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

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

Muhammed S. Hammad a,∗, Mai S. Mabrouk b, Walid I. Al-atabany c,d, Vidan F. Ghoneima a

a Biomedical Engineering Department, Helwan University, Egypt
b Biomedical Engineering Department, Misr University for Science and Technology (MUST), Egypt
c Center for Informatics Science, Nile University, Egypt
d Department of Biomedical Engineering, Helwan University, Egypt

Received 27 February 2022; revised 25 July 2022; accepted 15 August 2022
Available online 22 August 2022

KEYWORDS
Human coronavirus diseases; Genome sequences; Genomic image processing techniques; Statistical features; Machine learning classifiers

Abstract Coronavirus (CoV) disease 2019 (COVID-19) is a severe pandemic affecting millions worldwide. Due to its rapid evolution, researchers have been working on developing diagnostic approaches to suppress its spread. This study presents an effective automated approach based on genomic image processing (GIP) techniques to rapidly detect COVID-19, among other human CoV diseases, with high acceptable accuracy. The GIP technique was applied as follows: first, genomic graphical mapping techniques were used to convert the genome sequences into genomic grayscale images. The frequency chaos game representation (FCGR) and single gray-level representation (SGLR) techniques were used in this investigation. Then, several statistical features were obtained from the images to train and test many classifiers, including the k-nearest neighbors (KNN). This study aimed to determine the efficacy of the FCGR (with different orders) and SGLR images for accurately detecting COVID-19, using a dataset containing both partial and complete genome sequences. The results recommended the fourth-order FCGR image as a proper genomic image for extracting statistical features and achieving accurate classification. Furthermore, the results showed that KNN achieved an overall accuracy of 99.39% in detecting COVID-19, among other human CoV diseases, with 99.48% precision, 99.51% sensitivity, 99.47% specificity, 0.99 F1-score, and 0.99 Matthew’s correlation coefficient.

© 2022 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Coronaviruses (CoVs) are a broad group of viruses that cause various illnesses, ranging from the common cold to more severe diseases. These viruses infect mammals and animals and have single-stranded ribonucleic acid (RNA). CoVs are
divided into four variants: gamma-CoV (γCoV), delta-CoV (δCoV), beta-CoV (βCoV), and alpha-CoV (αCoV). Mammmalian infections of CoVs are caused by the αCoV and βCoV, whereas avian infections of CoVs are caused by the δCoV and γCoV [1,2].

The αCoV is classified into two groups: human CoV 229E (HCoV-229E) and human CoV NL63 (HCoV-NL63). The βCoV is classified into four groups: human CoV OC43 (HCoV-OC43), Middle East respiratory syndrome CoV (MERS-CoV), severe acute respiratory syndrome CoV 1 (SARS-CoV-1), and human CoV HKU1 (HCoV-HKU1). The HCoV-229E, HCoV-NL63, HCoV-HKU1, and HCoV-OC43 variants of CoVs can cause mild respiratory infections, whereas the MERS-CoV and SARS-CoV-1 variants of CoVs can cause fatal respiratory infections [3,4].

COVID-19 is a severe pandemic caused by a novel βCoV known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [5,6]. Globally, as of July 22, 2022, COVID-19 had resulted in 565.21 million confirmed cases, with 6.37 million deaths as reported by the World Health Organization [7].

SARS-CoV-2 shares roughly 50% similarity with MERS-CoV and 79% similarity with SARS-CoV-1 [8]. Therefore, it is difficult to detect SARS-CoV-2 from other human CoV strains. The most prevalent COVID-19 symptoms include fever, breath shortness, headache, and myalgia [9]; however, most of these symptoms are also common flu symptoms. Thus, detecting COVID-19 at an early stage is a difficult task. Once detected, it is crucial to treat the positive cases of COVID-19 because it rapidly spreads and threatens public health.

1.1. Background and related work

Medical imaging scans are among the most effective tools for detecting different diseases (e.g., COVID-19), as the images produced by these scans are analyzed using artificial intelligence and deep learning approaches [10–12]. Using these approaches to build a computer-based diagnostic system helps in the early detection of COVID-19 [13–15], thus reducing healthcare workers’ stress.

Aswathy et al. [13] developed a system to detect COVID-19 from computed tomography (CT) images using a pre-trained ResNet-50 convolution neural network (CNN) model. When the system’s output was a COVID-19, they classified the severity of the COVID-19 into low, medium, and high levels using the deep features obtained from the DenseNet-201 and ResNet-50 CNN models. The optimized features obtained from the principal component analysis algorithm were fed to a back-propagation neural network for the classification process. The proposed system detected COVID-19 with an accuracy of 98.5% and predicted the severity of COVID-19 with 97.48% accuracy.

Aswathy et al. [14] presented a three-dimensional (3D) UNet design for extracting the COVID-19 infection area of the patient’s lung using CT images. The design is divided into two stages: the first segments the lung tissue, and the second segments the infected 3D volumes from the CT volume input. The approach allows clinicians to analyze the infected area of the patient’s lung without labeling the lung parenchyma for each new patient. The proposed approach resulted in an accuracy of 98.07%.

Gunraj et al. [15] presented a system based on CNN models to detect COVID-19 among pneumonia and normal control from chest CT images. They used four pre-trained CNN networks: EfficientNetB0, SqueezeNet, NASNet-A-Mobile, and MobileNetV2. They also developed two new CNN models: COVID-Net CT L and COVID-Net CT S. The COVID-Net CT S model achieved an accuracy of 98.4%. However, the EfficientNetB0 and MobileNetV2 models detected COVID-19 with 99% accuracy. The COVID-Net CT S model has fewer parameters and faster processing than other models.

