Multifaceted Computational Framework for COVID-19 Variant Classification using Advanced Machine Learning, Signal Processing, and High-Dimensional Feature Reduction Techniques

Love Fadia
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Canada

Vatsal Shah
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Canada

Mohammad Hassanzadeh
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Canada

Majid Ahmadi
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Canada

Jonathan Wu
Department of Electrical and Computer Engineering
University of Windsor
Windsor, Canada

ABSTRACT

The coronavirus pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), had an extensive global impact, causing widespread disruptions to public health. The early and accurate identification of the virus and its various strains is imperative for safeguarding lives. Over the past few years, multifarious machine learning and deep learning techniques were used to classify genomic sequences . However, existing methods face several limitations. Many approaches struggle with dataset imbalance, leading to biased and unreliable models. Traditional neural network-based methods are computationally intensive, requiring significant time and resources. Moreover, existing techniques often fail to achieve consistently high classification accuracy across properly balanced datasets. To address these gaps, this article presents an efficient method for classifying coronavirus variants' DNA sequences using a combination of machine learning and signal processing. The DNA sequences are first converted into numbers using Electron-Ion Interaction Potential, Numeric, and Complex coding techniques. After that signal processing methods; Discrete Cosine Transform II, Discrete Cosine Transform III, Fast Fourier Transform, Haar Wavelet Transform, and Coiflet Wavelet Transform are applied to extract features from the coded data. The high dimensionality is reduced using Linear Discriminant Analysis and Principal Component Analysis. For the classification task, machine learning models such as Decision Tree, Support Vector Classifier, and a fusion of Light-Gradient Boosting Machine, AdaBoost, and Random Forest are employed. The proposed approach achieves an impressive accuracy of 99.8%, which surpasses the state of the art using a different combination of transformations with Numeric coding and Voting Classifier.

Keywords

Genomic Sequence Analysis, Signal Processing, Dimensionality Reduction, Machine Learning.

1. INTRODUCTION

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the virus that caused the COVID-19 pandemic, has evolved fast through various genetic alterations, resulting in the formation of several unique viral strains known as variants. These genetic variants in the SARS-CoV-2 genome have a significant impact on essential viral features such as transmissibility, immune response resistance, and the severity of infection. This virus is made up of Ribonucleic acid (RNA) which is transcribed into Deoxyribonucleic acid (DNA) for stable sequencing, enabling consistent mutation tracking and insights into viral evolution [1]. Among the identified variants, five are of considerable public health importance because of increased transmissibility and immune evasion: Alpha, Delta, Omicron, Beta, and Gamma [2]. Here Alpha, Delta, and Omicron variants have been used for classification because they have made a remarkable impact on human health and have influenced public health strategy.Various deep learning-based techniques have identified these variants by analyzing complex genetic data, hence providing automatic scalable identification from sequence data [3, 4]. Most such models are usually very demanding computationally and need a great amount of computational resources in terms of processing units and memory, hence very hard to be deployed in resource-sensitive environments. Also, as much as pre-trained models [5] reduce the burden of training a model from scratch, their high demand with regard to computation renders their usage infeasible at large scale or in real time applications. In this regard, effective classification approach was introduced that combines signal processing techniques with dimensionality reduction and machine learning classifiers, which will result in accurate variant classification with less computational cost. Five different signal processing techniques are utilized to extract meaningful features from DNA sequence data, namely: Discrete Cosine Transform II (DCT II), Discrete Cosine Transform III (DCT III), Fast Fourier Transform (FFT), Haar wavelet, and Coiflet wavelet. These methods capture the key sequence patterns by transforming the DNA data into the frequency domain where key features can be highlighted much better. Furthermore the efficiency was increased in the classification process by performing a reduction in dimensions with a view to optimize the dataset while maintaining only critical variant information with minimal complexity.The described approach of combining signal processing with dimensional reduction, compared to other traditional deep learning models, provides an economically efficient solution; hence, large-scale and real-time applications are feasible. Proposed methodology has ensured that SARS-CoV-2 variants are reliably classified, especially in resource-constrained environments, by significantly reducing processing time and memory utilization.

