Detection of COVID-19 Using Genomic Image Processing Techniques

Muhammed S. Hammad[1], Vidan F. Ghoneim[1], and Mai S. Mabrouk[2]
[1]Biomedical Engineering Department, Helwan University, Egypt.
[2]Biomedical Engineering Department, MUST University, Egypt.
Emails: muhammedsayed@h-eng.helwan.edu.eg, vidanfathighoneim@h-eng.helwan.edu.eg, msm_eng@yahoo.com

2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES) | 978-1-6654-2157-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/NILES53778.2021.9600525

***Abstract*—Novel Coronavirus Disease 2019 (COVID-19) is a new pandemic that appeared at the end of March 2019 in Wuhan city, China, which affected millions worldwide. COVID-19 is caused by the novel severe acute respiratory syndrome coronavirus 2 (SARSCoV-2) epidemic. Also, several viral epidemics have been listed in the last two decades, like the middle east respiratory syndrome coronavirus (MERSCoV) and the severe acute respiratory syndrome coronavirus 1 (SARSCoV-1), which cause MERS, and SARS diseases, respectively. Detection of these viral epidemics is a difficult issue because of their genetic similarity. In this paper, an effective automated system was developed to classify these viral epidemics using their complete genomic sequences via the genomic image processing techniques to facilitate the diagnosis and increase the detection accuracy in a short time. Results achieved an overall accuracy of 100% using two classifiers: SVM and KNN. However, the KNN classifier shows a privilege over the SVM in the execution time performance.**

Keywords—COVID-19, genomic image processing techniques, first-order features, SVM classifier, KNN classifier.

I. Introduction

Coronaviruses (CoVs) are a large family of viruses that cause illness ranging from a simple cold to more severe diseases. CoVs are viruses with single-stranded ribonucleic acid that infect mammals and animals [1]. The family is classified into four groups of viruses that include alpha-coronavirus (αCoV), beta-coronavirus (βCoV), delta-coronavirus (δCoV), and gamma-coronavirus (γCoV). The αCoV and βCoV infect mammals, while the δCoV and γCoV infect birds. NL63CoV and 229ECoV are the types of αCoV that result in simple respiratory marks like that in the common cold. On the other hand, HKU1CoV, OC43CoV, SARSCoV-1, and MERSCoV are the types of βCoV that result in mild to dangerous respiratory tract infections [1, 2].

A new βCoV coronavirus has appeared in China by the end of March 2019, transmitting from animals (Bats) to mammals. This virus was named a COVID-19 by the World Health Organization (WHO) in February 2020. The WHO announced the outbreak of COVID-19 as a global pandemic due to its spreading in many countries causing many deaths, at the same time [3]. Globally, as of 17 Sep 2021, there have been over 226.84 million confirmed cases of COVID-19, including over 4.66 million deaths, reported to WHO [4].

The SARSCoV-2 has about 79% similarity to the SARSCoV-1 and about 50% similarity to the MERSCoV. Therefore, detecting the SARSCoV-2 epidemic from other human coronavirus types is a difficult issue because of their genetic similarity [5]. Also, fever, headache, myalgia, shortness of breath, and dry cough are the most common signs of COVID-19 [6]. Most of these signs like those of the common flu. Therefore, diagnosing SARSCoV-2 at an early stage is a complex problem. Rapid identification of such positive cases is essential because the disease spreads quickly and poses a hazard to the public system.

X-ray and computed tomography (CT) scans are types of medical imaging modalities that are considered among the most effective approaches for COVID-19 identification. Recently, machine and deep learning models, two significant areas of artificial intelligence, have become very popular in medical applications and can be used for COVID-19 detection through the processing of X-ray and CT images [7-11].

