Interdisciplinary Sciences: Computational Life Sciences (2022) 14:504–519
https://doi.org/10.1007/s12539-021-00465-0

ORIGINAL RESEARCH ARTICLE

Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses

Imran Ahmed[1] · **Gwanggil Jeon**[2]

Received: 11 April 2021 / Revised: 18 July 2021 / Accepted: 23 July 2021 / Published online: 6 August 2021

Abstract

Recent pandemic of COVID-19 (Coronavirus) caused by severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) has been growing lethally with unusual speed. It has infected millions of people and continues a mortifying influence on the global population's health and well-being. In this situation, genome sequence analysis and advanced artificial intelligence techniques may help researchers and medical experts to understand the genetic variants of COVID-19 or SARS-CoV-2. Genome sequence analysis of COVID-19 is crucial to understand the virus's origin, behavior, and structure, which might help produce/develop vaccines, antiviral drugs, and efficient preventive strategies. This paper introduces an artificial intelligence based system to perform genome sequence analysis of COVID-19 and alike viruses, e.g., SARS, middle east respiratory syndrome, and Ebola. The system helps to get important information from the genome sequences of different viruses. We perform comparative data analysis by extracting basic information of COVID-19 and other genome sequences, including information of nucleotides composition and their frequency, tri-nucleotide compositions, count of amino acids, alignment between genome sequences, and their DNA similarity information. We use different visualization methods to analyze these viruses' genome sequences and, finally, apply machine learning based classifier support vector machine to classify different genome sequences. The data set of different virus genome sequences are obtained from an online publicly accessible data center repository. The system achieves good classification results with an accuracy of 97% for COVID-19, 96%, SARS, and 95% for MERS and Ebola genome sequences, respectively.

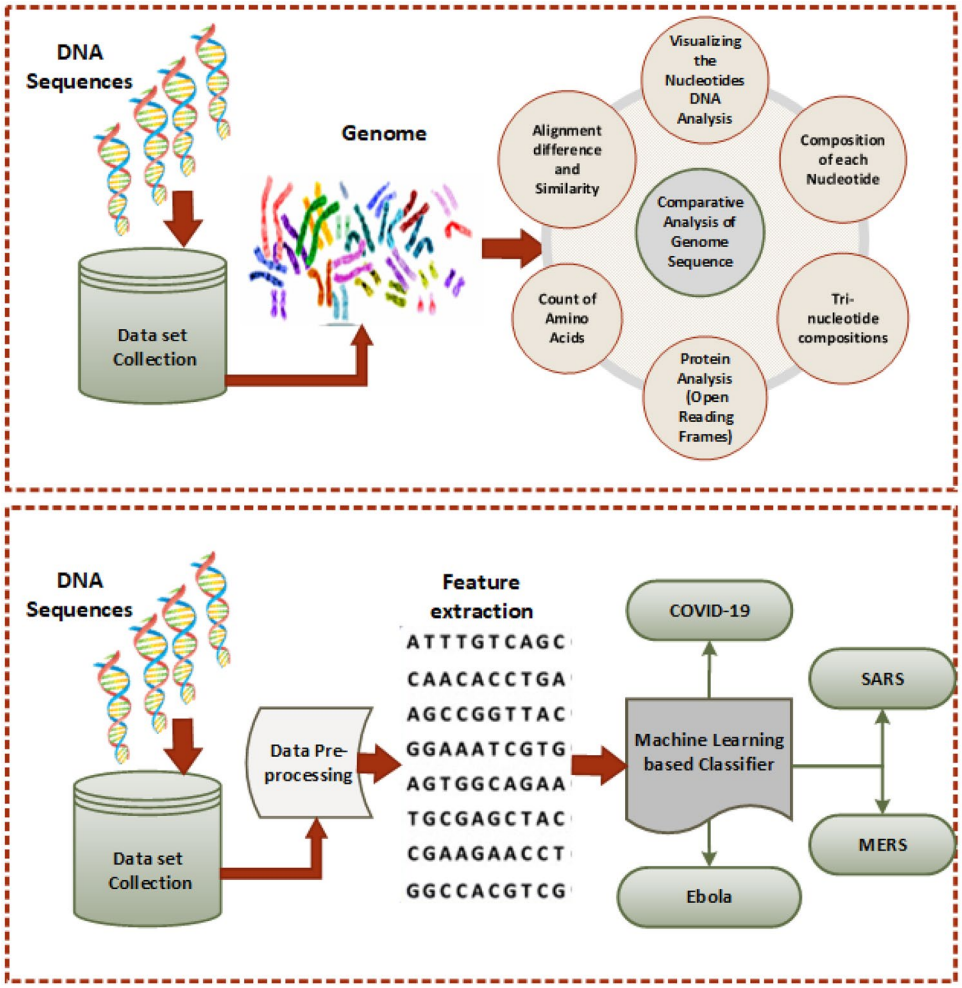
✉ Gwanggil Jeon
gjeon@inu.ac.kr

Imran Ahmed
imran.ahmed@imsciences.edu

[1] Center of Excellence in IT, Institute of Management Sciences, Hayatabad, Peshawar 25000, Khyber Pakhtunkhwa, Pakistan

[2] Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea

Graphic Abstract



Keywords Genome sequence analysis · Artificial intelligence · Machine learning · SVM · COVID-19

1 Introduction

Current evolution of the novel coronavirus diseases, also known as COVID-19 or SARS-CoV-2, originated from China, causes a global health emergency with its rapid worldwide spread. On 11 March 2020, the WHO (World Health Organization) publicly reported it as a global pandemic. Until now, in July 2021, the COVID-19 pandemic affected over 200 territories and countries with more than 188,655,968 verified cases including 4,067,517 losses.[1] The worldwide scientific response is unprecedented to the deadly infection or virus which killed thousands of people globally. Governments and health bodies practice and suggest preventive and quarantine measures to overcome the deadly virus's transmission, spread, and reproduction rate. Researchers focus on understanding the genome of the COVID-19, its functionalities, and its behavior to develop/produce an effective vaccine or drug; that provides long-term protection.

On 10th January 2020, 9 days after the first suspected COVID-19 case was identified, the virus's first genome sequence was shared publicly by [1]. Since then, tens of thousands of samples have been sequenced worldwide [2]. Genomics, which is concerned with the genetic substance of an organism, is one of the most promising fields of research for COVID-19. Genome sequence analysis uses an organism's genome information to guide clinical experts to give more personalized therapeutic or diagnostic decision-making strategies. By unlocking the virus's genetic information

[1] https://covid19.who.int/.

and the most badly affected hosts (patients), experts might hope to inform public health, make better decisions and find effective treatments. An organism's genome comprises four nucleotide bases (Adenine-A, Guanine-G, Cytosine-C, and Thymine-T) stored as an encoded sequence that forms its nucleic acids. The novel COVID-19 genome sequence is an enclosed single-stranded positive-sense RNA (Ribonucleic acid) from a long family called Coronavirus, approximately 30 kB long, categorized under three groups, in which two are responsible for infections in mammals, such as bat SARS-CoV like; (MERS-CoV) [3].

Genome sequence analysis is the identification of the nucleotides in a genome sequence. Until now, several organizations around the globe sequenced the COVID-19 or SARS-CoV-2 genome, which informed about different strains of the virus [4]. Genome features identification supports biomedical experts to provide predictions about these features' impact on the population's disease implications. Even though it is usually a tedious and resource-intensive procedure that mostly depends on field expertise. Various strains of the initial genome sequencing of COVID-19 did not set timely actionable discernments, yet many features of disease behavior are unknown [5].

One of the main objectives of genome sequence analysis is to develop genomics medication by stimulating the description and interpretation of disease and remedial associated with genetic modifications. According to this endeavor, significant attempts have been made to combine sequence data of population-level and genomics data with phenotypic knowledge, clinical reports, and other types of multi-omics datasets (e.g., transcriptomics, proteomics, and metabolomics) [6]. In contrast, significant to promote distinct biomedical data for personalized medicine, this integrative research of omics and clinical data creates complex scientific and computational demands for researchers, medical experts, and clinical services. Due to this, there is a need for computational techniques that facilitates the evaluation of extensive high-dimensional and heterogeneous data sets (i.e., those consist of distinct attributes) and systems that might give easier, cheaper, extensible, and comprehensive experimental and analytical solutions. For instance, such data sets may produce together: e.g., molecular data (genes, proteins, physiological evaluations), including measures of critical organ systems, medical image data (consist of CT, MRI scans, etc.).