The previously described classification systems based on medical imaging scans [13–15] are highly accurate and play a significant role in treating the COVID-19 pandemic. However, they have significant drawbacks, such as exposing the patient to a high radiation dose, which can cause serious health complications, especially in pregnant women.

Besides medical imaging scans, Suma et al. [16] presented a system based on the clinical data from 65,500 patients from different countries to detect and predict the severity of COVID-19 into mild, moderate, and severe levels. The clinical data consists of 26 features, including personal information and disease symptoms. The optimized features were obtained from the artificial bee colony optimization algorithm to detect COVID-19. These features were fed into several classifiers such as the support vector machine (SVM), naive Bayes (NB), and random forest. The SVM resulted in the best performance with 96% accuracy. The severity features were fed into the logistic regression (LR) to detect the severity of COVID-19. The LR resulted in a receiver operating characteristic of 0.96, 0.85, and 0.78 for the mild, moderate, and severe classes, respectively.

Moreover, molecular techniques, such as the reverse transcription–polymerase chain reaction (RT–PCR) tests, are the gold standard approaches for COVID-19 detection [17]. However, insufficient resources for conducting the RT–PCR tests reduce the speed and efficiency of screening suspected cases, which becomes a difficult issue, especially with a large patient population. Several studies [18,19] have shown that RT–PCR examinations have high false-positive and false-negative rates [20].

Besides the medical imaging scans, clinical data, and molecular techniques, various studies have detected COVID-19 from its genome sequences (SARS-CoV-2). Arslan and Arslan [21] extracted CpG-based features directly from the complete genome sequences of human CoVs to detect COVID-19, among other human CoV diseases. These features were fed into the k-nearest neighbors (KNN) classifier. They implemented the KNN with 19 distance metrics categorized into five classes: the inner product, L1, L2, vicisitude, and other metrics. With the L1 distance metric, the proposed approach resulted in an accuracy of 98.4%.

Adetiba et al. [22] presented a system based on CNN models to detect COVID-19 among MERS and SARS diseases. They used a genomic signal processing (GSP) technique based on a genomic signal mapping technique called the Z-curve to convert the genome sequences into genomic signals. These signals were converted into 3D matrices that were transformed into Z-curve images, which were used for the classification process. The system resulted in 98.33% accuracy.

Naeem et al. [23] developed a multiclass classification system based on GSP techniques to classify the SARS, COVID-19, and MERS diseases using supervised classification algo-
gorithms, including the KNN. First, they converted the complete genome sequences of human CoVs into genomic signals using the electron–ion interaction pseudopotential representation technique. Thereafter, they extracted nine features from these signals using the seven-moment invariants, discrete cosine transform, and discrete Fourier transform methods. The KNN classifier gave the highest accuracy of 100% using the train–test split strategy.

This study uses genomic image processing (GIP) techniques to detect COVID-19, among other human CoV diseases. GIP is a branch of bioinformatics that converts the genome sequences from the string form (G, T, C, and A) into genomic images using various genomic graphical mapping approaches [24,25]. It then processes these images using digital image processing tools to gain biological knowledge and translate that knowledge into value systems that detect various genetic diseases (e.g., COVID-19) [26].

In 2021, the same working group of the current work presented a system to classify the SARS, COVID-19, and MERS diseases using their complete genome sequences [26]. First, the genome sequences of human CoV diseases were transformed into two-dimensional (2D) images. After that, different first-order features were obtained from these 2D images and fed into the KNN and SVM algorithms. Both classifiers resulted in an average accuracy of 100%. However, the KNN was preferred because it outperforms the SVM in terms of system execution time. Table 1 shows the results of the related work in terms of the study, best approach, dataset, and accuracy of each study.

### 1.2. Main contribution

Previous research on COVID-19 [21–23,26] used only complete genome sequences in their studies. Therefore, these systems have limitations in analyzing partial genome sequences of various human CoV diseases. For instance, if the system input is a partial COVID-19 sequence, the system cannot predict whether the patient has COVID-19 or not. However, such a system works well with only complete COVID-19 sequences. In addition, the used datasets excluded some types of human CoV diseases that are genetically like COVID-19.

To avoid the drawbacks and limitations of the previous COVID-19 studies [21–23,26], an effective automated approach was developed to detect COVID-19, among other human CoV diseases, using both partial and complete genome sequences of the different human CoV diseases. The study’s contributions are as follows:

- First, GIP techniques that link bioinformatics and digital image processing tools are used to determine the best approach required for extracting statistical features to detect COVID-19, among other human CoV diseases, with high acceptable accuracy.