Ozturk et al. [7] presented a multi-classification system to classify COVID-19, healthy, and pneumonia cases using a deep learning model (DarkNet model) and chest X-ray images. The accuracy of the system was 87.02%. Another multi-classification system was developed by Elasnaoui et al. [8] to classify coronavirus, COVID-19, bacterial pneumonia, and normal cases. They used the deep learning model to extract features from X-ray and CT images and the multilayer perceptron to classify the images. They reached an accuracy of 92.18%. Jain et al. [9] used different deep learning models (VGG-16, DenceNet121, ResNet101, and ResNet18) to extract features and classify COVID-19 from bacterial pneumonia, viral pneumonia, and normal cases using chest X-ray images. ResNet101 model resulted in the highest accuracy of 98.93%. Barstugan et al. [10] used 150 abdominal CT images of different sizes to detect COVID-19 from other viral pneumonia types. They used five feature extraction methods with the SVM classifier giving an accuracy of 99.6%. Based on the domain extension transfer learning model, different characterized features were extracted from chest X-ray images by Basu et al. [11] to classify COVID-19, pneumonia, other diseases, and normal cases. Their system achieved an accuracy of 90.13%.

Although these imaging techniques play a vital role in controlling the COVID-19 pandemic and are accurate and large extent, they have many drawbacks. These techniques give the patient a dose of radiation, which has many hazards, especially for pregnant women.

Hilan [12] presented a binary classification system based on extracting some features from complete genomic sequences to classify COVID-19, among other types of human coronaviruses. These features are used as input for different supervised classifiers. Their experimental results proposed that the decision tree achieved the highest accuracy of 93%.

On the other hand, the reverse transcription-polymerase chain reaction (RT-PCR) test is the optimum approach for COVID-19 detection. However, the lack of resources and strict test environment requirements make it difficult to screen suspected cases quickly and effectively, especially when the patient population is large. Also, in some cases, RT-PCR examination results in false-negative rates [13, 14].

Recently, genomic signal processing (GSP) techniques are used for COVID-19 detection. They convert the genetic sequence from G (Guanine), A (Adenine), (Cytosine), and T (Thymine) characters into one dimensional vector using various mapping methods [15, 16].

Randhawa et al. [15] developed a system to classify SARSCoV-2, αCoV, βCoV, and δCoV cases using a GSP technique with machine learning models. The used models were SVM (using linear and quadratic kernels), KNN (Fine, and subspace), subspace discriminant, and linear discriminant analysis (LDA). They used 29 COVID-19 cases and 20 cases for each one of the other types. The LDA classifier achieved an accuracy of 100%. The main limitation of their approach is that they used δCoV genomes. However, the δCoV mainly infect bird hosts rather than human. Also, they used a small number of genome sequences in their research. Therefore, the system accuracy may decrease with large dataset size.

Naeem et al. [16] proposed an approach to classify SARSCoV-2, MERSCoV, and SARSCoV-1 sequences using a GSP technique with two supervised classifiers. Seventy-six complete genomic sequences were used in their study for each type of human coronaviruses. Their system achieved an accuracy of 100% with the KNN classifier. The small number of genome sequences is the main limitation of their study.

Nowadays, genomic image processing (GIP) techniques [17, 18] start to be used to detect different genetic diseases to avoid the drawbacks of the previously described diagnostic methods. GIP is an engineering field that has been defined as the processing of the images extracted from the genetic sequences using different mapping techniques to obtain biological knowledge and the translation of that knowledge into systems that have great consideration in the detection of the various genetic diseases [17-20]. Our approach mainly depends on employing a GIP technique with machine learning models.

In this study, an effective automated system was developed to identify the three most dangerous epidemics (SARSCoV-2, MERSCoV, and SARSCoV-1) of humans that resulted in many deaths in the last two decades using a GIP technique. The remainder of the paper is organized as follows; Section II explains the database, the used genomic mapping technique, and the features that are extracted from the genomic images. Section III represents the results and the discussion of the proposed approach. Lastly, Section IV represents the paper's conclusion and the future work.

II. Methedology

A block diagram for the main stages of the proposed approach was shown in Fig.1. First, complete genomic sequences of the different viral diseases are downloaded. Second, the downloaded sequences are transformed into a two-dimensional image using the gray level representation technique. Third, first-order features are extracted from the genomic images which are used to train and test different machine learning models based on the selected features. Finally, different effective parameters are used to evaluate the performance of the proposed approach.

A. Dataset

In this study, complete genomic sequences (about 30,000 nucleotides) of human coronaviruses are downloaded from the National Center for Biotechnology Information (NCBI) [21] and Virus Pathogen Database and Analysis Resource (ViPR) [22] websites. We used 300, 258, 57 genome sequences of SARSCoV-2, MERSCoV, and SARSCoV-1 epidemics, respectively. We note that all available complete genomic sequences of human coronaviruses related to MERSCoV, and SARSCoV-1 epidemics are downloaded and used in the study.

