Soft Computing
https://doi.org/10.1007/s00500-023-09577-9

DATA ANALYTICS AND MACHINE LEARNING

Application of genomic signal processing as a tool for high-performance classification of SARS-CoV-2 variants: a machine learning-based approach

Subhajit Kar[1] · Madhabi Ganguly[1]

Accepted: 10 December 2023

Abstract
From the beginning of COVID-19 pandemic, numerous mutants of SARS-CoV-2 have since been evolved owing to high transmissibility and virulence. Due to the limited effectiveness of previously imposed vaccines and preventive therapies, these strains are still causing concern. This paper proposes comparative evaluation of three novel genomic signal processing-based methods employing discrete wavelet decomposition with lifting (DWT), discrete Fourier transform (DFT), and singular value decomposition (SVD) for the classification of emerging SARS-CoV-2 variants utilizing feature extraction from collected SARS-CoV-2 variants acquired from the NCBI virus database. The efficiency and accuracy of the proposed alignment-free algorithms have been tested using three Coronavirus datasets including human Coronavirus (HCoV), SARS-CoV-2 variants (CoV-Variants and Omicron). The viral nucleotide sequences which are converted into numerical representation leveraging purine-pyrimidine mapping, DNA walk & Z-curve are fed into DWT, SVD, & DFT processors, respectively. In the approach with DWT, the second-generation wavelet transform employs two best wavelet bases Daubechies (Db) and Biorthogonal (Bior) based on the validation of the HCoV dataset for the feature extraction of the CoV-Variants dataset. Various machine learning algorithms, such as Support Vector Machine, K-nearest neighbors, and ensemble, are used to classify the virus strains and evaluate the efficacy of the algorithm. Finally, hyper-parametric tuning is done utilizing the Bayesian optimization technique to select the best fit model for KNN and SVM. The proposed algorithm has successfully classified the CoV-Variants dataset with an average accuracy of 98.76% utilizing the DWT, DFT, and SVD, while the best-achieved accuracy for this dataset is 98.9% using the DWT technique employing purine–pyrimidine mapping. The best-achieved accuracy rate for predicting Omicron is 99.8% using SVD-based technique. The best-obtained accuracy for HCoV dataset is 100% resulted in all three methods.

Keywords Genomic signal processing · Fourier transform · Wavelet transform · Machine learning · SARS-CoV-2

1 Introduction:

A specific type of RNA virus called a Coronavirus, which is a member of the distinguished Coronaviridae family, can cause respiratory and gastrointestinal problems in mammals, including humans. The genetic blueprint of Coronaviruses spans a range from 26 to 32 K, encompassing a remarkable breadth. Although the seasonal flu and COVID-19 manifestations occasionally exhibit similarities, it is worth noting that the latter boasts a mortality rate approximately four times higher than its counterpart (Abdelrahman et al. 2020). The global devastation caused by the COVID-19 virus persists, owing to the periodic emergence of malicious strains. Consequently, to effectively combat the perils of COVID-19, comprehensive tracking of demographic and clinical particulars, in conjunction with strain information, becomes an imperative necessity.

RNA virus genomes possess a proclivity for frequent mutations, resulting in a plethora of diverse variants readily available in the natural world. This phenomenon has catalyzed the emergence of novel species capable of

✉ Madhabi Ganguly
ray_madhabi@yahoo.co.in

Subhajit Kar
subhajitkar.wbsu@gmail.com

[1] Department of Electronics, West Bengal State University, Kolkata, 126, India

Published online: 22 January 2024

infiltrating previously untrodden hosts. The recent discovery of the highly mutable SARS-CoV-2 has shed light on its extraordinary genome length, surpassing that of other RNA viruses. Consequently, this viral entity showcases unparalleled flexibility in accommodating and modifying genes (Woo et al. 2009). It is noteworthy that this adaptability has given rise to various strains, such as Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and Omicron (B.1.1.529), which are rapidly proliferating across the globe (Hirotsu and Omata 2021). Therefore, it becomes imperative to undertake real-time classification of the SARS-CoV-2 variants by meticulously comparing copious sequences that are regularly uploaded onto international databases like NCBI, GISAID, NGDC, and Virus-host DB. By comprehensively analyzing the entire genome of diverse SARS-CoV-2 strains, researchers can harness knowledge from closely related variants as a foundation for designing antiviral drugs and/or pioneering vaccine innovation endeavors.

2 Literature survey

Classification of novel Coronavirus from similar types of other viruses like Influenza, Parainfluenza, Respiratory syncytial virus, Rhinovirus, and Adenovirus is important as Coronavirus disease is severe compared to other viruses which mainly responsible for the seasonal common cold. Novel Coronavirus also known as SARS-CoV-2 is very much similar to SARS-CoV-1 and MERS in terms of infectivity and virulence compared to other human Coronaviruses to which the human body has a good immune response. Although the symptoms of SARS-CoV-2 and other common cold viruses are the same at the initial stage, novel Coronavirus (COVID-19 disease) cannot get decimated within a week, but the symptoms suddenly increase advancing every part of the human body. As an effect, human lungs get infected by pneumonia and severe breathing distress may occur. Similarly, other vital organs are also affected and sometimes stop their normal functioning requiring comprehensive medication, hospitalization, and critical support to patients. Therefore, early detection of SARS-CoV-2 is important to provide essential support and boost the immune response to fight the disease. Also, it is important to discriminate SARS-CoV-2 from other human Coronaviruses as they are less symptomatic and recessed easily. As the pandemic progresses various lineages and sub-lineages are also circulated as the virus mutates to survive vaccine and antiviral therapies. Classification of various lineages is important as infectivity and severity are different for different viruses and their variants. To detect SARS-CoV-2 infection various machine learning and deep learning methods are implemented at the classification stage after efficient feature extraction by various means. Proper feature extraction is an important stage in machine learning-based techniques, whereas deep learning framework has various types of filters for effective feature capturing. Diagnostics features can be extracted from CT scans, X-rays, laboratory findings of infected persons, and genome sequences of viruses collected from blood samples. Following are the discussions about the current state-of-the-art of the proposed work.

(i) Deep learning (DL)-based classification

Deep learning algorithms have the capability to extract important features from the input data fed into them. The DL network is generally designed to take two-dimensional matrices such as images as input. Based on the extracted features, the DL network is able to classify objects, in our case Coronavirus sequences. For genomic sequences to be classified using DL, they must be converted into images by some mathematical transform. The DL method was applied by Ahsan et al. to classify SARS-CoV-1, SARS-CoV-2, and MERS (Ahsan et al. 2021). For that, they considered genomic sequences from the 29,000th nucleotide position to the end of the genomes as this hot zone is responsible for encoding nucleoprotein and spike protein of Coronavirus. The Coronavirus sequences are converted into binary images before application to the convolutional neural network. Ullah et al. employed a temporal convolutional neural network for accurate classification of SARS-CoV-2 variants (Ullah et al. 2022). The genomic sequences of virus variants were digitalized using the integer label Binarize method. The technique achieved 88.36% accuracy using a temporal convolutional network classifying eight different variants of SARS-CoV-2 including Omicron. Another convolutional neural network-based DL classification model was proposed by Camara et al. to classify viruses and other organisms using genome sequences (Câmara et al. 2022). The study compared 1553 SARS-CoV-2 and 14,684 other virus sequences for classification. The image representation of viral genomes was done using the k-mer method where $k = 6$. The model classified AlphaCoronaviruses, BetaCoronaviruses, and DeltaCoronaviruses with good accuracy scores. The genomic signal processing-based feature is employed in de Souza’s study to classify SARS-CoV-2 from other human Coronaviruses (Souza et al. 2023). The method used chaos game representation succeeded by discrete Fourier transform to build unique signatures of each genome to be classified by convolutional neural network (CNN). Azevedo et al. recently designed a technique to differentiate between SARS-CoV-2 and other human Coronaviruses with high accuracy (Azevedo et al. 2023). They employed viral genomic sequences to CNN having four convolutional layers and four fully connected layers. Before application to the CNN network, the raw genome sequences were digitized using a one-hot encoding technique. Apart from genomic sequences, Laboratory data, X-ray images, and CT-scan images were also used to be classified by DL models. One of the major drawbacks of CNN is it cannot provide the orientation of objects in the image. Therefore, its accuracy in image classification is less. To overcome this shortcoming, Das and Toraman introduced a capsule network to discriminate SARS-CoV-2 and healthy gene sequences. As the size of input sequences to the capsule network is not the same, these sequences are divided into 100 unit sections using the sliding window technique to be compared by the capsule network (Das and Toraman 2023). A study done by Ucar and Korkmaz classified normal, pneumonia, and COVID-19 based on the X-ray image of a patient (Ucar and Korkmaz 2020). The method is known as COVIDiagnosis-Net which made use of a CNN deep learning classifier for COVID-19 detection. Additionally, the method employed hyper-parameter tuning to get the best parameter selected for the DL model. Ghaderzadeh et al. presented a technique for early detection of COVID-19 using X-ray and CT-scan images (Ghaderzadeh et al. 2022). The method was augmented by hyper-parameter optimization and transfer learning. The method was designed in such a way that if the X-ray image of a patient was adequate for classification of COVID-19, then it recommends no further CT-scan; otherwise, it sends the patient for a CT-scan. The images of CT scans were used at the 2nd stage of classification, so that the overall accuracy of the method was high. Instead of CNN, a recurrent neural network (RNN) framework was proposed by Goreke et al. for COVID-19 detection using laboratory findings (Göreke et al. 2021). The model utilized an artificial bee colony algorithm to optimize the preweighting vector of the RNN model to give the best result. The method achieved 95% accuracy.

(ii) Machine learning (ML)-based classification

Machine learning algorithms work well with precise feature extraction and generally produce greater accuracy in a short time compared to deep learning algorithms. Genomic signal processing-based methods, such as Fourier transform, wavelet transform, and digital filter, capture effective features from SARS-CoV-2 and non-SARS-SARS-CoV-2 sequences for their classification. In this regard, Kindhi proposed an automated method for the classification of SARS-CoV-1, SARS-CoV-2, MERS, and Ebola viruses using machine learning methods (Al Kindhi 2020). Four conventional ML models, decision tree, discriminant analysis, KNN, and SVM, were further optimized for this purpose. In Ahmed and Jeon (2022), SARS-CoV-1, SARS-CoV-2, MERS, and Ebola were classified using features generated from the genome sequences. The hand-crafted features are information about nucleotide composition and their frequency, tri-nucleotide compositions, count of amino acids, alignment between genome sequences, and their DNA similarity. For visual interpretation of each virus category, simple dot plots were considered. The SVM classifier achieved 97% accuracy using this method. Instead of using genomic sequences, Afify and Zanaty examined protein sequences to distinguish SARS-CoV-2 and HIV-1 (Afify and Zanaty 2021). During the feature extraction step, the conjoint triad (CT) approach was used to transform the amino acids in each sequence into integers based on their side-chain dipoles and volumes. However, the practical application of the method will be limited as the symptoms of SARS-CoV-2 and HIV are not the same. Das devised a method for classification of SARS-CoV-2 sequences from healthy sequences taken from Homo Sapiens (Das 2022). The technique employed a genomic image processing method leveraging STFT, DWT, and SVD to extract statistical features from generated histograms. Out of 135 generated features, only 94 discriminative features were selected with the 'ReliefF' technique. Finally, SVM and KNN classifiers were used for discrimination purposes. Apart from the GSP features Abadi proposed a new feature extraction procedure named PC-mer based on k-mer and the Physico-chemical properties of nucleotides (Abadi and S. et al.. 2023). Compared to the traditional k-mer-based profiling method, this method reduces the encoded data size by approximately 2^k times. The method was designed to discriminate SARS-CoV-2 from other Human Coronaviruses using SVM and KNN classifiers. The limitation of considering only the full genome expression of Coronavirus sequences was eliminated in Hammad et al. where whole as well as partial genome sequences were classified with the same degree of accuracy (Hammad et al. 2023). The method utilized the frequency chaos game representation (FCGR) technique to convert genome sequences of HCOVs into genomic greyscale images. An Alexnet CNN model was applied to extract deep features from the FCGR images. Before application to the ML-based classifier models such as decision trees and KNN, the most significant features were selected using 'ReliefF' and LASSO algorithms. The aforementioned technique was able to classify SARS-CoV-2 sequences apart from other HCOVs with 99.71% accuracy. Classification of SARS-CoV-2 from the Influenza virus is important as infectivity and severity of the former are very high, although their symptoms are same. In this context, the study done by Khodadei et al. is worth mentioning (Khodaei et al. 2023). They applied genomic signal processing tools including linear predictive coding and singular value decomposition to extract features to be classified by SVM. The method achieved 99% accuracy using 3-D Z-curve-based numerical mapping of virus sequences. Very recently, Lin et al. proposed a sequence similarity method based on stationary discrete wavelet transform and k nearest-neighbor classifier. The method was evaluated using DNA, protein, and Next Generation sequences (Lin et al. 2018). Huang et al. developed an alignment-free classification method based on linear and quadratic discriminant analysis with discrete wavelet packet transform resulting satisfactory performance on mammals, Influenza, and Corona datasets (Huang and Girimurugan 2019; Huang et al. 2018).

Genomic signal processing (GSP) techniques, which rely on tools such as the Fourier transform, wavelet transform, S-transform, and digital filters, have also been employed for bio-computational purposes, including the prediction of splice sites in the human genome, the forecast of lethality in potential novel Coronaviruses, the identification of cancerous genes, and the prediction of intron–exon regions (Meher et al. 2019; Yin et al. 2020; Khodaei et al. 2020a; Daş et al. 2020). These methods have demonstrated their ability to yield accurate results within a remarkably brief time frame by utilizing supervised machine learning algorithms. Consequently, the application of GSP approaches, combined with machine learning, for the classification of SARS-CoV-2 lineages, could potentially deliver rapid and precise outcomes.

With the emergence of newly evolved virus strains of Coronavirus, it is essential to track the transmissibility of a particular strain in any region. As each of the SARS-COV-2 variants has their own characteristics their virulence and affectivity also deviate. For example, the 'Delta' variants of SARS-COV-2 are more destructive compared to 'Omicron' as the mortality rate of the latter is very low. However, 'Omicron' is more transmissible than 'Delta' (Wolter et al. 2021). Another important factor associated with virus strains is that each and every variant cannot be protected by vaccines as they have multiple mutations in spike protein. Therefore, identification of SARS-CoV-2 variants is very important, so that each detected virus sample can be classified accordingly within a very short span of time. The treatment of patients infected with different variants may also differ according to the severity of the strains. The current process of identifying a particular strain leveraging genome sequencing is very much efficient but it requires lengthy processing. Also, the cost of sequence analysis is very high. The proposed work can be used to classify any SARS-CoV-2 sequence according to its lineages. The proposed method contributes to the current virology in the following ways:

(a) A novel machine learning-based model has been developed for the classification of previously identified SARS-CoV-2 virus variants. The method can be utilized to predict the class of a new sequence. If a new sequence from any of the lineages previously trained in the ML processor is provided to the algorithm, then the method will be able to classify it according to the "Variant of concern" in the case of Omicron or "Variant being monitored" for other variants. Therefore, the model can be used to classify any 'Omicron' strain from other variants.

(b) The proposed model is very much efficient in classifying all the seven human Coronaviruses found to be affecting humans. These are known as H-CoVs (OC43, HKU1, 229E, NL63, MERS, SARS-CoV-1, and SARS-CoV-2). All these HCoVs are different in nature and severity. Some of them cause mild infections, whereas others are responsible for causing severe respiratory problems. Therefore, it is very much essential to classify all the HCoV sequences which can be persuaded using proposed DSP-based methods.

(c) Second-generation lifting scheme wavelet transform-based feature extraction is proposed along with various supervised machine learning algorithms for accurate and fast classification of novel Coronavirus strains. In the Lifting scheme, the signal decomposition can be accomplished by simple filtering steps reducing the number of arithmetic operations up to 50% compared to the first-generation wavelet transform. Other two digital signal processing-based methods employing Fourier transform and singular value decomposition are fast and accurate for the classification of SARS-CoV-2 strains.

