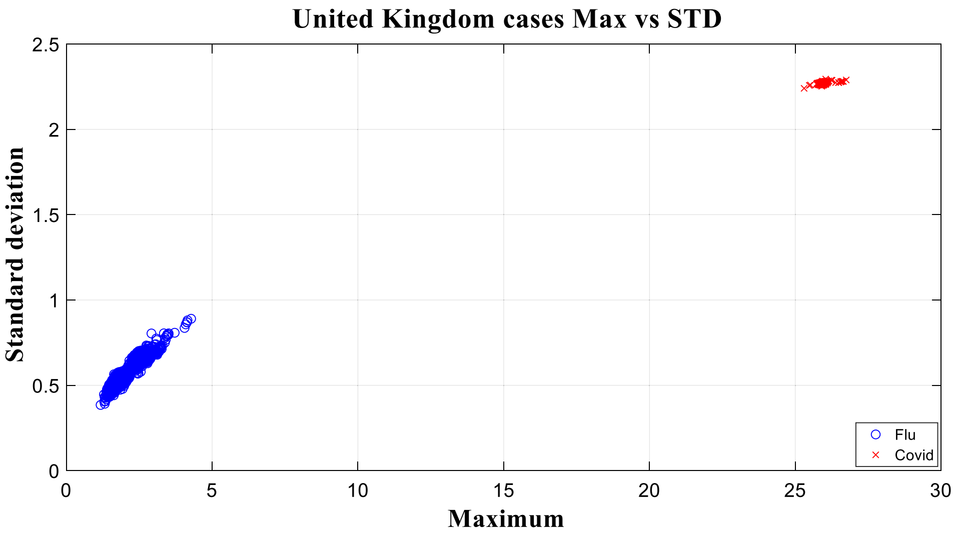
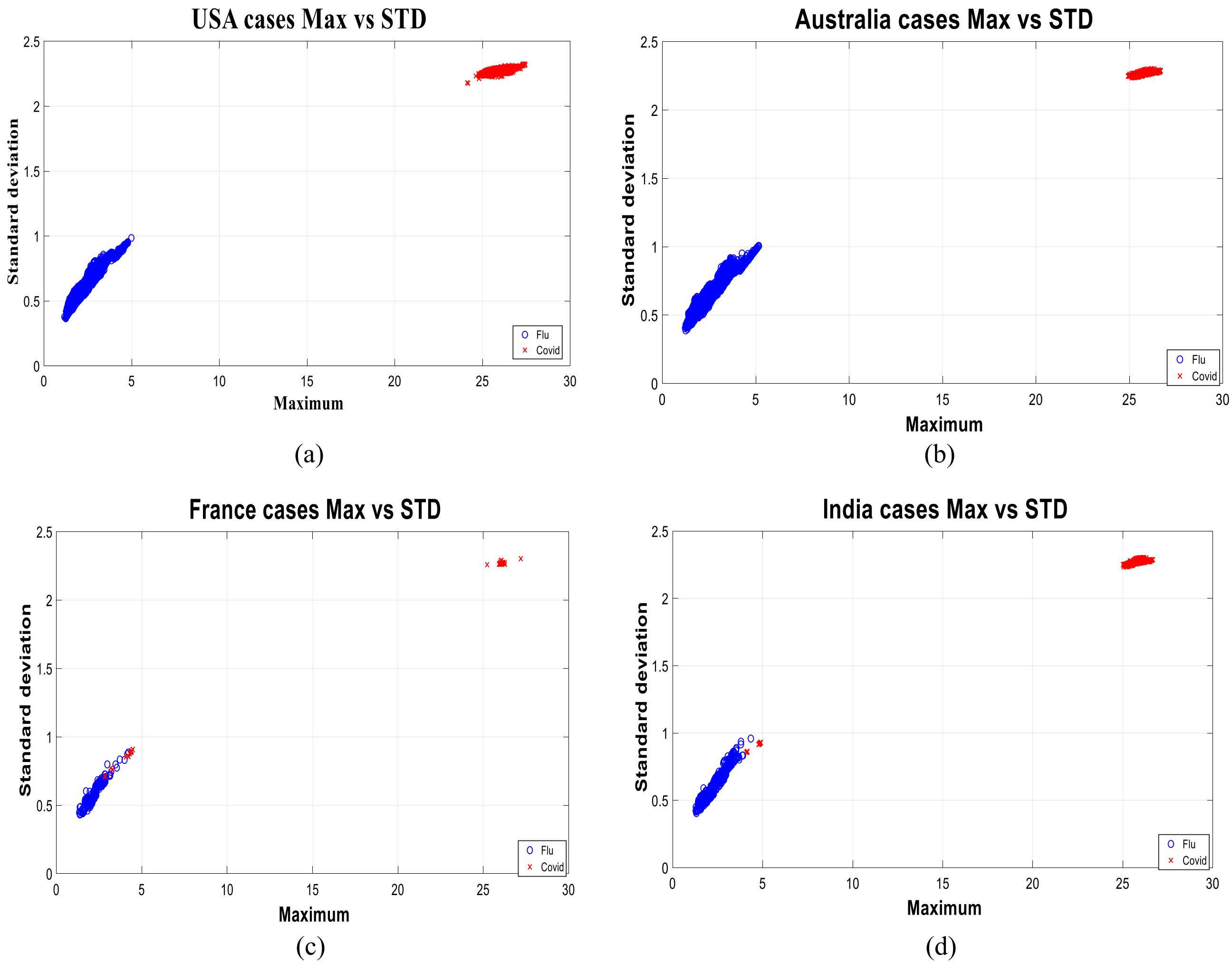