B. Genomic Mapping Technique

The genomic mapping technique is the process that converts the DNA sequence from characters into a two-dimensional image. In this paper, we used the single nucleotide gray level representation technique [18].

The genomic image is created by converting the characters of the genetic sequence (A, T, C, and G) into numbers. The characters of the DNA sequence are represented by values that are equally spaced between the gray level values from 1 to 255 for A, T, C, and G letters, respectively [18]. In this study, the image width is chosen to be 200 pixels. Therefore, the first 200 characters of the sequence represent the first row of the image, the second 200 characters represent the second row, and so on till the last base in the sequence. In this way, the genetic sequence is converted to a grayscale image with a width equal to 200 pixels, and with height changes according to the sequence length. The conversion of a COVID-19 sequence (partial genome (1539 nucleotides)) from FASTA format to a genomic grayscale image was shown in Fig. 2.

C. Feature Extraction and Selection

First-order statistical features are estimated directly from the pixel's values of the original image, regardless of the spatial relationship between them. The image is a function of two variables h, and w, where h represents the rows of the image, h=0, 1, 2, 3, …. X-1, X is the row length, while w represents the columns of the image, w=0, 1, 2, 3, …. Y-1, Y is the column length. The function has intensity values of G = 0, 1, 2, …. T-1, where T is the number of intensity levels of the image [23].

The histogram of each grey level H(G) is calculated by dividing the number of pixels with grey level (G) by the total number of image pixels (XY), as shown in the following equation [23]:

$$H(G) = \frac{Number\ of\ Pixels\ with\ intensity\ (G)}{Total\ number\ of\ pixels\ (XY)} \tag{1}$$

The first-order statistical features are extracted from the genomic images using various parameters, called the central moments that characterize the DNA sequences of the different viral diseases [23]. The most used central moments are the mean (μ), variance (μ_2), skewness (μ_3), kurtosis (μ_4), and entropy (E) given by the following equations [23]:

$$\mu = \sum_{G=0}^{T-1} G\, H(G) \tag{2}$$

$$\sigma^2 = \mu_2 = \sum_{G=0}^{T-1} (G-\mu)^2\, H(G) \tag{3}$$

$$\mu_3 = \sigma^{-3} \sum_{G=0}^{T-1} (G-\mu)^3\, H(G) \tag{4}$$

$$\mu_4 = [\sigma^{-4} \sum_{G=0}^{T-1} (G-\mu)^4\, H(G)] - 3 \tag{5}$$

$$E = -\sum_{G=0}^{T-1} H(G)\, log_2[H(G)] \tag{6}$$

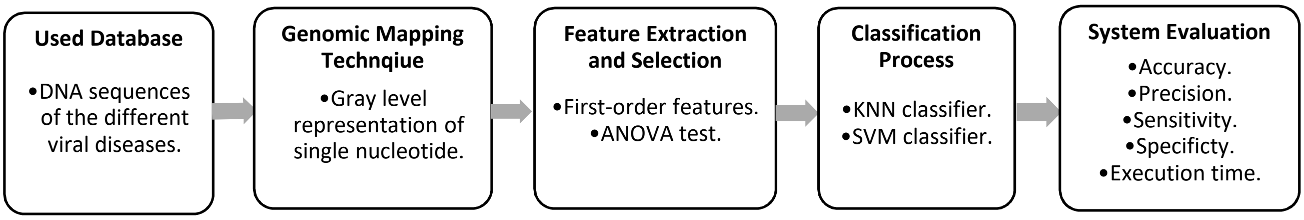
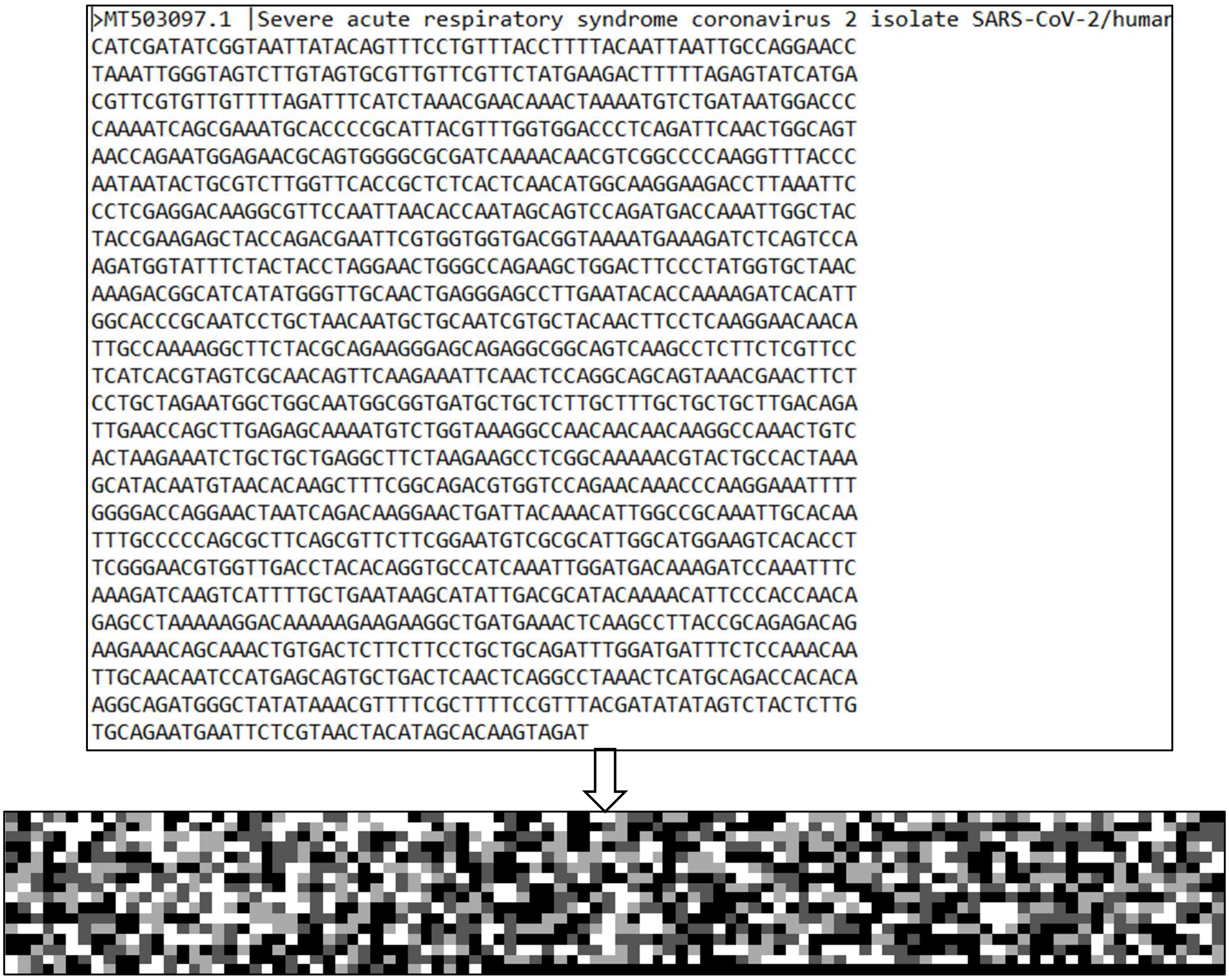


Fig. 1. Block diagram of the proposed approach.



```
>MT503097.1 |Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/humar
CATCGATATCGGTAATTATACAGTTTCCTGTTTACCTTTTACAATTAATTGCCAGGAACC
TAAATTGGGTAGTCTTGTAGTGCGTTGTTCGTTCTATGAAGACTTTTTAGAGTATCATGA
CGTTCGTGTTGTTTTAGATTTCATCTAAACGAACAAACTAAAATGTCTGATAATGGACCC
CAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGT
AACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCC
AATAATACTGCGTCTTGGTTCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTC
CCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAATTGGCTAC
TACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCA
AGATGGTATTTCTACTACCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAAC
AAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAATACACCAAAAGATCACATT
GGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACA
TTGCCAAAAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCC
TCATCACGTAGTCGCAACAGTTCAAGAAATTCAACTCCAGGCAGCAGTAAACGAACTTCT
CCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGACAGA
TTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTC
ACTAAGAAATCTGCTGCTGAGGCTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAA
GCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAACCCAAGGAAATTTT
GGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAA
TTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCT
TCGGGAACGTGGTTGACCTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAATTTC
AAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACA
GAGCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAG
AAGAAACAGCAAACTGTGACTCTTCTTCCTGCTGCAGATTTGGATGATTTCTCCAAACAA
TTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAAACTCATGCAGACCACACA
AGGCAGATGGGCTATATAAACGTTTTCGCTTTTCCGTTTACGATATATAGTCTACTCTTG
TGCAGAATGAATTCTCGTAACTACATAGCACAAGTAGAT
```