(d) The proposed method provides an alternative way to the traditional alignment-based multiple sequence classification method. Furthermore, the current genome sequencing methods to classify SARS-CoV-2 variants require a lot of processing and time. Our model can be used for instant classification as it provides a simple, cost-effective, and efficient alternative to MSA and genome sequencing programs.

(e) The method can be extended to classify viruses like Influenza, Corona, Rhino, Adeno, and Entero which generate similar kinds of symptoms like common cold at the initial phase. However, Coronavirus can pose a serious threat to vulnerable people. Similarly, Adenovirus can produce severe symptoms in children. All the other viruses are generally known to evoke mild symptoms which can be mitigated by our own immune system. Therefore, early identification of these viruses is very important for proper treatment. Current virus identification methods are accurate but time-consuming and costly. The proposed method can be extended to classify various viruses in fast and accurate way and hence can pave the way for cost-effective virus classification.

It can be concluded that genomic signal processing-based feature extraction procedures can be a significant factor in the automatic feature extraction and classification of various SARS-CoV-2 sequences. These methods can handle large datasets and produce results within a few seconds. Therefore, the proposed method can have a high impact on the detection of COVID-19 strains in the human body. In this paper, three alignment-free techniques based on digital signal processing instigated by machine learning are proposed to classify various lineages of SARS-COV-2 virus. In the 1st technique, known as PPDWT, the virus genomes are converted into 1-D purine-pyrimidine representation, and then, detailed coefficients are computed with the help of second-generation wavelet transforms. Detail coefficients are then used as the similarity measure of the virus sequences. In the 2nd method, known as DNA-walkWavelet, 2-D DNA walks of all the virus sequences are computed to be processed by singular value decomposition for comparison. A 3-D Z-curve-based discrete Fourier transform algorithm is proposed in the 3rd method known as ZcurveFFT. An even scaling method had to be applied to overcome the length heterogeneous problem of the virus sequences. Finally, Euclidean distance matrices are formed independently from each of the techniques and applied to SVM, KNN, and ensemble classifier to classify the Coronavirus sequences into corresponding biological groups. Using this method, we achieved greater accuracy compared to other state-of-art alignment-free methods. The schematic diagram of the proposed method is shown in Fig. 1.

3 Materials and methods

3.1 Dataset

The selection of an appropriate dataset is a part of important scientific findings. In our research work, we used three datasets, namely, HCoV, CoV-Variant, and Omicron. The source of all the datasets is the NCBI which may be accessed through the gateway at https://www.ncbi.nlm.nih.gov/labs/virus/. Users of the website have the capacity to carry out a detailed search for several types of viral genomes, including the most recent SARS-CoV-2 virus sequences. Additionally, users have access to a variety of other filtering options, including host, isolate sequence length, and many others. The first collection of sequences, known as “HCoV” or “Human Coronavirus dataset,” contains the sequences for the seven human Coronaviruses OC43, HKU1, 229E, NL63, MERS, SARS-CoV-1, and SARS-CoV-2. A total of 66 sequences are included in the dataset belonging to the different HCoV categories (OC43:10, HKU1: 9, 229E: 8, NL63: 8, MERS: 8, SARS-CoV-1: 8, SARS-CoV-2: 15). The dataset is carefully selected to validate the most effective wavelet basis for DNA sequence comparison, enabling the use of these wavelets for SARS-CoV-2 strain categorization.

The Coronavirus is an RNA virus and its associated nucleotides are A (Adenine), G (Guanine), C (Cytosine), and U (Uracil). The RNA template can be converted into cDNA via reverse transcription and then uploaded into the NCBI repository. The virus cDNA sequence is composed of a quartet of nucleotide bases which are Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The downloaded Coronavirus sequences are stored in FASTA format to be processed by the proposed algorithm in MATLAB environment. The second dataset (CoV-variants) is obtained from the NCBI SARS-CoV-2 data hub. The data hub provides various options to select desired SARS-CoV-2 sequences based on the Accession number and Pango lineage. A total of 640 Coronavirus sequences belonging to four SARS-CoV-2 strains B.1.1.7, B.1.2, B.1.526, and P.1 are included in the dataset. While selecting the sequences, only the complete nucleotide sequences which do not consist of any ambiguous characters are considered. The total numbers of sequences for each of the variants are B.1.1.7-168, B.1.2-377, B.1.526-38, and P.1-57. The third dataset is curated from the NCBI SARS-CoV-2 data hub by creating a randomized subset of five variants including Omicron. A total of 2000 SARS-CoV-2 sequences are randomly selected from BA.1 (Omicron), B.1.429, P.1.10, B.1.351, and B.2.525 Pango lineages to classify Omicron variants from others. The random selection of virus sequences by NCBI removes any bias while creating the dataset. The details of the datasets used in this study are described in Table 1.

3.2 Feature extraction methods

Since the SARS-CoV-2 lineages have almost identical genomes with just minor differences in nucleotide composition, feature extraction is the most crucial step in classifying them. Various techniques apart from GSP are available for this purpose (Fiscon et al. 2016; Lebatteux et al. 2019). In this study, we have applied Fourier transform, wavelet transform, and singular value decomposition for the classification of SARS-CoV-2 mutants. Coronavirus genomes cannot be manipulated by digital signal processing-based techniques unless they are converted into discrete-model-based data. For this purpose, the genomic sequences are first converted into a numeric model (Nair and Sreenadhan 2006; Vaegae 2020; Voss 1992; Das and Turkoglu 2018; Hoang et al. 2016; Kar et al. 2021, 2022).

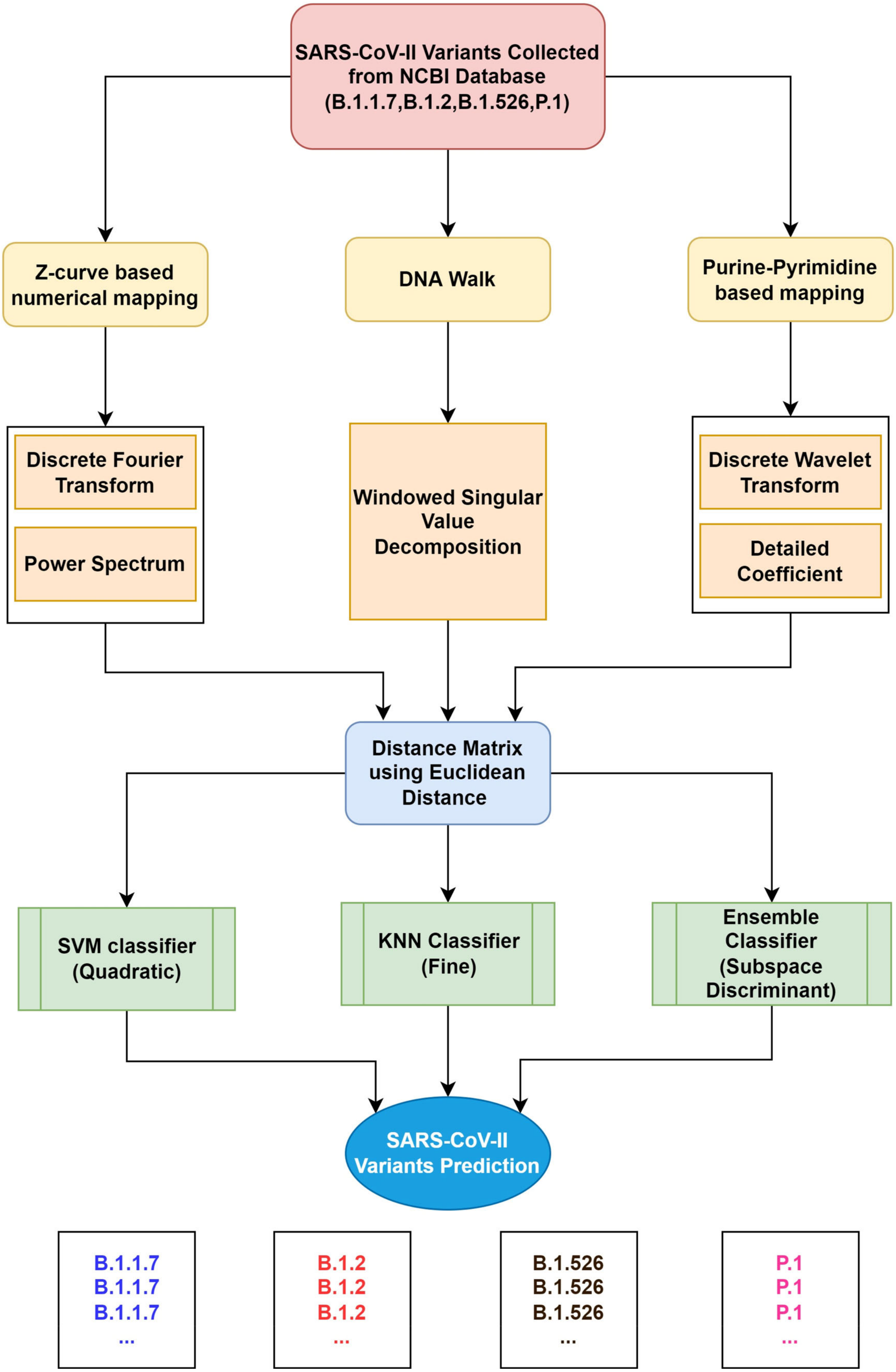

SARS-CoV-II Variants Collected from NCBI Database (B.1.1.7,B.1.2,B.1.526,P.1)
Z-curve based numerical mapping
DNA Walk
Purine-Pyrimidine based mapping
Discrete Fourier Transform
Power Spectrum
Windowed Singular Value Decomposition
Discrete Wavelet Transform
Detailed Coefficient
Distance Matrix using Euclidean Distance
SVM classifier (Quadratic)
KNN Classifier (Fine)
Ensemble Classifier (Subspace Discriminant)
SARS-CoV-II Variants Prediction
B.1.1.7
B.1.1.7
B.1.1.7
...
B.1.2
B.1.2
B.1.2
...
B.1.526
B.1.526
B.1.526
...
P.1
P.1
P.1
...

Fig. 1 Pipeline of descriptors calculated by proposed ZcurveFFT, DNAwalkSVD, and PPDWT methods

The converted sequences are then subjected to techniques based on digital signal processing. To provide data that are appropriate for classification purposes, we have used three approaches for extracting features from signals in this work: PPDWT, DNAwalkSVD, and ZcurveFFT. Machine learning algorithms of various types were used to complete the classification.

3.2.1 Purine–pyrimidine-based discrete wavelet transform (PPDWT)

Purine (A and G) and pyrimidine (T and C) are the two families of nitrogenous bases that make up nucleic acids. In purine–pyrimidine-based mapping, a nucleotide sequence S = ‘ATTCGTTG’ will be replaced by ‘−1 1 1 1 −1 1 1 −1’ (Berger et al. 2004). It is a one-dimensional mapping and thus computationally more efficient compared to 2-D mappings like CGR, and 4-D mapping like Voss. The mapping can be mathematically expressed as

$$X[n] = \begin{cases} 1, & \text{for nucleotide C and T} \\ -1, & \text{for nucleotide A and G} \end{cases}. \quad (1)$$

The numerically encoded signal is decomposed with the help of discrete wavelet transform. The wavelet transform provides a time–frequency analysis of a given signal and allows multi-resolution feature extraction of the signal. For the computation of the detail and approximate coefficients, we employed a second-generation lifting scheme that provides additional advantages compared to FFT and conventional DWT. The traditional first-generation wavelet transform uses a two-band transform scheme. In each step, the signal is fed into a high pass and low passband and then again sub-sampled. On low pass units, there is a risk of occurring recursion. The effectiveness of the lifting scheme enables a sequence of convolution-accumulate operations to be conducted on both the odd and even sequences. As a result, the speed of the second-generation wavelet transform is increased by a factor of 2 (Guntoro and Glesner 2008). Additionally, the second-generation wavelet transform performs the in-place computation of wavelet coefficients, making it an effective tool for clustering and classifying sequences in real time. In this paper, the detail coefficients are

Table 1 Demographics of the three datasets used in this study

Name of dataset	Coronavirus types	Description	Nos of sequences
HCoV	OC43	It is a type of beta Coronavirus that generally infect cattle and human. The virus is responsible for mild colds and fever	10
	HKU1	Type of BetaCoronavirus. It causes upper respiratory disease	9
	229E	It belongs to the AlphaCoronavirus category and infects humans and bats. It is a single-stranded RNA virus	8
	NL63	It belongs to the AlphaCoronavirus genus. The virus causes upper and lower respiratory tract infections	8
	MERS	The Middle East Respiratory Syndrome Coronavirus. The virus can cause severe respiratory disease. The fatality rate of the MERS virus is $\sim$ 30%	8
	SARS-CoV-1	Severe Acute Respiratory Syndrome Coronavirus which belongs to the BetaCoronavirus genus. Emerged from Bats in the year of 2003	8
	SARS-CoV-2	Also known as the COVID-19 virus. Type of BetaCoronavirus which causes severe respiratory disease in humans	15
CoV-Variant	B.1.1.7	Also known as Alpha variants. Emerged in the United Kingdom in November 2020	168
	B.1.2	Emerged in Brazil in February 2021	377
	B.1.526	Also known as Iota variants. The virus emerged from the USA in early 2021	38
	P.1	Popularly known as Gamma variants. The variant was first observed in Brazil in January 2021	57
Omicron	BA.1	The Omicron variant of SARS-CoV-2. Currently marked as a variant of concern[a] by CDC (Centre for Disease Control and prevention)	1329
	B.1.351	Beta variant of SARS-CoV-2. Emerged in South Africa in May, 2020. Currently marked as VBM by CDC	32
	B.1.429	Belongs to the Epsilon variants. Currently marked as variant being monitored by CDC	530
	B.1.525	Commonly known as Eta mutant. The present status is VBM as per CDC	29
	P.1.10	Sub-lineage of Gamma variant. The current status is VBM as per CDC	80

[a] CDC status of all the variants belongs to Omicron Dataset is taken on September, 2023

computed using Swelden's method which is given in the following equation (Sweldens 1998):

$$D[i] = X_o[i] - P \times X_e[i+2], \tag{2}$$

where P is the predict operator, and X_o and X_e are even and odd polyphase components of the signal and are generated by splitting the original numerical sequence $X[n]$. It can be obtained by delay and downsampling of the input signal. The z-transform of even and odd polyphase components are given by

$$X_e(Z) = \sum_n X(2n) z^{-n} \tag{3}$$

$$X_o(Z) = \sum_n X(2n+1) z^{-n}. \tag{4}$$

The detail coefficients can efficiently capture distinctive properties of each Coronavirus variants, because they comprise high-frequency components of the input signal. In this aspect, selecting an appropriate mother wavelet that would produce the greatest results is crucial. In this work, 7 widely used second-generation wavelet families, including Biorthogonal (bior), Coiflets (coif), Daubechies (db), Reverse Biorthogonal (rbio), Symlets (sym), Cohen-Daubechies–Feauveau (cdf), and Haar, were tested on the HCoV dataset. The decomposition level is an essential factor that must be taken into account when developing the algorithm. The maximum decomposition level of the input signal can be found using the formula

$$L < \log_2 \frac{N}{F-1} + 1, \tag{5}$$

where N is the length of the input signal and F is the length of the filter deployed (Chen et al. 2017). For a Coronavirus sequence which is $\sim$ 30 K, the maximum decomposition level is 14.

The length of all the output coefficients is equated to the coefficient generated by the longest sequence of the dataset to get equal dimensionality of wavelet coefficients. For this purpose, we have implemented a uniform scaling technique suggested by Yin and Yau (2015), whereby the shorter sequences are elongated to the longest series by taking consecutive data elements from it. The even scaling method is applied to all the feature extraction methods at the length adjustment stage. As a result, the feature sequences are generated for every input sequence within the dataset. Finally, Euclidean distances between all the feature sequences are computed to build a distance matrix which is then fed to a classifier.