Adoption of artificial intelligence and related technologies, including data analysis [7], machine [8–10], and deep learning [11–20] can speed up the process of identifying actionable insights and yet lead to a better worldwide response. There is a developing demand for computational methods that enable the analysis of large complex and high-dimensional genome data sets. Artificial intelligence can promote new findings in these data sets without indicating specific rules and interests in different genomic data pipeline stages. Mostly developed work in artificial intelligence with genome data sets has been taking place within the analysis and perception stage. Artificial intelligence has been devoting significant incremental developments in clinical genome analysis, such as phenotyping in rare syndromes, cancer, its variants investigation, and explanation. Although, mostly artificial intelligence-based work in the field of genomics is within the research phase. The demand of machine and deep learning techniques for functional genomics analysis is rising. Most aspects of genome analysis have been traced by these techniques, from genome sequencing, phenotyping, and variant identification, to downstream analysis. Now improvements in computing, artificial intelligence, and the increase in biomedical data sets enable advances to existing fields of service. Simultaneously, these advancements in open access research and open-source tools make artificial intelligence use and prosperity across various types of genome studies. Moreover, publicly available resources, good software distributors incorporate machine learning methods within their genome analysis tools and services.

Inspired from the above discussion and the advancement of artificial intelligence in genome sequence analysis, this paper aims to present an artificial intelligence based system for genome sequence analysis of COVID-19 and alike viruses, including SARS, MERS, and Ebola. We perform comparative analysis to study the basic patterns of the genome sequence of these viruses and further utilized a machine learning algorithm for classification. More precisely, the following main contributions are made to achieve the objective of the work:

- To present, an artificial intelligence based system for genome sequence analysis of COVID-19, SARS, MERS, and Ebola.
- To perform comparative analysis using different types of data analysis and visualization techniques to find interesting patterns in genome sequences.
- For classification, state of the art machine learning classifier SVM is applied using different genome sequences.
- To compare the results of SVM classifier with other machine learning algorithms.

The rest of the paper is arranged as follows. Section 2 presented related work based on various artificial intelligence based methods used for analyzing the COVID-19 pandemic. Section 3 introduces the artificial intelligence based system used for genome sequence analysis of COVD-19 and alike viruses. This section also gives comparative analysis performed to determine nucleotides and their frequent patterns, relationships between different genome sequences, and a machine learning classifier used to classify genome sequences of alike viruses. Evaluation results and discussion

of the presented system are made in Sect. 4. Finally, in Sect. 5, the paper is concluded with some future directions.

2 Literature Review

Artificial intelligence and data analytics techniques help researchers to understand variables, behaviors, and various data trends of different situations, problems, and diseases. Early researches and studies practiced various methods like statistical, data mining, and machine learning techniques for different kinds of investigations such as forecasting and risk analysis [21] of various diseases. This section discusses different artificial intelligence based techniques practiced by researchers to predict, analyze, and detect/identify the COVID-19. Researchers used different data sets, including clinical textual data sets, genomics data sets, medical image data sets like X-ray images, CT scan images, and some time vision-based data set to provide efficient solutions to control COVID-19 pandemic situations.

Ahmad et al. [22], presented a detailed study of machine-learning based techniques applied for analysis and prediction of the outbreak. Pashazadeh et al. [23] presented a survey on machine learning based methods and big data tools used in different healthcare treatments. Authors in [24] applied the machine learning method and presented a prognostic forecast algorithm for predicting people's death risk. In [21], authors classified textual clinical data set into four different classes of the virus by utilizing traditional and ensemble machine learning algorithms. Jiang et al. [25] provided a machine learning system that predicts an infected person having COVID-19 and has the probability of spreading or suffering from ARDS (Acute Respiratory Distress Syndrome) using clinical data set. Rao et al. [26] produced a method to identify subjects with COVID-19 utilizing mobile devices data. Finally, Chamola et al. [27] provided a survey on machine learning algorithms used for pandemic and disaster management.

Peng et al. [28] provided analysis of the COVID-19 epidemic using dynamical modeling. Their investigation mainly centered on devising a conceptual framework for various machine learning based applications, using various data mining approaches. Authors in [29] produced a regression paradigm to predict the accelerated transmission of COVID-19 depending on the number of patients reported outside of China. Ahmed et al. [7], demonstrated a health monitoring framework using clinical textual data set based on big data analytics and the Internet of Things for the analysis of the COVID-19 pandemic. Authors in their work performed different types of analysis, namely descriptive, diagnostic, predictive, and prescriptive analysis utilizing different infection symptoms.

Authors in [30], gave a taxonomy that groups developed methods into four classes. Further, they presented the challenges and provided recommendations to the machine learning specialists to enhance the techniques utilized to predict the COVID-19 cases. [31] suggested utilizing machine and deep learning methods to recognize exponential response and predict the expected spread of the COVID-19 across the countries. Rustam et al. [32], demonstrated the capacity of machine learning to determine the figure of expected patients affected by COVID-19. Ahmed et al. [13, 33] used a surveillance data set with deep learning models to monitor social distance and control infection transmission of COVID-19. Some researchers, e.g., [5] also used genome sequence data sets with artificial intelligence based method e.g., sequential pattern mining for COVID-19 genome analysis.

Mateos et al. [34] presented a deep learning technique for the classification of SARS-CoV-2 and co-infecting RNA viruses. In [35], authors used an convolutional neural network for classification and accurate detection of SARS-CoV-2. In [36], authors provided a comprehensive analysis of the pandemic. Authors discussed the role of artificial intelligence, drone cameras, the Internet of Things, blockchain, and 5G in controlling and monitoring the COVID-19. Rohet et al. [37] discussed different artificial intelligence based methods for monitoring of COVID-19. Further, they also discussed different prospectus and challenges of artificial intelligence based methods.

From the above discussion, it is concluded that researchers employed various artificial intelligence, machine learning, and deep learning based methods to predict, analyze, diagnose, and detect the COVID-19 using various kinds of data sets. Inspired by previous work, we also presented an artificial intelligence based method for analysis and classification of the COVID-19 genome sequence from alike viruses.

3 Artificial Intelligence Based Genome Sequence Analysis

In this paper, an artificial intelligence based system is presented for genome sequence analysis and classification of COVID-19, SARS, MERS, and Ebola. The details of the overall method are illustrated in Fig. 1. First, we performed comparative data analysis using different types of interpretation and visualization techniques to find inside details of the genome of these viruses, including length of genome sequences, visualizing the nucleotides, nucleotides frequency in the DNA, tri-nucleotides composition, GC percentage showing which genome sequence has the most stable DNA, count of Amino acids, similarity or alignment between different genome sequences.