### Table 1 Related work results.

| Study               | Best approach                                                                 | Dataset                                    | Accuracy (%) |
|---------------------|-------------------------------------------------------------------------------|--------------------------------------------|--------------|
| Medical imaging scans and clinical data |                                                                           |                                            |              |
| Aswathy et al. [13] | COVID-19 detection: pre-trained ResNet-50 CNN model for feature extraction and classification | COVID-19: 760 Non-COVID-19: 736 | Detection: 98.5, Severity: 97.48 |
|                     | Severity detection: DenseNet-201 and ResNet-50 CNN models with back propagation neural network |                                            |              |
| Aswathy et al. [14] | Two stages of three-dimensional U-Net architecture                           | COVID-19: 20                               | 98.07        |
| Gunraj et al. [15]  | MobileNetV2 CNN for feature extraction and classification                    | COVID-19: 60,083 Pneumonia:40,291 Normal:100,729 | 99           |
| Suma et al. [16]    | Personal information and disease symptoms features                          | 65,500 patients                            | 96           |
|                     | Artificial bee colony algorithm Support vector machine                       |                                            |              |
| Complete genome sequences of human coronavirus diseases |                                                                           |                                            |              |
| Arslan et al. [21]  | CpG-based features with the KNN classifier implemented with L1 distance     | COVID-19: 1000 MERS-CoV: 258 HCoV-229E: 27 | 98.4         |
|                     | MERS-CoV: 140 HCoV-NL63: 61 HCoV-HKU1: 18                                      |                                            |              |
|                     | Adetiba et al. [22] CNN model for feature extraction and classification stages | COVID-19: 20 MERS-CoV: 20 SARS-CoV-1: 20 | 98.3         |
| Naeem et al. [23]   | Seven-moment invariants, discrete cosine transform, and discrete Fourier transform features with the KNN classifier | COVID-19: 76 MERS-CoV: 76 SARS-CoV-1: 76 | 100          |
| Hammad et al. [26]  | First-order features include the mean, variance, skewness, kurtosis, and entropy with the KNN classifier. | COVID-19: 300 MERS-CoV: 258 SARS-CoV-1: 57 | 100          |

CNN, Convolution neural network; KNN, K-nearest neighbors.
Second, the frequency chaos game representation (FCGR) and single gray-level representation (SGLR) techniques are used to convert the genome sequences into genomic grayscale images. To the best of our knowledge, none of the earlier COVID-19 studies used the FCGR images (with different orders) to extract first- and second-order statistical features to detect COVID-19.

Third, a perfect dataset containing both partial and complete genome sequences of human CoV diseases is constructed, thereby demonstrating the strength and efficiency of the proposed approach.

Fourth, the study’s dataset contains almost all types of human CoV diseases that are genetically like COVID-19.

Finally, the proposed approach can rapidly diagnose COVID-19 with high performance while avoiding the drawbacks and limitations of the traditional diagnostic approaches.

The remaining sections of the paper are divided into three sections. Section 2 describes the dataset’s preparation and the proposed approach’s flowchart. Section 3 shows and discusses the obtained results and compares the proposed approach to the earlier COVID-19 studies. The last section gives the conclusions of the study.

2. Materials and methods

Fig. 1 represents the flowchart of the proposed approach. First, the genome sequences of human CoV diseases were downloaded and processed to remove any bad nucleotide (e.g., N bases) from the genome sequences, leaving only the basic nucleotides (G, T, C, and A bases). Second, the processed sequences were converted into genomic grayscale images using the FCGR and SGLR techniques. Third, different first- and second-order statistical features were obtained from the grayscale images to train and test several classifiers. Finally, the performance of the proposed approach was evaluated using several effective parameters. The following sections explain each stage in detail.

2.1. Dataset

The genome sequences of the seven human CoVs were obtained from the National Center for Biotechnology Information [27] website: SARS-CoV-2, MERS-CoV, HCoV-229E, HCoV-NL63, SARS-CoV-1, HCoV-HKU1, and HCoV-OC43. SARS-CoV-2 sequences were labeled as COVID-19 Pos, and other human CoV sequences were labeled as COVID-19 Neg. The partial and complete genome sequences of human CoV diseases other than SARS-CoV-2 were downloaded and used in this study. As a result, a total of 7331 genome sequences were collected, resulting in 3700 COVID-19 sequences and 3631 non-COVID-19 sequences.

The downloaded genome sequences of human CoV diseases are complementary deoxyribonucleic acid (cDNA) sequences. The cDNA sequence is DNA synthesized from the single-stranded RNA CoV in a reaction catalyzed by the enzyme reverse transcriptase. The specifications of these cDNA sequences were described as follows: sequence data (nucleotides), virus (e.g., SARS-CoV-2 for COVID-19), nucleotide completeness (partial and complete), length (ranged from 20,000 to 30,000 base pairs), host (homo sapiens), and geographic region (all). Table 2 lists the properties of human CoV sequences.

2.2. Genomic graphical mapping techniques

Genomic graphical mapping techniques transform the genome sequence from the string form (A, T, C, and G bases) into a 2D genomic image. In this study, the FCGR and SGLR techniques were used.

2.2.1. Frequency chaos game representation technique

The FCGR approach offers a 2D matrix indicating the frequency of the k-mers obtained from the genome sequences [28–30]. The term k-mer represents the subsequences of length k in a given genome sequence. The sequence TGCAT, for example, contains five 1-mers (T, G, C, A, and T), four 2-
mers (TG, GC, CA, and AT), three 3-mers (TGC, GCA, and CAT), two 4-mers (TGCA and GCAT), and one 5-mer (TGCAT). The FCGR matrix was derived directly from the genome sequence by counting the frequency of each k-mer and then placing that frequency into its appropriate location in the FCGR matrix [30].

Each k-mer is positioned in the FCGR matrix according to the following algorithm: for each character in a k-mer, the FCGR matrix is subdivided into four quadrants, with A in the bottom left, T in the bottom right, C in the top left, and G in the top right. Each quadrant was further split according to the same principle for the next character in the k-mer [30], recursively, as shown in Fig. 2.

