scientific reports

OPEN

# Multifractal analysis and support vector machine for the classification of coronaviruses and SARS-CoV-2 variants

J. P. Correia[1,4]✉, L. R. da Silva[1,2] & R. Silva[1,3]

**This study presents a novel approach for the classification of coronavirus species and variants of SARS-CoV-2 using Chaos Game Representation (CGR) and 2D Multifractal Detrended Fluctuation Analysis (2D MF-DFA). By extracting fractal parameters from CGR images, we constructed a state space that effectively distinguishes different species and variants. Our method achieved $100\%$ accuracy in species classification, with a notable $76\%$ accuracy for SARS-CoV-2 variants despite their genetic similarities. Using a Support Vector Machine (SVM) as a classifier further enhanced the performance. This approach, which requires fewer steps than most existing methods, offers an efficient and effective tool for viral classification, with implications for bioinformatics, public health, and vaccine development.**

**Keywords** *Coronaviridae*, RNA, Fractal, CGR, Viral, SVM

The rapid evolution of viruses, particularly RNA viruses like coronaviruses, has posed significant challenges to public health and global economies[1–3]. Viruses such as SARS-CoV-2 and its related species have demonstrated high mutation rates, leading to the emergence of new variants that may evade immune responses and reduce the efficacy of treatments and vaccines[4–6]. Accurate and timely classification of virus species is crucial for understanding viral pathogenesis, monitoring transmission patterns, and developing practical diagnostic tools[7–9]. Moreover, precise classification is fundamental in designing targeted therapeutic interventions, informing public health policies, and preventing future outbreaks[10]. Traditional classification methods based on morphological features and genetic sequence analysis, while effective to some extent, often fail to capture the complexity and non-linear dynamics inherent in viral genomes[11,12]. As viruses evolve, small genetic changes may not always be detectable with standard approaches, necessitating more sophisticated methods[3,13].

This study proposes a novel approach to virus classification that leverages advanced computational techniques. We employ chaos game representation (CGR), which maps complex sequences onto fractal images to extract meaningful geometric features from viral genomes[14–16]. Additionally, we utilize two-dimensional multifractal detrended fluctuation analysis (2D MF-DFA), in the images generated by CGR, to characterize the genomic sequences' scaling behavior and multifractal properties[17–19]. The 2D MF-DFA method is an extension of the traditional MF-DFA applied to two-dimensional data, such as images, and aims to identify multifractal behavior in such systems[20–22]. We obtained several multifractal parameters from the 2d MF-DFA and constructed a state space with the most relevant fractal parameters. These same parameters were used as the features to feed the support vector machine algorithm and distinguish between the different samples. To facilitate the understanding of our work, in Fig. (1), we show the flowchart of the method.

By combining CGR, MF-DFA, and machine learning, we aim to develop a robust and accurate method for classifying virus species, focusing on coronaviruses. Unlike traditional approaches that apply these methods separately, our methodology integrates them into a cohesive framework, enabling a comprehensive analysis of fractal properties and their use in distinguishing species and variants. This unified approach not only simplifies the workflow but also improves the interpretability of the results, leveraging the strengths of each technique to provide a more holistic view of genomic and evolutionary patterns. Our approach offers the potential to provide valuable insights into the evolutionary relationships and functional diversity of viruses, ultimately contributing

[1]Department of Theoretical and Experimental Physics, Federal University of Rio Grande do Norte, 59072-970 Natal-RN, Brazil. [2]National Institute of Science and Technology of Complex Systems, Brazilian Center for Physics Research, 22290-180 Rio de Janeiro-RJ, Brazil. [3]Department of Physics, Rio Grande do Norte State University, 59610-210 Mossoró-RN, Brazil. [4]Department of Technology and Data Science, Getúlio Vargas Foundation, 01313-902 São Paulo, Brazil. ✉email: jonathan.pessoa@fisica.ufrn.br

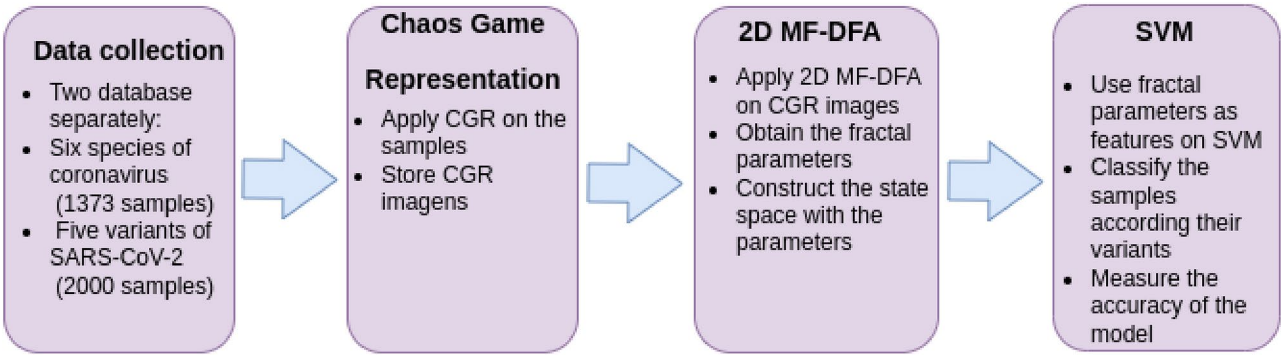

Data collection
Two database separately:
Six species of coronavirus (1373 samples)
Five variants of SARS-CoV-2 (2000 samples)
Chaos Game Representation
Apply CGR on the samples
Store CGR imagens
2D MF-DFA
Apply 2D MF-DFA on CGR images
Obtain the fractal parameters
Construct the state space with the parameters
SVM
Use fractal parameters as features on SVM
Classify the samples according their variants
Measure the accuracy of the model

**Fig. 1**. Graphical summary of workflows.

to better disease prevention and control. To test the feasibility of the application, we selected six species of coronavirus (HCoV-OC43, HCoV-HKU1, HCoV-NL63, HCoV-229E, MERS-CoV, and SARS-CoV-2) and five variants of SARS-CoV-2 (Alpha, Beta, Delta, Gamma, and Omicron) and applied the procedure above.

Our results indicate the fractal nature in the CGRs of all analyzed species. SARS-CoV-2 and MERS-CoV exhibited significantly higher fractal complexity, reflecting their unique genomic characteristics. Furthermore, the fractal parameters presented distinct values for each species, allowing a clear separation between them. Using these multifractal parameters as a basis for classification, our machine-learning models achieved high accuracy in differentiating coronavirus species. We obtained good performance for classifying SARS-CoV-2 variants, indicating a promising path for this approach. These findings suggest that the approach used is adequate for classifying different virus species and has excellent potential for classifying emerging coronavirus variants, offering a promising tool for epidemiological monitoring and control.

## Background and related work

The classification of virus species is a fundamental task in virology, contributing to the understanding of the functional diversity and transmission patterns of these pathogens. Although effective, traditional methods, such as sequence alignment and phylogenetic analysis, face limitations when dealing with highly mutated or recombined genomic regions[23,24]. In this context, approaches that combine graphical representations, mathematical transforms, and machine learning algorithms have shown promise for studying sequences from various species.[18,19]

The conversion of genome sequences into images, such as Chaos Game Representation (CGR) and Single Gray-Level Representation, allows capturing two-dimensional patterns that reflect structural and compositional properties of DNA and RNA, aiding in the identification of evolutionary and functional characteristics in coronavirus species[25,26]. Recently, advances in deep learning models, such as Vision Transformers (ViT), have enabled the detection of COVID-19 from X-ray images[27,28]. Furthermore, methods based on fractal analysis complement these tools by providing insights into the complexity of genome sequences and practical approaches for species classification[29]. Although some analyses use these techniques in isolation, more advanced approaches integrate methods such as the Discrete Fourier Transform (DFT) to extract magnitude spectra from CGR images[30], Discrete Wavelet Decomposition (DWT)[31] and Singular Value Decomposition (SVD)[32] , expanding the potential of genomic analysis.

Recently, several machine-learning approaches have explored clinical data and genome sequences of the coronavirus. Some studies have highlighted the use of classification algorithms to predict COVID-19 infections from clinical features[33], while others have employed convolutional neural networks to analyze SARS-CoV-2 sequences[34] directly. In addition, artificial intelligence-based methods have been used for diagnosis based on hematological markers[35,36], and autonomous approaches have been developed to detect and classify the virus[37].

In this context, the use of machine learning techniques to classify the coronavirus based on the spike region has shown to be a promising approach due to the importance of this region in viral infection and its interaction with human receptors. Several recent studies have applied classification algorithms, such as Convolutional Neural Networks (CNN), to identify patterns in genomic sequences related to the spike protein, aiming to differentiate between variants of the virus[38–40]. A distinction of our approach is using the entire genome to distinguish between virus variants.

Furthermore, in the current context of generative artificial intelligence development, advancing generative models based on deep learning has enabled new approaches for analyzing and predicting the evolution of SARS-CoV-2. The GenSLMs model proposes adapting language models for genomic sequences, demonstrating the ability to rapidly identify emerging variants by learning evolutionary patterns from millions of viral genomes[41]. Similarly, SARITA uses a generative language model explicitly trained on the S1 subunit of the Spike protein, being able to predict future mutations present in variants such as Delta and Omicron[42]. Complementing these efforts, SpikeGPT2 stands out by applying generative models to predict future mutations in the Spike protein, achieving high accuracy in predicting amino acid substitutions that impact the virus's transmissibility[43]. These studies demonstrate how artificial intelligence models can improve the identification and classification of variants, contributing to epidemiological monitoring and developing containment strategies.

Combinations of these genomic analysis methods have shown promising results in classifying genomic sequences (with model accuracy exceeding 90% overall). Our work contributes to this advance by integrating 2D multifractal MF-DFA analysis with CGR image transformations to study coronavirus genomic sequences. By combining these approaches with machine learning, we could easily classify species, achieving high accuracy when using SVM to distinguish coronavirus variants. This connection leads to a significant result, considering the high genetic similarity between the variants belonging to the same species.