3.2.2 DNA walk-based singular value decomposition (DNAwalkSVD)

The DNA walk provides a simple way of viewing and comparing various genomes by generating a graphical representation of DNA sequences. A DNA walk of any genomic sequence can be computed only after it is converted to the Purine-Pyrimidine model. A simple cumulative summation approach is then applied to draw the DNA walk of a genome (Berger et al. 2004). The DNA walk is calculated using the following expression:

$$Y[i] = \sum_{n=1}^{N} X(n), \tag{6}$$

where $X(n)$ is the equivalent numerical sequence generated by purine–pyrimidine coding of the virus genome and N is the length of the sequence. A two-dimensional DNA walk provides useful information about the change in nucleotide composition at any base location. Therefore, the representation will suit the cause of classification of Coronavirus strains as they contain multiple mutations in spike protein. The windowed SVD technique is applied to every encoded sequence to compute feature sequences. We have employed a window of length 81 to reshape the given sequence into small matrixes. Finally, the maximum value of each SVD is taken into consideration. In this way, the resulting sequence length would be the same as the input sequence. SVD is computationally simple and has a very strong mathematical background in linear algebra (Akhtar et al. 2008). In SVD, rectangular matrix A can be broken down into the product of three matrices—an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V. The theorem is formulated in Eq. (7) as

$$A_{mn} = U_{mm} S_{mn} V_{nn}^{T}, \tag{7}$$

where $U^T U$ = I, VTV = I, the columns of U are orthonormal eigenvectors of AA^T, the columns of V are orthonormal eigenvectors of $A^T A$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order (Zhang et al. 2012). After the computation of all the feature sequences using SVD, their length must be equated to the largest available sequence in the dataset to find the pairwise Euclidean distances. A distance matrix thus computed is used to classify the Coronavirus sequences using machine learning models.

3.2.3 Z-curve-based discrete Fourier transform (ZcurveFFT)

Z-curve is a three-dimensional numerical representation of a genomic sequence. Previous research findings have demonstrated that the Z-curve procedure exhibits robustness and computational efficiency in comparison to alternative numerical encoding procedures when it comes to representing a nucleotide sequence (Zhang and Wang 2000). The three-dimensional components of the Z-curve are independent and contain all the information of a genomic sequence. Furthermore, the Z-curve mapping is one-to-one or bidirectional. Hence the original sequence can be easily obtained from the generated Z-curve. The three-dimensional Z-curve of a genomic sequence can be formulated as

$$X_n = \begin{cases} -1, & \text{for nucleotide C or T} \\ 1, & \text{for nucleotide A or G} \end{cases} \tag{8}$$

$$Y_n = \begin{cases} -1, & \text{for nucleotide T or G} \\ 1, & \text{for nucleotide A or C} \end{cases} \tag{9}$$

$$Z_n = \begin{cases} -1, & \text{for nucleotide C or G} \\ 1, & \text{for nucleotide A or T} \end{cases}. \tag{10}$$

In this way, Z-curve describes the distribution of the bases of purine/pyrimidine, amino/keto, and strong/weak hydrogen bonds along with three sequences (Khodaei et al. 2020b). Once the three vectors are calculated, the discrete Fourier transform is applied to all of the vectors. The DFT of input sequence X_n can be computed as

$$X_k = \sum_{n=0}^{N-1} X_n e^{j2\pi nk/N}, 0 \le k \le N-1. \tag{11}$$

In a similar way, we can find Y_k and Z_k, respectively. To decrease the time complexity of DFT, Fast Fourier Transform (FFT) can be implemented in an MATLAB environment. It has been previously seen that Z-curve representation and FFT work well together (Yan et al. 1998). After the Fourier computation is over, the Power spectral density (PSD) plot is obtained using the formula (Tiwari et al. 1997)

$$P_{xyz} = \{\mathrm{DFT}(X_n)\}^2 + \{\mathrm{DFT}(Y_n)\}^2 + \{\mathrm{DFT}(Z_n)\}^2. \tag{12}$$

The length normalization process is used to address the problem of varied Coronavirus sequence lengths. Finally, the pairwise Euclidean distances are computed to find the distance matrix. The matrix will work as an input of machine learning algorithms for classification.

3.3 Classification approaches

In this stage, we tested various classifiers for the classification of HCoV, CoV-Variants, and Omicron datasets. Machine learning classifiers that are extensively employed for classification include Decision tree (DT), Random Forest (RF), Support Vector Machine (SVM), K nearest neighbors (KNN), and Naïve-Bayes (NB) are subjected to evaluation on the three chosen datasets. In our paper, we consider SVM, KNN, and Ensemble classifiers as the preferred choices due to their high accuracy measures.

3.3.1 Support Vector Machine

SVM is a supervised machine learning algorithm that was developed by Vapnik. It can be used for binary and multiclass classification problems (Press 2007; Osuna 1998). The SVM can perform linear as well as non-linear classification using kernel tricks by mapping the features into higher dimensions. It classifies the input points by drawing various hyperplanes in the feature space. Based on the kernel function, various types of SVM can be performed which include linear SVM, quadratic SVM, cubic SVM, and Gaussian SVM. In this study, the selection of a quadratic Support Vector Machine (SVM) is based on its demonstrated reliability in achieving consistent classification accuracy across various types of features.

3.3.2 K-nearest neighbors

KNN is a non-parametric machine learning algorithm that can be used in supervised or non-supervised classification. The method was developed by Fix and Hodges. KNN algorithm finds observations in the feature space that are similar to the new observation (Cover and Hart 1967; Duda and Hart 2001). These are called neighbors (K). The number of neighbors can be varied to get better classification accuracy. It classifies new input points by observing the K neighbors. To choose K-nearest neighbors various distance metrics including city block, Chebyshev, correlation, cosine, Euclidean, Hamming, Jaccard, and Mahalanobis can be considered. The optimum values of various parameters can be adjusted using hyper-parameter tuning during the training process. The paper examines the fine KNN model due to its proficient and expeditious processing.

3.3.3 Ensemble

The ensemble model combines various individual learning algorithms to increase the classification accuracy of constituent learning algorithms (Ho 1998; Dey et al. 2020). The time complexity of the algorithm is quite high, since it combines various individual ML algorithms. There are mainly three types of ensemble techniques: Bagging, Boosting, and Subspace. In this paper, we have examined the subspace discriminant model due to its superior performance in SARS-CoV-2 variants classification compared to others. The model is used to increase the accuracy of discriminant analysis. Among all the ensemble methods subspace ensembles require less memory.

4 Performance evaluation

To assess the performance of the proposed methods, various evaluation parameters are considered including sensitivity (Sn), specificity (Sp), accuracy (Acc.), F1 scores, and Matthews correlation coefficient (MCC). In parallel, the time complexity of each of the three algorithms is also taken into account. The definitions of these parameters are as per the below equations

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{14}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{15}$$

$$F1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{16}$$

$$\text{MCC} = \frac{(\text{TN} \times \text{TP}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}}. \tag{17}$$

Here, TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively. These parameters can be found in the generated confusion matrix at each classification stage. For multiclass classification, the average value of sensitivity can be computed by summing up all the sensitivity values and then dividing them by the total number of classes. Similarly, mean values of other matrices can be derived.

4.1 Validation checking

The performance of each classifier is evaluated using fivefold cross-validation. In this process, all the data points are divided into 5 subsets, and then, the average accuracy is calculated. In each classification step, 80% of the data are used for training purposes and the remaining 20% are used for testing. The aforementioned procedure is superior to that of hold-out validation due to its capability to mitigate the inherent bias and variance of the data points. In the case of hold-out validation, all the data are subdivided into two subsets for training and testing, respectively. The proportion of training and testing data can be varied. However, in this way, each and every data point cannot be taken for training purposes. For practical cases, 50–80% of data points are selected randomly for training purposes and the remaining for testing. Since the hold-out method involves a single run, this process is to be repeated many times and the final accuracy can be computed by averaging. In this way, the overall process will become time-consuming as the results are to be simulated many times for averaging. To ease the process, fivefold or tenfold cross-validation is preferred which can evaluate the model with similar efficacy but with less time complexity. Furthermore, we have conducted hyper-parameter tuning employing a Bayesian optimizer to select the optimal parameters for both SVD and KNN models. We refrained from conducting hyper-tuning on the ensemble classifier due to its relatively high accuracy, and the introduction of hyper-tuning would result in a slight delay in the process.

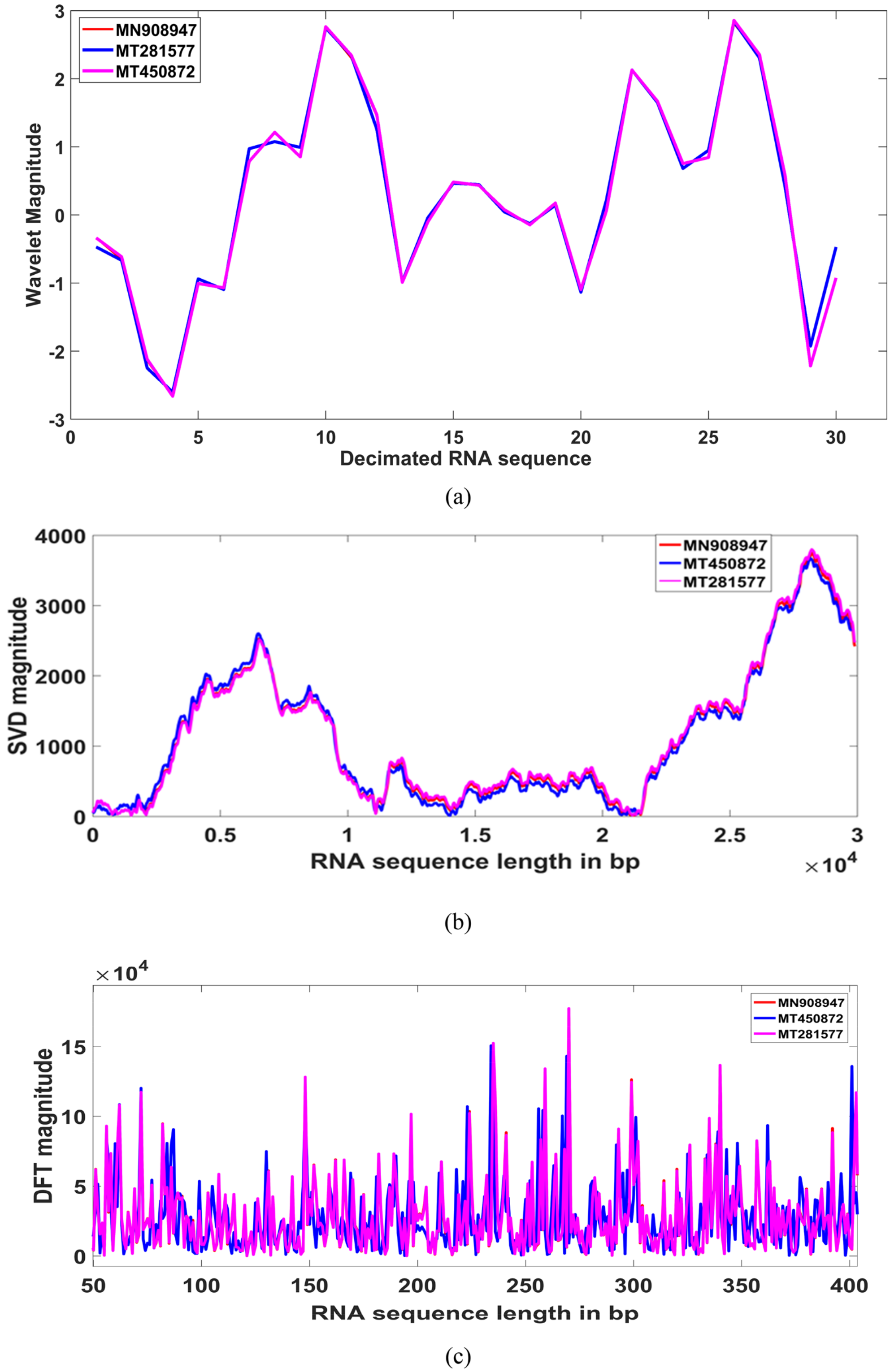
5 Results and discussion

In this section, we have analyzed the results of the proposed alignment-free approach for categorizing different Coronavirus strains utilizing methods from digital signal processing. Due to its affordability, precision, and speed of processing, the signal processing technique has become quite popular for the comparison and classification of genomic and proteomic sequences. This analysis focuses on three feature selection methodologies, namely, PPDWT, DNAwalkSVD, and ZcurveFFT. The features generated from the proposed technique for three SARS-CoV-2 sequences can be visualized using Fig. 2.

The diagrams in Fig. 2 depict the visual interpretation of the three GSP-based features when computed on three SARS-CoV-2 sequences. The Genbank accession numbers of the sequences are MN908947, MT450872, and MT281577. The diagrams clearly indicate that DFT, DWT, and SVD could extract sophisticated features from SARS-CoV-2 to discriminate them. Figure 2a shows the DFT magnitudes corresponding to the first 60 bases for three SARS-CoV-2 sequences. Similarly, Fig. 2b, c provides SVD and DWT magnitudes corresponding to every base position. However, Fig. 2c provides the truncated feature plot up to 400 base pair of the whole SARS-CoV-2 sequences. Every SARS-CoV-2 strain having specific mutations in spike protein can easily be traced down by these signal processing methods. Finally, these distinguishable features are fed into the ML classifier for variant identification.

5.1 Selection of best wavelet in PPDWT feature for classification of SARS-CoV-2 virus variants by validating HCoV dataset

In contrast to the PPDWT approach, which requires the choice of two key parameters, the feature selection process in DNAwalkSVD and ZcurveFFT algorithms is remarkably simple. In the case of PPDWT, the selection of appropriate mother wavelet and decomposition levels is essential, since the classification accuracy varies vastly according to these parameters. To do that, we have validated the HCoV



(a)

(b)

(c)

Fig. 2 The magnitude feature plots of SARS-CoV-2 viruses calculated using various GSP features. **a** PPDWT, **b** DNAwalkSVD, **c** ZcurveFFT

dataset which comprises 66 human Coronavirus sequences. To choose the optimum mother wavelet and decomposition level, the HCoV dataset tests all seven second-generation mother wavelets at 14 distinct decomposition levels. KNN, SVM, and Ensemble methods are used at the classifier stage to act upon 66 × 66 distance matrices computed using DWT detail coefficients for the classification of various Human Coronaviruses. Tables 2, 3 and 4 show the results of the KNN, SVM, and Ensemble classifier, respectively, while using PPDWT feature.

To analyze the results, in a certain wavelet family, if several levels achieved the same accuracy, then the minimum level is noted in the tables. Referring the Tables 2, 3 and 4, it is evident that Bior wavelet could achieve the highest accuracy in identifying the human Coronavirus types. Most Bior wavelets produced the best accuracy of 100% using all three classifiers. Another highly effective mother wavelet that has the capacity to yield a superior outcome in the classification of genomes is predicted as Db, as substantiated by the data obtained from the computation of the best and average accuracy values. This result can be verified from Fig. 3 where average accuracy for each mother wavelet is plotted for all the classifiers. Henceforth, we have selected Bior and Db for feature selection in the CoV-variants dataset.

5.2 Selection of best window length in DNAwalkSVD feature

Although SVD is a straightforward decomposition method of the matrix, the present study implemented a windowed version of SVD to classify SARS-CoV-2 lineages. The value of window length can change the accuracy and runtime of the feature selection process. Since DNA is constituted of codons, a sequence of three nucleotides responsible for forming a unit of genomic information encoding a particular amino acid, we choose the window length as a multiple of three. This will enable the algorithm to analyze each codon and therefore capture better attributes from SARS-CoV-2 variants. Furthermore, a wide range of window lengths is examined using the CoV-Variant dataset for choosing the best option in terms of time requirement and preciseness. The result obtained using various window lengths is given in Table 5.