Fig. 8. Comparison of the obtained maximum vs standard deviation values on United Kingdom dataset.



(a)

(b)

(c)

(d)

Fig. 9. Maximum diagram vs standard deviation feature space on some countries datasets (a) USA (b) Australia (c) France (d) India.

and flu samples were limited in the range [0.2,1.2].

A review of previously published articles on the coronavirus sequences data, can highlight the role of the number of the case studies on the obtained results. Table 3 lists the published articles in this scope. In Table 3, the Database column indicates general information about the number and types of the examined genomic data described. The Methods column also contains information about the utilized algorithms and techniques, and the Evaluation column also summarize the obtained

Table 3
Comparative analysis of the proposed method and previously published methods.

#	Database	Number of data	Methods	Evaluation
[27]	**Various types of humans coronaviruses (Alpha CoV, BetaCov-1, MERS-CoV, NL63-** CoV, HKU1-CoV and SARS-COV-2)	**592 (Multi-Class)**	CpG island feature selection + KNN classifier	98 %
[28]	**complete genomes of COVID-19, SARS-CoV** and MERS-CoV sequences	**76**	Combinatorial of DFT, DCT, and Moment Invariants techniques + KNN classifier	100 %
[18]	COVID-19 and three types of Influenza viruses	594	**cockroach optimized** deep neural network	99 %
[25]	**DNA sequences from 24 virus families and SARS-CoV**	**347,363 (Multi-Class)**	Pseudo-convolutional method + Random Forest and MLP classifier	99 %
[23]	**SARS-CoV-2,** **MERS-CoV, HCoV-NL63,** **HCoV-OC43, HCoV-229E, HCoV-HKU1, and SARS-CoV full genome**	**553 (Multi-Class)**	CNN Deep learning	98 %
Proposed method	**coronavirus and influenza virus sequences**	107,000	Sliding window technique on LPC model + SVM classifier	99 %

results. Each row of this table refers to the specific research, and the last row is dedicated the proposed approach of our research.

It is clear that in some recent studies, the number of case studies have not been enough for machine learning purpose. However, in this study several data from different geographic regions tested. Numerous experiments were performed in this study to deal with the randomness of the results. Another noteworthy point about the previously researches is the use of methods and algorithms with high time complexity, such as deep learning techniques.

These experiments and researches that have already been conducted in the genomic signal processing field show that the use of DSP-based approaches can be effective in solving or improving the genetic problems. Also. these methods digitize the information of nucleotide sequences. It can be effective idea in quantifying medical information and concepts. Another advantage that may remain hidden is the ineffectiveness of the nucleotide sequences length of on the obtained results. It can be concluded that the proposed model can compress the information of genomic sequences to recognize and differentiate nucleotide sequences discriminative features.