2. RELATED WORK

This section discusses literature combining machine learning classifiers and signal processing techniques for SARS-CoV-2 classification. Khodel *et al.* [6] used Singular Value Decomposition, linear predictive feature extraction, and z-curve mapping to achieve 99% accuracy with Support Vector Machine (SVM), showcasing the potential of combining linear algebra techniques with machine learning. Naeem *et al.* [7] transformed DNA sequences into the frequency domain using Discrete Cosine and Fourier Transforms, achieving 98.89% accuracy with a KNN model, highlighting the efficacy of frequency-based feature extraction. Patel *et al.* [8] employed wavelet transformation and statistical analysis to distinguish COVID-19-infected genes, providing approach based on genomic signal analysis. Meng *et al.* [9] analyzed Wavelet Transform applications in DNA sequences for cancer studies, suggesting broader applicability of wavelet techniques across various domains. Yadav *et al.* [10] mapped DNA sequences to complex numbers, applying Short Time Ramanujan Fourier Transform for pattern extraction, enabling the identification of intricate sequence structures. Chalco *et al.* [11] implemented Modified Gabor Wavelet Transform for coding region identification, which is critical for understanding functional regions in viral genomes. Randhawa *et al.* [12] combined supervised machine learning and digital signal processing, achieving 98.1% accuracy in COVID-19 variant classification, demonstrating the synergy between classical signal processing and advanced machine learning algorithms. Muhammad *et al.* [13] achieved 99.2% accuracy with eXtreme Gradient Boosting (XGB) and faster computation with Light Gradient Boosting Machine (LGBM), emphasizing the balance between accuracy and computational efficiency. Hammad *et al.* [14] used Frequency Chaos representation and AlexNet for feature selection, achieving 99.71% accuracy with KNN and Decision Trees, illustrating the integration of deep learning architectures with traditional classifiers. Saha *et al.* [15] developed COVID-DeepPredictor, delivering 100% accuracy for classification with class imbalance issue, representing an innovation in predictive modeling for SARS-CoV-2 analysis.Eldosuky *et al.* [16] classified COVID-19 and influenza with 99% accuracy using optimized deep neural networks, underlining the versatility of neural networks in cross-disease classification.

3. METHODOLOGY

Our proposed methodology begins with data collection from the NCBI Virus database, focusing on genomic information [17]. For this study, 1,000 DNA sequences for each of three SARS-CoV-2 variants was selected. After removing ambiguous sequences to ensure data quality, the DNA sequences are encoded using Complex, EIIP, and Numeric coding techniques to structure the data for analysis. These encoded sequences are then transformed into the frequency domain using methods such as DCT II, DCT III, FFT, Haar Wavelet Transform, and Coiflet Wavelet Transform, which help capture patterns and features. Subsequently, dimensionality reduction is applied using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to refine features relevant to accurate predictions. The processed dataset is split into 70% training and 30% testing sets, and various machine learning models are trained and evaluated. Accuracy assessments of predictions are conducted to compare the methodologies and their effectiveness in identifying SARS-CoV-2 variants. Fig 1. illustrates the workflow.

3.1 Coding Methods

3.1.1 EIIP coding. The method that converts DNA nucleotide sequences into numbers based on Electron Potential is known as EIIP coding. The method for these values' calculation is the atoms' potentials for electron-ion interactions [18]. The nucleotides of DNA have the following EIIP values:

- Adenine (A): 0.1260
- Thymine (T): 0.1335
- Guanine (G): 0.0806
- Cytosine (C): 0.1340

3.1.2 Complex coding. In this method, we assign 4 nucleotides of DNA sequence to complex numbers, which involves complementary characteristics [19]:

- Adenine (A): $1 + i$
- Thymine (T): $-1 - i$
- Guanine (G): $-1 + i$
- Cytosine (C): $1 - i$

3.1.3 Numeric coding. In Numeric coding, we transform DNA nucleotides into integers and it is a very fast and efficient approach. Here's how we can achieve it. [20].