For a sequence S with the frequency of each k-mer (A, T, C, and G bases) represented by F_k, the first-order (k = 1) FCGR matrix is given by the following equation [30]:

\[
FCGR_1(S) = \begin{pmatrix}
F_C & F_G \\
F_A & F_T
\end{pmatrix}
\]  

(1)

Moreover, the kth-order FCGR matrix can be estimated by substituting each F_k element with the four elements (C, G, A, and T bases), using the following equation [30]:

\[
FCGR_k(S) = \begin{pmatrix}
F_{CC} & F_{GC} & F_{CG} & F_{GG} \\
F_{AC} & F_{TC} & F_{AG} & F_{GT} \\
F_{CA} & F_{GA} & F_{CT} & F_{GT} \\
F_{AA} & F_{TA} & F_{AT} & F_{TT}
\end{pmatrix}
\]  

(2)

Therefore, the second-order (k = 2) FCGR matrix is given by Eq. (3), and higher-order FCGR matrices can be sequentially computed [30].

\[
FCGR_2(S) = \begin{pmatrix}
F_{CC} & F_{GC} & F_{CG} & F_{GG} \\
F_{AC} & F_{TC} & F_{AG} & F_{GT} \\
F_{CA} & F_{GA} & F_{CT} & F_{GT} \\
F_{AA} & F_{TA} & F_{AT} & F_{TT}
\end{pmatrix}
\]  

(3)

From Eqs. (1) and (3), it can be seen that the FCGR matrix of order k is a 2^k x 2^k matrix, and it contains 4^k occurrences of the oligonucleotides with a length of k (total possible k-mers). The frequencies of the FCGR matrix were normalized in the range of 0 and 255 using Eq. (4) to build a genomic grayscale image with different gray levels from black (least frequent) to white (most frequent).

\[
NF_k = \frac{F_k - MinF}{MaxF - MinF} \times 255,
\]

(4)

where F_k and NF_k are the original and normalized frequencies of each k-mer in the genome sequence, respectively, and MinF and MaxF represent the minimum and maximum frequencies, respectively.

The genome sequences are converted into uniform-sized images with the FCGR technique [32]. The resolution of the converted grayscale FCGR image is a function of order k. The image dimension is 2^k x 2^k. For example, if k = 4, the image resolution will be 2^4 x 2^4 = 16 x 16 pixels [30].

2.2.2. Single gray-level representation technique

The SGLR technique transforms the genome sequences into genomic grayscale images by representing the bases of the genome sequences (A, T, C, and G bases) by values that are evenly distributed between the gray levels from 1 to 255 [26]. The width of the genomic image was represented as W pixels. Therefore, the first row of the image represents the first W bases of the genome sequence. The second row represents the subsequent W bases, and the process is repeated until the last base of the genome sequence. In this manner, the SGLR image will have a fixed width equal to W and a variable height depending on the genome sequence’s length. In this study, the W value was selected as 200. For example, if the genome sequence length is 30,000 base pairs, the image dimension will be 200 x 150 pixels.

This study evaluated the efficiency of the FCGR and SGLR images for extracting statistical features to detect COVID-19, among other human CoV diseases. The FCGR technique was applied with different orders ranging from second to sixth (2 ≤ k ≤ 6). This study stopped at the sixth order (k = 6) due to the large matrix dimension of the FCGR image at the seventh order (k = 7; 128 x 128 pixels) and beyond being computationally expensive in terms of the time required to convert the genome sequence into an FCGR image and the classification task with no benefits to the system performance. Figs. 3 and 4 show the genomic grayscale images created with the FCGR and SGLR techniques for a COVID-19 sequence.

| Human coronaviruses         | Label          | Number of sequences |
|-----------------------------|----------------|--------------------|
| SARS-CoV-2 (COVID-19)       | COVID-19 Pos   | 3700               |
| MERS-CoV                    | COVID-19 Neg   | 732                |
| HCoV-229E                   |                | 635                |
| HCoV-NL63                   |                | 64                 |
| SARS-CoV-1                  |                | 405                |
| HCoV-HKU1                   |                | 1335               |

Table 2 List of human coronaviruses, their assigned labels, and the number of sequences used in this study.
2.3. Feature extraction and selection

2.3.1. First-order features

The statistical features of first-order were estimated directly from the original image. The image is a function of Y and X parameters, representing the columns and rows of the image, respectively. Y = 0, 1, ..., C - 1, and X = 0, 1, ..., R - 1, where C and R represent the number of columns and rows in the image, respectively. The intensity values of the image are given by I = 0, 1, ..., L - 1, where L is the number of gray levels in the image [25,31–32]. The histogram \( H(I) \) of each gray-level (I) in the image is estimated as follows:

\[
H(I) = \frac{P(I)}{RC},
\]

where \( P(I) \) is the number of pixels with an intensity value equal to I and RC is the number of pixels in the image.

Various central moments are used to estimate first-order features from the genomic images created with the FCGR and SGLR techniques. These moments describe the genome sequences of human CoV diseases. The following equations represent the central moments used in this study: mean (\( \mu \)), variance (\( \mu_2 \)), skewness (\( \mu_3 \)), kurtosis (\( \mu_4 \)), and entropy (\( E \)). These central moments are given as follows:

\[
\mu = \sum_{I=0}^{L-1} IH(I),
\]

\[
\sigma^2 = \mu_2 = \sum_{I=0}^{L-1} (I - \mu)^2 H(I),
\]

\[
\mu_3 = \sigma^3 \sum_{I=0}^{L-1} (I - \mu)^3 H(I),
\]

\[
\mu_4 = \left[ \sigma^4 \sum_{I=0}^{L-1} (I - \mu)^4 H(I) \right] - 3,
\]

\[
E = - \sum_{I=0}^{L-1} H(I) \log_2[H(I)],
\]

where \( H(I) \) represents the histogram of each gray-level (I), and \( \sigma \) is the standard deviation of the image.

The mean represents the average intensity value of the image, and the variance is how far the grayscales are from the mean. The skewness measures the image’s symmetry. The flatness of the histogram is measured using kurtosis [25,32].