From Table 5, it is evident that the best accuracy of 98.6% is obtained by employing a window length of 81 for the ensemble classifier. Also, the accuracy scores are 97 and 95.6% utilizing SVD and KNN classifiers, respectively. Although these accuracies are not the highest in the case of corresponding classifiers, they are relatively close to the highest mark. However, the time complexity of the DNAwalkSVD technique increases with an increase in window length. This is because the matrix dimension expands as the window length gets larger. The input matrix of SVD is formed using the windowed data, such that the number of rows will be three and the number of columns is of the size of (Window length/3). The largest singular value of the input matrix is taken into account. Therefore,

Table 2 Wavelet member, best decomposition level and corresponding accuracy, average accuracy obtained using Fine KNN classifier employing HCoV dataset

Dataset	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc
HCoV	**Db1**	**97/11**	**91.56**	Cdf4.2	89.4//7	82.47	**Bior3.5**	**100/13**	**91.98**
	Db2	**93.9/10**	**87.54**	Cdf4.4	84.8/1	81.05	**Bior3.7**	**100/13**	**95.78**
	Db3	**95.5/9**	**87.32**	Cdf4.6	89.4/4	78.02	**Bior3.9**	**100/13**	**95.57**
	Db4	**100/12**	**92.87**	Sym2	95.5/12	87.86	**Bior4.4**	**100/12**	**91.55**
	Db5	**100/12**	**91.33**	Sym3	100/12	92.21	**Bior5.5**	**89.4/1**	**73.91**
	Db6	**90.9/1**	**85.71**	Sym4	97/9	90.58	Rbio1.1	97/13	90.58
	Db7	**97/12**	**91.45**	Sym5	95.5/1	88.20	Rbio1.3	97/13	91.67
	Db8	**92.4/7**	**86.15**	Sym6	97/10	87.44	Rbio1.5	97/12	91.35
	Haar	97/11	90.80	Sym7	98.5/9	89.34	Rbio2.2	93.9/8	86.47
	Cdf1.1	97/11	91.56	Sym8	100/12	91.56	Rbio2.4	93.9/10	89.16
	Cdf1.3	95.5/11	91.45	**Bior1.1**	**100/13**	**92.10**	Rbio2.6	95.5/10	87.6
	Cdf1.5	97/13	91.12	**Bior1.3**	**100/13**	**92.52**	Rbio2.8	93.9/8	89.10
	Cdf2.2	93.9/10	86.35	**Bior1.5**	**100/13**	**91.77**	Rbio3.1	93.9/7	86
	Cdf2.4	93.9/10	88.62	**Bior2.2**	**100/13**	**93.28**	Rbio3.3	93.9/8	87.32
	Cdf2.6	93.9/8	87.44	**Bior2.4**	**100/13**	**91.71**	Rbio3.5	95.5/8	87.65
	Cdf3.1	92.4/1	84.07	**Bior2.6**	**100/14**	**90.58**	Rbio3.7	87.9/7	82.13
	Cdf3.3	92.4/9	87.11	**Bior3.1**	**100/13**	**92.98**	Rbio3.9	89.4/7	82
	Cdf3.5	93.9/8	87.76	**Bior3.3**	**100/13**	**92.31**	Rbio4.4	92.4/9	84.61

Bold font wavelets (Db and Bior) are considered for feature extraction in subsequent studies as it produced good accuracy scores over other wavelets

Table 3 Wavelet member, best decomposition level and corresponding accuracy, average accuracy obtained using Quadratic SVM classifier employing HCoV dataset

Dataset	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc
HCoV	**Db1**	**97/12**	**90.48**	Cdf4.2	87.9/3	81.16	**Bior3.5**	**100/13**	**93.09**
	Db2	**92.4/12**	**87.2**	Cdf4.4	84.8/1	81.15	**Bior3.7**	**100/13**	**96.86**
	Db3	**93.9/8**	**86.56**	Cdf4.6	90.9/4	80.40	**Bior3.9**	**100/13**	**97.3**
	Db4	**100/12**	**92.74**	Sym2	95.5/10	87.44	**Bior4.4**	**98.5/12**	**91.92**
	Db5	**100/12**	**92.2**	Sym3	100/12	92.21	**Bior5.5**	**90.9/1**	**74**
	Db6	**93.9/1**	**83.97**	Sym4	97/11	91.67	Rbio1.1	97/13	90.04
	Db7	**97/12**	**90.91**	Sym5	95.5/1	87.12	Rbio1.3	97/13	91.22
	Db8	**93.9/1**	**83.32**	Sym6	95.5/10	86.04	Rbio1.5	98.5/12	91.67
	Haar	93.9/11	88.72	Sym7	98.5/9	90.16	Rbio2.2	97/8	86.15
	Cdf1.1	97/12	90.26	Sym8	100/12	92.54	Rbio2.4	93.9/10	88.18
	Cdf1.3	98.5/8	90.57	**Bior1.1**	**98.5/12**	**91.23**	Rbio2.6	97/10	87
	Cdf1.5	95.5/13	89.93	**Bior1.3**	**98.5/13**	**90.95**	Rbio2.8	95.5/8	89.16
	Cdf2.2	92.4/10	86.8	**Bior1.5**	**100/14**	**91.95**	Rbio3.1	84.8/11	80.51
	Cdf2.4	93.9/8	88.73	**Bior2.2**	**100/12**	**93.30**	Rbio3.3	92.4/8	84.51
	Cdf2.6	93.9/10	87.97	**Bior2.4**	**100/12**	**93.72**	Rbio3.5	93.9/9	86.67
	Cdf3.1	89.4/1	80.83	**Bior2.6**	**100/14**	**92.54**	Rbio3.7	86.4/7	74.12
	Cdf3.3	92.4/9	84.17	**Bior3.1**	**100/13**	**93.52**	Rbio3.9	89.4/3	81.26
	Cdf3.5	93.9/9	87.10	**Bior3.3**	**100/13**	**93.5**	Rbio4.4	95.5/8	85.27

Bold font wavelets (Db and Bior) are considered for feature extraction in subsequent studies as it produced good accuracy scores over other wavelets

Table 4 Wavelet member, best decomposition level and corresponding accuracy, average accuracy obtained using subspace discriminant ensemble classifier employing HCoV dataset

Dataset	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc	Wavelet	Best (Acc/Lev)	Avg Acc
HCoV	**Db1**	**97/11**	**92.65**	Cdf4.2	92.4/9	87.12	**Bior3.5**	**100/13**	**93.17**
	Db2	**93.9/9**	**91.22**	Cdf4.4	92.4/3	85.92	**Bior3.7**	**100/13**	**97.29**
	Db3	**97/7**	**92**	Cdf4.6	92.4/5	84.75	**Bior3.9**	**100/13**	**97.2**
	Db4	**100/12**	**94.59**	Sym2	97/11	92.30	**Bior4.4**	**100/12**	**93.83**
	Db5	**100/12**	**93.3**	Sym3	100/12	93.84	**Bior5.5**	**90.9/1**	**79.7**
	Db6	**92.4/12**	**88.74**	Sym4	97/10	93.72	Rbio1.1	98.5/14	92.75
	Db7	**98.5/12**	**93.62**	Sym5	97/10	92.1	Rbio1.3	98.5/13	94.37
	Db8	**95.5/3**	**89.10**	Sym6	98.5/10	91.37	Rbio1.5	98.5/12	93.18
	Haar	97/11	92.86	Sym7	100/9	92.20	Rbio2.2	95.5/8	90.7
	Cdf1.1	97/11	92.65	Sym8	100/12	94.27	Rbio2.4	97/12	92
	Cdf1.3	97/14	92.96	**Bior1.1**	**100/13**	**93.29**	Rbio2.6	97/10	91.32
	Cdf1.5	97/13	93.72	**Bior1.3**	**100/13**	**93.94**	Rbio2.8	95.5/10	91.85
	Cdf2.2	95.5/4	89.67	**Bior1.5**	**100/13**	**93.62**	Rbio3.1	95.5/7	89.17
	Cdf2.4	97/10	92.34	**Bior2.2**	**100/13**	**94.26**	Rbio3.3	95.5/9	90.25
	Cdf2.6	93.9/9	90.25	**Bior2.4**	**100/13**	**93.94**	Rbio3.5	95.5/9	91.56
	Cdf3.1	92.4/4	87	**Bior2.6**	**100/14**	**92.31**	Rbio3.7	90.9/8	85.3
	Cdf3.3	95.5/8	90.67	**Bior3.1**	**100/14**	**94.81**	Rbio3.9	93.9/2	85.39
	Cdf3.5	97/11	91.65	**Bior3.3**	**100/13**	**93.62**	Rbio4.4	95.5/8	89.92

Bold font wavelets (Db and Bior) are considered for feature extraction in subsequent studies as it produced good accuracy scores over other wavelets

for $\sim$ 30 K base-pair long SARS-CoV-2 sequences, we will have 30 K singular values corresponding to each frame which will generate the feature vector for a single sequence. The computation of SVD requires matrix

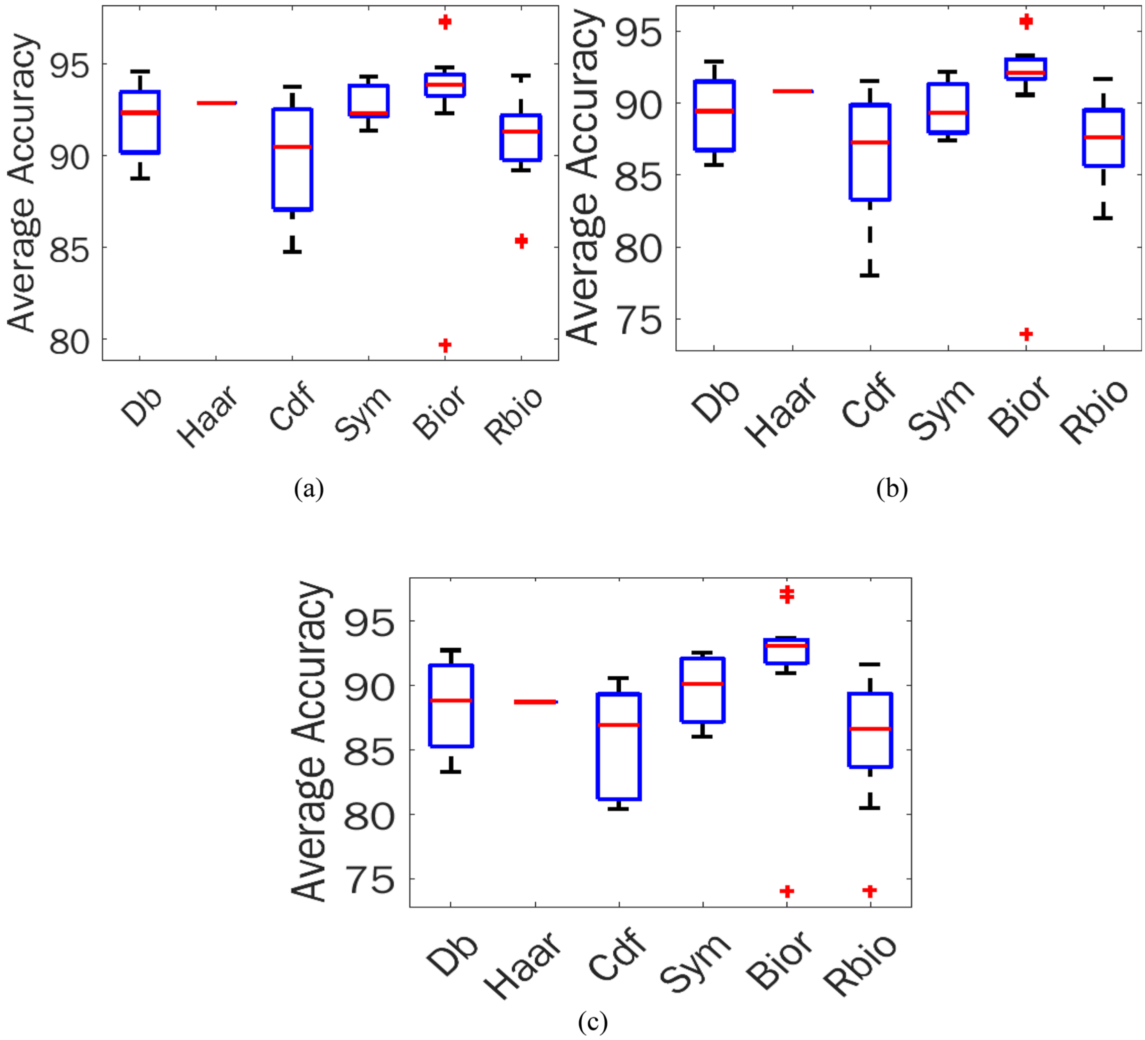


Fig. 3 Box plot for the selection of best wavelet in case of classification of SARS-CoV-2 virus engaging HCoV dataset. **a** Ensemble **b** KNN, and **c** SVM

Table 5 Variation of accuracy using SVD of different window lengths when tested on CoV-Variant dataset

Window length	Classifier			Run time in s
	KNN	SVD	Ensemble	
30	96.1	96.9	97.3	220
60	95.6	96.9	97.2	228
81	95.6	97	98.6	236
120	95.6	96.9	97.2	240
150	96.6	97	98	243
210	95.2	95.8	96.7	254
270	95.8	97.3	97.2	260
360	94.8	96.1	97.5	273
450	95.6	96.7	96.9	423
540	95.2	97.2	97	480
600	96.3	96.7	97	490

multiplication and solving polynomial equations to find the singular values. The general time complexity of matrix multiplication of $m \times n$ matrix by an $n \times p$ matrix is $O(mnp)$. Therefore, time complexity will increase for higher dimension matrix multiplication. Also, higher order polynomial equation must be solved to find singular values in case of higher dimensional matrix. Thus the window length of the DNAwalkSVD feature is selected as 81 and used in subsequent studies.

5.3 Classification of CoV-Variant dataset

This dataset is made of four SARS-CoV-2 lineages, namely, B.1.1.7, B.1.2, B.1.526, and P.1. In the case of PPDWT feature selection, we have applied Bior and Db mother wavelets up to 14 decomposition levels for feature selection. A total of 21 separate mother wavelets belonging to the Biorthogonal and Daubechies families are

Table 6 The result of the application of Db and Bior wavelets in CoV-Variant dataset

Dataset	Wavelet	SVM: quadratic best (Acc/level)	KNN: fine Best (Acc/level)	Ensemble: subspace discriminant Best (Acc/level)
CoV-Variant	Db1	95.8/12	93.8/12	95.3/9
	Db2	95.6/5	93.3/5	97.5/9
	Db3	95.2/7	93.1/7	95.5/7
	Db4	96.6/10	94.5/10	97.5/11
	Db5	96.7/12	95/12	97.3/9
	Db6	95.6/5	94.1/5	95.9/7
	Db7	95.5/6	93.9/6	95.3/5
	Db8	95.3/5	93.8/5	93.6/6
	Bior1.1	96.3/12	94.1/14	94.8/14
	Bior1.3	95.5/12	93.9/8	95.2/14
	Bior1.5	96.3/12	93.9/8	95.6/13
	Bior2.2	96.6/11	95.2/11	98/10
	Bior2.4	97.2/10	94.7/12	98.1/9
	Bior2.6	96.7/11	95.5/13	97.8/11
	Bior3.1	**97.7/10**	**94.4/10**	**98.9/10**
	Bior3.3	96.4/12	96.4/12	98/9
	Bior3.5	97.5/12	97/12	98.4/9
	Bior3.7	96.1/10	94.7/10	97/9
	Bior3.9	95.6/9	94.4/12	96.3/9
	Bior4.4	96.4/10	93.8/5	97.8/10
	Bior5.5	94.7/5	93.3/5	93.8/5

Bior 3.1 wavelet highlighted in bold font indicates the chosen wavelet for PPDWT feature

considered for the computation of accuracy. Three popular machine learning algorithms SVM, KNN, and Ensemble learner are utilized at the classification level. Specifically, for advanced model selection, quadratic SVM, fine KNN, and subspace discriminant models are chosen as they resulted in best accuracy values. The result can be found in Table 6.