## Theoretical background

### Chaos game representation

Chaos Game Representation (CGR) is a technique used to represent DNA sequences in the form of two-dimensional images, providing a visual way to analyze patterns present in symbolic chains[44,45]. The graphical representation generated by CGR is based on the theory of dynamical systems and allows the analysis of fractal and statistical properties of symbolic sequences. This section describes this method and its application in a DNA sequence[44–47].

Let $\mathcal{S} = (s_1, s_2, \ldots, s_L)$ be a sequence composed of symbols belonging to a finite alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, where each $s_i \in \mathcal{A}$. For the case of DNA sequences, $\mathcal{A} = \{A, T, C, G\}$, representing the four nucleotide bases, namely Adenine, Thymine, Cytosine, and Guanine, respectively. The CGR of a sequence $\mathcal{S}$ is constructed inside a square with vertices $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$, which are associated to the bases $\{A, T, C, G\}$.

The initial position $r_0$ within the square is typically defined as the center, i.e., $r_0 = (0, 0)$. The CGR representation is then constructed iteratively along the sequence $\mathcal{S}$, following the rule that for each symbol $s_i \in \mathcal{A}$ , the new position $r_i = (x_i, y_i)$ is computed as the midpoint between the current position $r_{i-1} = (x_{i-1}, y_{i-1})$ and the vertex corresponding to symbol $s_i$.

Formally, the position $r_i$ is given by:

$$r_i = \frac{1}{2}(r_{i-1} + V(s_i)), \tag{1}$$

where $V(s_i)$ is the coordinate of the vertex associated with the symbol $s_i$. For a DNA sequence, we have the vertices:

$$\begin{aligned} V(A) &= (0, 0), \\ V(T) &= (1, 0), \\ V(C) &= (0, 1), \\ V(G) &= (1, 1). \end{aligned}$$

Chaos Game Representation (CGR) provides a visual way to represent DNA sequences and allows us to calculate the frequency of subsequences of length *k*, known as *k*-mers. For example, for a value of $k = 4$, the CGR image will be a 16x16 pixel matrix, where each pixel represents a unique combination of four symbols. The frequency with which each *k*-mer appears in the sequence is reflected by the number of times the CGR trajectory hits the same pixel.

Pixels of the most frequent subsequences appear in darker tones in the image generated by CGR, while pixels corresponding to rare subsequences appear in lighter tones. In addition, the image generated by CGR can be analyzed for its fractal properties using techniques such as Multifractal Analysis (2D MF-DFA). This allows the identification of complex patterns in the distribution of subsequences, which may not be evident by counting frequencies alone. Thus, CGR, in conjunction with multifractal analysis, offers a powerful tool for exploring the structure of large genomic sequences, revealing both frequent and rare patterns.[48,49]

### Análise multifractal detrended fluctuation 2D (2D MF-DFA)

Consider a self-similar (or self-affine) surface, denoted by a two-dimensional array *X*(*i*,*j*), where $i = 1, 2, 3, \ldots, M$ , and $j = 1, 2, 3, \ldots, N$. The 2D MF-DFA method is defined by[17,50,51]:

1. The surface is partitioned into $M_s \times N_s$ disjoint square segments of the same size $s \times s$, where $M_s = \text{int}(M/s)$ and $N_s = \text{int}(N/s)$. Each segment can be denoted by $X_{v,w}(i, j) = X(l_1 + i, l_2 + j)$ for $1 \leq i$ and $j \leq s$,where $l_1 = (v - 1)s$ and $l_2 = (w - 1)s$. We define the profile

$$u_{v,w}(i, j) = \sum_{k_1=1}^{i} \sum_{k_2=1}^{j} X_{v,w}(k_1, k_2), \tag{2}$$

2. For each subsurface $u_{v,w}$ we obtain the local trend $\tilde{u}_{v,w}$ using a bivariate polynomial function. In this paper, we choose:

$$\tilde{u}_{v,w}(i, j) = ai + bj + c, \tag{3}$$

where *a*, *b* and *c* are free parameters to be determined and $1 \leq i$ and $1 \leq s$ . These parameters can be obtained through a matrix operation derived from the least squares method.

3. We obtain the residual matrix

$$y_{v,w}(i,j) = u_{v,w}(i,j) - \tilde{u}_{v,w}(i,j). \tag{4}$$

The variance of the residual matrix for each surface is given by

$$F^2(v,w,s) = \frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s} y^2_{v,w}(i,j) \tag{5}$$

4. Subsequentemente, definimos a 2D $q$th-order fluctuation function

$$F_q(s) = \left\{\frac{1}{M_s N_s}\sum_{v=1}^{N_s}\sum_{w=1}^{M_s}[F(v,w,s)]^q\right\}^{1/q}, q \neq 0 \tag{6}$$

and

$$F_0(s) = \exp\left\{\frac{1}{M_s N_s}\sum_{u=1}^{N_s}\sum_{v=1}^{M_s}\ln\left[F(v,w,s)\right]\right\}, q = 0 \tag{7}$$

5. Vary the value of s ranging from 6 to $\min(M,N)/4$. If there is a long-range power-law correlation for large values of s, then

$$F_q(s) \sim s^{h(q)}, \tag{8}$$

where $h(q)$ is the generalized Hurst exponent of 2D surfaces. This allows us to obtain the scaling exponent $h(q)$ via linearly regressing $\ln F_q(s)$ vs $\ln s$.

If we vary the value of $q$ in the range from $-10$ to 10, we can determine the strength of the multifractality, we calculate the difference between the asymptotic values of $h(q)$, that is, $\Delta h = h(q_{min}) - h(q_{max})$, here $q_{max} = \max\{q, q \in [-10,10]\}$ and $q_{min} = \; min\{q, q \in [-10,10]\}$. Here $\Delta h(q)$ quantitatively measures the deviation from monofractal behavior[29].

The multifractal scale exponent $\tau(q)$ of the following form can be used to understand the dependency on $q$ in the multifractal situation

$$\tau(q) = qh(q) - D_f. \tag{9}$$

In this context, $D_f$ represents the fractal dimension of the system. For two-dimensional images, $D_f$ equals 2, a fixed value for this type of system. However, the precise estimation of the fractal dimension can be influenced by the generalized Hurst exponent $h(q)$, which describes the scaling behavior of fluctuations in the system. If the image is monofractal, $\tau(q)$ is linearly connected to $q$. Otherwise, the image is multifractal, with a nonlinear relationship between $\tau(q)$ and $q$. Therefore, the properties of multifractals are more robust, which makes the nonlinear relationship stronger[29].

The other two indicators that characterize the strength of the singularity of the multifractal surface are the Hölder exponent and the singularity spectrum $(\alpha, f(\alpha))$, which is related to the multifractal scale spectrum $\tau(q)$ through a first-order Legendre transformation. If $\tau(q)$ is sufficiently smooth, the singularity's strength $\alpha$ , is given by

$$\alpha = \frac{d\tau(q)}{dq} = h(q) + qh?(q) \;, \tag{10}$$

from which the singularity spectrum $f(\alpha)$ can be constructed

$$f(\alpha) = q\alpha - \tau(q) = q[\alpha - h(q)] + 2 \;. \tag{11}$$

The exponent $\alpha$ characterizes the local singularity of an image texture, and $f(\alpha)$ measures the global singularity of an image texture[50]. Varying the value of q in the range from $-10$ to 10 we can determine two other multifractal parameters $\Delta\alpha$ and $\Delta f$ to describe an image

$$\Delta\alpha = \alpha_{max} - \alpha_{min} \tag{12}$$

$$\Delta f = f(\alpha_{max}) - f(\alpha_{min}) \tag{13}$$

where $\alpha_{max} = \max\{\alpha(q), q \in [-10, 10]\}$ and $\alpha_{min} = \min\{\alpha(q), q \in [-10, 10]\}$. Note que $\Delta\alpha$ is considered an indicator to measure the absolute magnitude of grayscale volatility. The higher the value of $\Delta\alpha$, the less uniform the distribution of the probability measure and the higher the expected image surface roughness. The index $\Delta f$ is the Hausdorff dimension of the measurement object, which measures the degree of confusion[50].

## Support vector machine (SVM)

The Support Vector Machine (SVM) algorithm is a supervised learning method for classification and regression. Its main goal is to find a hyperplane that separates the data into different classes with the most significant possible margin. The SVM constructs a hyperplane in a high-dimensional feature space for binary classification, maximizing the margin between the two classes[52–54].

Consider a training dataset $\{\mathrm{x}_i, y_i\}_{i=1}^{N}$, where $\mathrm{x}_i \in \mathbb{R}^d$ represents a feature vector of dimension $d$ and $y_i \in \{-1, 1\}$ is the label associated with the sample $\mathrm{x}_i$. The objective of the SVM is to find a hyperplane $f(\mathrm{x}) = \mathrm{w}^T\mathrm{x} + b = 0$ that best separates the samples of the two classes, where w is the weight vector and $b$ is the bias term.

The margin of the hyperplane is given by the distance between the closest points of the two classes and the hyperplane. To maximize this margin, we need to minimize $||w||$, subject to the constraint that all samples are correctly classified, which can be written as:

$$y_i(\mathrm{w}^T\mathrm{x}_i + b) \geq 1 \forall i = 1, 2, \ldots, N \tag{14}$$

The constraint says that we want all points to be on the correct side of the decision boundary with a margin of at least 1. For this reason, we say that an SVM is an example of a large margin classifier[55,56].

# Experiment materials

To demonstrate our classification method, we used six species of coronaviruses that infect humans: HCoV-HKU1, HCoV-OC43, HCoV-NL63, HCoV-229E, MERS-CoV, and SARS-CoV-2. The species HCoV-HKU1, HCoV-NL63, HCoV-229E and HCoV-OC43 cause mild respiratory infections such as fever, headache, sore throat and runny nose. While MERS-CoV, associated with Middle East Respiratory Syndrome, causes severe symptoms such as fever, cough, and difficulty breathing, with high mortality. SARS-CoV-2, which causes COVID-19, presents a wide range of symptoms, from mild to severe cases that can lead to death[57,58].