All previous parameters were used to create the vector \( V_1 \) representing the first-order statistical features. Thus, \( V_1 \) is expressed as follows:

\[
V_1 = [\mu, \mu_2, \mu_3, \mu_4, E]
\]

2.3.2. Second-order features

The statistical features of second-order provide information on the relationship between the neighboring pixel values of the original image. This study used the gray-level co-occurrence matrix (GLCM) technique to extract second-order features from the genomic images created with the FCGR and SGLR techniques to classify human CoV diseases.

Haralick et al. [33] proposed the GLCM to extract different texture features from a grayscale image. The GLCM is a square matrix \( \mathbf{C} \) of size \( S \times S \), where \( S \) represents the number of the normalized intensity levels of the original image \( I \). The columns and rows of matrix \( C \) represent the reference pixel values \( r \) and the neighboring pixel values \( n \), respectively. Each element of the matrix \( C(r, n) \) denotes how often a gray-level \( r \) is adjacent to a gray-level \( n \).

GLCM is a function of two parameters, \( d \), and \( \theta \), representing the relative distance between the reference and neighboring pixels (measured in pixels) and their relative orientation (measured in degrees), respectively. The value of orientation, \( \theta \) can be 0°, 45°, 90°, or 135°, as shown in Fig. 5 (right).
Energy is given as follows:

\[
\text{Homogeneity} = \frac{\sum_{i=0}^{S-1} \sum_{n=0}^{S-1} C(r, n)}{1 + |r - n|},
\]

where \(C(r, n)\) represents the GLCM elements derived from the original image; \(r\) and \(n\) represent the reference and neighbor pixels, respectively; \(\mu\) and \(\sigma^2\) represent the mean and variance of the image, respectively; and \(S\) is the number of normalized intensity levels of the original image.

The contrast measures the intensity contrast between the reference and neighbor pixels over the image. The contrast of 0 for an image represents identical grayscale values, and contrast is greater than 0 when there are significant differences in gray levels. The correlation measures the relationship between the reference pixel and its neighbor over the image.

Its value is 1 or –1 for perfect correlation and 0 for no correlation.

The energy measures the textural uniformity of the image. Its value is high when the image has either a repeating pattern of values or a constant value. The homogeneity measures the image's smoothness, which is high for an image with a slight difference in gray levels [33]. All previous parameters were used to create the vector \(V_2\) representing the second-order statistical features. Thus, \(V_2\) is expressed as follows:

\[
V_2 = [\text{Contrast, Correlation, Energy, Homogeneity}]
\]

Besides the \(V_1\) and \(V_2\) feature vectors, a \(V_3\) vector that combines both \(V_1\) and \(V_2\) vectors was also used. The significance of the features inside \(V_1\), \(V_2\), and \(V_3\) feature vectors was estimated using the analysis of variance (ANOVA) test.

2.3.3 Analysis of variance test

The one-way ANOVA test is a statistical approach used to select or discard features based on calculating the variation between and within classes. The ANOVA partitions the total variance in the data into two parts: the sum of squares within classes (SSW), which represents the variation of observations in each class from their class mean, and the sum of squares between classes (SSB), which represents the variation of class means from the overall mean. SSW and SSB are calculated using Eqs. (17) and (18), respectively.

\[
SSW = \sum_i \sum_j \left( Y_{ij} - \bar{Y}_j \right)^2,
\]

\[
SSB = \sum_j N_j \left( \bar{Y}_j - \bar{Y} \right)^2,
\]

where \(Y_{ij}\) represents the Ith observation of the Jth class, \(\bar{Y}_j\) is the mean of the Jth class, \(N_j\) is the number of observations in the Jth class, and \(\bar{Y}\) is the overall mean of classes.

Then ANOVA calculates the ratio of the variation between classes to the variation within classes to calculate the F-statistic parameter, which indicates the significance of the fea-

![Fig. 5 GLCM (center) derived from the original image (left) in the horizontal direction 0° with d = 1.](image-url)
The following equations were used to calculate the F-statistic of the feature.

\[ F = \frac{MSB}{MSW} \quad (19) \]

\[ MSB = \frac{SSB}{N - X} \quad (20) \]

\[ MSW = \frac{SSW}{X - 1} \quad (21) \]

where \( F \) is the computed F-statistic value, \( MSB \) is the mean square between classes, \( MSW \) is the mean square within classes, \( N \) is the total number of observations, and \( X \) is the number of classes.

Finally, ANOVA estimates the p-value, which is the probability (P) that the F-statistic can take a value greater than the computed F-statistic value, \( P(F_{\text{statistic}} > F_{\text{computed}}) \). The cumulative distribution function of the F-distribution is used to derive this probability (p-value). If the p-value of the feature is less than the significance level, then the feature is considered significant and can be used in the classification process. Otherwise, the feature is discarded from the subset. 0.05 is the most common significance level.

### 2.4. Classification process

Several machine learning models, such as NB, SVM using linear (LSVM) and Gaussian (GSVM) kernels, and KNN, were used in this study. For the KNN classifier, the neighborhood size (\( k \)) was increased from 1 to 19 with a step of 2, and it was discovered that the accuracy of the proposed approach was unaltered or slightly decreased. Consequently, the \( k \) value was set to 1. The performance of the proposed approach was evaluated using the average of 20 runs based on a tenfold cross-validation technique, such that all human CoV sequences were used for training and testing, as shown in Fig. 6.