Results obtained from Table 6 showed that Bior wavelet can perform consistently to generate superior accuracies. It is also evident that most of the best accuracies come from a certain decomposition level band, i.e., 10–12. As the decomposition level increases from 1 to 14, classification accuracy is improved further, but that in turn increases the computation cost. However, in some cases, the accuracies declined at level 14. Therefore, while selecting the level, the lowest level is considered if the same accuracy is generated at different decomposition levels for a particular mother wavelet. Though all the mother wavelets utilized for feature generation in CoV-Variant dataset produced over 90% accuracy, the best one is Bior3.1 at level 10 generating the highest accuracy of 98.9% using an ensemble classifier. Subsequently, we used the attribute generated by Bior3.1 at level 10 as a PPDWT feature for our experiment. The average running times at the classification stage of CoV-Variant dataset are 1.9, 2.8, and 20 s, respectively, for KNN, SVM, and Ensemble classifier. Therefore, to gain better accuracy, time complexity is also increasing. To ascertain the best accuracy values of KNN and SVM learners, we applied the Bayesian optimization technique which considers an extensive array of designing parameters to shuffle between and provide the best outcomes. Application of the tuned parameters increased the classifier performance, and thus, greater accuracies are obtained. The resulting accuracies before and after hyper-parameter tuning are mentioned in Table 7. Additional information about Bayesian optimization is provided in supplementary file.

The table presents the calibrated parameters that will yield optimal outcomes for the KNN and SVM algorithms. Bayesian optimization is not considered in the case of ensemble classifier as its time complexity is quite high and optimization will further increase the execution time. Results showed that measurement accuracy was greatly enhanced with tuned parameters in the case of the SVM classifier. The SVM model achieves an accuracy of 98.6% through the utilization of the PPDWT technique following the process of hyper-parameter tuning. However, the results of the KNN classifier do not increase significantly even after the optimization. The best accuracy yielded using the KNN classifier is 95.8% post-Bayesian

Table 7 Results of application of Bayesian optimization on SVM and KNN classifier

Method	SVM			KNN			Ensemble
	Optimized hyper-parameter	Acc (before)	Acc (after)	Optimized hyper-parameter	Acc (before)	Acc (after)	Acc (without optimization)
PP-DWT	Box constraint level: **8.11** Multiclass method: **One-vs-all** Standardize data: **false**	97.7	98.6	No. of neighbor: **1** Distance metric: **Spearman** weight: **Inverse** **Standardize** data: **true**	94.4	95.6	98.9
DNAwalk SVD	Box constraint level: **7.55** Multiclass method: **One-vs-one** Standardize data: **false**	96.6	97.3	No. of neighbor: **1** Distance metric: **Minkowski** weight: **Inverse** data: **true**	95.6	95.8	98.6
ZcurveFFT	Box constraint level: **11.16** Multiclass method: **one-vs-all** Standardize data: **false**	97.5	98.1	No. of neighbor: **1** Distance metric: **Euclidean** Weight: **Squared inverse** Data: **false**	94.8	95.3	98.8

Best parameters are indicated by bold font

The optimization is not considered in the case of ensemble classifier owing to higher time complexity

optimization. From the results, it is evident that all the three designed algorithms are equally effective in classifying SARS-CoV-2 lineages. Before the application of Bayesian optimization, separate histograms are generated to illustrate the mean accuracies of ensemble, KNN, and SVM models derived from Table 6 to assess the effectiveness of different classifiers. The results are given in Fig. 4

From the result, it can be seen that the mode of best accuracy obtained using the ensemble classifier is 98%, whereas it is 94% for the KNN classifier and 96.5% for the SVM classifier. Therefore, the ensemble classifier produces the best result in terms of accuracy. The performance of the ensemble (subspace discriminant) classifier to discriminate variants of SARS-CoV-2 is assessed by evaluation parameters like sensitivity, specificity, F1-score, and Mathew's Correlation Coefficient. The results are given in Tables 8, 9, and 10.

The tables provided above present a thorough analysis of the effectiveness of three proposed techniques in relation to the classification of SARS-CoV-2 variants when exposed to subspace discriminant ensemble classifier. The accuracy of each category is above 99% for most of the cases. The potential cause for the relatively reduced accuracy of 98.91% discovered in Table 9 in relation to the B.1.526 variant might be attributed to the limited abundance of this specific variant. This finding indicates that although machine learning techniques are often employed with large amounts of data, they may be applied in situations where obtaining large and labeled datasets may be burdensome. The results of comparative analysis of proposed methods employing various evaluation parameters are provided in Fig. 5.

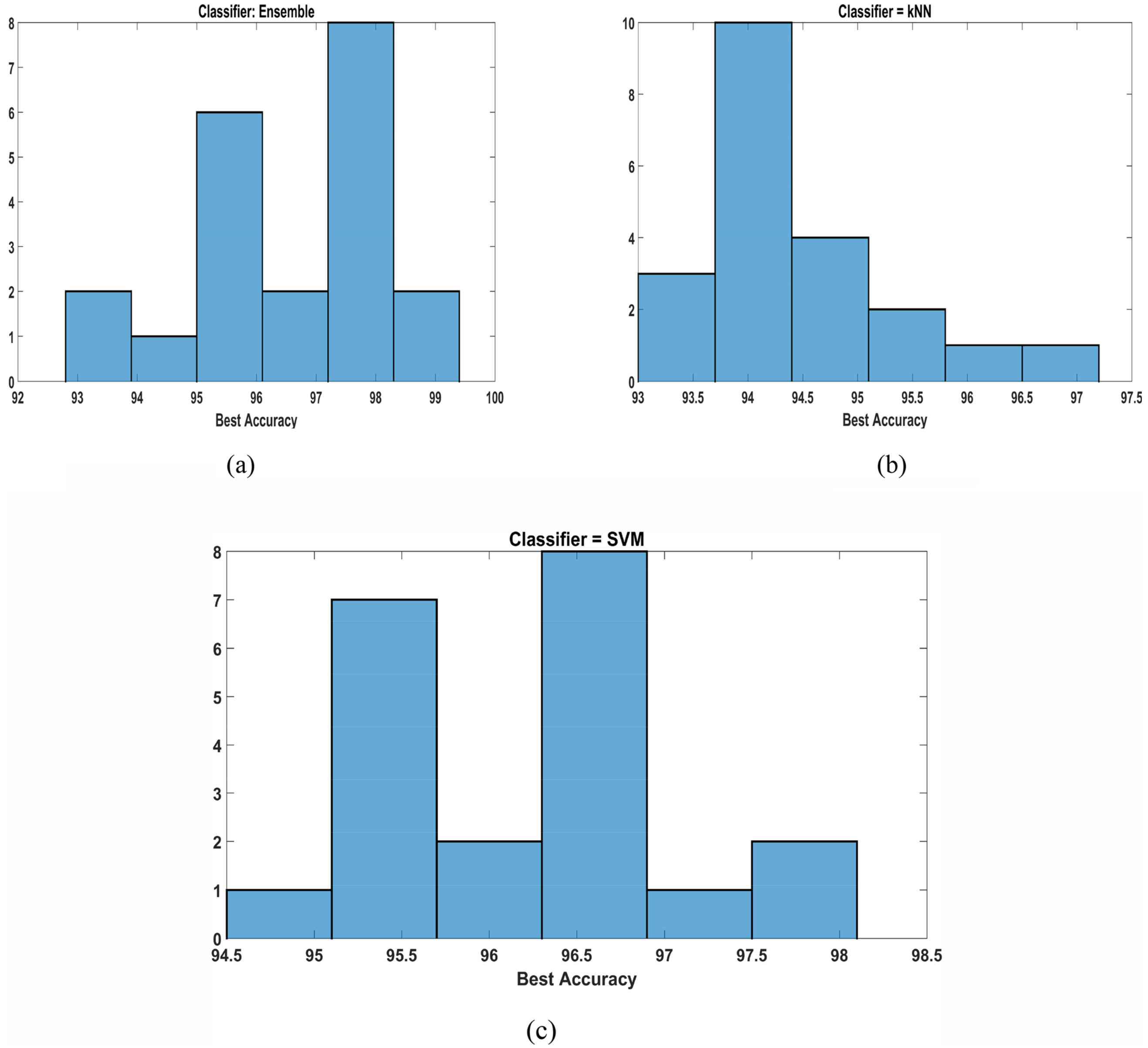
The average computation time of each of the three feature extraction methods is noted in Table 11. The average accuracy in terms of correct classifications is obtained using the expression below

$$\text{Average Accuracy} = \frac{\text{Numbers of Correct Classifications}}{\text{Total Numbers of Samples}}. \tag{18}$$

Total number of misclassification yielded during the run of each method is also given in Table 11.

We have applied three signal processing-based feature extraction methods to classify the SARS-CoV-2 variants and obtained close to 99% accuracy for the proposed three feature extraction methods. Fivefold cross-validation was performed on the CoV-variant dataset to obtain the results. Analyzing the results, it can be said that the PP-DWT method produced the best classification with an accuracy value of 98.9% while using the ensemble classifier. The graphical representation of the results is provided in Fig. 6.

The PPDWT methodology offers a diverse range of mother wavelets and decomposition levels for the purpose of extracting distinctive characteristics. This, in turn, generates a plethora of possibilities for experimentation, thereby significantly increasing the likelihood of achieving a high level of accuracy. Whereas Fourier transform (FT) captures different frequency components of the signal at a



(a)

(b)

(c)

Fig. 4 Histograms for showing the distribution of best accuracies using PPDWT feature when assessed using **a** ensemble, **b** KNN, and **c** SVM classifier

Table 8 Various evaluation metrics obtained using the ensemble classifier

SARS-CoV-2 lineages	Sensitivity	Specificity	F1 score	MCC	Accuracy
B.1.1.7	98.80	99.15	98.21	97.58	99.06
B.1.2	100	99.62	99.87	99.68	99.84
B.1.526	92.11	99.83	94.59	94.30	99.38
P.1	98.25	100	99.12	99.03	99.84
Average	97.29	99.65	97.94	97.64	99.53

The result is obtained using PP-DWT method

single level; hence, feature extraction method is limited to only one possible way. On the other hand, the Fourier transform has zero temporal resolution, so the signal cannot be analyzed in the joint time–frequency domain. Another major problem of FT is that any discontinuity in the input signal cannot be represented appropriately. FT in

Table 9 Various evaluation metrics obtained using the ensemble classifier

SARS-CoV-2 lineages	Sensitivity	Specificity	F1 score	MCC	Accuracy
B.1.1.7	99.40	99.58	99.11	98.8	99.53
B.1.2	100	99.24	99.74	99.36	99.7
B.1.526	99.74	99.17	91.14	90.63	98.91
P.1	89.47	100	99.44	94.11	99.06
Average	97.15	99.5	97.35	95.65	99.3

The result is obtained using DNA-Walk SVD method

Table 10 Various evaluation metrics obtained using the ensemble classifier

SARS-CoV-2 lineages	Sensitivity	Specificity	F1-score	MCC	Accuracy
B.1.1.7	98.21	99.36	98.21	97.58	99.06
B.1.2	100	99.62	99.87	99.68	99.84
B.1.526	92.11	100	95.89	95.73	99.53
P.1	96.49	99.31	94.83	94.33	99.06
Average	96.70	99.57	97.2	96.83	99.37

The result is obtained using Z-Curve FFT method

combination with three-dimensional Z-curve representation measured 98.8% accuracy for the CoV-Variant dataset which is less than the PPDWT method. The time complexity of both the PPDWT and ZcurveFFT methods are relatively the same. The computational complexity is high in the case of DNAwalkSVD as it runs a window through the genomic sequence and then SVD is calculated. The average time taken by DNAwalkSVD is 235 s for 10 consecutive runs. The average specificity of the proposed techniques varies from 99.5% to 99.65% as derived from Tables 8, 9, and 10 implying a minimal number of misclassifications. Also, the mean sensitivities are 97.29, 96.70, and 95.90%, respectively, for PPDWT, ZcurveFFT, and DNAwalkSVD. The high values of F1 and MCC suggest that all four variants have been appropriately classified by the proposed methods. Various confusion matrix obtained using PPDWT features are given in Fig. 7 to validate the results.

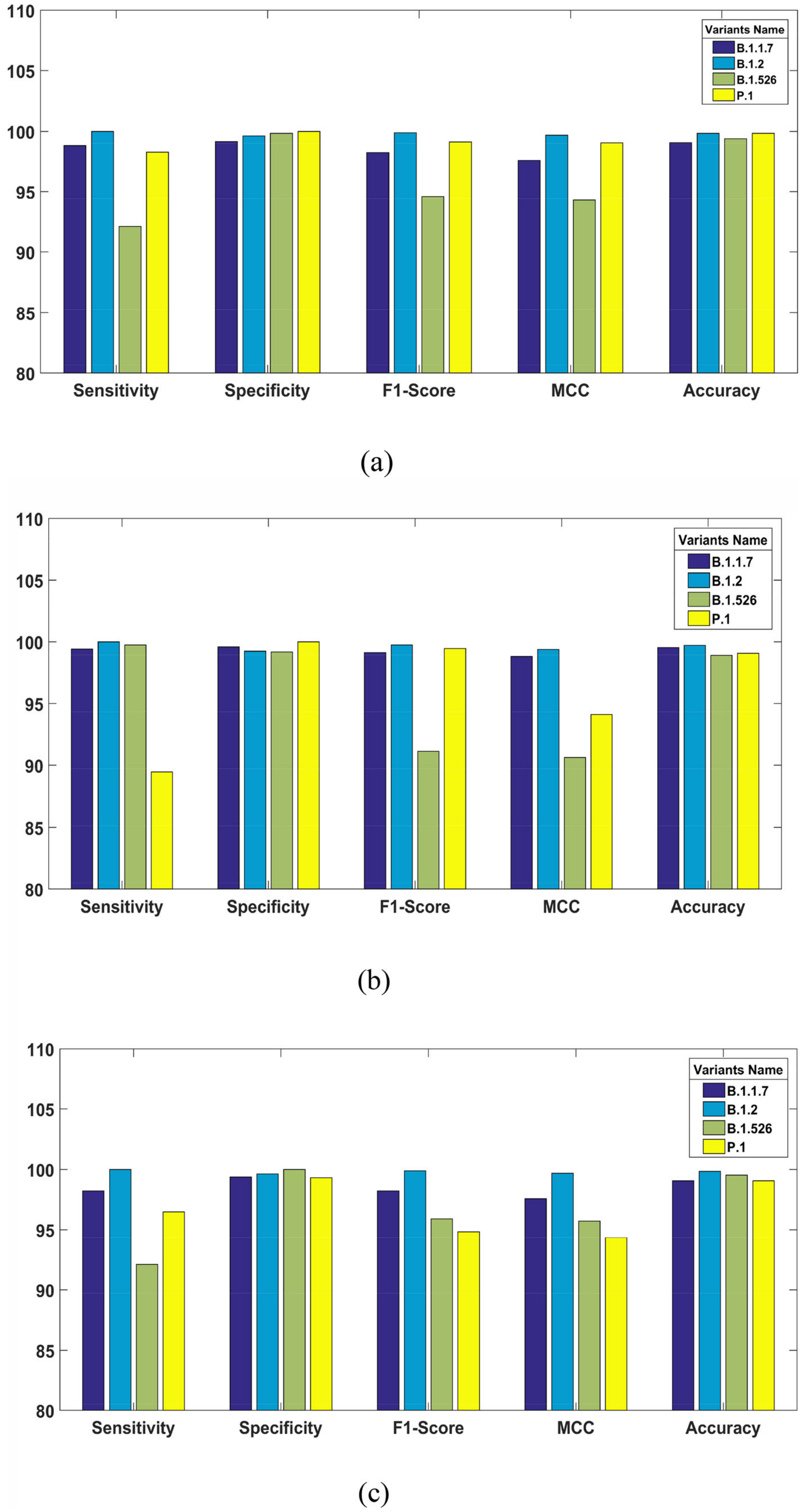
Figure 7a provides numbers of correctly and wrongly classified variants from each category. Figure 7b, c provides additional information, such as true-positive rate, false-negative rate, positive predictive value, and false discovery rate of the classification. The confusion matrices tell the misclassification rates for all the four variants of Coronavirus investigated in our study. The proposed algorithm classified lineage B.1.2 with 100% accuracy, whereas there are a total of four misclassifications out of 168 in the case of B.1.1.7, two misclassifications out of 38 in the case of B.1.526, and three misclassifications out of 57 in the case of P.1. The positive predictive values (PPV) obtained as 97.6, 99.5, 97.3, and 96.4%, respectively, for the variants B.1.1.7, B.1.2, B.1.526, and P.1. When we calculated the confusion matrix for the ZcurveFFT and DNAwalkSVD techniques, we observed similar outcomes. However, best results were achieved when utilizing the PPDWT method. The scatter plots for the first two features (D1 and D2) of the distance matrix for all the three methods are given in Fig. 8.