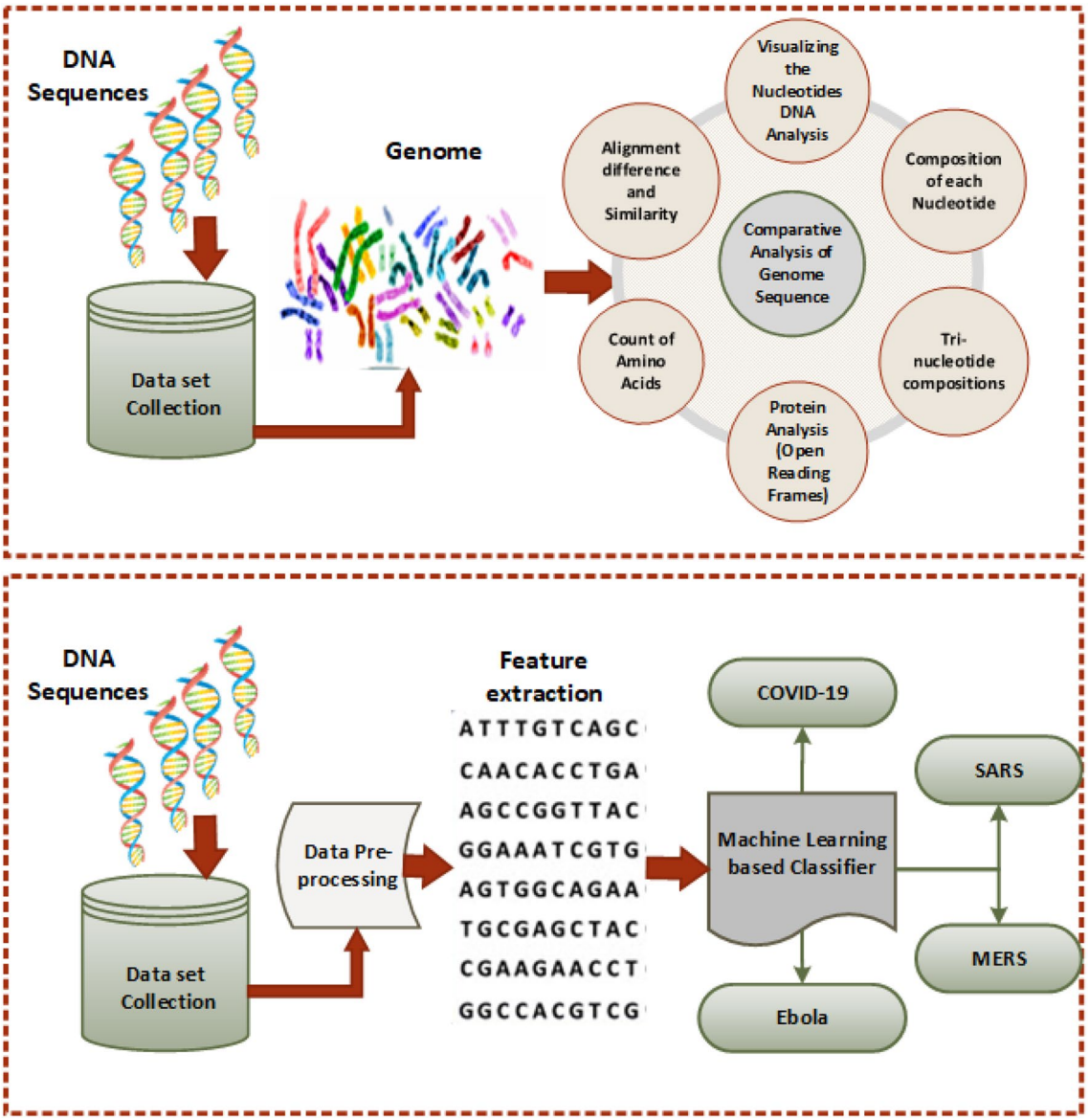


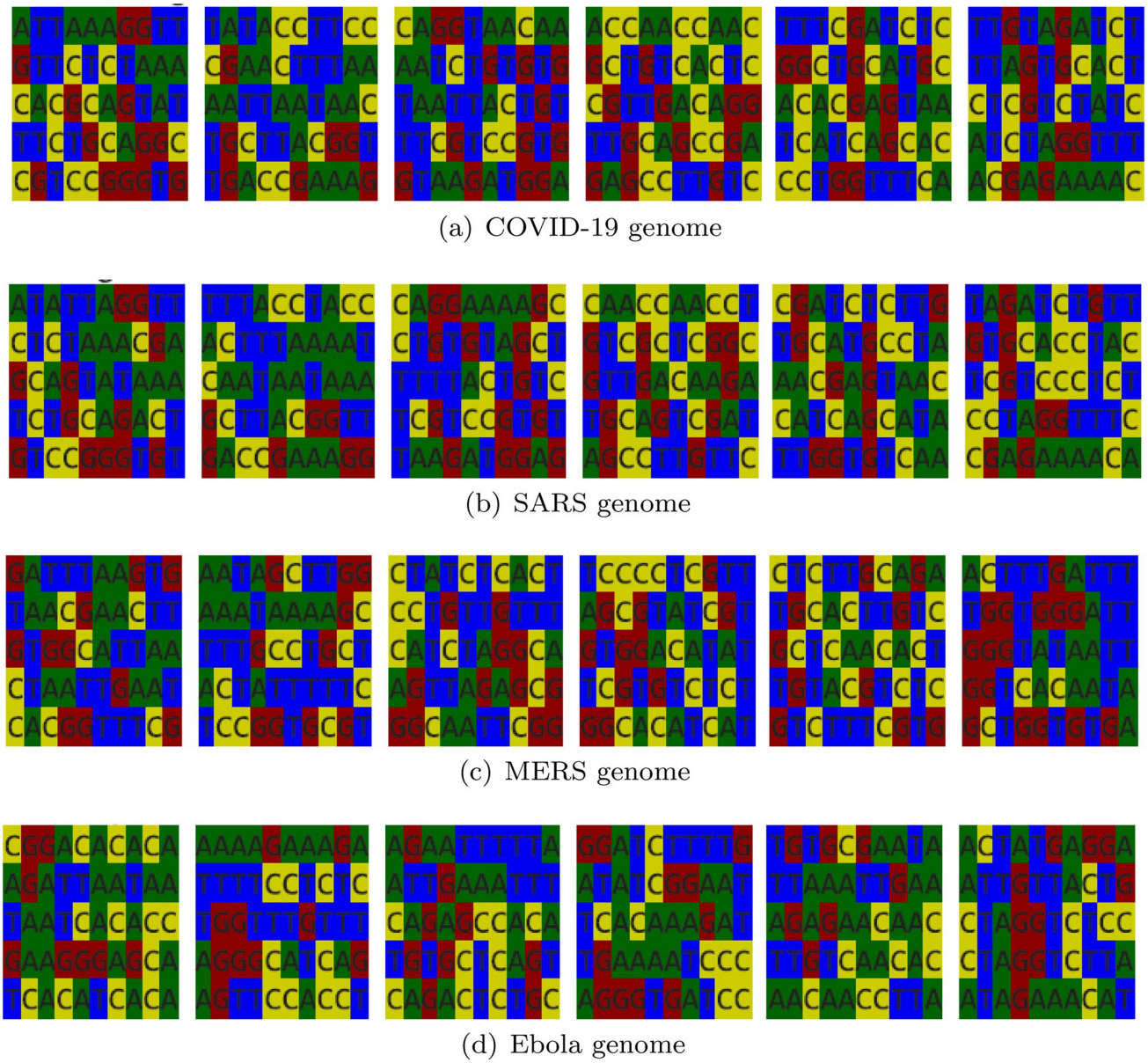
Fig. 1 Artificial intelligence based genome sequence analysis and classification of COVID-19 and alike viruses. The presented system first performs comparative data analysis and then used a machine learning based classifier to classify genome sequences of different viruses

In addition, we used different types of graphs to provide details about genome sequence information. Finally, we applied a machine learning based classifier to classify different genome sequences.

3.1 Data Set

The genome sequences of different viruses have been collected in file extension of .fasta and .gb from GenBank.[2] We used total of 300 different genome coding sequences of all four types of viruses. GenBank is one of the familiar online open-access databases of nucleotide sequences that additionally promotes biological annotation. It is supported by the National Center for Biotechnology Information (NCBI). GenBank has evolved exponentially in the last 2 decades, with increasing sequence files approximately every 18 months. It supports research analysts worldwide to instantly evaluate any specific viral structure and function of the genome sequence. The genome sequenced data for viruses gathered from an online database is still essential in international efforts to produce vaccines, antiviral drugs, and exclusively false sensitive diagnostic tests.

[2] http://www.ncbi.nlm.nih.gov.



(a) COVID-19 genome

(b) SARS genome

(c) MERS genome

(d) Ebola genome

Fig. 2 Visualization of nucleotides in the DNA sequence of four types of viruses

3.2 Genome Sequence Analysis of COVID-19 and Alike Viruses

For comparative analysis of genome sequences, we used Biopython.[3] Genome sequence analysis has frequently become an essential tool for studying disease outbreaks. The COVID-19 and other viruses, including SARS, MERS, and Ebola genomes, are used in this study. We start our analysis by reading the DNA sequence; by doing this, we extracted the nucleotides information or length of the genome sequence, the length of the COVID-19 genome sequence is 29,903, SARS genome sequence is 2975, MERS genome sequence is 30,119, and Ebola genome sequence length is 18,959. The visualization results of the nucleotides of each genome sequence DNA are shown in Fig. 2. DNA is the hereditary material in organisms found in the cell's nucleus (where it is called nuclear DNA). The DNA information is saved as a code formed from four chemical bases: Adenine-A, Guanine-G, Cytosine-C, and Thymine-T, as shown in Fig. 2. It can be seen that the composition of nucleotides in each genome sequence is different. The nucleotides in DNA have been shown in different colors. For example, we just plotted the first three hundred nucleotides of each genome

[3] https://biopython.org.