4. Conclusion

In this study, a pattern recognition model was proposed for the identification and differentiation of coronavirus samples in the form of genomic sequences. In the feature extraction phase, LPC model led to the analysis of any genomic sequence with any length. Another effective approach in the high precision analysis of biological sequences is the windowing technique. This approach causes no region of biological sequences to be assumed to be unaffected. The proposed approach, was tested on two viruses case studies including revealed variants of covid virus and flu virus family, separated the two categories with high accuracy. The use of the SVM kernel functions was also well able to classify and model the existing non-linear feature space. Unlike the initial feature space, the obtained feature space did not have a high dimension.

If we look at the issue from a medical point of view, this virus is not the last virus and biological challenge of humanity against epidemics. In the not-too-distant future, viruses or other diseases originating from germs, bacteria, fungi, etc. may also be prevalent in the world and pose a threat to human life and even other beings. Therefore, studying viruses and pandemics from different dimensions can play an important role in emergency management approaches to coping with these conditions. As mentioned above, many diseases such as corona, influenza, and colds have many similarities in terms of clinical symptoms Analysis of the biological phenomenon in the form of genomic data such as DNA and RNA sequences has certain problems. One of the problems and challenges of analyzing this data is the large and different size of genomic sequences. To solve this problem, artificial intelligence and signal processing approaches can help to analyze big data. The high accuracy of these methods, on the one hand, and the analysis of this data format efficiency on the manufacture of drugs and vaccines against this virus, on the other hand, has analyzed this data format valuable.

CRediT authorship contribution statement

Amin Khodaei: Visualization, Resources, Methodology, Writing – original draft, Writing – review & editing, Data curation, Conceptualization, Software. **Parvaneh Shams:** Validation, Investigation, Writing – original draft. **Hadi Sharifi:** Writing – original draft, Writing – review & editing. **Behzad Mozaffari-Tazehkand:** Investigation, Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

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.

Data availability

Data will be made available on request.

References

[1] "World o meters," 2020. [Online]. Available: https://www.worldometers.info.

[2] Y. Liu, R. Kuo, S. Shih, COVID-19: The The first documented coronavirus pandemic in history, Biomed. J. 43 (4) (2020) 328–333.

[3] F. Khanam, I. Nowrin, M.R.H. Mondal, Data Visualization and Analyzation of COVID-19, J. Sci. Res. Reports 26 (3) (2020) 42–52.

[4] H. Kumar, Anuradha, A. K. Solanki, and S. Tanwar, "Machine Learning-Based Scheme to Identify COVID-19 in Human Bodies," Stud. Syst. Decis. Control, vol. 324, pp. 35–56, 2021.

[5] W.M. Shaban, A.H. Rabie, A.I. Saleh, M.A. Abo-Elsoud, Detecting COVID-19 patients based on fuzzy inference engine and Deep Neural Network, Appl. Soft Comput. 99 (2021), 106906.

[6] C.P. West, V.M. Montori, P. Sampathkumar, COVID-19 Testing: The Threat of False-Negative Results, Mayo Clin. Proc. 95 (6) (2020) 1127–1129.

[7] Y.Z. Zhang, E.C. Holmes, A Genomic Perspective on the Origin and Emergence of SARS-CoV-2, Cell 181 (2) (2020) 223–227.

[8] L.S. Sturman, K.V. Holmes, The molecular biology of coronaviruses, Adv. Virus Res. 28 (C) (1983) 35–112.

[9] B.W.J. Mahy, The Molecular Biology of Coronaviruses, Mol. Basis Viral Replication 6 (1987) 239–254.

[10] P. S. Masters, The Molecular Biology of Coronaviruses, vol. 65. 2006.

[11] A. Vabret, T. Mourez, S. Gouarin, J. Petitjean, F. Freymuth, An outbreak of coronavirus OC43 respiratory infection in Normandy, France, Clin. Infect. Dis. 36 (8) (2003) 985–989.

[12] P.C.Y. Woo, Y. Huang, S.K.P. Lau, K.Y. Yuen, Coronavirus genomics and bioinformatics analysis, Viruses 2 (8) (2010) 1805–1820.