The genome sequences of these species were obtained from the *Nation Center of Biotechnology Information*-NCBI[59] database, and we extracted 1373 samples from the six species of interest. The number of samples for each species and other information is shown in Table (1).

To select the samples, we used specific filters in the NCBI database. Only sequences that met the following criteria were extracted: Human host, the maximum allowed number of ambiguous characters (represented by "N" in the nucleotide sequence) of 500 base pairs (bp), and the selected sequence should have a minimum length of 27,000 bp. At NCBI, sometimes only a few regions of the species genome are added to the database. We use this criterion to ensure that we are extracting the entire genome sequence of the species. This size criterion was applied exclusively to nucleotide sequences, regardless of protein sequences. Thus, all samples from the six species that remained after applying these filters were used for our study. We emphasize that the SARS-CoV-2 species has more samples due to the pandemic that began in 2019. Because of this, in order to maintain each species with a similar number of samples, we randomly selected 350 SARS-CoV-2 samples. We selected only these SARS-CoV-2 samples to avoid bias in the classification model.

Due to our interest in the SARS-CoV-2 species, we tested this method to classify the Alpha, Beta, Delta, Gamma, and Omicron variants. We selected 400 samples of each variant, totaling 2,000 samples obtained from the GISAID database[60] and a file on Github[61]. To extract these samples, we applied the filters: "complete" and "High Coverage" and host: "Human." Considering that the GISAID database does not allow random collection of the remaining samples, we selected each variant's first 400 available samples.

Although sequential collection may introduce some bias, choosing 400 samples per variant is representative enough to capture the genetic diversity of each group, minimizing the impact of possible correlations. This strategy aims to ensure that our analyses are robust and reflect the variations among the variants. Remembering

| Specie | Samples | Size (bp) | CG content (%) |
|---|---|---|---|
| HCoV-HKU1 | 64 | 29,926 | 32 |
| HCoV-OC43 | 342 | 30,741 | 37 |
| HCoV-NL63 | 165 | 27,553 | 34.5 |
| HCoV-229E | 125 | 27,317 | 38.5 |
| HCoV-MERS | 327 | 30,119 | 41 |
| SARS-CoV-2 | 350 | 29,903 | 38 |

**Table 1.** Characteristics of the coronavirus species. In the first and second columns, we have the species and number of samples extracted from NCBI for each of them. In NCBI, one sample of each species is identified as a reference sample. Thus, in the third and fourth column, we present, respectively, the size (in base pairs) and CG content (in the percentage of the sequence) of the samples identified as reference on NCBI.

that we used the complete sequences of the SARS-COV-2 variants and therefore, their properties are similar to those shown in the Table (1).

## Results and discussion

### Chaos game representation

We constructed the images generated by the CGR method for all 1373 samples, with pixelation degree $k = 6$ . The code to construct them is available on GitHub[61], and we present the results for some samples in Fig. (2). In NCBI, a sample of the species is identified as a reference sample. In Fig. (2), we plot the CGR for the sample identified as the reference for that species.

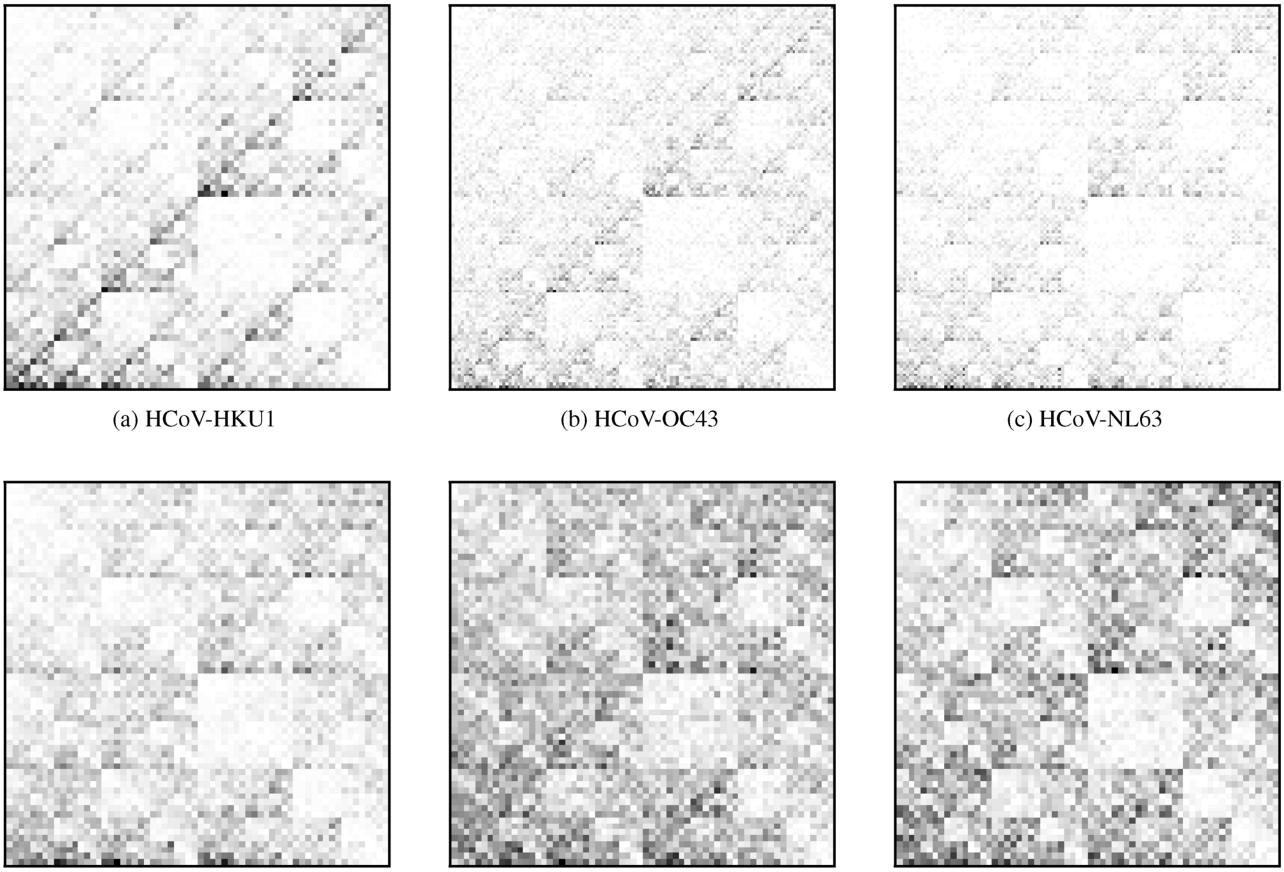
An important point when applying CGR is to consider the appropriate scale $k$ because if the value of $k$ is too large, many of the elements of the FCGR matrix may have zeros, making it difficult to identify patterns in the sequences. To avoid such a situation, the maximum value of $k$ can be calculated using the equation

$$k_{max} = \text{int}\left(\frac{\ln N}{2\ln 2} - 1\right) \tag{15}$$

where N is the total length of the sequence[62]. Using this Eq. (15) and the sizes $N$ from Table (1), then we use $k = 6$ to generate the CGR images, both coronavirus species and variants.

Our Frequency Game Representation (FCGR), employing multiple scales, revealed empty regions shaped like squares (self-similarity) across all samples. Samples of each species present a CGR pattern similar to the reference sample of the species.

Comparative analysis of CGR representations of different coronavirus species revealed a striking visual similarity between the genetic sequences of these species, see Fig. (2). In particular, we observed that some CGR images, such as those generated from the SARS-CoV-2 and MERS-SARS sequences, exhibit sharper and more defined geometric patterns, suggesting regularity and repetition of subsequences (k-mers). These patterns may indicate the presence of conserved regions in the genome, such as essential genes or regulatory sequences that are less prone to mutations, conferring functional stability to the species. The five selected SARS-CoV-2 variants, Alpha, Beta, Delta, Gamma, and Omicron, presented visual patterns similar to that shown in Fig. (2a).


(a) HCoV-HKU1

(b) HCoV-OC43

(c) HCoV-NL63

(d) HCoV-229E

(e) HCoV-MERS

(f) SARS-CoV-2

**Fig. 2**. Chaos game representation for coronavirus species: HCoV-HKU1, HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-MERS, SARS-CoV-2. We used the samples identified as reference sequences in NCBI.