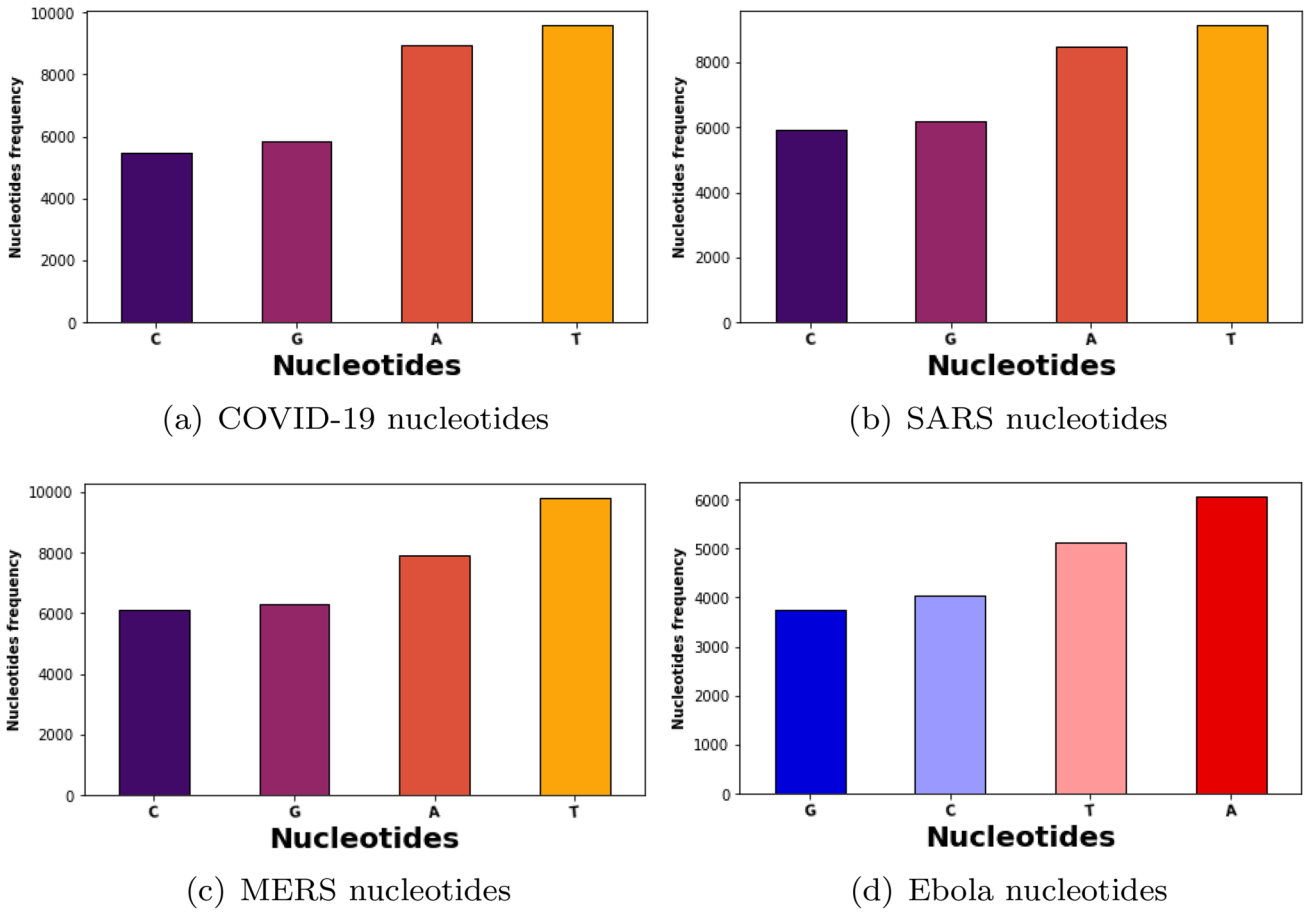


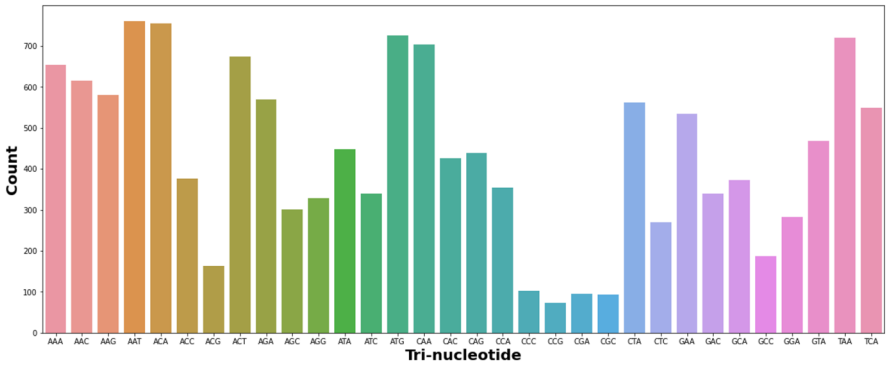
Fig. 3 Nucleotides frequency in the DNA sequence of four types of viruses

sequence; it can be easily observed that the distribution of nucleotides is varying.

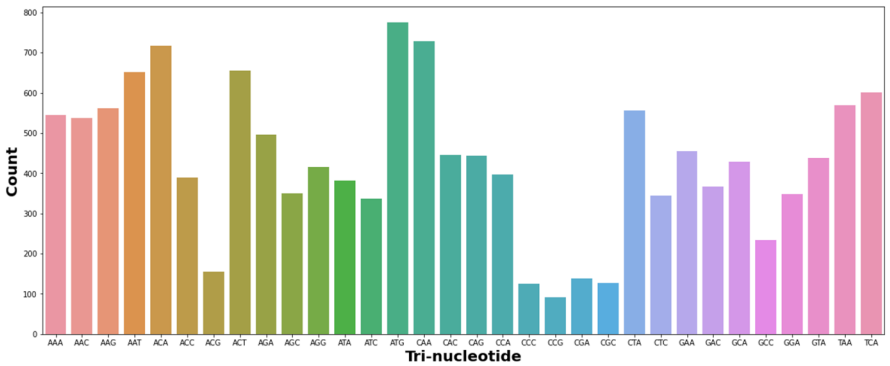
In Fig. 3, we have shown the composition of each nucleotide in the DNA of different genome sequences. It can be seen that the frequency of the nucleotides A and T are higher than the frequency of C and G. Due to the nucleotides pairing, the most present nucleotide in COVID-19, SARS, and MERS is T, while the most present nucleotide in Ebola is A. It is concluded that A and T are the most present nucleotides in all of the genome sequences; the distribution of other nucleotides still varies. We can observe that there is a big difference between the amount of Cytosine–Guanine (GC) and Adenine–Thymine (AT) in both COVID-19 and SARS. Using the nucleotides frequency, we can also determine the GC% for each genome sequence, which gives important information regarding the stability of the DNA. The amount of GC is pretty much close to AT, which will lead us to assume that the GC percentage of the Ebola virus genome sequence has the highest GC percentage estimated as 45.50%, which means it has the most stable DNA. Second, comes the SARS and MERS Virus genome sequence with 40.76% and 41.76% of GC percentage, respectively, and the genome sequence with the least stable DNA is COVID-19 with 37.97%. This information is important for researchers as the stability of DNA is essential to resist change.

Further in Fig. 4, we show the tri-nucleotide compositions; in bioinformatics, it is sub-sequences of the range included within a biological sequence. It is principally applied as computational genomics and sequence analysis, made of basic nucleotides including (i.e., A, T, G, and C), increase heterologous gene representation, distinguish species in metagenomic samples, and helps to produce attenuated vaccines.

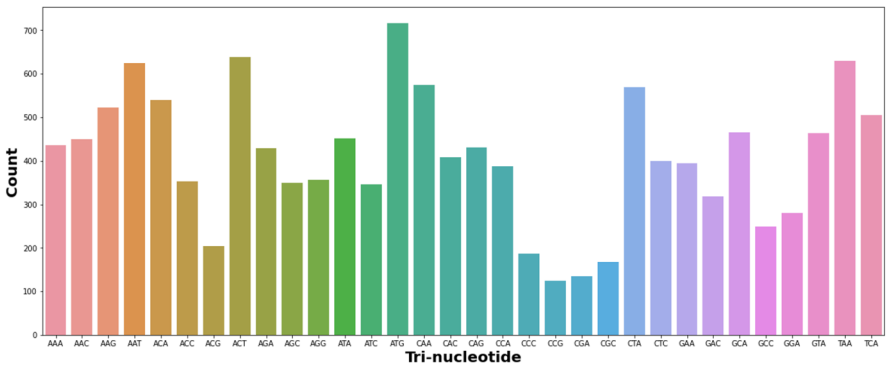
The genome sequence contains all encoded basic information about the virus. Understanding genetic information is the key to obtain cures and vaccines. The process called gene expression: data from a gene is used to synthesize an operative gene product. These products are often proteins. Thus we performed transcription in which DNA is copied out into a messenger RNA (mRNA), and using translation, mRNA is translated into amino acids. It is a translation from one code (nucleotide A T C G sequence) to another code (amino acid sequence). Not all amino acid sequences are proteins. Only the sequences with more than 20 amino acids code