- Adenine (A): 2
- Thymine (T): 0
- Guanine (G): 3
- Cytosine (C): 1

3.2 Signal Processing Methods

In our study, we use five linear transformations to take our data to the frequency domain.A key tool for compressing digital signals is the Discrete Cosine Transform II (DCT II). By concentrating signal energy in a small number of coefficients, particularly in the lower frequency components, it can express signals more succinctly. The equation yields the (DCT II). [21]:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot \cos\left(\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right) \tag{1}$$

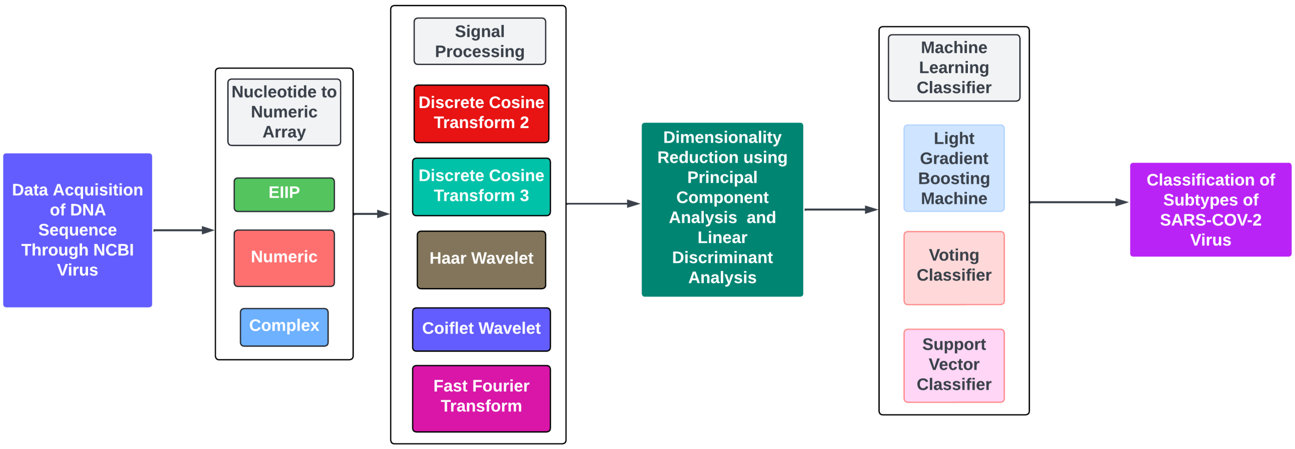
Data Acquisition of DNA Sequence Through NCBI Virus
Nucleotide to Numeric Array
EIIP
Numeric
Complex
Signal Processing
Discrete Cosine Transform 2
Discrete Cosine Transform 3
Haar Wavelet
Coiflet Wavelet
Fast Fourier Transform
Dimensionality Reduction using Principal Component Analysis and Linear Discriminant Analysis
Machine Learning Classifier
Light Gradient Boosting Machine
Voting Classifier
Support Vector Classifier
Classification of Subtypes of SARS-COV-2 Virus

Fig. 1: Proposed Methodology

Here, X_k is the DCT coefficient at index k, x_n is the input sequence, N is the length of the sequence, and k ranges from 0 to $N-1$. The inverse of Discrete Cosine Transform II (DCT II) is Discrete Cosine Transform III (DCT III) given by the equation [22]

$$X_k = \sum_{n=0}^{N-1} x_n \cdot \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right) \tag{2}$$

.

An algorithmic method for quickly calculating the Discrete Fourier Transform (DFT) and its inverse is the Fast Fourier Transform (FFT). The DFT computation is substantially sped up by the FFT. The FFT significantly speeds up the computation of the DFT, The FFT is defined by the equation [23]

$$X[k] = \&FFT[x[n]] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-j(2\pi/N)kn} \tag{3}$$

.

Haar Wavelet is a simple wavelet function that is popularly used in the field of Signal and Image processing due to piecewise linear functionality. Haar Wavelet equation is given by [24]:

$$\psi(x) = \begin{cases} 1, & \text{if } 0 \le x < \frac{1}{2}, \\ -1, & \text{if } \frac{1}{2} \le x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $\psi(x)$ represents the associated scaling function.

Coiflet wavelet is a mathematical function that is designed to have both properties; Vanishing Moments and Orthogonality. Vanishing Moments make sure that the wavelet follows a polynomial trend without affecting detailed coefficients. and Orthogonal property ensures that the inverse of this wavelet is proper. The equation is given as [25].

$$\psi(x) = \begin{cases} (1/\sqrt{2}) \times (\phi(x) - \phi(x-3)), & \text{if } 0 \le x < 3, \\ (1/\sqrt{2}) \times (-\phi(x-1) + \phi(x-4)), & \text{if } 3 \le x < 4, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where $\psi(x)$ represents the associated scaling function.