Fig.2. Conversion of a COVID-19 sequence from FASTA format into a genomic image using the gray level representation technique.

The average intensity value of the image is the mean. The variance refers to the deviation of the grayscales from the mean. The skewness is a measure of image symmetry. The kurtosis is the histogram flatness. Finally, the entropy is a measure of the image's randomness [18, 23, 24]. The feature vector (V) was designed using the previous parameters. Therefore, V is represented in the following manner:

$$V = [\mu, \mu_2, \mu_3, \mu_4, E] \quad (7)$$

The one-way analysis of variance (ANOVA) test is used to estimate the significance of the extracted features among the different viral diseases. The statistically significant features have been used as an input to the classifiers. We used in this paper two classifiers which are KNN (K=1) and SVM (using linear kernel). A cross-validation of tenfold is used to split the dataset into ten train and test datasets, and the mean of twenty runs is reported to evaluate the performance of the proposed approach.

III. Results and Discussion

All the work done in this paper is implemented on a laptop with a 2.5 GHz Intel Core i5 Processor and 16 GB RAM under Windows operating system. The conversion of genetic sequences into genomic images, feature extraction and selection, and classification process stages were performed under MATLAB - R2020a programming language.

To evaluate the performance of the multi-class classification system, some evaluation parameters are calculated at a macro averaging level. These parameters are the accuracy (A), precision (P), sensitivity (Se), and specificity (Sp) given by the following equations [25]. The execution time of feature extraction and selection, and classification process stages was also estimated.

$$A = \frac{\sum_{i=1}^{C} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{C} \; x100 \;\; \% \quad (8)$$

TABLE I. RESULTS OF SVM AND KNN CLASSIFIERS.

Evaluation Parameter	SVM	KNN
A (%)	100	100
P (%)	100	100
Se (%)	100	100
Sp (%)	100	100
Execution Time (s)	19.95	10.93

$$P = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}}{C} \quad x100 \quad \% \tag{9}$$

$$Se = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}}{C} \quad x100 \quad \% \tag{10}$$

$$Sp = \frac{\sum_{i=1}^{C} \frac{TN_i}{TN_i + FP_i}}{C} \quad x100 \quad \% \tag{11}$$

Where C is the number of classes (C=3), TN and TP represent the True Negative and True Positive values of correctly identified classes, respectively, and FN and FP represent the False Negative and False Positive values of incorrectly identified classes, respectively.

The ANOVA test showed that all first-order features were statically significant (p-value <0.05). Therefore, all features were used to train and test the SVM and KNN classifiers. The system performance was reported in Table I.

As shown from Table I, all the results of SVM and KNN classifiers are acceptable and satisfying, but the KNN results are perfect for the developed classification system in terms of obtaining the best efficient diagnosing technique with a short execution time.

IV. CONCLUSIONS

COVID-19 is a new βCoV pandemic that had infected over 226.84 million people worldwide, and killed over 4.66 million, making it the most serious global health crisis since the 1918 influenza pandemic. The detection of SARSCoV-2 epidemic, which is responsible for COVID-19 from other human viral epidemics, is a complicated issue because of their genetic similarity.

In this study, an effective automated system was developed to detect human SARSCoV-2 epidemic among two types of human coronaviruses which are SARSCoV-1 and MERSCoV epidemics using a GIP technique with machine learning models. The proposed system can assist in the fast diagnosis of COVID-19 with reliable results achieving 100% accuracy while avoiding the drawbacks of the previously described diagnostic methods. In future work, we aim to develop a complete classification system to classify the SARSCoV-2 cases among the other six types of human coronaviruses, which are 229ECoV, NL63CoV, HKU1CoV, OC43CoV, SARSCoV-1, and MERSCoV.