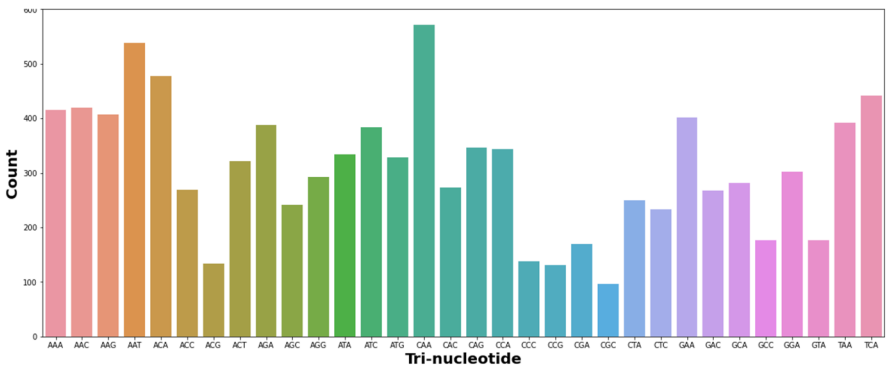
(a) COVID-19 Tri-Nucleotides



(b) SARS Tri-Nucleotides



(c) MERS Tri-Nucleotides



(d) Ebola Tri-Nucleotides

Fig. 4 Tri-nucleotides frequency in the DNA

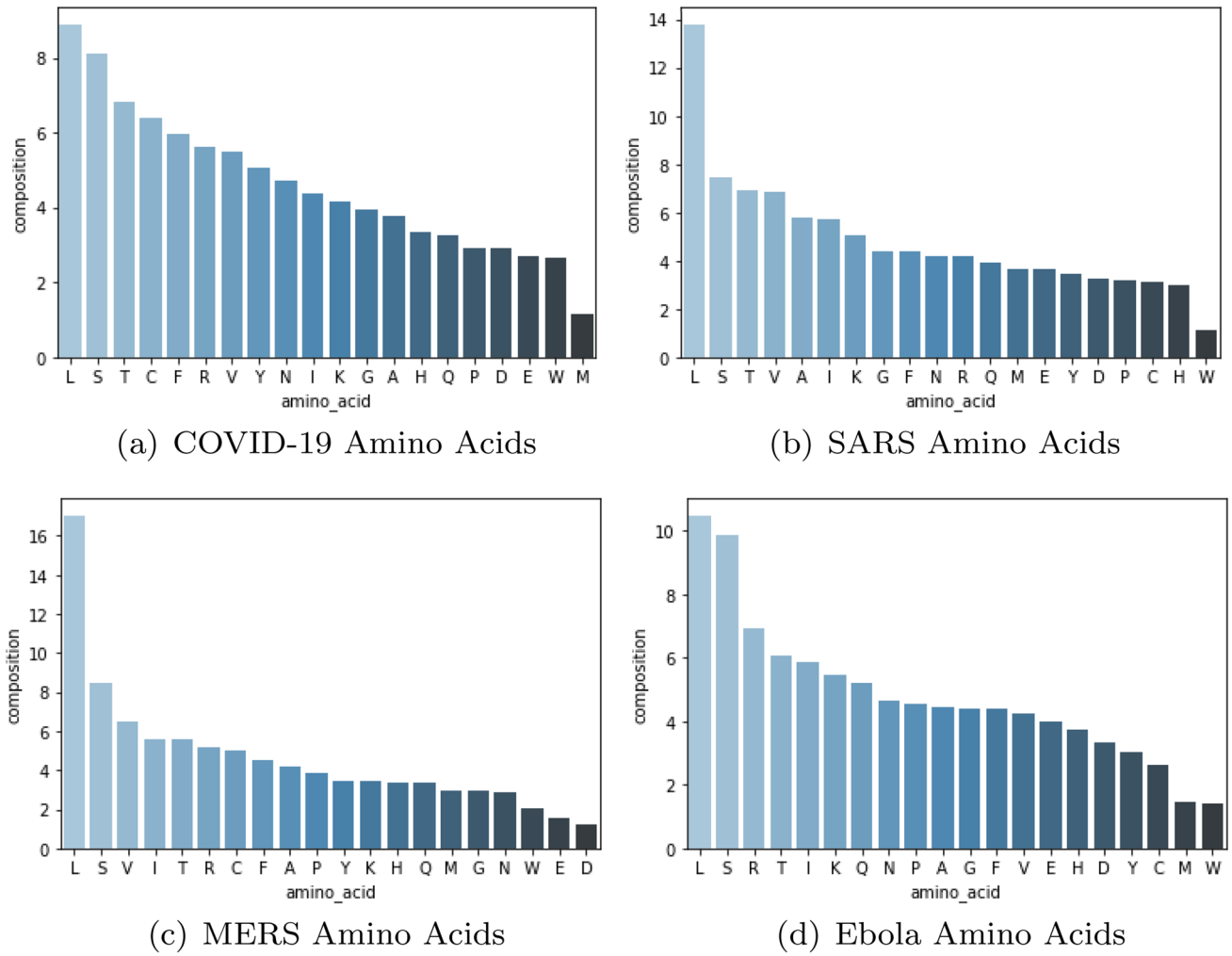


Fig. 5 Composition of amino acids

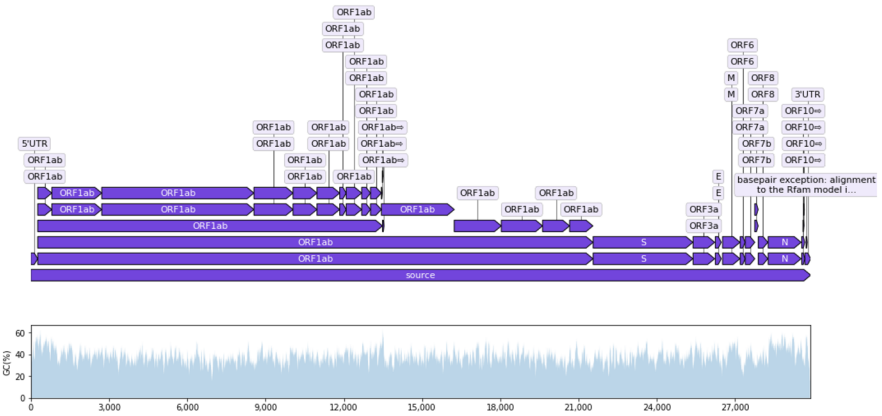
are functional proteins. In Fig. 5, we show the frequency of different amino acids; it can be seen that the distribution of amino acids varies in different genome sequences. The figure shows that the composition of Leucine (L) is high in all sequences, while the composition of the least amino acid is varying. The two dominant amino acids are Leucine (L) and Serine (S). The *y*-axis count is in 1000 in Fig. 5.

Figure 6 shows the open reading frames (ORFs) in the COVID-19, SARS, MERS, and Ebola genome. After converting the DNA into amino acids, as now we have the protein sequences, we used an open reading frame, which is the portion of a reading frame that can be translated. It is a portion of a DNA molecule that carries no-stop codons when changed into amino acids. The genetic code shows DNA sequences in combinations of three base partners, which determines that a double-stranded DNA molecule can be interpreted in any of six probable reading frames, like three in reverse and three in the forward direction. A long reading frame is an acceptable component of a gene.[4] It usually starts with an origin codon (normally AUG) and terminates at a stop codon (ordinarily UAA, UAG, or UGA). We used the transcription and translation to genome sequences with (Genbank format). It shows the ORFs in the genome sequence and the GC% content. The graph on the *x*-axis showing the GC%; it is important to understand where the coding regions are in the genome sequence. It can be seen that COVID-19 and SARS are more similar, mainly with ORF1ab, ORF3a, E, M, S, and N. While the ORF of MERS and Ebola is slightly different than COVID-19.

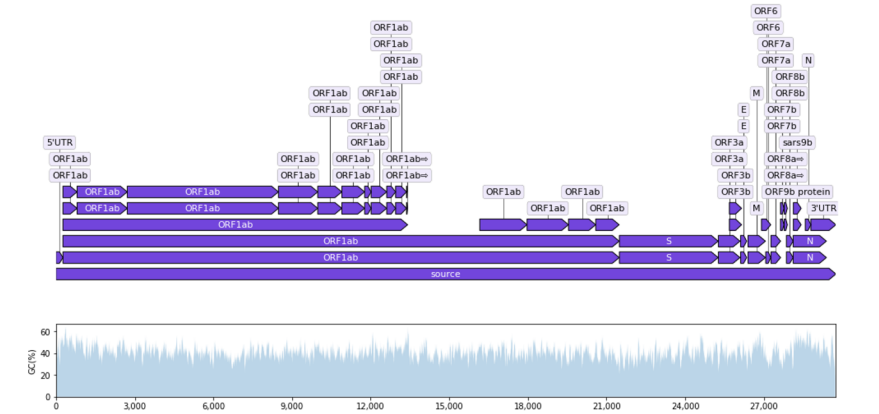
The gene coding region, also recognized as the CDS (coding sequence), is the part of a gene's DNA or RNA that is coded for protein.[5] To find the coding regions CDS in the

[4] https://www.genome.gov/genetics-glossary/Open-Reading-Frame.

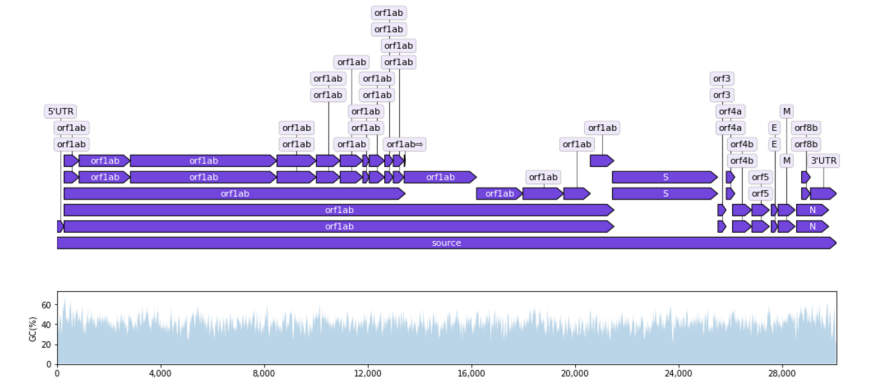
[5] https://web.archive.org/web/20070328214808/, http://genome.wellcome.ac.uk/doc_WTD020755.html.