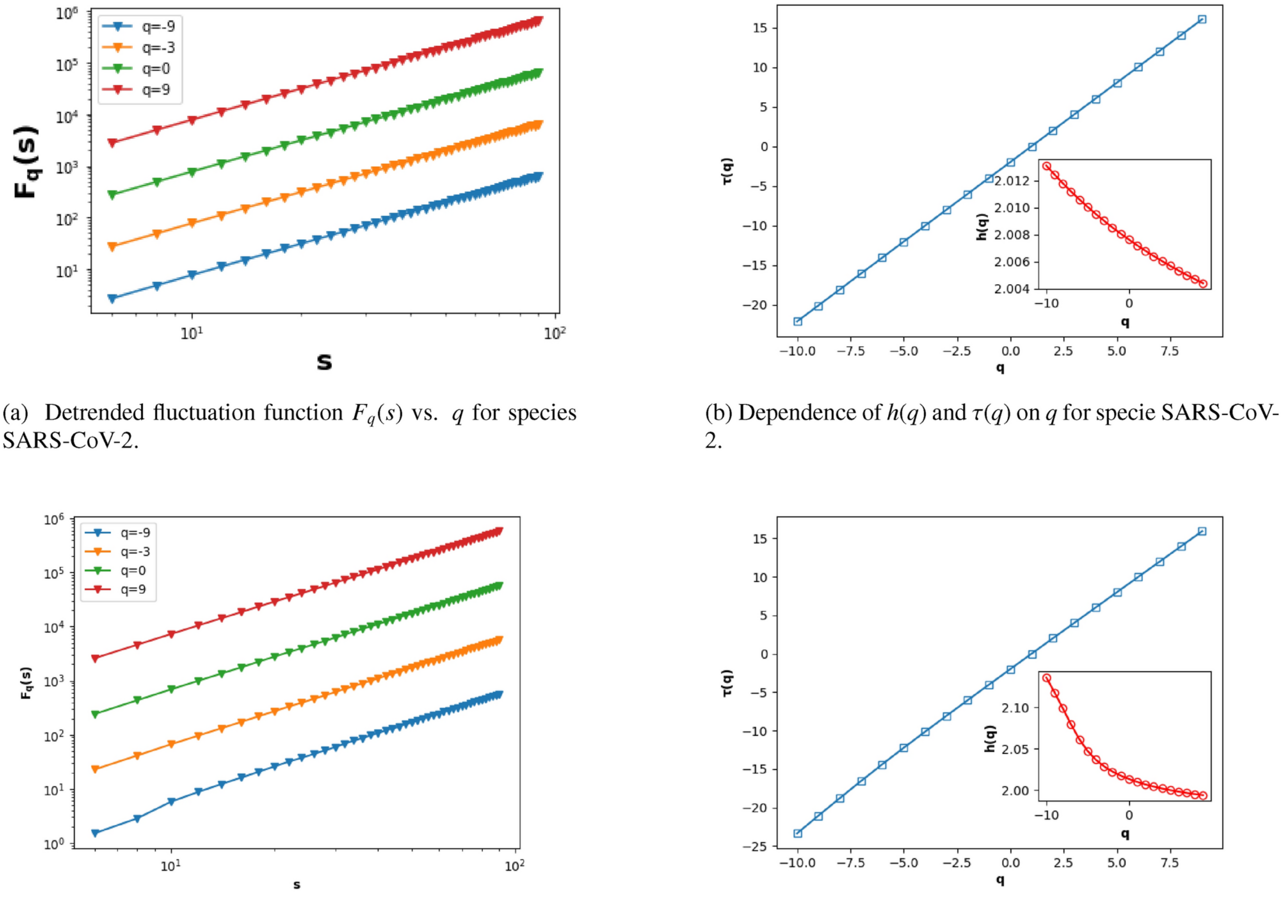
In addition, the empty regions observed in the CGR representations of all species indicate the underrepresentation of specific patterns, such as CG dinucleotides. This phenomenon is likely associated with the hypermutability of cytosine, which, due to its tendency to undergo spontaneous deamination and conversion to uracil, results in C-G to T-A mutations during replication. This process is one of the main factors responsible for reducing the frequency of CG dinucleotides, creating the characteristic empty regions in CGR representations, as seen in SARS-CoV-2. These observations are corroborated by the CG contents shown in Table (1), suggesting that the coronavirus sequence's high mutability rate contributes to these empty regions' emergence.

Thus, the difference in sharpness in CGR representations may be directly influenced by the nature of the subsequences present in each species, with more conserved genomes presenting more regular patterns and more diverse genomes exhibiting more diffuse patterns. These results are consistent with the hypothesis that the fractal organization of genetic sequences is associated with coronavirus species' functional stability and evolution[63].

## 2d MF-DFA

After generating CGR images for all 1373 samples of the six species, we applied the 2d MF-DFA method to determine the fractal parameters of the images. We performed the same procedure for the 2000 samples of the Alpha, Beta, Delta, Gamma, and Omicron variants. The code for this procedure is on GitHub[61]. Each image is stored as a 2D matrix in 256 grey levels, and we vary $s$ from 4 to $\max(M, N)/4$.

The results of the multifractal analysis of the CGR images of the coronavirus species and the variants of the SARS-CoV-2 species show that all the images have a multifractal nature. The multifractal nature observed in the CGR images refers to the inherent fractality of the genetic sequences captured by the CGR method and to the nonlinearity of the parameters calculated by the 2D MF-DFA. Expressly, the multifractality is confirmed by the fact that the function $h(q)$ is not constant for different values of $q$, indicating the presence of several complexity scales within the analyzed sequences. Fig. (3) demonstrates the multifractal nature of a CGR image of the reference sequence of the SARS-CoV-2 species and the randomly chosen Delta variant. The panels on the left illustrate the dependence of the fluctuation function $F_q(s)$ as a function of the scale $s$ for different $q$ for the



(a) Detrended fluctuation function $F_q(s)$ vs. $q$ for species SARS-CoV-2.

(b) Dependence of $h(q)$ and $\tau(q)$ on $q$ for specie SARS-CoV-2.

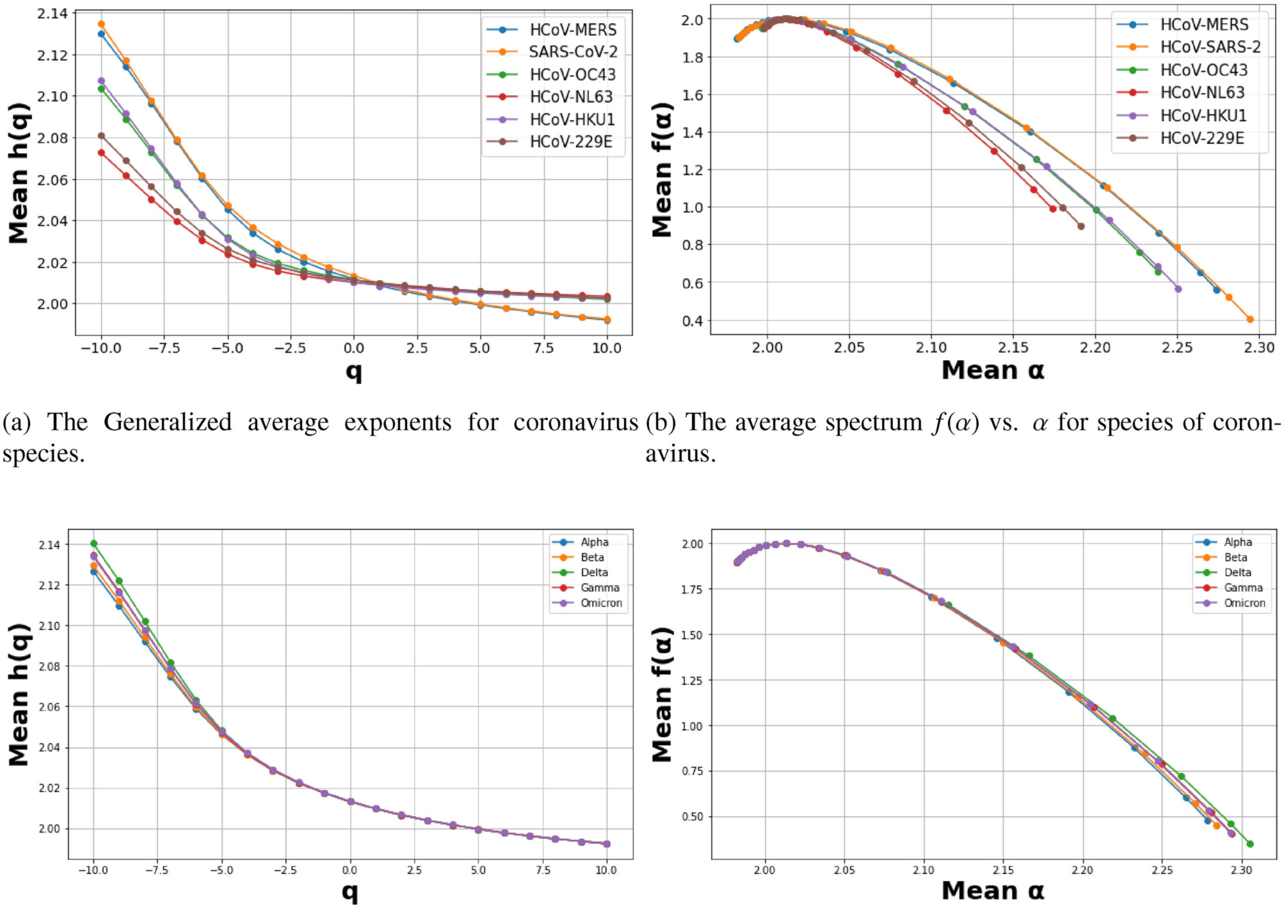
(c) Detrended fluctuation function of $F_q(S)$ vs. $s$ for Delta variant of the specie SARS-CoV-2.

(d) Dependence of $h(q)$ and $\tau(q)$ on $q$ for Delta variant from specie SARS-CoV-2.