3.3 Dimensionality Reduction

Our original data has a maximum length of 30,000 base pairs and there are a total of 1000 DNA sequences for each variant. It can be seen that the size of the data is very large, i.e., $1000 \times 30,000$ columns for each variant class. Since there are a total of 3 different variants, resulting in 3 categories for classification. During machine learning operations, it was observed that computationally it takes much more time to compute predictions due to the huge size of the data. To speed up the computation we use two dimensionality reduction techniques.

3.3.1 Principal Component Analysis. Principal Component Analysis (PCA) is a dimensionality reduction technique that reduces the size of a dataset while preserving as much variance as possible. It basically transforms the raw data into a set of uncorrelated variables which are known as principal components, This is how it simplifies the complexity of dimension data. It is observed that the first principal component has the highest amount of variations which goes on decreasing while moving further [26]. To calculate PCA, the eigenvalues and eigenvectors of the covariance matrix χ has been calculated. The eigenvalues ultimately represent the amount of variations explained by each principal component, while the eigenvectors represent the direction of these components. In our study, 60 principal components are used for classification.

3.3.2 Linear Discriminant Analysis (LDA). Linear Discriminant Analysis is used to maximize the ratio of between-class variance to the within-class variance in any particular dataset, which leads to maximal separability among classes. The Scatter matrices play a vital role in this analysis which is defined as: [27].

$$S_W = \sum_{i=1}^{k} \Sigma(x-\mu_i)(x-\mu_i)^T \tag{6}$$

Here, x belongs to the particular class i.

$$S_B = \sum_{i=1}^{k} N_i(\mu_i-\mu)(\mu_i-\mu)^T \tag{7}$$

where N_i is the number of samples in class i, μ_i is the mean vector of class i, and μ is the overall mean vector of the dataset. The last step of LDA is to find a projection matrix W that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix [27]:

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{8}$$

The optimal projection matrix W can be found by solving the generalized eigenvalue problem [27]:

$$S_W^{-1} S_B W = \Lambda W \tag{9}$$

where Λ is a diagonal matrix whose entries are the eigenvalues. The eigenvectors corresponding to the largest eigenvalues form the columns of the projection matrix W.

3.4 Machine Learning Methods

3.4.1 Voting Classifier. A voting classifier is an ensemble method that combines the predictions of multiple machine learning models to produce a single, more accurate prediction by leveraging the strengths of diverse classifiers and mitigating their individual weaknesses. We have used three models in our study to create a voting classifier, which are Light Gradient Boosting Machine(LGBM), Random Forest, and AdaBoost.

LGBM is a highly robust and efficient model that utilizes a technique called gradient boosting to construct an ensemble of decision trees. It combines the results or predictions of multiple weak learners to form a strong learner sequentially, and with each iteration, the trees learn from the errors of the previous ones [28]. The final prediction is the sum of the output of all the trees. Random Forest is a strong supervised ensemble learning approach that is frequently used for classification tasks. The working principle of Random Forest is to create a large number of decision trees and combine all of their predictions to increase accuracy and simultaneously decrease overfitting, and improve generalization. [29]. Adaboost short for Adaptive Boosting is also one of the best algorithms from the ensemble family. The main difference between the Adaboost and other decision tree algorithms is that; Adaboost uses stumps; which are nodes with two leaves to create the tree [30].

3.4.2 Decision Trees. Decision Tree is a supervised learning machine learning method that uses a tree-like structure and nodes for options and leaves for outcomes, that produce decision rules based on data properties. Data is divided in a binary way at each node in accordance with predetermined rules in order to a construct decision tree. While building a decision tree a common practice is to utilise the CART (Classification and Regression Trees) algorithm. CART utilizes Gini impurity in classification issues to select the best node for splitting. The number of times a randomly chosen element from the dataset would be incorrectly classified if it were labeled at random using the label distribution within that subset is measured by the Gini impurity which is defined as [31]:

$$I_G(t) = 1 - \sum_{i=1}^{J} p_i^2 \tag{10}$$

where p_i is the probability of class i at node t and J is the number of classes.