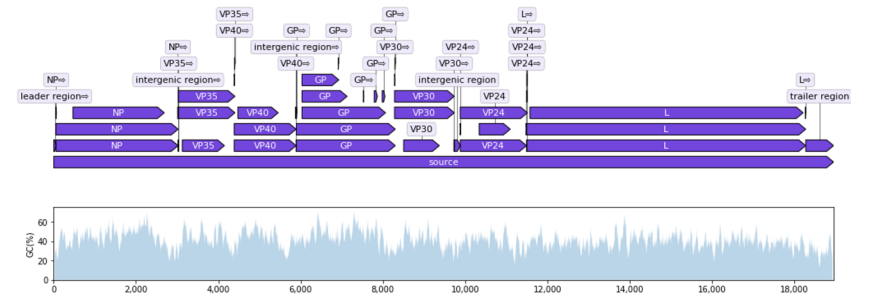
(a) COVID-19 Open Reading Frames



(b) SARS Open Reading Frames

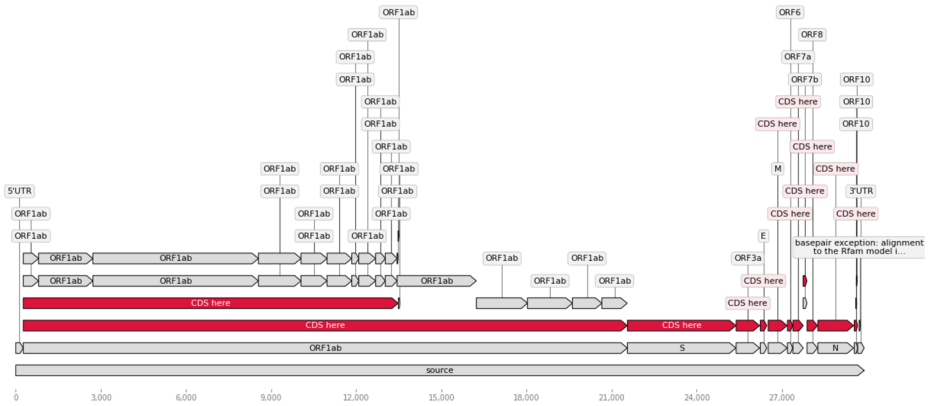


(c) MERS Open Reading Frames

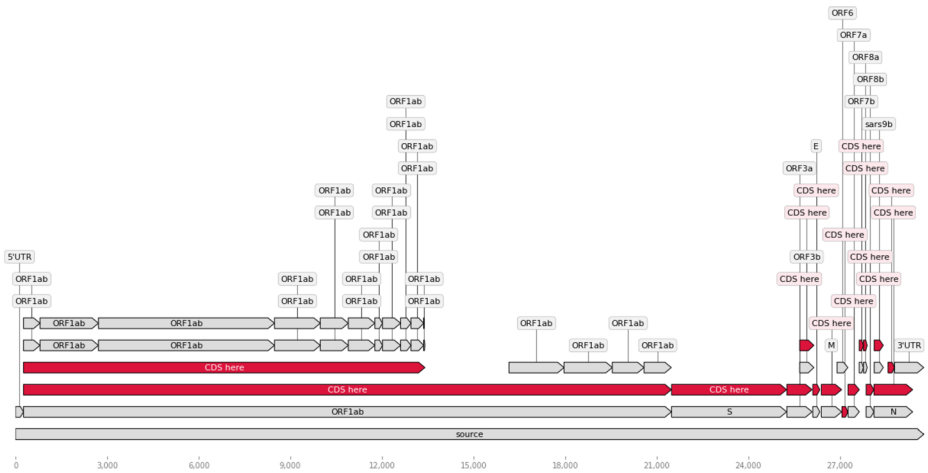


(d) Ebola Open Reading Frames

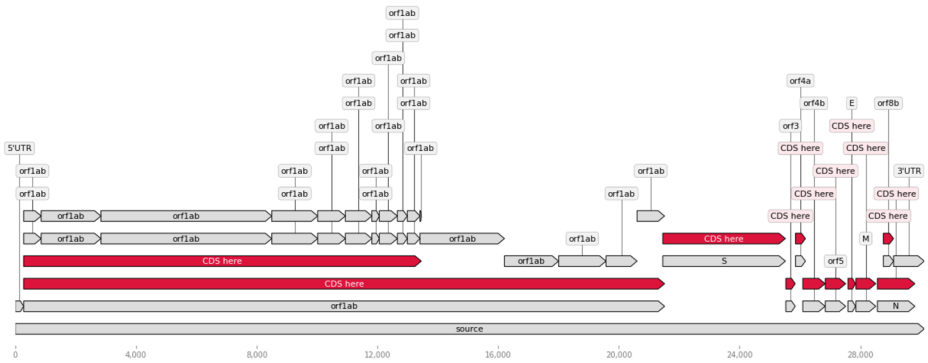
Fig. 6 Sequence of nucleotide triplets of all four types of genome sequences



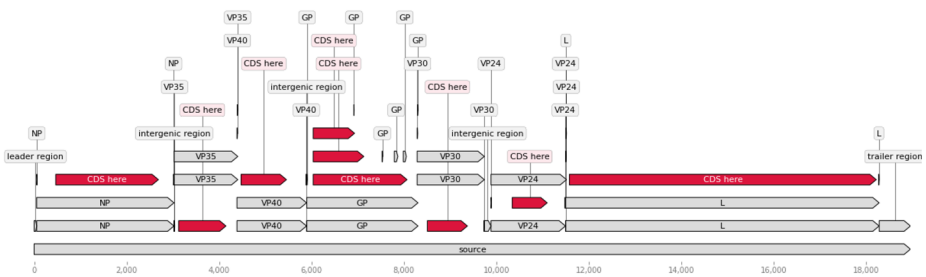
(a) COVID-19 CDS



(b) SARS CDS



(c) MERS CDS



(d) Ebola CDS

Fig. 7 Coding sequence CDS in genome sequences of different genome sequences

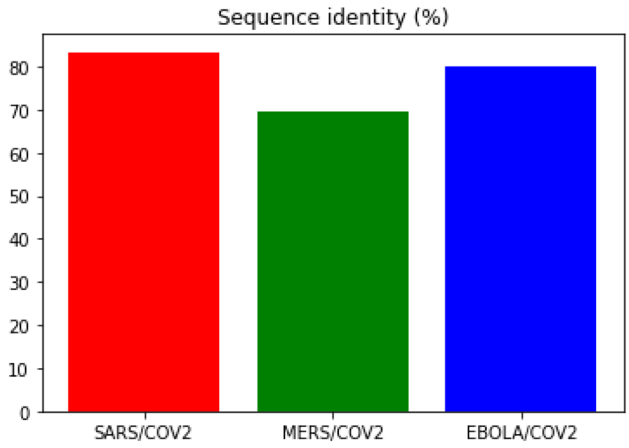


Fig. 8 Alignment similarity between genome sequences

genome sequences, which is an essential step for functional annotation of genes, we plotted the CDS graphs as depicted in Fig. 7. CDS is also known as the sequence of nucleotides that resembles the sequence of amino acids in a protein. A common CDS begins with ATG and terminates with a stop codon. The code in Fig. 7 highlights the coding regions CDS in red. The main CDS are among the ORFs already found in Fig. 6, which include the ORF1ab, ORF3a, S protein, M protein, and N protein. It can be seen that the DNA structure of COVID-19 and SARS is nearly identical, while MERS and Ebola's DNA is slightly different.

In Fig. 8, we show the alignment similarity of COVID-19 with other genome sequences . We used the sequence alignment method to analyze the similarity between four types of DNA sequences. The sequence alignment provides two or more sequences (of DNA, RNA, or protein sequences) in a specific order that helps to recognize the region of similarity among them. Recognizing a similar region allows to understand information like what features are conserved among species, how several species genetically are close and how species grow, etc. The pairwise sequence alignment correlates only two sequences and gives the most reliable feasible sequence alignments. It is an easy and good method to interpret and to determine from the resulting sequence alignment.

To provide a correlation between two biological genome sequences and to recognize the regions of the close similarity among them, we used a dot plot as presented in Fig. 9. It is the easiest method, places a dot where sequences are identical. It compares two sequences by coordinating one sequence on the x-axis and other on y-axis. When both sequences' excesses simultaneously resemble the plot, a dot is marked at the corresponding position/location. It is helpful and can also be applied to visually examine sequences for inverted or direct repeats of sequences. Furthermore, it is also used to investigate areas with low sequence complexity, similar regions or areas, replicated sequences, rearrangements of genome sequences, RNA structures, and gene order.