### 2.5. System evaluation

Different assessment parameters, including accuracy, precision, specificity, sensitivity, \( F_1 \)-score, and Matthew’s correlation coefficient (MCC), were estimated to evaluate the performance of the proposed approach [34]. These parameters are given as follows:

- **Accuracy**
  \[ \text{Accuracy} = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \times 100\%, \quad (22) \]

- **Precision**
  \[ \text{Precision} = \frac{Tp}{Fp + Tp} \times 100\%, \quad (23) \]

- **Specificity**
  \[ \text{Specificity} = \frac{Tn}{Tn + Fp} \times 100\%, \quad (24) \]

- **Sensitivity**
  \[ \text{Sensitivity} = \frac{Tp}{Tp + Fn} \times 100\%, \quad (25) \]

- **\( F_1 \)-score**
  \[ \text{\( F_1 \)-score} = \frac{2Tp}{2Tp + Fp + Fn} \quad (26) \]

- **MCC**
  \[ \text{MCC} = \frac{(Tp \times Tn) - (Fp \times Fn)}{\sqrt{(Tp + Fp)(Fp + Tn)(Tn + Fn)(Fn + Tp)}} \quad (27) \]

where \( Tp \) and \( Tn \) are the true-positive and true-negative values of successfully identified classes, and \( Fp \) and \( Fn \) are the false-positive and false-negative values of wrongly identified classes.

### 3. Results and discussion

#### 3.1. Results of the proposed approach

MATLAB-R2020a software was used to implement the proposed approach using a laptop computer with 16 GB RAM and a 2.5 GHz Intel Core i5 CPU. Based on the results of the ANOVA test, \( V_1 \), \( V_2 \), and \( V_3 \) feature vectors were observed to be statistically significant at a p-value less than 0.05. Consequently, \( V_1 \), \( V_2 \), and \( V_3 \) feature vectors were passed directly to the different proposed classification algorithms.

Figs. 7 and 8 show the accuracy of the different classifiers using \( V_1 \), \( V_2 \), and \( V_3 \) feature vectors extracted from the FCGR and SGLR images, respectively. Figs. 9 and 10 show the maximum accuracy of the different classifiers among \( V_1 \), \( V_2 \), and \( V_3 \) feature vectors extracted from the FCGR and SGLR images, respectively. In Fig. 7, among the different classification algorithms, it is observed that the system accuracy increases as the order of the FCGR image increases from the second to the fourth-order. Afterward, the accuracy decreases with higher orders above the fourth-order. Therefore, the research results recommended the fourth-order FCGR image.
as an appropriate genomic image to extract various statistical features to detect COVID-19, among other human CoV diseases.

Fig. 8 shows that the maximum accuracy obtained when using the SGLR images was 94.69%. However, the fourth-order FCGR images achieved an accuracy of 99.39%. Therefore, the study recommended the fourth-order FCGR images for extracting statistical features to detect COVID-19, among other human CoV diseases, compared to the SGLR images.

Obtaining high system performance parameters indicates that the classifier employed in the experiment is efficient at recognizing the required target with the fewest possible errors.

**Fig. 7** Classifier accuracy using FCGR images.
According to the research findings, the KNN and GSVM classifiers provide high system performance: high values of all previously mentioned evaluation parameters, compared with both NB and LSVM classifiers, as shown in Figs. 9 and 10. Therefore, the KNN and GSVM classifiers can distinguish COVID-19 from other human CoV diseases. However, the KNN classifier outperforms the GSVM classifier, as shown in Figs. 11 and 12, which show the maximum accuracy of the GSVM
and KNN classifiers among $V_1$, $V_2$, and $V_3$ feature vectors extracted from the FCGR and SGLR images, respectively.

Furthermore, the KNN classifier is preferred because it outperforms the GSVM in terms of system execution time. The execution time when using the FCGR images was 42.33 and 106.94 s for the KNN and GSVM classifiers, respectively. However, the execution time when using the SGLR images was 95.66 and 158.69 s for the KNN and GSVM classifiers, respectively. Therefore, the FCGR images outperform the SGLR images in terms of system performance and execution time. Fig. 13 shows the performance parameters of the KNN classifier obtained when using the FCGR and SGLR images.

Finally, the best approach was obtained by extracting statistical features from the fourth-order FCGR images and using the KNN classifier in the classification process. The proposed approach detected COVID-19, among other human CoV diseases, with 99.39% accuracy, 99.48% precision, 99.31% sensitivity, 99.47% specificity, 0.99 $F_1$-score, and 0.99 MCC. The success of using GIP approaches (FCGR) to detect COVID-19, among other human CoV diseases, demonstrates that the proposed approach provides new scope for CoV research.

3.2. Comparison with the previous COVID-19 studies

The following section compares the proposed approach with earlier studies that detected COVID-19 using the CoV sequences [21–23,26]. Arslan and Arslan [21] extracted CpG-based features from the complete genome sequences of human CoVs to detect COVID-19 versus other human CoV diseases. These features were fed to the KNN classifier implemented with different distance metrics. Their method resulted in an accuracy of 98.4%. The main limitation of this system is that its dataset excluded SARS-CoV-1 sequences, which have 79% similarity to the COVID-19 sequences [8]. Therefore, the system’s overall accuracy may decrease if these sequences were added to the dataset.
Fig. 11  Maximum accuracy of the GSVM and KNN classifiers using FCGR images.

Fig. 12  Maximum accuracy of the GSVM and KNN classifiers using SGLR images.

Fig. 13  Performance of the KNN classifier using the SGLR and FCGR images.
Naeem et al. [23] converted the complete genome sequences of SARS, COVID-19, and MERS diseases into genomic signals and then extracted useful features from these signals. Although their approach resulted in 100% accuracy, it has many limitations. First, their dataset size was small, consisting of only 76 sequences for each epidemic. Second, they used the train and test split strategy to evaluate the approach, although this strategy gives accurate results with large dataset size. Finally, their dataset included only three variants of human CoV diseases, limiting their approach in analyzing the other variants of human CoV diseases such as HCoV-NL63, HCoV-229E, HCoV-HKU1, and HCoV-OC43.