The distance matrix has a total of 640 distance columns which are used as a feature in the classification stage. The four lineages are marked with separate colors in the scatter plot. To examine the significance of the generated features, we conducted one one-way ANOVA test. It is a filter-based feature selection method used to find the most relevant features using the metric p value. If the feature generates a p value less than 0.05 then it is significant and contributes to determining the result. The proposed method generates a feature matrix comprising 640 features for each method. In our case, every feature is important, since it determines the Euclidean distances between every virus. The results of the ANOVA test also established this fact as most of the features generated p values below the threshold level of 0.05. Therefore, we have considered each and every feature for the classification of SARS-CoV-2 variants. Any reduction in features could reduce the overall accuracy of each and individual distance having equal significance. The result of ANOVA test can be found in supplementary file.

5.4 Classification of Omicron dataset

The Omicron variant of SARS-CoV-2 is characterized by the highest transmissibility rate among all SARS-CoV-2 lineages. Therefore, the Centers for Disease Control and Prevention (CDC) designated it as a variant of concern (VOI). The proposed method is employed successfully to classify the Omicron variant from other SARS-CoV-2

Fig. 5 Various evaluation parameters obtained employing the proposed method: **a** PPDWT, **b** DNAwalkSVD, and **c** ZcurveFFT while subjected to ensemble classifier



(a)

(b)

(c)

Table 11 Various evaluation parameters yielded from CoV-Variants dataset while using ensemble classifier

Method	Average accuracy	No of misclassification	Time (s)
PP-DWT	98.9	6	66.68
Z-Curve FFT	98.8	8	66.36
DNA-Walk SVD	98.6	9	240

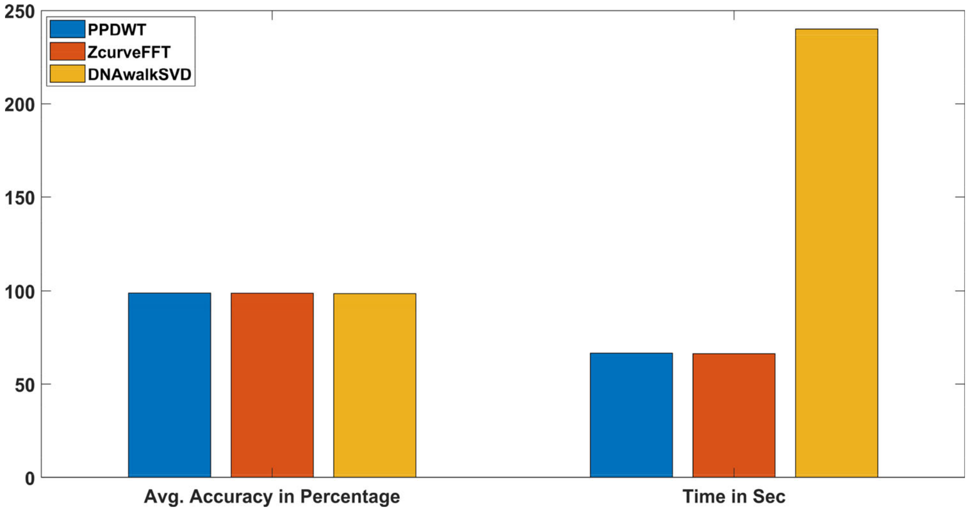


Fig. 6 Accuracies and time comparison of PPDWT, ZcurveFFT, and DNAwalkSVD methods using ensemble classifier

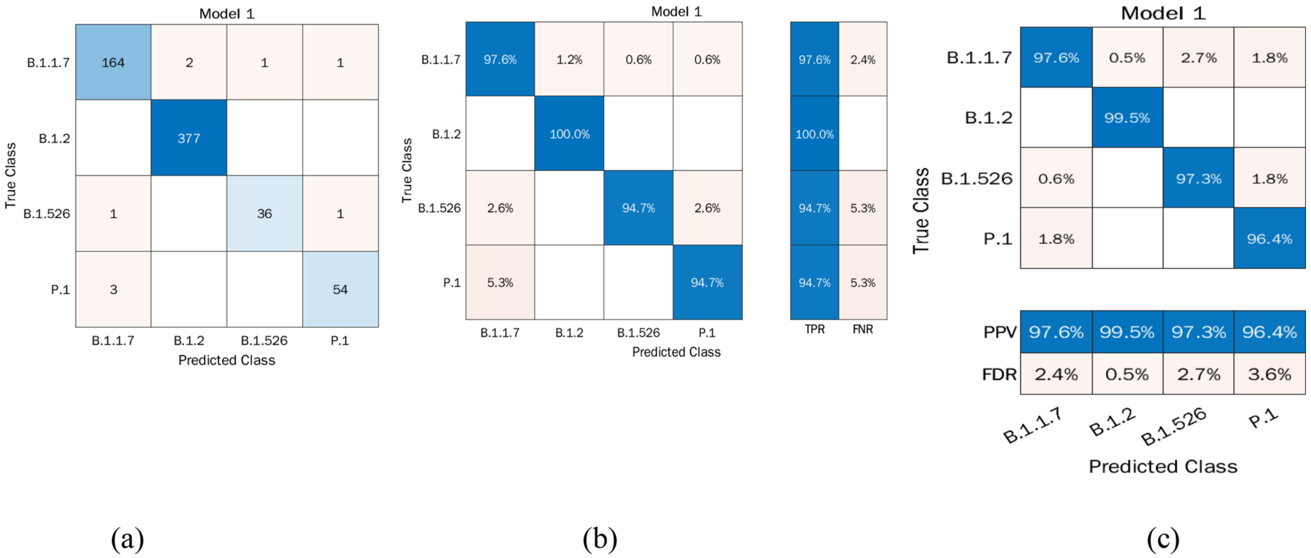
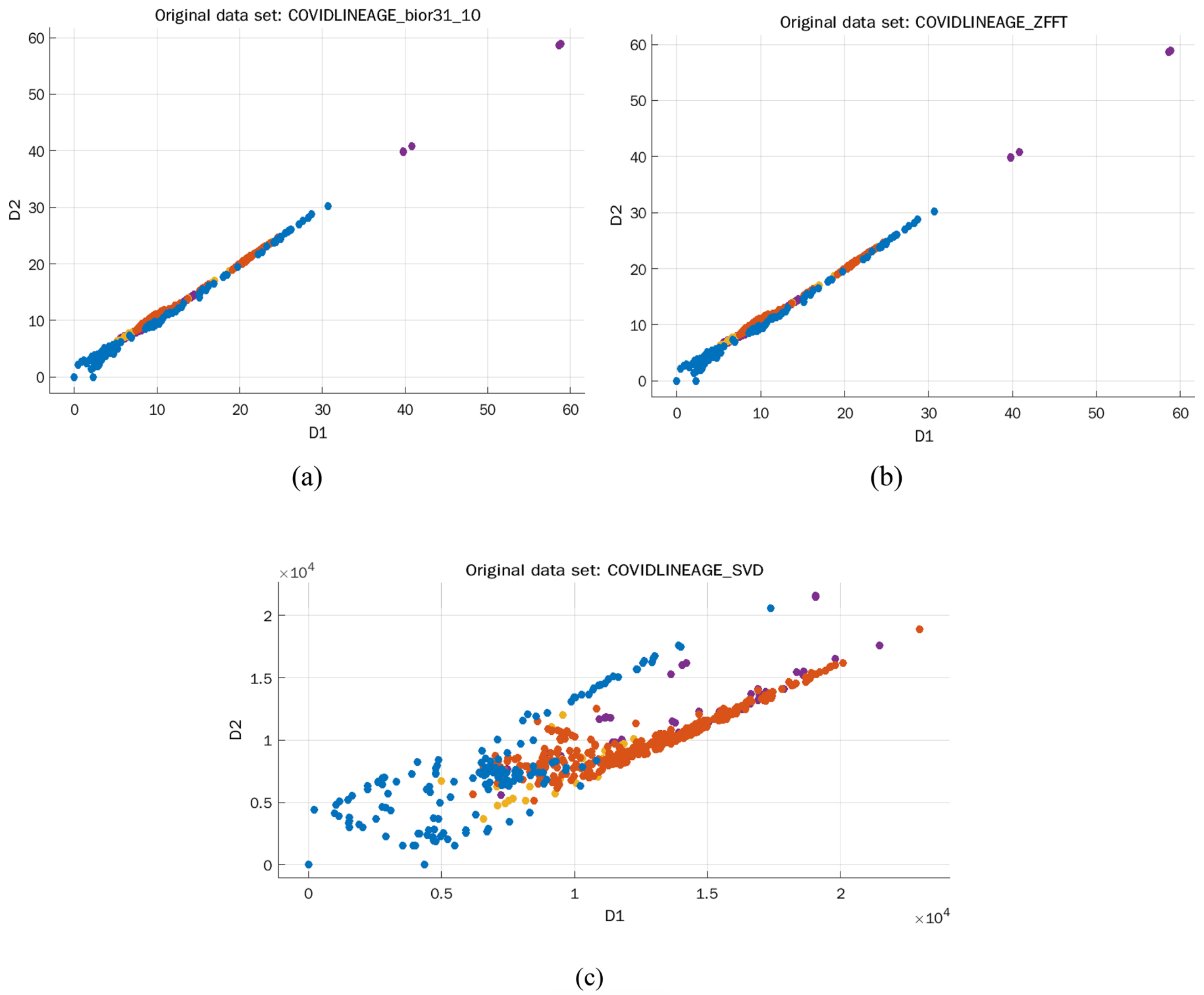


Fig. 7 Confusion matrixes obtained while using Ensemble classifier on PPDWT features (dataset: CoV-Variant)

variants. The Omicron variant BA.1 is still infecting people with a high infection rate. To verify the effectiveness of the proposed method, it is applied to classify Omicron variant BA.1 from B.1.429, P.1.10, B.1.351, and B.1.525 variants which are marked as variants being monitored (VBM). The Omicron dataset comprised 2000 sequences taken from the NCBI database. The Omicron variant BA.1 has a 66.45% share in the dataset. Since the ensemble classifier emerged as the best classifier in the previous discussion, we employed it in the classification stage of the Omicron dataset. The classification result is provided in Table 12.

Gathered information from Table 12 shows that all three GSP features are efficient in extracting discriminating features of Omicron from other variants of SARS-CoV-2. The obtained accuracy values are more than 99% for every case. The other evaluation parameter values also suggest the method is efficient in classifying Omicron variants. Thus, the method can be utilized to predict a new Omicron sequence.



(a)

(b)

(c)

Fig. 8 Scatter-plot of all the three proposed genomic signal processing-based features. **a** PPDWT (Bior3.1 at level 10), **b** ZcurveFFT, and **c** DNAwalkSVD generated employing CoV-Variant dataset

Table 12 Classification result of Omicron dataset using ensemble (subspace discriminant) classifier

Feature	Accuracy	Sensitivity	Specificity	F1-score
PP-DWT	99.1	99.4	98.5	99.3
Z-curve-FFT	99.8	99.7	100	99.9
DNA-walk-SVD	99.7	99.7	99.6	99.7

6 Comparison and discussion

The study highlighted three genomic signal processing-based feature extraction methods to classify SARS-CoV-2 variants with the help of machine learning classifiers. GSP techniques are capable of capturing distinguishing features of SARS-CoV-2 lineages in a fast and convenient way. To demonstrate the effectiveness of the proposed techniques, phylogenetic trees are computed for (i) simple integer encoded sequences without feature extraction and (ii) all three feature selection procedures using the HCoV dataset. The phylogenetic trees are given in Figs. 9 and 10.

The phylogenetic tree given in Fig. 9 is attained using simple integer encoded sequences of the HCoV dataset. No GSP feature is utilized to construct the trees. The tree fails to cluster all the six Human Coronaviruses, and hence, the classification accuracy is poor. However, when GSP-based features are employed to draw the phylogenetic tree, it is able to generate accurate results. All six Human Coronaviruses are marked by separate colors in Fig. 10 to show the classification of HCoV dataset using PPDWT, DNA-walkSVD, and ZcurveFFT. The comparative results of

GSP-based feature extraction and without feature extraction method in the variant analysis of SARs-CoV-2 are provided in Table 13.

The above table portrays clear evidence that the application of GSP generated comprehensive features to classify various lineages of SARS-CoV-2 including Omicron. For all the datasets, accuracy values obtained without any feature extraction fall below 90% indicating erroneous classification.

To establish the preeminence of the suggested approach, it is imperative to conduct a comparative analysis with the modern state-of-the-art technique employed for the categorization of the SARS-CoV-2 virus. Our method proposed novel ways to classify various strains of SARS-CoV-2 using signal processing-based features and machine learning tools. Many previous works of literature have studied the classification problem of Coronavirus sequences with the help of alignment-free machine learning approaches hybridized with digital signal processing. However, there is very few literature studying the SARS-CoV-2 variants and classifying them using fast and cost-effective genomic signal processing methods. In the majority of machine learning approaches that are based on genomic signal processing, the genomic sequences are transformed into numerical sequences rather than images. Subsequently, the process of feature extraction occurs, which is then followed by the classification phase. Table 14 gives the comparative result of SARS-CoV-2 classification methods. Although the datasets used to evaluate the algorithms were not the same, the table provides relative accuracy improvement of the proposed method in comparison to others.