3.3 Classification of COVID-19 Genome and Alike Viruses Using SVM

For the classification of different genome sequences, we practiced a machine learning classifier, as shown in Fig. 1. As discussed in Sect. 2, researchers utilized various machine learning algorithms for classification purposes. We used SVM, a popular and efficient supervised machine learning classifier used for regression and classification problems. We used multiple genome sequences of all four types of viruses, performed some pre-processing, manually assigned class labels, and extracted useful features provided to SVM classifiers. We randomly splitted the collected genome coding sequences into training and testing samples at ratio of 80% and 20%, respectively. At the output, the classifier

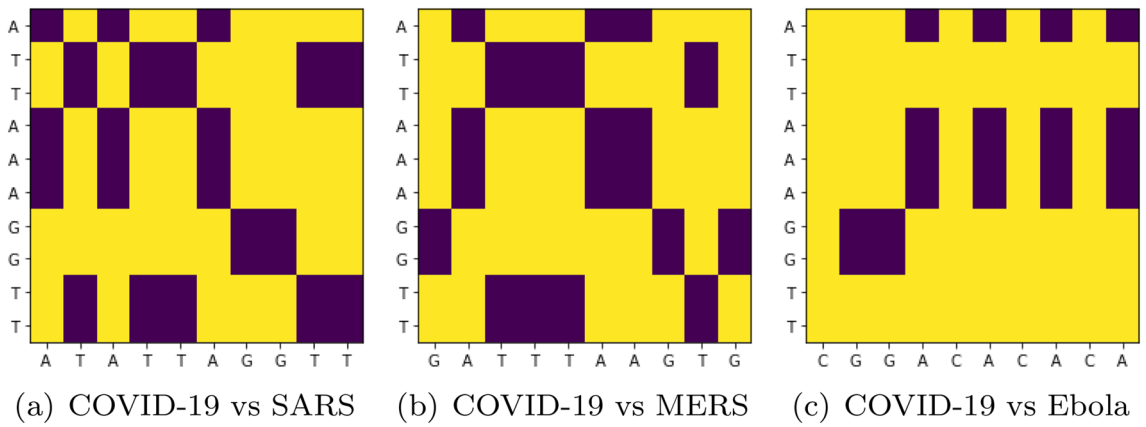


Fig. 9 Dot Plot showing difference between different genome sequences

classifies genome sequences of different viruses as depicted in Fig. 1. The SVM classifier determines a hyper-plane in feature space that best segregates the sequence data into four classes. The further the data points fall from the hyper-plane, the higher the chances are of being classified correctly. The data points that are nearest to the hyper-plane are referred to as support vectors. If these support vectors are eliminated, the hyperplane's position will alter; thus, they are considered the crucial elements of the data set. The distance between either side of the hyperplane and the support vectors is referred to as the margin. The target is to select a hyper-plane with the largest margin between each point in the training set and the hyper-plane to correctly classify the new data. Since SVM are binary, multi-class problems need to be reduced to several binary classification problems. In this work, we used linear SVM. The objective function of SVM classification is provided as follows;

$$L_{(w)} = \sum_{i=1} |\max(0, 1 - y_i[w^{\mathrm{T}} x_i + b]) + \lambda ||w^2||. \tag{1}$$

In Eq. (1), we have two terms: one is loss, and other is regularization. The first term is used to penalize misclassifications. It estimates the error because of misclassification, while the second one is used for regularization to avoid over-fitting. The λ represents the regularization coefficient.

4 Results and Discussions

This section provides the performance results of the artificial intelligence based system used for genome sequence classification. Mostly, classification accuracy is applied to estimate the system's performance; though, it is not enough to correctly/accurately evaluate the algorithms' performance. In this work, we applied different evaluation metrics to decide the performance of the above discussed system. We used different classification metrics such as True Positives, True Negatives, False Positives, and False Negatives defined as follows;

- True Positive (TP) for accurately predicted positive classes.
- False Positive (FP) for inaccurately predicted positive classes.
- True Negative (TN) for accurately predicted negative classes.
- False negative (FN) for inaccurately predicted negative classes.

Using the above matrices the following parameters are determined:

- Classification accuracy is the ratio of the total number of input samples and correct predictions. It operates properly if there are the same number of samples related to all classes. The classification accuracy is evaluated as follows:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

- Precision: It is determined as the fraction of important samples (True Positives) among all of the samples predicted to belong in a particular class give as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

- Recall is described as the fraction of samples that are predicted to belong to a class concerning to all of the samples that truly belong in the class calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

- True Positive Rate (TPR), also called (Sensitivity) is describe as proportion of positive samples that are correctly classified as positive. It is defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

- True negative rate TNR, is also known as (Specificity), the proportion of negative samples that are correctly classified as negative. It is mathematically given as:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6}$$

- False positive rate FPR, represents the number of negative samples mistakenly classified as positive. It is defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{7}$$

- F1-Score is defined as the harmonic mean between Recall and Precision. It mainly determines how accurate the classifier is (how many samples are correctly classified) and its robustness. The higher precision and lower recall values provide a very accurate, but it then misses many samples that are hard to classify. The higher value of the F1 Score, the better is the performance of the algorithm. It usually tries to maintain the balance between recall and precision. Mathematically, it can be estimated as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

We calculated precision, recall, accuracy, and F1-score, as shown in Fig. 10, We can see that the SVM classifier

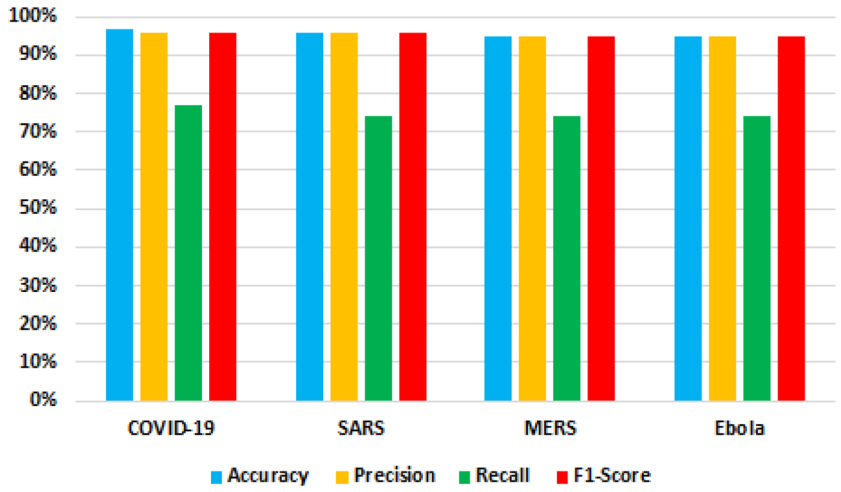


Fig. 10 Accuracy, Precision, Recall and F1 Score of classification method used of different genome sequences (COVID-19 and alike viruses)

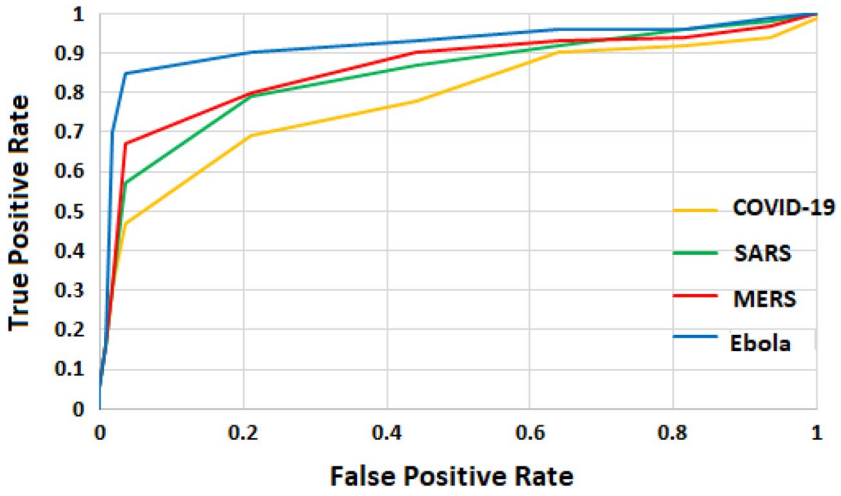


Fig. 11 Classification results using SVM (TPR vs FPR)

Table 1 Comparative analysis of different machine learning algorithms

S. no	Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
1	Extreme gradient boost	92	77	94	92
2	Random forest	94	76	93	96
3	Naive Bayes	94	70	92	96
4	K-nearest neighbour	95	77	94	95
5	Logistic regression	95	70	94	96
6	Decision tree	95	77	95	96
7	Support vector machine	97	77	97	97

achieves good classification accuracy results for all types of genome sequences, including COVID-19 with 97%, SARS with 96%, MERS, and Ebola with 95%, respectively. The Precision, Recall, and F1-Score is 96 , 77%, 96% for COVID-19, 96%, 74%, 96% for SARS, 95%, 74%, 95% for MERS and 95%, 74 % 95% for Ebola, respectively.