3.4.3 Support Vector Classifier. Another kind of classification technique in machine learning, Support Vector Classifiers(SVCs) is well-known in the field for their efficiency when used for high-dimensional data. SVCs try to find the best hyperplane in the feature space to divide the classes. The hyperplane that maximizes the margin between the nearest data points of any class, or the support vector, is the optimal one. Furthermore, in difficult classification tasks, SVCs are particularly useful due to their ability to maximize margin on top of that, they can also perform non-linear classification effectively with the use of kernel strategies.

4. EXPERIMENTAL RESULTS

4.1 Results comparision for PCA

Table 1. : Experimental Results for Principal Component Analysis

Sr No	Coding Techniques	Transformation	ML Techniques	Accuracy
1	EIIP	DCT II	Voting Classifier	96.2%
2	EIIP	DCT III	Voting Classifier	95.8%
3	EIIP	FFT	Voting Classifier	98.65%
4	EIIP	Haar Wavelet	Voting Classifier	86.86%
5	EIIP	Coiflet Wavelet	Voting Classifier	95.06%
6	Numeric	DCT II	Voting Classifier	95.95%
7	Numeric	DCT III	Voting Classifier	95.95%
8	Numeric	FFT	Voting Classifier	98.31%
9	Numeric	Haar Wavelet	Voting Classifier	99.21%
10	Numeric	Coiflet Wavelet	Voting Classifier	95.7%
11	Complex	DCT II	Voting Classifier	96.18%
12	Complex	DCT III	Voting Classifier	97.19%
13	Complex	FFT	Voting Classifier	97.5%
14	Complex	Haar Wavelet	Voting Classifier	80%
15	Complex	Coiflet Wavelet	Voting Classifier	80%
16	EIIP	DCT II	Decision Tree	64%
17	EIIP	DCT II	SVC	56%
18	Numeric	DCT II	Decision Tree	67%
19	Numeric	DCT II	SVC	64%
20	Complex	DCT II	Decision Tree	72%
21	Complex	DCT II	SVC	70%
22	EIIP	DCT III	Decision Tree	75%
23	EIIP	DCT III	SVC	89%
24	Numeric	DCT III	SVC	62%
25	Numeric	DCT III	Decision Tree	59%
26	Complex	DCT III	SVC	71%
27	Complex	DCT III	Decision Tree	70%
28	EIIP	FFT	Decision Tree	88%
29	EIIP	FFT	SVC	94%
30	Numeric	FFT	Decision Tree	77%
31	Numeric	FFT	SVC	93%
32	Complex	FFT	Decision Tree	81%
33	Complex	FFT	SVC	95%
34	EIIP	Haar Wavelet	Decision Tree	61%
35	EIIP	Haar Wavelet	SVC	50%
36	Numeric	Haar Wavelet	Decision Tree	86%
37	Numeric	Haar Wavelet	SVC	84%
38	Complex	Haar Wavelet	Decision Tree	57%
39	Complex	Haar Wavelet	SVC	50%
40	EIIP	Coiflet Wavelet	Decision Tree	65%
41	EIIP	Coiflet Wavelet	SVC	52%
42	Numeric	Coiflet Wavelet	Decision Tree	65%
43	Numeric	Coiflet Wavelet	SVC	63%
44	Complex	Coiflet Wavelet	Decision Tree	62%
45	Complex	Coiflet Wavelet	SVC	67%

Table 1 summarizes the results on PCA for dimensionality reduction, highlighting the effect that machine learning algorithms, different transformations, and coding methods have on classification accuracy. Using EIIP coding, for instance, the Voting Classifier achieved maximum accuracy with FFT (98.65%), followed by DCT II (96.2%) and a low accuracy value of 86.86% for Haar Wavelet. Decision Tree and SVC showed lower accuracies, lying in ranges from 61-88% and from 50–94%, respectively. For Numeric coding, the Voting Classifier with Haar Wavelet gives the highest accuracy of 99.21%, Coiflet Wavelet gives 95.57%, and both DCT II and DCT III give consistently approximately 95.95%. Decision Tree and SVC yield accuracies in the range of 65-77% and 62-93%, respectively. With Complex coding, the Voting Classifier with FFT gives the highest accuracy of 97.5%, followed by DCT III with 97.19%, while Haar Wavelet yields the lowest accuracy of 80%. While Decision Tree and SVC give rather poor results with accuracies ranging from 57-81% and 50-95%, respectively. These results indicate the large impact of transformation, coding, and choice of algorithm on the accuracy in classification.