Table 14 reveals that Randhawa et al. achieved the highest average accuracy of 100% while classifying the Coronavirus sequences (Randhawa et al. 2020). However, the study was conducted on limited numbers of Coronavirus sequences. Thus, if more numbers of sequences are considered, then the maximum achieved accuracy may fall due to the versatility of each and individual sequence. Also, many machine learning algorithms require more numbers of data points in the training dataset to be accurate in the classification of SARS-CoV-2 variants. In Das (2022), a GSP method integrating DWT and SVD is adopted to extract statistical features for the classification of healthy and COVID-19 patients. The accuracy of the method is

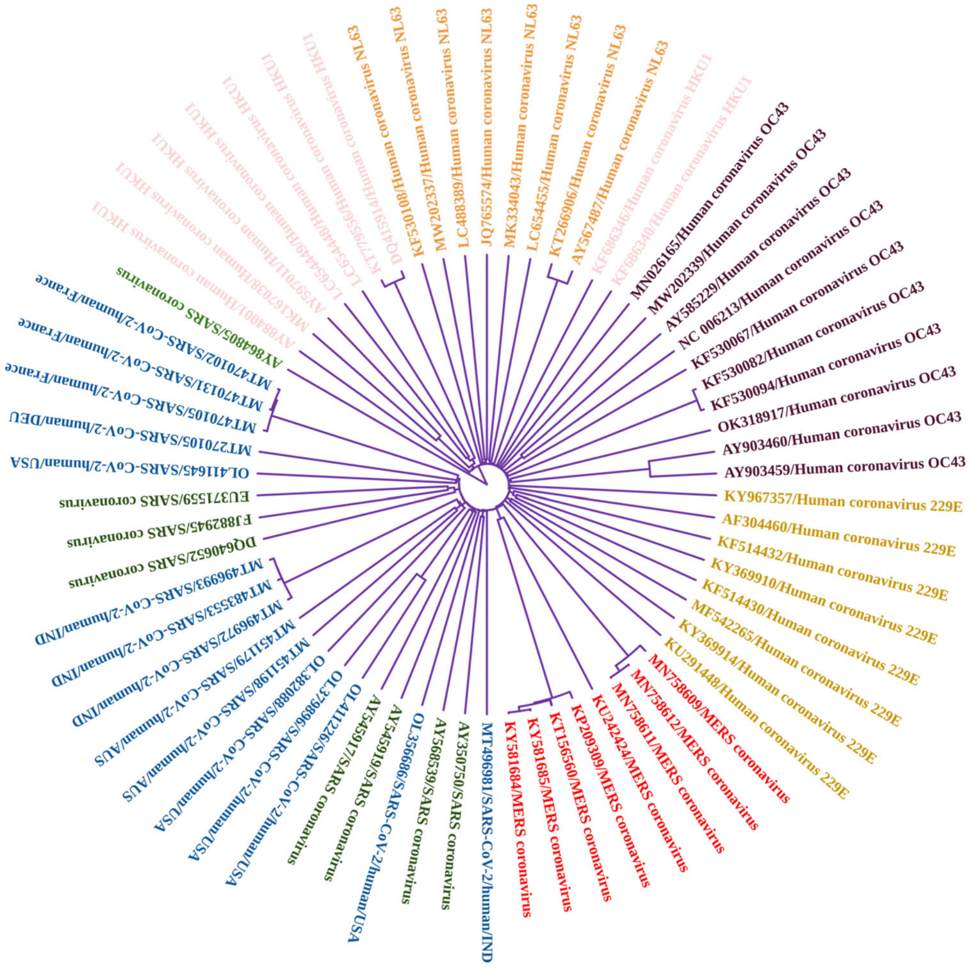

AY884001/Human coronavirus HKU1
MK167038/Human coronavirus HKU1
AY597011/Human coronavirus HKU1
LC654449/Human coronavirus HKU1
LC654448/Human coronavirus HKU1
KT779556/Human coronavirus HKU1
DQ415914/Human coronavirus NL63
KF530108/Human coronavirus NL63
MW202337/Human coronavirus NL63
LC488389/Human coronavirus NL63
JQ765574/Human coronavirus NL63
MK334043/Human coronavirus NL63
LC654455/Human coronavirus NL63
KT266906/Human coronavirus NL63
AY567487/Human coronavirus NL63
KF686346/Human coronavirus HKU1
KF686340/Human coronavirus HKU1
MN026165/Human coronavirus OC43
MW202339/Human coronavirus OC43
AY585229/Human coronavirus OC43
NC 006213/Human coronavirus OC43
KF530067/Human coronavirus OC43
KF530082/Human coronavirus OC43
KF530094/Human coronavirus OC43
OK318917/Human coronavirus OC43
AY903460/Human coronavirus OC43
AY903459/Human coronavirus OC43
KY967357/Human coronavirus 229E
AF304460/Human coronavirus 229E
KF514432/Human coronavirus 229E
KY369910/Human coronavirus 229E
KF514430/Human coronavirus 229E
MF542265/Human coronavirus 229E
KY369914/Human coronavirus 229E
KU291448/Human coronavirus 229E
MN758609/MERS coronavirus
MN758612/MERS coronavirus
MN758611/MERS coronavirus
KU242424/MERS coronavirus
KP209309/MERS coronavirus
KT156560/MERS coronavirus
KY581685/MERS coronavirus
KY581684/MERS coronavirus
MT496981/SARS-CoV-2/human/IND
AY350750/SARS coronavirus
AY568539/SARS coronavirus
OL356686/SARS-CoV-2/human/USA
AY545919/SARS coronavirus
AY545917/SARS coronavirus
OL411226/SARS-CoV-2/human/USA
OL379896/SARS-CoV-2/human/USA
OL382088/SARS-CoV-2/human/USA
MT451198/SARS-CoV-2/human/AUS
MT451179/SARS-CoV-2/human/AUS
MT496972/SARS-CoV-2/human/IND
MT483553/SARS-CoV-2/human/IND
MT496993/SARS-CoV-2/human/IND
DQ640652/SARS coronavirus
FJ882945/SARS coronavirus
EU371559/SARS coronavirus
OL411645/SARS-CoV-2/human/USA
MT270105/SARS-CoV-2/human/DEU
MT470105/SARS-CoV-2/human/France
MT470131/SARS-CoV-2/human/France
MT470102/SARS-CoV-2/human/France
AY864805/SARS coronavirus

Fig. 9 Phylogenetic tree of HCoV dataset computed using simple integer mapping ($T = 0$, $C = 1$, $A = 2$, and $G = 3$) and without GSP feature

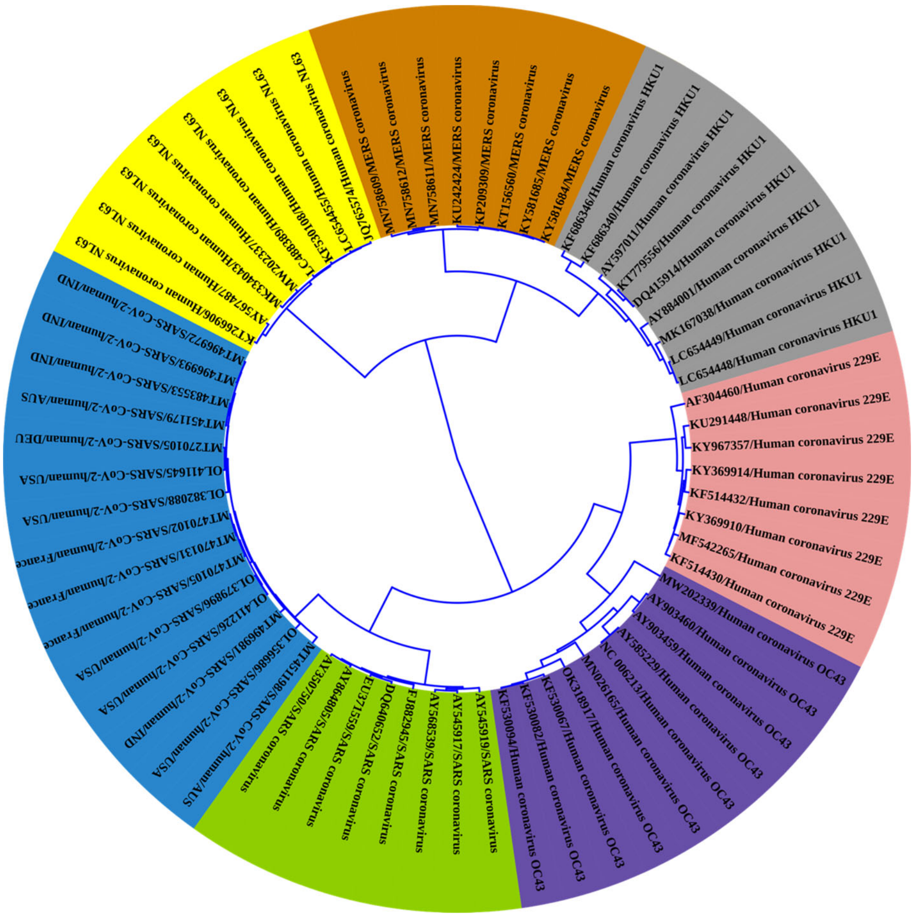

JQ765574/MERS coronavirus
MN758609/MERS coronavirus
MN758612/MERS coronavirus
MN758611/MERS coronavirus
KU242424/MERS coronavirus
KP209309/MERS coronavirus
KT156560/MERS coronavirus
KY581685/MERS coronavirus
KY581684/MERS coronavirus
KF686346/Human coronavirus HKU1
KF686340/Human coronavirus HKU1
AY597011/Human coronavirus HKU1
KT779556/Human coronavirus HKU1
DQ415914/Human coronavirus HKU1
AY884001/Human coronavirus HKU1
MK167038/Human coronavirus HKU1
LC654449/Human coronavirus HKU1
LC654448/Human coronavirus HKU1
AF304460/Human coronavirus 229E
KU291448/Human coronavirus 229E
KY967357/Human coronavirus 229E
KY369914/Human coronavirus 229E
KF514432/Human coronavirus 229E
KY369910/Human coronavirus 229E
MF542265/Human coronavirus 229E
KF514430/Human coronavirus 229E
MW202339/Human coronavirus OC43
AY903460/Human coronavirus OC43
AY903459/Human coronavirus OC43
AY585229/Human coronavirus OC43
NC 006213/Human coronavirus OC43
MN026165/Human coronavirus OC43
OK318917/Human coronavirus OC43
KF530067/Human coronavirus OC43
KF530082/Human coronavirus OC43
KF530094/Human coronavirus OC43
AY545919/SARS coronavirus
AY545917/SARS coronavirus
AY568539/SARS coronavirus
FJ882945/SARS coronavirus
DQ640652/SARS coronavirus
EU371559/SARS coronavirus
AY864805/SARS coronavirus
AY350750/SARS coronavirus
MT451198/SARS-CoV-2/human/AUS
OL356686/SARS-CoV-2/human/USA
MT496981/SARS-CoV-2/human/IND
OL411226/SARS-CoV-2/human/USA
OL379896/SARS-CoV-2/human/USA
MT470105/SARS-CoV-2/human/France
MT470131/SARS-CoV-2/human/France
MT470102/SARS-CoV-2/human/France
OL382088/SARS-CoV-2/human/USA
OL411645/SARS-CoV-2/human/USA
MT270105/SARS-CoV-2/human/DEU
MT451179/SARS-CoV-2/human/AUS
MT483553/SARS-CoV-2/human/IND
MT496993/SARS-CoV-2/human/IND
MT496972/SARS-CoV-2/human/IND
KT266906/Human coronavirus NL63
AY567487/Human coronavirus NL63
MK334043/Human coronavirus NL63
MW202337/Human coronavirus NL63
LC488389/Human coronavirus NL63
KF530108/Human coronavirus NL63
LC654455/Human coronavirus NL63

Fig. 10 The phylogenetic tree of the HCoV dataset computed using the all the three GSP-based features

Table 13 Comparison of accuracy values obtained employing all the datasets when subjected to feature extraction and without any feature extraction

Dataset	With GSP feature	Without feature extraction
HCoV	PPDWT: 100% ZcurveFFT: 100% DNAwalkSVD: 100%	89.4%
CoV-Variant	PPDWT: 98.9% ZcurveFFT: 98.8% DNAwalkSVD: 98.6%	89.8%
Omicron	PPDWT: 99.1% ZcurveFFT: 99.8% DNAwalkSVD: 99.7%	88.6%

Classifier used is ensemble

comparable to the proposed method, but the time complexity is higher as it used DWT and SVD simultaneously to calculate the feature vector. Also, the method dealt with non-homogeneous sequences due to which its practical implication is limited. The approach developed by Naeem and colleagues achieved a peak accuracy of 100%. However, this method also examined a limited number of genomic sequences for the purpose of classification (Naeem et al. 2021). The computational complexity of this signal processing-based method is on the higher side as they have calculated nine distinct features including seven-moment invariant features for classification purposes. Singh et al. suggested a digital filtering method for discrimination of SARS-CoV-2 and non-SARS-CoV-2 requiring only 0.31 s to compare 1582 sequences (Singh et al. 2021). Very recently, Arslan proposed a CpG island and similarity features-based ML method which produced comparable results to our method (Arslan 2021a). However, they worked with a binary classification problem comprising two classes which are SARS-CoV-2 and non-SARS-CoV-2 sequences. It is arduous to classify the various lineages of Coronavirus by identifying the smoothest traits. Our algorithm performed admirably in that respect, as it successfully discriminated between all the various strains of Coronavirus simultaneously. The method is useful to classify Omicron variant of Coronavirus. Basu et al. applied deep learning methods to classify various Coronavirus variants, but the maximum achieved accuracy for their method was 92.5%, while the average accuracy was found to be a mere 73.25% (Basu and Campbell 2021). The purpose of their study is similar to the proposed method as they have studied 20 different lineages of

Table 14 Comparison with existing state-of-art classification techniques of SARS-CoV-2 variants

Study	Method	Dataset	Number of sequences studied	Accuracy (avg.)	Accuracy (best)
Das (2022)	DWT features extracted from STFT with SVM and KNN classifier	COVID-19 infected patient (156) Healthy patient (104)	260	97.67%	99.17%
Hammad et al. (2023)	Deep features extracted from eight-order FCGR images with KNN classifier	COVID-19 (3700) HCoV-HKU1 (412) HCoV-NL63 (637) MERS-CoV (734) HCoV-OC43 (1351) SARS-CoV-1 (64) HCoV-229E (465)	7363	99%	99.71%
Khodaei et al. (2023)	Sliding window technique on LPC model with SVM classifier	SARS-CoV-2 (344) Influenza (706)	1050	99%	99.4%
Randhawa et al. (2020)	Fast Fourier Transform with LDA, SVM, KNN, Ensemble	SARS-CoV-2 (29) Sarbecovirus (47)	76	100%	100%
Rui et al. (2020)	Chaos-Game-Representation with LR, RF, KNN, NN, CNN, RNN	SARS-CoV-2 (1638) SARS-CoV-1 (351) MERS-CoV (321)	2310	96.7%	99.9%
Singh et al. (2021)	Digital filters with KNN, DT, RF, SVM	SARS-CoV-2 (615) and non-SARS-CoV-2 (967)	1582	97.47%	98.78%
Basu and Campbell (2021)	K-mer based Long short-Term Memory with RNN	SARS-CoV-2 variants (20 lineages)	20 lineages	73.25%	92.5%
Naeem et al. (2021)	Discrete Fourier Transform, Discrete Cosine Transform, Seven moment invariants with KNN and NN	SARS-CoV2 (76) SARS-CoV-1 (76) MERS-CoV (76)	228	99.45%	100%
Arslan (2021a)	SVM, Naïve Bayes, KNN, RF with CpG island based features	SARS-CoV-2 (1000) and non-SARS-CoV-2 (331)	1331	90%	93%
Arslan and Arslan (2021)	KNN with L1 type metrics with CPG based features	SARS-CoV-2 (1000) Non SARS-CoV-2 (592)	1592	94.92%	98.4%
Arslan (2021b)	KNN, SVM, DT, AdaBoost, MLP,RF with CpG island and similarity features	SARS-CoV-2 (1000) AlphaCov (92) BetaCoV (523) RaTG BatCoV (1)	1616	99.56%	99.8%
Proposed	KNN, SVD & Ensemble with DFT, DWT, and SVD computed features	SARS-CoV-2 variants [B.1.1.7 (168) B.1.2 (377) B.1.526 (38) P.1 (57)]	640	99.15%	99.8%
		SARS-CoV-2 variants [B.1.351 (32) B.1.429 (530) B.1.525 (29) BA.1 (1329) P.1.10 (80)]	2000		

SARS-CoV-2 to classify them using deep learning models. However, the main disadvantage of the method is that it obtained very low values of accuracy when classifying the chosen variants of SARS-CoV-2 sequences. There are several alternative approaches outlined in Table 14 that have produced lower accuracy results compared to the suggested method (Hammad et al. 2023; Khodaei et al. 2023; Arslan and Arslan 2021; Arslan 2021b). Some deep learning methods based on X-ray or CT-scan images were also reported in the literature to detect the SARS-CoV-2 virus (Ozturk et al. 2020; Jain et al. 2020; Asnaoui and Chawki 2020; Apostolopoulos and Mpesiana 2020). A deep learning technique proposed by Lopez-Rincon used various human Coronavirus genome sequences to distinguish the SARS-CoV-2 virus. The dataset for their experiment contained 553 sequences, including MERS-CoV, HCoV-NL63, HCoV-OC43, HCoV-229E, HCoV-HKU1, and SARS-CoV-1. They obtained an average accuracy of 98.75% which is very close to our results (Lopez-Rincon et al. 2020a). In another experiment, Lopez-Rincon et al. classified Coronavirus lineage B.1.1.7 from other lineages using a convolutional neural network. The dataset for this experiment contained 8923 numbers of sequences taken from GISAID. The method yielded an average accuracy of 99% (Lopez-Rincon et al. 2020b). Another study conducted by Adetiba et al. developed a DeepCovid-19 tool to distinguish SARS-CoV-2, SARS-CoV-1, and MERS viruses. They produced Z-curve images of Coronaviruses for the purpose of implementing them in deep learning models to classify each group. The best-obtained accuracy in this study is 90% obtained by employing the transfer learning-based CNN model (Adetiba et al. 2022). It is worth mentioning that deep learning (DL) models take a long time to train datasets, and the memory requirement of DL models is also very high compared to ML models.