Using the above evaluation parameters, we have plotted ROC curve as shown in Fig. 11. The curve is plotted using the false positive rate FPR values versus the TPR for a defined cutoff value. The higher the ROC curve (i.e., SVM closer to the line $y = 1$), the better the fit. We can see from the Figure that the performance of all algorithms is better and nearly equals 95%, but the performance of SVM is better among all.

The comparative results of different machine learning algorithms is also shown in Table 1, it can be observed that

the performance of SVM classifier is good as compared to other machine learning algorithms.

5 Conclusion and Future Directions

In this paper, an artificial intelligence based system is presented to perform genome sequence analysis of COVID-19, SARS, MERS, and Ebola viruses. First, we performed comparative sequence analysis by extracting essential information of COVID-19 and alike viruses, including information of nucleotides composition and their frequency, tri-nucleotide compositions, count of amino acids, alignment between genome sequences, and their DNA similarity information. Second, we used different visualization methods to analyze these viruses' genome sequences. The comparative data analysis helps to get important information from the genome sequences of different viruses. Finally, we applied a machine learning based classifier, i.e., SVM, to classify different genome sequences. The data set of different virus genome sequences are obtained from National Center for Biotechnology Information Data Center repository. The proposed system achieves good classification results with an accuracy of 97% for COVID-19, 96% for SARS, and 95% for MERS, and Ebola 95%, respectively. In the future, we might extend this work for the analysis and classification of other genome sequences. Researchers are encouraged to use other artificial intelligence based techniques for the analysis and classification of genome sequences.

Acknowledgements This work was supported by Incheon National University Research Concentration Professors Grant in 2019.

References

1. Marquez S, Prado-Vivar B, Guadalupe JJ, Gutierrez B, Jibaja M, Tobar M, Mora F, Gaviria J, Garcia M, Espinosa F et al (2020) Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19 in Ecuador. medRxiv. https://doi.org/10.1101/2020.06.11.20128330
2. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri M, Hakmi M, Essabbar A, Laamarti M, Hlali H, Bendani H, Boumajdi N et al (2020) Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. PLoS One 15(11):e0240345. https://doi.org/10.1371/journal.pone.0240345
3. Leila M, Sorayya G (2021) Genotype and phenotype of COVID-19: their roles in pathogenesis. J Microbiol Immunol Infect 54(2):159–163. https://doi.org/10.1016/j.jmii.2020.03.022
4. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395(10224):565. https://doi.org/10.1016/S0140-6736(20)30251-8
5. Nawaz MS, Fournier-Viger P, Shojaee A, Fujita H (2021) Using artificial intelligence techniques for COVID-19 genome analysis. Appl Intell 51:3086–3103. https://doi.org/10.1007/s10489-021-02193-w
6. Raza S (2020) Artificial intelligence for genomic medicine. Tech. rep, London. https://www.phgfoundation.org/report/artificial-intelligence-for-genomic-medicine
7. Ahmed I, Ahmad M, Jeon G, Piccialli F (2021) A framework for pandemic prediction using big data analytics. Big Data Res 25:100190. https://doi.org/10.1016/j.bdr.2021.100190
8. Ahmed I, Ahmad M, Adnan A, Ahmad A, Khan M (2019) Person detector for different overhead views using machine learning. Int J Mach Learn Cybern 10(10):2657. https://doi.org/10.1007/s13042-019-00950-5
9. Ahmed I, Ahmad M, Nawaz M, Haseeb K, Khan S, Jeon G (2019) Efficient topview person detector using point based transformation and lookup table. Comput Commun 147:188–197. https://doi.org/10.1016/j.comcom.2019.08.015
10. Ahmed I, Ahmad M, Ahmad A, Jeon G (2020) Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: within 5G infrastructure. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-020-01220-5
11. Ullah K, Ahmed I, Ahmad M, Rahman AU, Nawaz M, Adnan A (2019) Rotation invariant person tracker using top view. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01526-5
12. Ahmed I, Din S, Jeon G, Piccialli F (2019) Exploring deep learning models for overhead view multiple object detection. IEEE Internet Things J 7(7):5737. https://doi.org/10.1109/JIOT.2019.2951365
13. Ahmed I, Ahmad M, Rodrigues JJ, Jeon G, Din S (2021) A deep learning-based social distance monitoring framework for COVID-19. Sustain Cities Soc 65:102571. https://doi.org/10.1016/j.scs.2020.102571
14. Ahmed I, Ahmad A, Jeon G (2020) An IoT based deep learning framework for early assessment of COVID-19. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2020.3034074
15. Ahmad M, Ahmed I, Khan FA, Qayum F, Aljuaid H (2020) Convolutional neural network-based person tracking using overhead views. Int J Distrib Sens Netw 16(6):1550147720934738. https://doi.org/10.1177/1550147720934738
16. Ahmed I, Anisetti M, Jeon G (2021) An IoT-based human detection system for complex industrial environment with deep learning architectures and transfer learning. Int J Intell Syst. https://doi.org/10.1002/int.22472
17. Wasim M, Ahmed I, Ahmad J, Hassan MM (2021) A novel deep learning based automated academic activities recognition in cyber-physical systems. IEEE Access 9:63718. https://doi.org/10.1109/ACCESS.2021.3073890
18. Ahmed I, Jeon G, Chehri A, Hassan MM (2021) Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. Sustain Cities Soc 70:102908. https://doi.org/10.1016/j.scs.2021.102908
19. Ahmed I, Ahmad M, Khan FA, Asif M (2020) Comparison of deep-learning-based segmentation models: using top view person images. IEEE Access 8:136361. https://doi.org/10.1109/ACCESS.2020.3011406
20. Ahmed I, Ahmad M, Ahmad A, Jeon G (2021) IoT-based crowd monitoring system: using SSD with transfer learning. Comput Electr Eng 93:107226. https://doi.org/10.1016/j.compeleceng.2021.107226
21. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Din MMU (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731. https://doi.org/10.1007/s41870-020-00495-9
22. Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, Barukab OM (2020) The number of confirmed cases of COVID-19 by

using machine learning: methods and challenges. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-020-09472-8
23. Pashazadeh A, Navimipour NJ (2018) Big data handling mechanisms in the healthcare applications: a comprehensive and systematic literature review. J Biomed Inform 82:47. https://doi.org/10.1016/j.jbi.2018.03.014
24. Yan L, Zhang HT, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Li S, Zhang M et al (2020) Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv. https://doi.org/10.1101/2020.02.27.20028027
25. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T et al (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Comput Mater Continua 63(1):537. https://doi.org/10.32604/cmc.2020.010691
26. Rao ASS, Vazquez JA (2020) Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. Infect Control Hosp Epidemiol 41(7):826. https://doi.org/10.1017/ice.2020.61
27. Chamola V, Hassija V, Gupta S, Goyal A, Guizani M, Sikdar B (2020) Disaster and pandemic management using machine learning: a survey. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2020.3044966
28. Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020) Epidemic analysis of COVID-19 in China by dynamical modeling. medRxiv. https://doi.org/10.1101/2020.02.16.20023465
29. Li Y, Liang M, Yin X, Liu X, Hao M, Hu Z, Wang Y, Jin L (2021) COVID-19 epidemic outside China: 34 founders and exponential growth. J Investig Med 69(1):52. https://doi.org/10.1101/2020.03.01.20029819
30. Khan SA, Khan MA, Song OY, Nazir M (2020) Medical imaging fusion techniques: a survey benchmark analysis, open challenges and recommendations. J Med Imaging Health Inform 10(11):2523. https://doi.org/10.1166/jmihi.2020.3222
31. Punn NS, Sonbhadra SK, Agarwal S (2020) COVID-19 epidemic analysis using machine learning and deep learning algorithms. medRxiv. https://doi.org/10.1101/2020.04.08.20057679
32. Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, Choi GS (2020) COVID-19 future forecasting using supervised machine learning models. IEEE Access 8:101489. https://doi.org/10.1109/ACCESS.2020.2997311
33. Ahmed I, Ahmad M, Jeon G (2021) Social distance monitoring framework using deep learning architecture to control infection transmission of COVID-19 pandemic. Sustain Cities Soc 69:102777. https://doi.org/10.1016/j.scs.2021.102777
34. Mateos PA, Balboa RF, Easteal S, Eyras E, Patel HR (2021) PACIFIC: a lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses. Sci Rep 11(1):1. https://doi.org/10.1038/s41598-021-82043-4
35. Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Mulders DG, Molenkamp R, Perez-Romero CA, Claassen E, Garssen J, Kraneveld AD (2021) Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Sci Rep 11(1):1. https://doi.org/10.1038/s41598-020-80363-5
36. Chamola V, Hassija V, Gupta V, Guizani M (2020) A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. IEEE Access 8:90225. https://doi.org/10.1109/ACCESS.2020.2992341
37. Rohmetra H, Raghunath N, Narang P et al (2021) AI-enabled remote monitoring of vital signs for COVID-19: methods, prospects and challenges. Computing. https://doi.org/10.1007/s00607-021-00937-7