**Fig. 3**. Power-law multifractal nature of coronavirus species CGR image. Some constants are subtracted to make the contrast between the different curves clearer in graphics of $F(q)$ vs. $q$. The straight lines are the best-fit lines whose slopes are shown in the legend.

| Parameters | HCoV-HKU1 | HCoV-OC43 | HCoV-NL63 | HCoV-229E | HCoV-MERS | SARS-CoV-2 |
|---|---|---|---|---|---|---|
| $h(-2)$ | $2.0145 \pm 0.00034$ | $2.0158 \pm 0.00026$ | $2.0132 \pm 0.00022$ | $2.0147 \pm 0.00036$ | $2.0200 \pm 0.00048$ | $2.0224 \pm 0.00040$ |
| $h(-1)$ | $2.0120 \pm 0.00020$ | $2.0133 \pm 0.00016$ | $2.0114 \pm 0.00014$ | $2.0127 \pm 0.00024$ | $2.0155 \pm 0.00027$ | $2.0174 \pm 0.00023$ |
| $h(0)$ | $2.0087 \pm 0.00005$ | $2.0114 \pm 0.00010$ | $2.0101 \pm 0.00008$ | $2.0111 \pm 0.00015$ | $2.0117 \pm 0.00013$ | $2.0132 \pm 0.00012$ |
| $h(1)$ | $2.0076 \pm 0.00001$ | $2.0098 \pm 0.00006$ | $2.0089 \pm 0.00004$ | $2.0098 \pm 0.00009$ | $2.0085 \pm 0.00005$ | $2.0096 \pm 0.00004$ |
| $h(2)$ | $2.2506 \pm 0.01336$ | $2.0084 \pm 0.00003$ | $2.0080 \pm 0.00002$ | $2.0086 \pm 0.00004$ | $2.0058 \pm 0.00011$ | $2.0066 \pm 0.00006$ |
| $\alpha_{max}$ | $1.9987 \pm 0.00022$ | $2.2381 \pm 0.00838$ | $2.1736 \pm 0.00666$ | $2.1912 \pm 0.00887$ | $2.2740 \pm 0.01256$ | $2.2942 \pm 0.01548$ |
| $\alpha_{min}$ | $0.1050 \pm 0.00728$ | $1.9970 \pm 0.00022$ | $1.9999 \pm 0.00021$ | $1.9978 \pm 0.00033$ | $1.9812 \pm 0.00070$ | $1.9824 \pm 0.00039$ |
| $\Delta h$ | $0.10176 \pm 0.00484$ | $0.1017 \pm 0.00484$ | $0.06923 \pm 0.00337$ | $0.07832 \pm 0.00501$ | $0.1379 \pm 0.00766$ | $0.14212 \pm 0.00889$ |
| $\Delta\alpha$ | $0.2519 \pm 0.01355$ | $0.2410 \pm 0.00853$ | $0.1736 \pm 0.00685$ | $0.1934 \pm 0.00918$ | $0.2928 \pm 0.01296$ | $0.3118 \pm 0.01576$ |
| $\Delta f$ | $-1.3963 \pm 0.06189$ | $-1.2945 \pm 0.03671$ | $-0.9736 \pm 0.03354$ | $-1.0533 \pm 0.03930$ | $-1.3329 \pm 0.05168$ | $-1.4963 \pm 0.06872$ |

**Table 2**. Average values of fractal parameters for coronavirus species.



(a) The Generalized average exponents for coronavirus species.

(b) The average spectrum $f(\alpha)$ vs. $\alpha$ for species of coronavirus.

(c) The Generalized average exponents for variants of SARS-CoV-2.

(d) The average spectrum $f(\alpha)$ vs. $\alpha$ for variants of SARS-CoV-2.

**Fig. 4**. Multifractal spectrum of the coronavirus species (above) and the variants of SARS-CoV-2 (below).

two samples shown. The well-fitted straight lines indicate the evident power-law scaling of $F_q(s)$ versus $s$. The right panel shows that $\tau(q)$ is nonlinear in $q$, indicated by the fact that $h(q)$ depends on $q$.

Once we have identified a multifractal nature among the CGR images, we are interested in determining the fractal parameters of each group of samples and comparing them. Therefore, for each species, we calculated the values of the parameters $\Delta h$, $h(-2), h(-1), h(0), h(1), h(2)$, $\alpha_{max}, \alpha_{min}$, $\Delta f$ and $\Delta\alpha$. The average values of these parameters for each species are shown in Table (2) and in the upper part of Fig. (4) we plot the average fractal spectra of $h(q)$ vs. $q$ and $f(\alpha)$ vs. $\alpha$. In the lower part of Fig. (4), we show the average spectra of the Alpha, Beta, Delta, Gamma, and Omicron variants. From Fig. (4), we can notice that the average spectra of $h(q)$ vs. $q$ and $f(\alpha)$ vs. $\alpha$ across species show more significant variability across species than across variants.

The parameters $\Delta h$ and $\Delta \alpha$ represent the fractal variability and the amplitude of the multifractality of the patterns present in the images. CGR images that present higher values of $\Delta h$ and $\Delta \alpha$, such as those generated from SARS-CoV-2 and MERS-CoV, would indicate a greater complexity and diversity of geometric patterns, suggesting that these species have genomic regions that vary significantly in their visual structure. This variability may indicate a greater diversity of functional elements or a less homogeneous organization. The more significant variability observed in the multifractal spectra of the different species means that the fractal parameters, both $h(q)$ as a function of $q$, and $f(\alpha)$ as a function of $\alpha$, are more distinct between the species. The graphs of these spectra distance themselves significantly from each other, whereas, when we compare the variants, the spectra are much closer to each other, suggesting a more remarkable similarity between the SARS-CoV-2 variants. As expected, the fractal spectrum between the variants is more similar since they share a joint genetic base. In contrast, the different coronavirus species show more pronounced variations in their fractal spectra, reflecting the more significant genetic divergence between them.

## Parameters space

We define the standard deviation of each multifractal parameter-$\Delta h$, $h(-2)$, $h(-1)$, $h(0)$, $h(1)$, $h(2)$, $\alpha_{max}$, $\alpha_{min}$, $\Delta f$, and $\Delta \alpha$-for each species $i$ as $\sigma_{in}(i)$. The intra-species variability, $\sigma_{in}$, is calculated as the average of these standard deviations across all species:

$$\sigma_{in} = \frac{1}{6}\sum_{i} \sigma_{in}(i). \tag{16}$$

where $i$ represents each species. This measure reflects the variation of the multifractal parameters within a single species.

Next, we define the inter-species variability, $\sigma_{bet}$, as the standard deviation of the six intra-species standard deviations ($\sigma_{in}(i)$) calculated for the six species. Specifically

$$I_0 = \frac{\sigma_{bet}}{\sigma_{in}} \tag{17}$$

In this case, $I_0$ represents a ratio of the between-species variability $\sigma_{bet}$ to the average within-species variability $\sigma_{in}$. This measure indicates how much the variability between species stands out compared to the variability within a single species.

A high value of $I_0$ indicates that the variability between species is much greater than the variability within each species, suggesting that the analyzed multifractal parameter is an good discriminant to differentiate species.

Based on the values of $I_0$ calculated for the parameters $h(-2), h(-1), h(0), h(1), h(2), \alpha_{min}, \alpha_{max}, \Delta\alpha, \Delta h, \Delta f$, we display the values in Table (3), select the four with the most significant values: $h(2)$, $\Delta f$, $\alpha_{max}h$ and $\alpha_{min}$ and plot these parameters in a scatter plot two by two, as shown in Fig. (5). We call the space formed by the axes of these four parameters ($h(2)$, $\Delta f$, $\alpha_{max}h$, and $\alpha_{min}$) the state space.

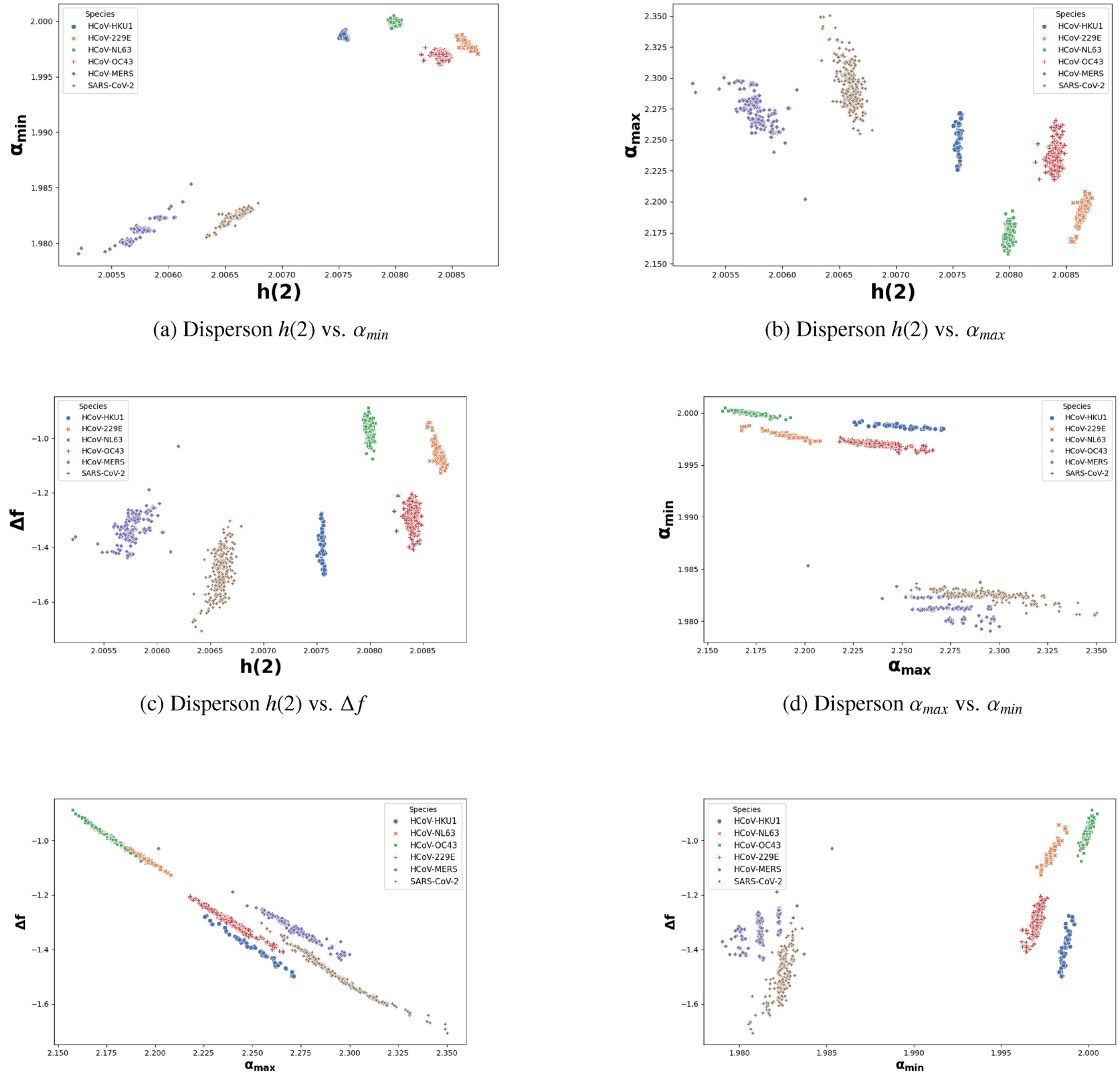
We observe that the coronavirus species are distributed in a dispersed manner in the scatter plots, allowing them to be separated by a straight line. See Fig. (5). This linear separability suggests that linear regression is an appropriate choice for classifying the species since a simple straight line can delimit specific regions occupied by each species. The most straightforward case is with the parameter $h(2)$ vs $(\alpha_{max}, \alpha_{min}, \Delta f)$ in which each species occupies a specific region and is easily distinguishable from the other species. The other scatter plots $(\alpha_{max} \times \alpha_{min})$, $(\alpha_{max} \times \Delta f)$ and $(\alpha_{min} \times \Delta f)$ also indicate that the species occupy specific regions of space but some intersections, especially between the SARS-CoV-2 and SARS-MERS species. Interestingly, the SARS-CoV-2 and MERS species are closer to each other in parameter space than the other species. This proximity reflects the more remarkable similarity between these viruses regarding the multifractal indicators analyzed, which may be related to genetic similarities.