Adetiba et al. [22] converted the genome sequences into Z-curve images to detect COVID-19 among MERS and SARS diseases. The CNN model detected COVID-19 with an accuracy of 98.3%. Their dataset included only 20 sequences for each disease. This approach had the same limitation as the Naeem et al. [23] study; the dataset size of the approach was insignificant, with only three variants of human CoV diseases.

Moreover, the datasets of earlier COVID-19 studies [21–23,26] include only complete genome sequences of human CoV strains. Therefore, these systems have limitations in analyzing the partial genome sequences. For example, if the system input is a partial COVID-19 sequence, the system cannot predict whether the patient has COVID-19 or not. However, such a system works well with only complete COVID-19 sequences. Table 3 compares the accuracy of the

| Study          | Classifier | Feature                          | Sequence type                                      | Dataset                  | Accuracy (%) |
|----------------|------------|----------------------------------|---------------------------------------------------|--------------------------|--------------|
| Arslan et al. [21] | KNN using the L1 distance metric | CpG-based features               | Complete genome sequences of human coronavirus diseases | COVID-19: 1000 MERS-CoV: 258 HCoV-229E: 27 βCoV: 140 HCoV-NL63: 61 HCoV-HKU1: 18 SARS-CoV-1: 20 | 98.4          |
| Adetiba et al. [22] | Convolution neural network model for feature extraction and classification stages |                                      |                                                                      | COVID-19: 100 MERS-CoV: 76 SARS-CoV-1: 76 | 98.3          |
| Naeem et al. [23] | KNN        | Seven-moment invariants          | Complete genome sequences of human coronavirus diseases | COVID-19: 100 MERS-CoV: 258 HCoV-229E: 460 HCoV-NL63: 635 SARS-CoV-1: 64 HCoV-HKU1: 405 | 100          |
| Hammad et al. [26] | KNN        | Mean, variance, skewness, kurtosis, and entropy | Complete genome sequences of human coronavirus diseases | COVID-19: 100 MERS-CoV: 258 HCoV-229E: 460 HCoV-NL63: 635 SARS-CoV-1: 64 HCoV-HKU1: 405 | 100          |
| Proposed approach | KNN        | First- and second-order statistical features | Complete genome sequences of human coronavirus diseases | COVID-19: 100 MERS-CoV: 732 HCoV-229E: 460 HCoV-NL63: 635 SARS-CoV-1: 64 HCoV-HKU1: 405 | 99.39         |

KNN, K-nearest neighbors.
proposed approach with that of other COVID-19 studies that used genome sequences to detect COVID-19.

The proposed approach overcomes the limitations of the previous COVID-19 studies [21–23,26] as follows: first, the dataset includes both partial and complete genome sequences of human CoVs, thereby demonstrating the strength and efficiency of the proposed approach. Second, the dataset includes all types of human CoV strains that are genetically like COVID-19. Finally, the dataset size was very large; a total of 7331 genome sequences were collected, resulting in 3700 COVID-19 sequences and 3631 non-COVID-19 sequences. The proposed approach can rapidly diagnose COVID-19 with high performance while avoiding the drawbacks and limitations of the traditional diagnostic approaches.

4. Conclusions

COVID-19 is a severe pandemic affecting millions worldwide. Molecular tests and medical imaging scans can detect COVID-19 with high acceptable accuracy levels, but these procedures have several drawbacks. The insufficient resources for conducting the RT–PCR tests reduce the speed and efficiency of suspected cases, which becomes a difficult issue, especially with a large patient population. In some cases, RT–PCR examinations may give false-positive or false-negative results. Furthermore, medical imaging scans expose patients to high radiation dosage, which can cause acute health problems, particularly in pregnant women.

This study proposed an effective automated approach to detect COVID-19, among other human CoV diseases. The approach is based on GIP techniques using two genomic graphical mapping approaches, the FCGR and SGLR. High performance was achieved using both partial and complete genome sequences of human CoV diseases. Yet, avoiding the drawbacks and limitations of the previously mentioned traditional diagnostic approaches.

The efficiency of the SGLR and FCGR images is assessed for accurately detecting COVID-19, among other human CoV diseases. The FCGR images outperform the SGLR images in terms of system performance and execution time. The 99.39% accuracy, 99.31% sensitivity, and 99.47% specificity resulting from the combination of the fourth-order FCGR images, V3 feature vector, and KNN classifier show the efficiency of the proposed approach as a potential accurate diagnostic tool for detecting COVID-19, among other human CoV diseases.

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.