4.1.1 Performance Analysis of Signal Processing and Machine Learning Models with EIIP coding for PCA . The accuracy of three distinct machine learning algorithms—Voting Classifier, Decision Tree, and SVC—when paired with EIIP coding and diverse signal processing transformations is assessed in this section. The Voting Classifier obtains a high accuracy of 96.2% with DCT II transformation, while the accuracies of Decision Tree and SVC are much lower, at 64% and 56%, respectively. The Voting Classifier continues to perform well for DCT III, recording 95.8%; SVC follows it closely with at 89%, while Decision Tree records 75%. The Voting Classifier gives the best accuracies (98.65%, SVC 94%, and Decision Tree 88%) when the FFT transformation is used, demonstrating the efficacy of FFT. Besides this, all classifiers, including the Voting Classifier (86.86%), Decision Tree (61%), and SVC (50%), have much worse accuracy when the Haar Wavelet transformation is applied. The Voting Classifier finally achieves 95.06% accuracy after applying Coiflet Wavelet transformation, outperforming Decision Tree and SVC at 65% and 52%, respectively, proving its supremacy.

4.1.2 Performance Analysis of Signal Processing and Machine Learning Models with Numeric coding for PCA. The accuracy of three machine learning algorithms—the Voting Classifier, Decision Tree, and SVC utilizing Numeric coding with different signal processing transformations—is compared in this section. The Voting Classifier attains 95.95% accuracy for the DCT II transformation, while Decision Tree and SVC fall short at 67% and 64%, respectively. The Voting Classifier remains at 95.95% with DCT III, while SVC falls to 62% and Decision Tree to 59%. The Voting Classifier produces the highest accuracy at 98.31%, SVC at 93%, and Decision Tree at 77% when using the FFT transformation. By comparison, the Voting Classifier achieves the maximum accuracy of 99.21% when using the Haar Wavelet transformation, while Decision Tree and SVC also demonstrate good performance, at 86% and 84%, respectively. Lastly, with Coiflet Wavelet transformation, the Voting Classifier achieves 95.7%, while Decision Tree and SVC perform moderately at 65% and 63%, respectively.

4.1.3 Performance Analysis of Signal Processing and Machine Learning Models with Complex coding for PCA. The accuracy of the Voting Classifier, Decision Tree, and SVC employing complex coding with different signal processing transformations is compared in this section. Voting Classifier scores 96.18% for DCT II, whilst Decision Tree and SVC perform worse at 72% and 70%, respectively. The Voting Classifier achieves 97.19% with DCT III, way ahead of SVC at 71% and Decision Tree at 70%. The Voting Classifier produces 97.5% accuracy, SVC 95%, and Decision Tree 81%; these results demonstrate the efficacy of the FFT transformation. On the other hand, accuracy is greatly decreased by Haar Wavelet transformation, with the Voting Classifier at 80%, Decision Tree at 57%, and SVC at 50%. The Voting Classifier, at 80%, is the last product of the Coiflet Wavelet transformation; Decision Tree and SVC, at 62% and 67%, respectively, perform moderately.