Our method is advantageous because it involves fewer steps to extract the features used by machine learning algorithms than most existing approaches. By constructing our state space using fractal parameters, we can easily distinguish species visually with just two parameters. Remarkably, these results are almost independent of the choice of parameters, as any pair of fractal parameters allows us to differentiate the regions occupied by each species. This consistent separability facilitates the use of simple classification algorithms and provides insights into the genomic similarity among coronavirus species. Species closer in the state space may share similar genomic characteristics, highlighting the potential of our approach to reveal underlying evolutionary or structural relationships.

In Fig. (6), we indicate the state space formed by the parameters $h(2), \Delta f$, and $\alpha_{min}$ for the SARS-CoV-2 variants. We use these parameters for the SARS-CoV-2 variants because they present the highest values of $I_0$

| Parameters | $h(-2)$ | $h(-1)$ | $h(0)$ | $h(1)$ | $h(2)$ | $\alpha_{max}$ | $\alpha_{min}$ | $\Delta\alpha$ | $\Delta h$ | $\Delta f$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma_{bet}$ | 0.00009 | 0.00005 | 0.00007 | 0.00001 | 0.00015 | 0.00524 | 0.00257 | 0.00316 | 0.00188 | 0.02600 |
| $\sigma_{in}$ | 0.00035 | 0.0002 | 0.00009 | 0.00005 | 0.00014 | 0.00580 | 0.00152 | 0.01114 | 0.00576 | 0.03723 |
| $I_0$ | 0.25714 | 0.2500 | 0.7770 | 0.2000 | 1.0714 | 0.90348 | 1.69078 | 0.28366 | 0.32638 | 0.69836 |

**Table 3.** Calculation of $\sigma_{bet}$, $\sigma_{in}$ and $I_0$ for 10 fractal parameters.



(a) Disperson $h(2)$ vs. $\alpha_{min}$

(b) Disperson $h(2)$ vs. $\alpha_{max}$

(c) Disperson $h(2)$ vs. $\Delta f$

(d) Disperson $\alpha_{max}$ vs. $\alpha_{min}$

(e) Disperson $\alpha_{max}$ vs. $\Delta f$

(f) Dispersion $\alpha_{min}$ vs. $\Delta f$

**Fig. 5**. Scatter plots of fractal parameters.

according to the Table (3). The samples are much more "mixed" and distributed much closer than samples of the coronavirus species. Thus, this is because the variants present a more significant genetic similarity between them and are, therefore, more difficult to distinguish from each other. Despite this clear separability between the coronavirus species, we notice a more significant overlap between the samples when analyzing specific variants, as in the case of the SARS-CoV-2 variants. In these cases, more robust methods such as the Support Vector Machine (SVM), which handles more complex boundaries and overlaps well, become more suitable to ensure efficient classification.

## SVM

We chose the Support Vector Machine (SVM) algorithm because of its effectiveness in classification problems, mainly when the data are well distributed in distinct regions, as observed in the state space generated by the fractal parameters of the CGR images for coronavirus species (Fig. 5). The SVM is a robust approach for relatively small datasets, such as the coronavirus samples used, and is capable of identifying hyperplanes that maximize the separation margin between classes[54].

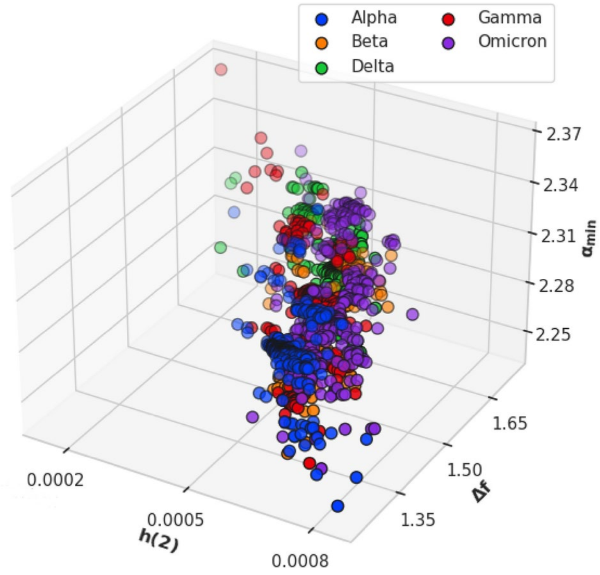


**Fig. 6**. State space constructed using fractal parameters ($h(2)$, $\Delta f$, $\alpha_{min}$ ) for SARS-CoV-2 variants. The apparent mixing of certain variants may reflect their evolutionary proximity or similarities in genomic features.

For the classification of SARS-CoV-2 variants (Fig. (6)), we observed that the samples present a more overlapping distribution, with less defined regions compared to the coronavirus species. However, we observed clustering tendencies among the variants, which justifies the use of the SVM to separate these classes, even if the overlap makes the task more challenging.

To ensure the robustness of the results, we used cross-validation through the Scikit-Learn "StratifiedShuffleSplit" function, dividing the data into five parts and maintaining the proportion of classes in each division. Here, $K$ represents the number of data splits into training and testing sets, with each split ensuring that $80\%$ of the data is used for training and $20\%$ for testing. In addition, we applied the "StandardScaler" to standardize the data, which is essential for optimal SVM performance. The classification model used was a Support Vector Machine (SVM) with RBF kernel (kernel="rbf"), a non-linear kernel that allows the identification of complex patterns in the data. The metric used to evaluate the model's performance was accuracy. The accuracies of each of the five divisions were calculated.

To test the algorithm, we applied it to the six coronavirus species, using as features the pairs of parameters and the configurations mentioned in the previous paragraph. As expected, for the state spaces $h(2)$ vs $(\alpha_{max}, \alpha_{min}, \Delta f)$, we obtained an accuracy of $100\%$, since in these spaces, the species are linearly separable. See Fig. (5). Furthermore, using as features ($\alpha_{max}$ vs. $\alpha_{min}$), ($\alpha_{max}$ vs. $\Delta f$), ($\alpha_{min}$ vs. $\Delta f$) and a space formed by the combination of the three parameters ($\Delta f \times \alpha_{min}, \times\alpha_{max}$). With the combination of two parameters, we obtained an accuracy higher than $97\%$, and with the combination of three parameters, it was possible to obtain an accuracy of $100\%$. See Fig. (7). This method indicates that the species are distinct in the shapes and textures of the CGR images. It shows once again that this method is effective in species classification.

This result reinforces the advantages of our method, as previously mentioned. The ability to achieve high accuracy with a minimal number of parameters highlights the efficiency of our approach. Additionally, the clear separability of species in the state spaces, even with different parameter combinations, demonstrates our framework's robustness and simplicity for the classification of genomic species.

In Table (4), we compare the performance of our method with the results of recent studies on the classification of coronavirus genome sequences. The table highlights the techniques, features extracted, classification algorithms employed, and accuracy obtained. Our method, which combines CGR with 2D MF-DFA and uses fractal parameters as features, obtained an accuracy of $100\%$, surpassing or equaling the results of other methods described in the literature, such as those mentioned in Table (4) .

For the SARS-CoV-2 variants, when applying the four parameters with the highest values of $I_0$ ($\Delta f$, $\alpha_{max}$ , $\alpha_{min}$ and $h(2)$) and we label each variant as follows: 0: Alpha, 1: Beta, 2: Delta, 3:Gamma, 4: Omicron. We obtained an average accuracy of approximately $76\%$. See Fig. (7). Although the separation between the variants is less pronounced due to their more remarkable genetic similarity, the SVM still proved effective in identifying patterns that allow the classification of the variants.

When SVM presents a lower accuracy, it can be attributed to the more significant genetic similarity between SARS-CoV-2 variants. Therefore, it results in similar CGR images and, consequently, similar fractal parameters. This similarity makes distinguishing variants more challenging. However, the obtained accuracy of approximately $76\%$ demonstrates that the SVM algorithm still effectively captures subtle patterns within the same species that allow the classification of these variants even if they are genetically close.