[13] Y. Zarai, et al., Evolutionary selection against short nucleotide sequences in viruses and their related hosts, DNA Res. 27 (2) (2020) 1–32.

[14] Á. Nagy, B. Ligeti, J. Szebeni, S. Pongor, and B. Gyrffy, "COVID outcome-estimating COVID severity based on mutation signatures in the SARS-CoV-2 genome," Database (Oxford)., vol. 2021, no. Cv, pp. 1–6, 2021.

[15] Y. Cao, et al., Comparative genetic analysis of the novel ACE2 in different populations, CellDiscov. 6 (2020) 4–7.

[16] S. Yan, G. Wu, Application of neural network to predict mutations in proteins from influenza A viruses - A review of our approaches with implication for predicting mutations in coronaviruses, J. Phys. Conf. Ser. 1682 (1) (2020) 1–7.

[17] K. Kuzmin, et al., "Machine learning methods accurately predict host specifi city of coronaviruses based on spike sequences alone," *Biochem. Biophys*. Res. Commun., no. xxxx, pp. 1–6, 2020.

[18] M.A. El-dosuky, M. Soliman, A.E. Hassanien, COVID-19 vs influenza viruses: A cockroach optimized deep neural network classification approach, Int. J. Imaging Syst. Technol. 31 (2) (2021) 472–482.

[19] M.S. Nawaz, P. Fournier-Viger, A. Shojaee, H. Fujita, Using artificial intelligence techniques for COVID-19 genome analysis, Appl. Intell. 51 (2021) 3086–3103.

[20] S. Ma, et al., Metagenomic analysis reveals oropharyngeal microbiota alterations in patients with COVID-19, Signal Transduct. Target. Ther. 6 (1) (2021).

[21] B.H. Kim, Y.C. Won, S.Y. Jeong, The first association study of single-nucleotide polymorphisms (SNPs) of the IFITM1 gene with influenza H1N1 2009 pandemic virus infection, Mol. Cell. Toxicol. 17 (2) (2021) 179–186.

[22] A.A. Tsonis, G. Wang, L. Zhang, W. Lu, A. Kayafas, K. Del Rio-Tsonis, An application of slow feature analysis to the genetic sequences of coronaviruses and influenza viruses, Hum. Genomics 15 (1) (2021) 1–10.

[23] A. Lopez-Rincon, et al., Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning, Sci. Rep. 11 (1) (2021) 1–12.

[24] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado, E. Claassen, J. Garssen, A. D. Kraneveld, Accurate identification of sars-cov-2 from viral genome sequences using deep learning, BioRxiv (2020).

[25] J.C. Gomes, et al., Covid-19 diagnosis by combining RT-PCR and pseudo-convolutional machines to characterize virus sequences, Sci. Rep. 11 (1) (2021) 1–28.

[26] R. Khan, M. Biswas, M. Uddin, Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model, Chaos, Solitons Fractals 138 (January) (2020) 1–7.

[27] H. Arslan, H. Arslan, A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier, Eng. Sci. Technol. an Int. J. 24 (4) (2021) 839–847.

[28] S.M. Naeem, M.S. Mabrouk, S.Y. Marzouk, M.A. Eldosoky, A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19, Brief. Bioinform. 22 (2) (2021) 1197–1205.

[29] T. Paul, S. Vainio, J. Roning, Clustering and classification of virus sequence through music communication protocol and wavelet transform, Genomics 113 (1P2) (2021) 778–784.

[30] M. Cohen-Mcfarlane, R. Goubran, F. Knoefel, Novel Coronavirus Cough Database: NoCoCoDa, IEEE Access 8 (2020) 154087–154094.

[31] J. Andreu-Perez, et al., A Generic Deep Learning Based Cough Analysis System from Clinically Validated Samples for Point-of-Need Covid-19 Test and Severity Levels, IEEE Trans. Serv. Comput. (2021) 1–13.

[32] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, V. Aharonson, SARS-CoV-2 Detection From Voice, IEEE Open J. Eng. Med. Biol. 1 (2020) 268–274.