7 Conclusion

After the emergence of the COVID-19 pandemic, its tendrils have permeated numerous nations and cast a profound impact upon them. Since that time, numerous iterations of the SARS-CoV-2 virus have traversed the globe periodically, infiltrating the innate immune system of human beings. Many of these variants have mutations in spike protein and can invade the attained antibodies produced in the human body after the vaccination. Omicron variants of SARS-CoV-2 are still spreading at a very fast pace all over the world. However, finding every sub-variant is difficult, because it calls for accurate diagnostic tests and a lot of processing time. Consequently, it is necessary to create novel technologies that can detect COVID-19 viral variations at the fundamental level. In this regard, the genomic signal processing-based approach can be life-saving, because it produces the best degree of classification accuracy for SARS-CoV-2 variations while requiring less time and money.

In this work, three signal processing-based methods are applied for the classification of three datasets HCoV, CoV-Variants, and Omicron. Unlike, the alignment-based method, the time complexity of this method is less. The use of a second-generation lifting algorithm employing Bior accelerated the overall process and facilitated the comparison of numerous genomic sequences measuring 30 K in length within a span of just one minute. We employed ML-based classifiers SVM, KNN, and Ensemble in the proposed method. The result through various evaluation parameters showed the highest accuracy using the ensemble classifier, but the time complexity of the ensemble method is quite high compared to SVM and KNN. To attain the highest level of accuracy within a limited span of time, it becomes imperative to optimize the design parameters of both the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifier. We used the Bayesian optimization technique and achieved 98.6% accuracy with the SVM classifier and PP-DWT feature. The overall performance and computational efficiency of the proposed machine learning-based model are very good compared to present alignment-free and alignment-based ML and DL methods.

The proposed method is extended to detect Omicron from other lineages and proved to be efficient. Second-generation discrete wavelet transform is utilized first time in COVID-19 detection and lineage identification. The method is found to be very much efficient in acquiring detailed information about a SARS-CoV-2 sequence at different decomposition levels. Windowed SVD and DFT tools are equally efficient in detecting the Omicron variant which is currently tagged as VOC by the CDC. The proposed method could play a pivotal role in distinguishing any virulent strain of SARS-CoV-2 in future pandemics. However, the classification of SARS-CoV-2 variants using the proposed method still lags in some areas that need to be addressed in future studies.

1. The proposed method depends on the classification accuracy of the supervised machine learning algorithm. All the supervised-based machine learning algorithms required adequate numbers of pre-labeled datasets. Also, the collection of large datasets comprising sequences of each and every variant in the same proportion is sometimes difficult due to their non-availability in online databases. Therefore, data imbalance might be an issue with the classification of SARS-CoV-2 virus sequences using a supervised machine learning algorithm. Data resampling techniques can be

adopted to address the issue. Therefore, stratified cross-validation techniques could be utilized for precise evaluation of the model. In the stratified k-fold cross-validation method, the dataset is partitioned into k-folds, such that the mean response value is somehow equal in all the partitions.

2. Genome sequencing programs provide information about each and every mutation taking place in the virus genome as well as the relative positions of such mutations in virus structure. Thus, it can provide detailed information on any SARS-CoV-2 sequence. It has the ability to identify new variants as well as classify a sequence as per previously available lineages. The proposed model has a limitation as it could not provide details about any new mutation taking place in the virus genome. However, it has the capability to differentiate SARS-CoV-2 sequences in terms of alteration in various base positions. Thus the method is suitable for the initial screening of a SARS-CoV-2 variant, since it takes a few seconds to generate the result.

Since the classification problem of SARS-CoV-2 variants using genomic signal processing methods is still new and evolving, there is a lot of scope for improvements. Also considering that various sub-variants are triggering Coronavirus waves, we will consider the protein sequences of SARS-CoV-2 variants for classification purposes in coming studies. Since most of the Pango lineages have mutations in spike protein, therefore, the study of protein sequences could reveal detailed information about every variant thus making the classification problem more accurate. Also, the total number of sequences can be further increased by obtaining other variants of concerns of SARS-CoV-2 to validate the developed model.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00500-023-09577-9.

Author contributions SK is responsible for all the research and development in this work as well as the drafting of the paper. MG is responsible for providing theoretical and development support as well as editing the draft.

Funding No funding is received for conducting this study.

Data availability All data generated or analyzed during this study are included in this article as supplementary files 1, 2, and 3 and can be found using the URL: https://data.mendeley.com/datasets/9jpxmvg5m6/1. The code used to implement the results can be obtained on request to authors.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical approval This article does not contain any studies with animals performed by any of the authors.

References

Abdelrahman Z, Li M, Wang X (2020) Comparative review of SARS-CoV-2, SARS-CoV, MERS-CoV, and influenza a respiratory viruses. Front Immunol 11:2309

Adetiba E, Abolarinwa JA, Adegoke AA, Taiwo TB, Ajayi OT, Abayomi A, Adetiba JN, Badejo JA (2022) DeepCOVID-19: a model for identification of COVID-19 virus sequences with genomic signal processing and deep learning. Cogent Eng 9(1):2017580

Afify HM, Zanaty MS (2021) A comparative study of protein sequences classification-based machine learning methods for COVID-19 virus against HIV-1. Appl Artif Intell 35(15):1733–1745

Ahmed I, Jeon G (2022) Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. Interdiscipl Sci Comput Life Sci 14(2):504–519

Ahsan R, Tahsili MR, Ebrahimi F, Ebrahimie E, Ebrahimi M (2021) Image processing unravels the evolutionary pattern of SARS-CoV-2 against SARS and MERS through position-based pattern recognition. Comput Biol Med 134:104471

Akbari Rokn Abadi S, Mohammadi A, Koohi S (2023) A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes. BMC Genomics 24(1):266

Akhtar M, Epps J, Ambikairajah E (2008) Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE J Sel Top Signal Process 2(3):310–321

Al Kindhi B (2020) Optimization of machine learning algorithms for predicting infected COVID-19 in isolated DNA. Int J Intell Eng Syst 13(4)

Apostolopoulos ID, Mpesiana TA (2020) COVID-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 43(2):635–640

Arslan H (2021a) Machine learning methods for COVID-19 prediction using human genomic data. In: Multidisciplinary digital publishing institute proceedings, vol 74, no 1, p 20

Arslan H (2021b) COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like Coronavirus. Comput Ind Eng 161:107666

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

Azevedo K, Souza L, Coutinho M, Barbosa R, Fernandes M (2023) Deep learning applied to the SARS-CoV-2 classification

Basu S, Campbell RH (2021) Classifying COVID-19 variants based on genetic sequences using deep learning models. Biorxiv

Berger JA, Mitra SK, Carli M, Neri A (2004) Visualization and analysis of DNA sequences using DNA walks. J Franklin Inst 341(1–2):37–53

Câmara GB, Coutinho MG, Silva LMD, Gadelha WVDN, Torquato MF, Barbosa RDM, Fernandes MA (2022) Convolutional neural network applied to SARS-CoV-2 sequence classification. Sensors 22(15):5730

Chen D, Wan S, Xiang J, Bao FS (2017) A high-performance seizure detection algorithm based on discrete wavelet transform (DWT) and EEG. PLoS ONE 12(3):e0173138

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Das B (2022) An implementation of a hybrid method based on machine learning to identify biomarkers in the COVID-19 diagnosis using DNA sequences. Chemom Intell Lab Syst 230:104680

Das B, Toraman S (2023) New Coronavirus 2 (SARS-CoV-2) detection method from human nucleic acid sequences using capsule networks. Braz Arch Biol Technol 66

Das B, Turkoglu I (2018) A novel numerical mapping method based on entropy for digitizing DNA sequences. Neural Comput Appl 29(8):207–215

Daş B, Toraman S, Türkoğlu İ (2020) A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks. Turk J Electr Eng Comput Sci 28(4):1932–1948

de Souza LC, Azevedo KS, de Souza JG, Barbosa RDM, Fernandes MA (2023) New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning. BMC Bioinform 24(1):1–19

Dey L, Chakraborty S, Mukhopadhyay A (2020) Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. Biomed J 43(5):438–450

Duda RO, Hart PE, Stork DG (2001) Pattern classification. Willey, New York

El Asnaoui K, Chawki Y (2021) Using X-ray images and deep learning for automated detection of Coronavirus disease. J Biomol Struct Dyn 39(10):3615–3626

Fiscon G, Weitschek E, Ciccozzi M, Bertolazzi P, Felici G (2016) A novel feature selection method to extract multiple adjacent solutions for viral genomic sequences classification. BMC Bioinform 17:207–208

Ghaderzadeh M, Eshraghi MA, Asadi F, Hosseini A, Jafari R, Bashash D, Abolghasemi H (2022) Efficient framework for detection of COVID-19 Omicron and delta variants based on two intelligent phases of CNN models. Comput Math Methods Med 2022

Göreke V, Sarı V, Kockanat S (2021) A novel classifier architecture based on deep neural network for COVID-19 detection using laboratory findings. Appl Soft Comput 106:107329

Guntoro A, Glesner M (2008) A lifting-based discrete wavelet transform and discrete wavelet packet processor with support for higher order wavelet filters. In: IFIP/IEEE international conference on very large scale integration-system on a chip, pp 154–173. Springer, Berlin, Heidelberg

Hammad MS, Ghoneim VF, Mabrouk MS, Al-Atabany WI (2023) A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques. Sci Rep 13(1):4003

Hirotsu Y, Omata M (2021) Discovery of a SARS-CoV-2 variant from the P.1 lineage harboring K417T/E484K/N501Y mutations in Kofu, Japan. J Infect 82(6):276–316

Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844

Hoang T, Yin C, Yau SST (2016) Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics 108(3–4):134–142

Huang HH, Girimurugan SB (2019) Discrete wavelet packet transform based discriminant analysis for whole genome sequences. Stat Appl Genet Mol Biol 18(2)

Huang HH, Hao S, Alarcon S, Yang J (2018) Comparisons of classification methods for viral genomes and protein families using alignment-free vectorization. Stat Appl Genet Mol Biol 17(4)

Jain G, Mittal D, Thakur D, Mittal MK (2020) A deep learning approach to detect COVID-19 Coronavirus with X-ray images. Biocybern Biomed Eng 40(4):1391–1405

Kar S, Ganguly M, Ganguly A (2022) Spectral analysis of DNA on 1-D hydration enthalpy-based numerical mapping using optimal filtering. In: Emerging technologies for computing, communication and smart cities: proceedings of ETCCS 2021. Springer Nature, Singapore, pp 137–149

Kar S, Ganguly M, Ghosal S (2021) Prediction of coding region and mutations in Human DNA by effective numerical coding and DSP technique. In: 2021 international conference on computing, communication, and intelligent systems (ICCCIS). IEEE, pp 180–185

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

Khodaei A, Feizi-Derakhshi MR, Mozaffari-Tazehkand B (2020b) A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods. Soft Comput 24:16315–16334

Khodaei A, Shams P, Sharifi H, Mozaffari-Tazehkand B (2023) Identification and classification of Coronavirus genomic signals based on linear predictive coding and machine learning methods. Biomed Signal Process Control 80:104192

Lebatteux D, Remita AM, Diallo AB (2019) Toward an alignment-free method for feature extraction and accurate classification of viral sequences. J Comput Biol 26(6):519–535

Lin J, Wei J, Adjeroh D, Jiang BH, Jiang Y (2018) SSAW: A new sequence similarity analysis method based on the stationary discrete wavelet transform. BMC Bioinform 19(1):1–11

Lopez-Rincon A, Tonda A, Mendoza-Maldonado L, Claassen E, Garssen J, Kraneveld AD (2020a) Accurate identification of SARS-CoV-2 from viral genome sequences using deep learning. Biorxiv

Lopez-Rincon A, Perez-Romero C, Tonda A, Mendoza-Maldonado L, Claassen E, Garssen J, Kraneveld AD (2020b) Design of specific primer set for detection of B. 1.1. 7 SARS-CoV-2 variant using deep learning. Biorxiv

Meher PK, Sahu TK, Gahoi S, Satpathy S, Rao AR (2019) Evaluating the performance of sequence encoding schemes and machine learning methods for splice sites recognition. Gene 705:113–126

Naeem SM, Mabrouk MS, Marzouk SY, Eldosoky MA (2021) A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19. Brief Bioinform 22(2):1197–1205

Nair AS, Sreenadhan SP (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). Bioinformation 1(6):197

Osuna EE (1998) Support vector machines: training and applications. Doctoral dissertation, Massachusetts Institute of Technology

Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR (2020) Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 121:103792

Press WH (2007) Numerical recipes 3rd edition: the art of scientific computing. Cambridge University Press, Cambridge

Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS ONE 15(4):e0232391

Rui YIN, Luo Z, Kwoh CK (2020) Alignment-free machine learning approaches for the lethality prediction of potential novel human-adapted Coronavirus using genomic nucleotide. Biorxiv

Singh OP, Vallejo M, El-Badawy IM, Aysha A, Madhanagopal J, Faudzi AAM (2021) Classification of SARS-CoV-2 and non-SARS-CoV-2 using machine learning algorithms. Comput Biol Med 136:104650

Sweldens W (1998) The lifting scheme: a construction of second generation wavelets. SIAM J Math Anal 29(2):511–546

Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R (1997) Prediction of probable genes by Fourier analysis of genomic sequences. Bioinformatics 13(3):263–270

Ucar F, Korkmaz D (2020) COVIDiagnosis-Net: deep Bayes-SqueezeNet based diagnosis of the Coronavirus disease 2019 (COVID-19) from X-ray images. Med Hypotheses 140:109761

Ullah W, Ullah A, Malik KM, Saudagar AKJ, Khan MB, Hasanat MHA, AlTameem A, AlKhathami M (2022) Multi-stage temporal convolution network for COVID-19 variant classification. Diagnostics 12(11):2736

Vaegae NK (2020) Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes. Biomed Signal Process Control 58:101859

Voss RF (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 68(25):3805

Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, Amoako DG, Everatt J, Bhiman JN, Scheepers C, Tebeila N (2021) Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa. Medrxiv

Woo PC, Lau SK, Huang Y, Yuen KY (2009) Coronavirus diversity, phylogeny and interspecies jumping. Exp Biol Med 234(10):1117–1127

Yan M, Lin ZS, Zhang CT (1998) A new fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics (oxford, England) 14(8):685–690

Yin C, Yau SST (2015) An improved model for whole genome phylogenetic analysis by Fourier transform. J Theor Biol 382:99–110

Yin R, Luo Z, Kwoh CK (2020) Alignment-free machine learning approaches for the lethality prediction of potential novel human-adapted Coronavirus using genomic nucleotide. Biorxiv

Zhang CT, Wang J (2000) Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res 28(14):2804–2814

Zhang W, Arvanitis A, Al-Rasheed A (2012) singular value decomposition and its numerical computations. Michigan Technological University, Houghton

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