To evaluate how our algorithm classified samples from each variant, we obtained the confusion matrix, See Fig. (8) and calculated the precision, recall, and F1-score measures for each class. The confusion matrix is a table that summarizes the performance of a classification model, showing the number of correct and incorrect

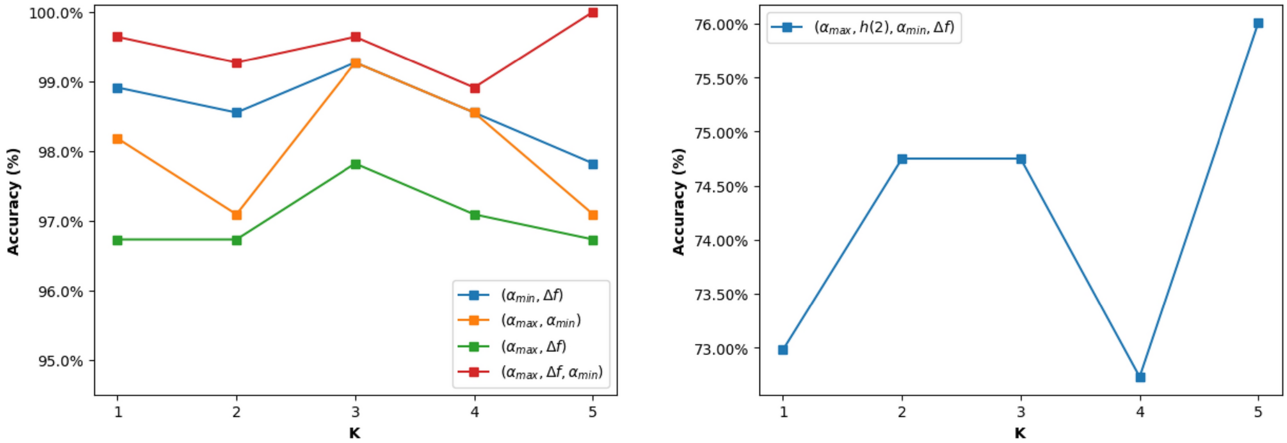
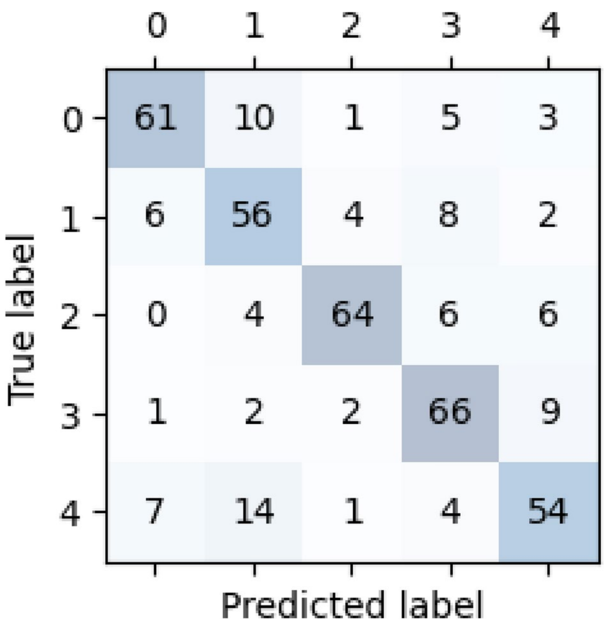


**Fig. 7**. Left: The accuracy of the six coronavirus species for the selected combinations with increasing K. Right: The accuracy of the five SARS-CoV-2 species variants for the selected combinations with increasing K.

| Study | Techniques Used | Features | classification algorithms | Accuracy (%) |
|---|---|---|---|---|
| This work | CGR + 2D MF-DFA | Fractal Parameters | SVM | 100 |
| Hammad et al[25] | CGR+ SGLR | Statistical features from the images | KNN | 99.39 |
| de Souza et al[26] | CGR + DFT | Viral signatures vector | CNN | 99.69 |
| Naeem et. al[66] | EIIP + DFT +DCT | Moment invariants | KNN | 100 |
| Arslan et. al[67] | CpG Islands | Ratio(C), ratio(G), ratio(CG) | KNN | 98.4 |

**Table 4**. Related works results. CGR: Chaos Game Representation, SGLR: Single gray-level representation, DFT: Discrete Fourier Transform, EIIP: Electron-ion-interaction pseudo potentials, DCT: Discrete cosine transform, SVM: Support Vector Machine, k-Nearest Neighbors, CNN: Convolutional neural Network.



|  | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 61 | 10 | 1 | 5 | 3 |
| 1 | 6 | 56 | 4 | 8 | 2 |
| 2 | 0 | 4 | 64 | 6 | 6 |
| 3 | 1 | 2 | 2 | 66 | 9 |
| 4 | 7 | 14 | 1 | 4 | 54 |

**Fig. 8**. Confusion Matrix. Each row represents the actual class, and each column represents the predicted class. The diagonal elements indicate correctly classified samples. Classes: 0 (Alpha variant), 1 (Beta variant), 2 (Delta variant), 3 (Gamma variant), 4 (Omicron variant).

predictions organized by each class. Each row represents the samples from the actual class, while each column represents the model's predictions, allowing us to observe where hits and misses occurred[64,65].

The confusion matrix, Fig. (8) shows the model predictions for five classes (SARS-CoV-2 variants). For each class, the model obtained the following metrics:

- Class 0 (variant Alpha): Precision of $81.33\%$, recall of $76.25\%$, and F1-score of $78.71\%$, indicating a good ability of the model to correctly identify examples of this class, although some errors still occur.
- Class 1 (Variant Beta): Precision of $65.12\%$, recall of $73.68\%$ and F1-score of $69.14\%$, suggesting that the model had a significant error rate in this class, possibly confusing it with other classes.
- Class 2 (variant Delta): Precision of $88.89\%$, recall of $80\%$ and F1-score of $84.21\%$, revealing that the model performed strongly in correctly identifying examples of this class.
- Class 3 (variant Gamma): Precision of $74.16\%$, recall of $82.5\%$ and F1-score of $78.11\%$, with a slight tendency to incorrectly classify this class, but with a high recall rate.
- Class 4 (variant Omicron): Precision of $72.97\%$, recall of $67.5\%$ and F1-score of $70.13\%$, which indicates a slightly greater difficulty in correctly classifying this class.

The model performed well with classes 0 and 2 (corresponding to the Alpha and Delta variants), which exhibited high precision and recall values. It indicates that the fractal parameters of these variants are more efficient in classifying them. One reason may be that the region these variants occupy in the state space is more defined than the others due to the genetic divergence between them.

While classes 1 and 4 (corresponding to the Beta and Omicron variants) presented a considerable amount of samples overlapping with other variants, suggesting a possible overlap of the multifractal parameters $(\alpha_{min}, \Delta f, \alpha_{max})$ with the other variants or difficulty of the model in distinguishing them adequately.

In general, we achieved an overall accuracy of $76\%$, showing that the model could classify a reasonable amount of samples correctly, but there is still room for improvement, especially in some classes. An avenue for further investigation is to test the same classification method on regions of the SARS-CoV-2 genomic sequence with higher mutation rates. For instance, instead of analyzing the entire SARS-CoV-2 genome, the method could be applied specifically to the Spike region, which has shown promise in achieving higher classification accuracies in related studies.

## Conclusion

In this work, we use Chaos Game Representation (CGR) and multifractal analysis (2D MF-DFA) to explore and classify different species of coronaviruses and variants of SARS-CoV-2. Using fractal parameters extracted from CGR images, we constructed a state space to distinguish coronavirus species efficiently. We observed a fractal nature in the CGR images of all coronavirus species. The clear separation between species evidenced in the space formed by these parameters, combined with the high accuracy of the Support Vector Machine (SVM) algorithms, which reached $100\%$ in some combinations of features, confirms the viability of the proposed method for biological classification problems.

The application of SVM in the classification of SARS-CoV-2 variants, although more challenging due to the more significant genetic similarity between the variants, obtained a satisfactory performance with an accuracy of approximately $76\%$. This result shows that, even in scenarios with high overlap between samples, the multifractal approach and SVM offer an effective solution for identifying patterns in complex data and the potential to classify coronavirus variants. A possible extension of this work could be to increase the variant database and use more robust machine learning algorithms, such as neural networks.

The proposed method proved effective for discriminating species and provided a solid basis for the analysis and classification of variants within a single species. Thus, this study contributes to the advancement of multifractal analysis techniques in bioinformatics and opens promising avenues for using CGR images and machine learning algorithms in future studies of the classification of organisms and their variants.

The dependence on the quality of the genomic data used in this work is an important limitation. Because the sequences were extracted from the NCBI database, which does not always provide complete genomes, rigorous filters were required to ensure the integrity and consistency of the data analyzed. Furthermore, although the method has demonstrated efficiency in separating species in state space, the biological interpretation of the fractal parameters based on the species-specific genomic characteristics is not yet fully elucidated, representing an opportunity for future studies that connect these patterns to specific molecular properties.

Although we focused on variants due to their relevance, we recognize that the field of virology has evolved, with the predominant Omicron sublineages. A possible extension of this work would be to apply the proposed method to classify these lineages, which could provide a more detailed view of viral evolution and contribute to the study of the most current variants. Furthermore, the proposed methodology may be helpful in other contexts, such as analyzing new variants or data from other viral families.

## Data availability

The datasets generated and/or analyzed during the current study and the code of programs are available in the GitHub repository[61]. The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request. The corresponding author's contact email is jonathan.pessoa@fisica.ufrn.br.