Table 2. : Experimental Results for Linear Discriminant Analysis

Sr No	Coding Techniques	Transformation	ML Techniques	Accuracy
1	EIIP	DCT II	Voting Classifier	99.6%
2	EIIP	DCT III	Voting Classifier	99.8%
3	EIIP	FFT	Voting Classifier	99.7%
4	EIIP	Haar Wavelet	Voting Classifier	99.5%
5	EIIP	Coiflet Wavelet	Voting Classifier	99.6%
6	Numeric	DCT II	Voting Classifier	99.8%
7	Numeric	DCT III	Voting Classifier	99.4%
8	Numeric	FFT	Voting Classifier	99.8%
9	Numeric	Haar Wavelet	Voting Classifier	99.8%
10	Numeric	Coiflet Wavelet	Voting Classifier	99.3%
11	Complex	DCT II	Voting Classifier	99.7%
12	Complex	DCT III	Voting Classifier	99.3%
13	Complex	FFT	Voting Classifier	99.6%
14	Complex	Haar Wavelet	Voting Classifier	99.2%
15	Complex	Coiflet Wavelet	Voting Classifier	99.4%
16	EIIP	DCT II	Decision Tree	99%
17	EIIP	DCT II	SVC	97.3%
18	Numeric	DCT II	Decision Tree	99%
19	Numeric	DCT II	SVC	98.65%
20	Complex	DCT II	Decision Tree	99%
21	Complex	DCT II	SVC	98.53%
22	EIIP	DCT III	Decision Tree	99%
23	EIIP	DCT III	SVC	97.52%
24	Numeric	DCT III	SVC	98.03%
25	Numeric	DCT III	Decision Tree	99%
26	Complex	DCT III	SVC	98.05%
27	Complex	DCT III	Decision Tree	99%
28	EIIP	FFT	Decision Tree	99%
29	EIIP	FFT	SVC	98.08%
30	Numeric	FFT	Decision Tree	99%
31	Numeric	FFT	SVC	97.97%
32	Complex	FFT	Decision Tree	99%
33	Complex	FFT	SVC	97.97%
34	EIIP	Haar Wavelet	Decision Tree	99%
35	EIIP	Haar Wavelet	SVC	97.86%
36	Numeric	Haar Wavelet	Decision Tree	99%
37	Numeric	Haar Wavelet	SVC	98.42%
38	Complex	Haar Wavelet	Decision Tree	99%
39	Complex	Haar Wavelet	SVC	96.74%
40	EIIP	Coiflet Wavelet	Decision Tree	99%
41	EIIP	Coiflet Wavelet	SVC	97.86%
42	Numeric	Coiflet Wavelet	Decision Tree	99%
43	Numeric	Coiflet Wavelet	SVC	98.65%
44	Complex	Coiflet Wavelet	Decision Tree	99%
45	Complex	Coiflet Wavelet	SVC	97.19%

4.2 Results comparision for LDA

Table 2 presents the results after LDA for dimensionality reduction. The accuracy of classification is greatly influenced by the choice of the machine learning algorithm, transformation, and coding method, as shown in the table. The Voting Classifier with EIIP coding shows consistent high accuracy, with all the transformations producing results within a narrow range of 99.5% to 99.8%. Most notably, the Voting Classifier with DCT-III produces the highest accuracy of 99.8%. Decision Tree and SVC show reduced accuracy with EIIP coding, especially SVC's, which ranges from 97.3% up to 98.65%. The Voting Classifier is pretty robust with the FFT: 99.7% for EIIP, 99.8% for Numeric, and 99.6% for Complex coding. Haar Wavelet, in turn, provides competitive performance of up to 99.8%, which was especially obtained with Numeric coding. Basically, the DCT-II and DCT-III transforms act very well for the various methods of encoding. This is why the Voting Classifier reaches 99.6% for EIIP, 99.8% for Numeric, and 99.7% for Complex encoding. The rest of the analyzed algorithms classify as follows: the Decision Tree achieves 99%, while the SVC shows 96.74 to a maximum of 98.65% on all other transformations and encoding methods. These results confirm the better generalization of the Voting Classifier in different scenarios, while Decision Tree and SVC are more sensitive to transform and coding strategies. In the end, only a correct choice of the used coding strategy, transformation, and machine learning algorithm can yield an optimal classification performance.

4.2.1 Performance Analysis of Signal Processing and Machine Learning Models with EIIP coding for LDA . Using EIIP coding with different signal processing transformations, the accuracy of three machine learning algorithms Voting Classifier, Decision Tree, and SVC is assessed in this section. 99.6% with DCT II, 99.8% with DCT III, 99.7% with FFT, 99.5% with Haar Wavelet, and 99.6% with Coiflet Wavelet are the highest performance levels regularly achieved by the Voting Classifier. The robustness of the Decision Tree in handling EIIP encoded data is demonstrated by its strong performance, which maintains 99% accuracy across all transformations. Although the SVC algorithm performs better with FFT (98.08%) and DCT II (97.3%), it performs worse with Haar and Coiflet Wavelets (97.86%). In comparison, the SVC algorithm exhibits greater variability. Comparing SVC to the other models, this variability indicates that SVC has difficulty capturing the complexity of EIIP transformations to the fullest. Though not as effectively as the Voting Classifier and Decision Tree, SVC still shows some promise in spite of these difficulties.