V. REFERNCES

[1] S. Ludwig and A. Zarbock, “Coronaviruses and SARS-CoV-2: a brief overview,” Anesthesia & Analgesia, vol. 131, no. 1, pp. 93–96, 2020.

[2] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, “The proximal origin of SARS-CoV-2,” Nature Medicine, vol. 26, no. 4, pp. 450–452, 2020.

[3] L. Pan et al., “Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study,” American Journal of Gastroenterology, vol. 115, no. 5, pp. 766–773, 2020.

[4] “Egypt: WHO coronavirus disease (COVID-19) dashboard with vaccination data,” World Health Organization. [Online]. Available: https://covid19.who.int/region/emro/country/eg. [Accessed: 17-Sep-2021].

[5] R. Lu et al., “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding,” The Lancet, vol. 395, no. 10224, pp. 565–574, 2020.

[6] L. Fu et al., “Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis,” Journal of Infection, vol. 80, no. 6, pp. 656–665, 2020.

[7] T. Ozturk et al., “Automated detection of COVID-19 cases using deep neural networks with X-ray images,” Computers in Biology and Medicine, vol. 121, p. 103792, 2020.

[8] K. El Asnaoui and Y. Chawki, “Using X-ray images and deep learning for automated detection of coronavirus disease,” Journal of Biomolecular Structure and Dynamics, vol. 39, no. 10, pp. 3615–3626, 2020.

[9] G. Jain, D. Mittal, D. Thakur, and M. K. Mittal, “A deep learning approach to detect COVID-19 coronavirus with X-Ray images,” Biocybernetics and Biomedical Engineering, vol. 40, no. 4, pp. 1391–1405, 2020.

[10] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning methods," arxiv:2003.09424, 2020.

[11] S. Basu, S. Mitra, and N. Saha, “Deep learning for screening COVID-19 using chest X-Ray images,” arXiv:2004.10507, 2020.

[12] H. Arslan, “Machine learning methods for COVID-19 prediction using human genomic data,” Proceedings, vol. 74, no. 1, p. 20, 2021.

[13] B. Udugama et al., “Diagnosing COVID-19: the disease and tools for detection,” ACS Nano, vol. 14, no. 4, pp. 3822–3835, 2020.

[14] A. Tahamtan and A. Ardebili, “Real-time RT-PCR in COVID-19 detection: issues affecting the results,” Expert Review of Molecular Diagnostics, vol. 20, no. 5, pp. 453–454, 2020.

[15] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study,” PLOS ONE, vol. 15, no. 4, 2020.

[16] S. M. Naeem, M. S. Mabrouk, S. Y. Marzouk, and M. A. Eldosoky, “A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19,” Briefings in Bioinformatics, vol. 22, no. 2, pp. 1197–1205, 2020.

[17] L. A. Santamaria, S. Zuuiga, I. H. Pineda, M. J. Somodevilla, and M. Rossainz., “DNA sequence recognition using image representation,” Research in Computing Science, vol. 148, no. 3, pp. 105–114, 2019.

[18] E. Delibas and A. Arslan, “DNA sequence similarity analysis using image texture analysis based on first-order statistics,” Journal of Molecular Graphics and Modelling, vol. 99, p. 107603, 2020.

[19] S. Mizuta, “Graphical representation of biological sequences,” Bioinformatics in the Era of Post Genomics and Big Data, 2018.

[20] M. Randic, M. Vracko, N. Lers, and D. Plavsic, “Novel 2-D graphical representation of DNA sequences and their numerical characterization,” Chemical Physics Letters, vol. 368, pp. 1–6, 2003.

[21] National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed: 12-Jul-2021].

[22] Virus Pathogen Database and Analysis Resource (ViPR). [Online]. Available: https://www.viprbrc.org/. [Accessed: 12-Jul-2021].

[23] A. Materk and M. Strzelecki, “Texture analysis methods a Review,” Report, Technical University of Lodz, Institute of Electronics, 1998.

[24] S. Alobaidli et al., “The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning,” The British Journal of Radiology, vol. 87, no. 1042, p. 20140369, 2014.

[25] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.