[33] S.S. Roy, S. Barman, A Non-invasive Cancer Gene Detection Technique using FLANN based Adaptive Filter, Microsyst. Technol. (2018) 1–16.

[34] J. Das, S. Barman, DSP based Entropy Estimation for Identification and Classification of Homo sapiens Cancer Genes, Microsyst. Technol. 23 (9) (2016) 4145–4154.

[35] A. Khodaei, M.R. Feizi-Derakhshi, B. Mozaffari-Tazehkand, A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods, Soft Comput. 24 (21) (2020) 16315–16334.

[36] S. M. Naeem, M. S. Mabrouk, and M. A. Eldosoky, "Detecting genetic variants of breast cancer using different power spectrum methods," ICENCO 2017 - 13th Int. Comput. Eng. Conf. Boundless Smart Soc., vol. 2018-Janua, pp. 147–153, 2018.

[37] A. Shoeibi, et al., "Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques: A Review," arXiv, 2020, pp. 1–20.

[38] D. Sobya, S. Manoj, Prediction and Exposure of Cancer Cells through Walsh Hadamard Transform and MATLAB R2017a Techniques, Mater. Today Proc. (2020) 1–9.

[39] L. Das, S. Nanda, J.K. Das, Hereditary disease prediction in eukaryotic DNA: an adaptive signal processing approach, Nucleosides Nucleotides Nucleic Acids 39 (8) (2020) 1179–1199.

[40] G.B. Rathod, V. Shah, N. MacWan, S.D. Jiteshkumar, N.H. Ashvinbhai, The statistical approach and overview in detection of cancer cells based on fft and dwt employing genomics signal processing techniques on DNA, Reliab. Theory Appl. 16 (60) (2021) 233–242.

[41] D.C. Dalwadi, V. Shah, H. Navadiya, Y. Mehta, Aids detection using genomics signal processing techniques on dna, Innovations in Electrical and, in: Electronic Engineering 661, Springer, Singapore, 2021, pp. 651–663.

[42] T. Roy, S. Barman, Performance analysis of network model to identify healthy and cancerous colon genes, IEEE J. Biomed. Heal. informatics 20 (2) (2016) 710–716.

[43] Y. Sun, et al., Identification of 12 cancer types through genome deep learning, Sci. Rep. 9 (1) (2019) 1–9.

[44] A. Khodaei, M.R. Feizi-Derakhshi, B. Mozaffari-Tazehkand, A Markov chain-based feature extraction method for classification and identification of cancerous DNA sequences, BioImpacts 11 (2) (2020) 87–99.

[45] T. Roy, P. Bhattacharjee, Performance analysis of melanoma classifier using electrical modeling technique, Med. Biol. Eng. Comput. 58 (10) (2020) 2443–2454.

[46] S.M. Naeem, M.S. Mabrouk, M.A. Eldosoky, A.Y. Sayed, Moment invariants for cancer classification based on electron–ion interaction pseudo potentials (EIIP), Netw. Model. Anal. Heal. Informatics Bioinforma. 9 (1) (2020) 1–5.

[47] A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, L. Zhang, Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA, Front. Bioeng. Biotechnol. 8 (September) (2020) 1–13.

[48] R. Zhang, C.-T. Zhang, Z curves, an intutive tool for visualizing and analyzing the DNA sequences, J. Biomol. Struct. Dyn. 11 (4) (1994) 767–782.

[49] D. O'Shaughnessy, Linear predictive coding, IEEE Potentials 7 (1) (1988) 29–32.

[50] A.R. Madane, Z. Shah, R. Shah, S. Thakur, Speech compression using Linear predictive coding, in: In proceedIngs International workshop on MachIne Intelligence Research MIR labs, 2009, pp. 119–122.

[51] S. Theodoridis, A. Pikrakis, K. Koutroumbas, D. Cavouras, Introduction to pattern recognition: a matlab approach, Academic Press, 2010.

[52] C. Cortes, V. Vapnik, Support vector machine, Mach. Learn. 20 (3) (1995) 273–297.

[53] "NCBI SARS-CoV-2 Data Hub," 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe acute respiratory syndrome coronavirus.