References

[1] S. Ludwig, A. Zarbock, Coronaviruses and SARS-CoV-2: a brief overview, Anesth. Analg. 131 (2020) 93–96.
[2] Y.R. Guo, Q.D. Cao, Z.S. Hong, Y.Y. Tan, S.D. Chen, H.J. Jin, K.S. Tan, D.Y. Wang, Y. Yan, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status, Mil Med Res. 7 (2020) 11.
[3] H.M. Ashour, W.F. Elkhatab, M.M. Rahman, H.A. Elshabrawy, Insights into the recent 2019 Novel Coronavirus (SARS-CoV-2) in light of past human coronavirus outbreaks, Pathogens. 9 (3) (2020) 186.
[4] K.G. Andersen, A. Rambaut, W.I. Lipkin, E.C. Holmes, R.F. Garry, The proximal origin of SARS-CoV-2, Nat. Med. 26 (4) (2020) 450–452.
[5] F. Wu, S.u. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. i. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y.i. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature. 579 (7798) (2020) 265–269.
[6] L. Pan, M. Mu, P. Yang, Y. Sun, R. Wang, J. Yan, P. Li, B. Hu, J. Wang, C. Hu, Y. Jin, X. Niu, R. Ping, Y. Du, T. Li, G. Xu, Q. Hu, L. Tu, Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study, Am. J. Gastroenterol. 115 (2020) 766–773.
[7] Egypt: WHO Coronavirus Disease (COVID-19) Dashboard with Vaccination Data, (n.d.). https://covid19.who.int/country/emro/country; (accessed July 22, 2022).
[8] B. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Yang, W. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E.C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet. 395 (2020) 565–574.
[9] L. Fu, B. Wang, T. Yuan, X. Chen, Y. Ao, T. Fitzpatrick, P. Li, Y. Zhou, Y.-F. Lin, Q. Duan, G. Luo, S. Fan, Y. Lu, A. Feng, Y. Zhan, B. Liang, W. Cai, L. Zhang, X. Du, L. Li, Y. Shu, H. Zou, Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis, J. Infect. 80 (2020) 656–665.
[10] A.S. El-Hossiny, W. Al-Atabany, O. Hassan, A.M. Soliman, S. A. Sami, Classification of thyroid carcinoma in whole slide images using cascaded CNN, IEEE Access. 9 (2021) 88429–88438.
[11] E.N. Eltayeb, N.M. Salem, W. Al-Atabany, Automated brain tumor segmentation from multi-slices FLAIR MRI images, Biomed. Mater. Eng. 30 (2019) 449-462.
[12] A.L. Aswathy, S.S. Vinod Chandra, Detection of brain tumor abnormality from MRI FLAIR images using machine learning techniques, J. Inst. Eng. India Ser. B 103 (4) (2022) 1097–1104.
[13] A. A.L., A.H. S., V.C. S.s., COVID-19 diagnosis and severity detection from CT-images using transfer learning and back propagation neural network, J. Infect. Public Health. 14 (10) (2021) 1435–1445.
[14] A.A. L., V.C.S., Cascaded 3D UNet architecture for segmenting the COVID–19 infection from lung CT volume, Sci. Rep. 12 (1) (2022).
[15] H. Gunraj, A. Sabri, D. Koff, A. Wong, COVID-net CT-2: enhanced deep neural networks for detection of COVID-19 from chest CT images through bigger, more diverse learning, Front. Med. 8 (2021) 729287.
[16] L.S. Suma, H.S. Anand, S.S.V. Chandra, Nature inspired optimization model for classification and severity prediction in COVID-19 clinical dataset, J. Ambient Intell. Humaniz. Comput. (2021) 1–13.
[17] A.Kinganayagam, M. Patel, A. Charlett, J. Lopez Bernal, V. Saliba, J. Ellis, S. Ladhani, M. Zambon, R. Gopal, Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020, Euro Surveill. 25 (2020).
Genomic image representation of human coronavirus sequences for COVID-19 detection

[18] A. Tahamtan, A. Ardebili, Real-time RT-PCR in COVID-19 detection: issues affecting the results, Expert Rev, Mol. Diagn. 20 (5) (2020) 453–454.

[19] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases, Radiology. 296 (2020) (2019) E32–E40.

[20] B. Udugama, P. Kadhiresan, H.N. Kozlowski, A. Malekjahani, M. Osborne, V.Y.C. Li, H. Chen, S. Mubareka, J.B. Gubbay, W.C.W. Chan, Diagnosing COVID-19: the disease and tools for detection, ACS Nano. 14 (2020) 3822–3835.

[21] 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.

[22] E. Adetiba, J.A. Abolarinwa, A.A. Adegoke, T.B. Taiwo, O.T. Ajayi, A. Abayomi, J.N. Adetiba, J.A. Badejo, DeepCOVID-19: a model for identification of COVID-19 virus sequences with genomic signal processing and deep learning, Cogent Eng. 9 (2022) 2017580.

[23] 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 (2021) 1197–1205.

[24] L.A. Santamaria, S. Zuniga, I.H. Pineda, M.J. Somodevilla, M. Rossainz, DNA sequence recognition using image representation, Res. Comput. Sci. 148 (2019) 105–114.

[25] E. Delibas, A. Arslan, DNA sequence similarity analysis using image texture analysis based on first-order statistics, J. Mol. Graph. Model. 99 (2020) 107603.

[26] M.S. Hammad, V.F. Ghoneim, M.S. Mabrouk, Detection of COVID-19 using genomic image processing techniques, in: in: 2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2021, pp. 83–86.

[27] National Center for Biotechnology Information, (n.d.). https://www.ncbi.nlm.nih.gov/ (accessed January 11, 2022).

[28] S. Pei, W. Dong, X. Chen, R.L. He, S.-T. Yau, Fast and accurate genome comparison using genome images: the extended natural vector method, Mol. Phylogenet. Evol. 141 (2019) 106633.

[29] J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, Bioinformatics. 17 (5) (2001) 429–437.

[30] Y. Wang, K. Hill, S. Singh, L. Kari, The spectrum of genomic signatures: from dinucleotides to chaos game representation, Gene. 346 (2005) 173–185.

[31] A. Materka, M. Strzelecki, Texture analysis methods—a review, Technical University of Lodz, institute of electronics, COST B11 Report, Brussels. 10 (1998) 4968.

[32] S. Aloabdil, S. McQuaid, C. South, V. Prakash, P. Evans, A. Nisbet, The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning, Br. J. Radiol. 87 (2014) 20140369.

[33] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. SMC-3 (6) (1973) 610–621.

[34] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (4) (2009) 427–437.