Received: 2 December 2024; Accepted: 10 April 2025
Published online: 29 April 2025

## References

1. Nicola, M. et al. The socio-economic implications of the coronavirus pandemic (covid-19): A review. *Int. J. Surg.* **78**, 185–193 (2020).
2. Hiscott, J. et al. The global impact of the coronavirus pandemic. *Cytokine Growth Factor Rev.* **53**, 1–9 (2020).
3. Drake, J. W. & Holland, J. J. Mutation rates among rna viruses. *Proc. Natl. Acad. Sci.* **96**(24), 13910–13913 (1999).
4. Zhao, Z. et al. Moderate mutation rate in the sars coronavirus genome and its implications. *BMC Evol. Biol.* **4**, 1–9 (2004).
5. Duffy, S. Why are rna virus mutation rates so damn high?. *PLoS biology* **16**(8), e3000003 (2018).
6. Holland, J. J. d., De La Torre, J. & Steinhauer, D. Rna virus populations as quasispecies, *Genetic diversity of RNA viruses* **176** 1–20, (1992).
7. Souf, S. "Recent advances in diagnostic testing for viral infections," *Bioscience Horizons: The International Journal of Student Research* **9** p. hzw010, (2016).
8. Flint, S. J., Racaniello, V. R., Rall, G. F., Hatziioannou, T. & Skalka, A. M. *Principles of virology, Volume 2: pathogenesis and control.* John Wiley & Sons, (2020).
9. Shors, T. *Understanding viruses*. Jones & Bartlett Publishers, (2017).
10. Cassedy, A., Parle-McDermott, A. & O'Kennedy, R. Virus detection: a review of the current and emerging molecular and immunological methods. *Front. Mol. Biosci.* **8**, 637559 (2021).
11. Simmonds, P. & Aiewsakun, P. Virus classification-where do you draw the line?. *Arch. Virol.* **163**, 2037–2046 (2018).
12. Murphy, F. A., Fauquet, C. M., Bishop, D. H., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A. & Summers, M. D. *Virus taxonomy: classification and nomenclature of viruses*, vol. 10. Springer Science & Business Media, (2012).
13. Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* **15**(3), 183–192 (2017).
14. Jeffrey, H. J. Chaos game visualization of sequences. *Computers & Graphics* **16**(1), 25–33 (1992).
15. Fiser, A., Tusnady, G. E. & Simon, I. Chaos game representation of protein structures. *J. Mol. Graph.* **12**(4), 302–304 (1994).
16. Joseph, J. & Sasikumar, R. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics.* **7**, 1–10 (2006).
17. Gu, G.-F. & Zhou, W.-X. Detrended fluctuation analysis for fractals and multifractals in higher dimensions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**(6), 061104 (2006).
18. Correia, J., Silva, R., Anselmo, D., Vasconcelos, M. & da Silva, L. Multifractal properties of human chromosome sequences. *Fractal Fract.* **8**(6), 312 (2024).
19. Correia, J. Multifractal analysis of maize and soybean dna. *Sci. Rep.* **14**(1), 10687 (2024).
20. Wang, J., Shao, W. & Kim, J. Automated classification for brain mris based on 2d mf-dfa method. *Fractals* **28**(06), 2050109 (2020).
21. Wang, J., Shao, W. & Kim, J. Combining mf-dfa and lssvm for retina images classification. *Biomed. Signal Process Control.* **60**, 101943 (2020).
22. Wang, J., Xu, H., Jiang, W., Han, Z. & Kim, J. A novel mf-dfa-phase-field hybrid mris classification system. *Expert Syst. Appl.* **225**, 120071 (2023).
23. Fuentes-Pardo, A. P. & Ruzzante, D. E. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Mol. Ecol.* **26**(20), 5369–5406 (2017).
24. Awadalla, P. The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* **4**(1), 50–60 (2003).
25. Hammad, M. S., Mabrouk, M. S., Al-atabany, W. I. & Ghoneim, V. F. Genomic image representation of human coronavirus sequences for covid-19 detection. *Alexandria Engineering Journal* **63**, 583-597 (2023).
26. de Souza, L. C., Azevedo, K. S., de Souza, J. G., Barbosa, R. d. M. & Fernandes, M. A. New proposal of viral genome representation applied in the classification of sars-cov-2 with deep learning. *BMC Bioinformatics.* **24**, 92 (2023).
27. Shome, D. et al. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *Int. J. Environ. Res. Public Health.* **18**(21), 11086 (2021).
28. Gao, X., Qian, Y. & Gao, A. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. arXiv preprint arXiv:2107.01682, (2021).
29. Kantelhardt, J. W. et al. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications* **316**(1–4), 87–114 (2002).
30. Randhawa, G. S. et al. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *Plos one*. **15**(4), e0232391 (2020).
31. Kar, S., Ganguly, M. & Sen, S. Lifting scheme-based wavelet transform method for improved genomic classification and sequence analysis of coronavirus. *Innovation and Emerging Technologies* **10**, 2350002 (2023).
32. Kar, S. & Ganguly, M. Application of genomic signal processing as a tool for high-performance classification of sars-cov-2 variants: a machine learning-based approach. *Soft Computing* **28**(4), 2891–2918 (2024).
33. Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N. & Peng, M. Predicting the covid-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimed. Tools Appl.* **80**, 11943–11957 (2021).
34. Câmara, G. B., Coutinho, M. G., Silva, L. M. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. d. M. & Fernandes, M. A. Convolutional neural network applied to sars-cov-2 sequence classification. *Sensors* **22**(15), 5730 (2022).
35. Chadaga, K. et al. Artificial intelligence for diagnosis of mild-moderate covid-19 using haematological markers. *Ann. Med.* **55**(1), 2233541 (2023).
36. Chadaga, K., Prabhu, S., Sampathila, N., Chadaga, R., Umakanth, S., Bhat, D. & GS, S. K. Explainable artificial intelligence approaches for covid-19 prognosis prediction using clinical markers. *Sci. Rep.* **14**(1), 1783 (2024).
37. Shahin, O. R., Alshammari, H. H., Taloba, A. I. & Abd El-Aziz, R. M. Machine learning approach for autonomous detection and classification of covid-19 virus. *Comput. Electr. Eng.* **101**, 108055 (2022).
38. Ali, S., Murad, T., Chourasia, P. & Patterson, M. Spike2signal: Classifying coronavirus spike sequences with deep learning, in *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 81–88, IEEE, (2022).
39. Ali, S. & Patterson, M. Spike2vec: An efficient and scalable embedding approach for covid-19 spike sequences, in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1533–1540, IEEE, (2021).
40. Rancati, S., Nicora, G., Prosperi, M., Bellazzi, R., Salemi, M. & Marini, S. Forecasting dominance of sars-cov-2 lineages by anomaly detection using deep autoencoders. *Brief. Bioinform.* **25**(6), bbae535 (2024).
41. Zvyagin, M. et al. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications* **37**(6), 683–705 (2023).
42. Rancati, S., Nicora, G., Bergomi, L., Buonocore, T. M., Czyz, D. M., Parimbelli, E., Bellazzi, R., Salemi, M., Prosperi, M. & Marini, S. Sarita: A large language model for generating the s1 subunit of the sars-cov-2 spike protein. *bioRxiv* bioRxiv:2024.12.10.627777 pp. 2024–12, (2024).
43. Dhodapkar, R. M. A deep generative model of the sars-cov-2 spike protein predicts future variants. *bioRxiv* bioRxiv:2023.01.17.524472 pp. 2023–01, (2023).
44. Gulick, D. & Ford, J. *Encounters with chaos and fractals*. Chapman and Hall/CRC, (2012).
45. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.* **18**(8), 2163–2170 (1990).
46. Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. Analysis of genomic sequences by chaos game representation. *Bioinformatics.*, (2001).
47. Löchel, H. F. & Heider, D. Chaos game representation and its applications in bioinformatics. *Comput. Struct. Biotechnol. J.* **19**, 6263–6271 (2021).

48. Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**(10), 1391–1399 (1999).
49. Löchel, H. F. & Heider, D. Chaos game representation and its applications in bioinformatics. *Comput. Struct. Biotechnol. J.* **19**, 6263–6271 (2021).
50. Wang, F., Liao, D.-W., Li, J.-W. & Liao, G.-P. Two-dimensional multifractal detrended fluctuation analysis for plant identification. *Plant methods.* **11**, 1–11 (2015).
51. Wang, F., Fan, Q. & Stanley, H. E. Multiscale multifractal detrended-fluctuation analysis of two-dimensional surfaces. *Physical Review E* **93**(4), 042213 (2016).
52. Somvanshi, M., Chavan, P., Tambade, S. & Shinde, S. A review of machine learning techniques using decision tree and support vector machine, in *2016 international conference on computing communication control and automation (ICCUBEA)*, pp. 1–7, IEEE, (2016).
53. Somvanshi, M., Chavan, P., Tambade, S. & Shinde, S. A review of machine learning techniques using decision tree and support vector machine, in *2016 international conference on computing communication control and automation (ICCUBEA)*, pp. 1–7, IEEE, (2016).
54. Suthaharan, S. & Suthaharan, S. "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, (2016).
55. Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, (2012).
56. Mehta, P. et al. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).
57. Liu, D. X., Liang, J. Q. & Fung, T. S. Human coronavirus-229e, -oc43, -nl63, and -hku1 (coronaviridae), in *Encyclopedia of Virology (Fourth Edition)* (D. H. Bamford and M. Zuckerman, eds.), pp. 428–440, Oxford: Academic Press, fourth edition ed., (2021).
58. Alabama public health. https://www.alabamapublichealth.gov/covid19/coronavirus.html. Accessed: 2024-07.
59. National library of medicine. https://www.ncbi.nlm.nih.gov/. Accessed: 2024-07.
60. Gisaid. https://gisaid.org/. Accessed: 2024-07.
61. Virus-classification-using-2d-mf-dfa-and-svm. https://github.com/jpcorreia96/Virus-classification-using-2D-MF-DFA-and-SVM. Accessed: (2024).
62. Pal, M., Satish, B., Srinivas, K., Rao, P. M. & Manimaran, P. Multifractal detrended cross-correlation analysis of coding and non-coding dna sequences through chaos-game representation. *Physica A: Statistical Mechanics and its Applications* **436**, 596–603 (2015).
63. Phillips, J. Synchronized attachment and the darwinian evolution of coronaviruses cov-1 and cov-2. *Physica A: Statistical Mechanics and its Applications* **581**, 126202 (2021).
64. Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **14**(1), 6086 (2024).
65. Dalianis, H. & Dalianis, H. Evaluation metrics and evaluation. *Clinical Text Mining: secondary use of electronic patient records*, pp. 45–53, (2018).
66. Naeem, S. M., Mabrouk, M. S., Marzouk, S. Y. & Eldosoky, M. A. A diagnostic genomic signal processing (gsp)-based system for automatic feature analysis and detection of covid-19. *Brief. Bioinform.* **22**(2), 1197–1205 (2021).
67. Arslan, H. & Arslan, H. A new covid-19 detection method from human genome sequences using cpg island features and knn classifier. *Engineering Science and Technology, an International Journal* **24**(4), 839–847 (2021).

## Author contributions

J. P. Correia wrote the original draft, collected the data, and performed the analyses. L.R. Silva developed the methodology and reviewed the manuscript. R. Silva wrote parts of the article, conducted the analyses, and reviewed the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.P.C.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.