4.2.2 Performance Analysis of Signal Processing and Machine Learning Models with Numeric coding for LDA . In this part, the accuracy of three machine learning algorithms SVC, Voting Classifier, and Decision Tree is compared using different transformations and Numeric coding. For Numeric DCT II, the Voting Classifier has the highest accuracy at 99.8%, followed by Decision Tree at 99% and SVC at 98.65%. The Voting Classifier outperforms the Decision Tree at 99% and SVC at 98.03% for Numeric DCT III, however it decreases somewhat to 99.4%. The Voting Classifier and Decision Tree with Numeric FFT attain 99.8% and 99%, respectively, whereas SVC achieves 97.97%, indicating the effectiveness of FFT in enhancing classification. The Voting Classifier retains 99.8%, the Decision Tree reaches 99%, and the SVC increases to 98.42% in the Numeric Haar Wavelet transformation. Last but not least, the Voting Classifier records 99.3%, Decision Tree keeps 99%, and SVC records 98.65% with Numeric Coiflet Wavelet, demonstrating a slight decline in performance for all models.

4.2.3 Performance Analysis of Signal Processing and Machine Learning Models with Complex coding for LDA . This section examines the accuracy of the Decision Tree, SVC, and Voting Classifier using Complex coding with various transformations. The Voting Classifier does better than the Decision Tree (99%) and SVC (98.53%) with a 99.7% accuracy rate for DCT II, exhibiting good model performance with little variations. While the Decision Tree maintains its 99% level and the SVC achieves 98.05%, the Voting Classifier for DCT III slightly declines to 99.3%. The voting classifier, decision tree, and SVC all perform exceptionally well in the complex FFT setup, with the voting classifier at 99.6%, decision tree at 99%, and SVC at 97.97%. This shows how powerful FFT is in enhancing classification accuracy.. After applying Haar Wavelet transformation, the Voting Classifier achieves 99.2%, the Decision Tree maintains its 99%, and SVC falls to 96.74%. Finally, the Voting Classifier scores 99.4% in the Coiflet Wavelet transformation, Decision Tree keeps scoring 99%, and SVC scores 97.19%. These results show some difficulties with complex feature extraction but overall good performance.

Table 3. : Coding Techniques Analysis for PCA and LDA

Technique	Dimensionality Reduction	Avg. Acc.	Std Dev	Worst Acc.	Best Acc.
EIIP	PCA	75.74	0.1777	50	98.65
Numeric	PCA	80.34	0.155651	59	99.21
Complex	PCA	76.39	0.150266	50	97.5
EIIP	LDA	98.78	0.008431	97.3	99.8
Numeric	LDA	99.01	0.005704	97.97	99.8
Complex	LDA	98.71	0.008629	96.74	99.7

4.3 Comparative analysis of Dimensionality Reduction Techniques

Table 3 examines three coding techniques EIIP, numeric, and complex for their average accuracy, standard deviation, worst accuracy, and highest accuracy across Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The Numeric method under PCA has the best average accuracy 80.34% and the lowest standard deviation 0.155651, with maximum and worst accuracy of 99.21% and 59%. EIIP has the lowest average accuracy 75.74% and the highest variability, with values ranging from 50% to 98.65%. The Complex technique has a comparable, but more consistent, average accuracy of 76.39%. All methods improve when combined with LDA. With less variability and an accuracy range of 97.3% to 99.8%, the average accuracy of EIIP rises to 98.78% . With the lowest standard deviation 0.005704 and the highest average accuracy 99.01%, the numeric coding technique performs consistently between 97.97% and 99.8%. With a complex approach, moderate variability and accuracy of 98.71% are achieved 0.008629. For all coding approaches combined, LDA greatly improves accuracy and consistency; Numeric performs particularly well in this regard.

Table 4 compares different signal processing methods with PCA and LDA, including DCT-II, DCT-III, FFT, Haar Wavelet, and Coiflet Wavelet, based on some evaluation criteria such as mean accuracy, standard deviation, minimum accuracy, and maximum accuracy. In the case of PCA, FFT obtained the highest mean accuracy of 89.35% with a low standard deviation of 0.078 and the accuracy ranging from 77% to 98.65%, which indicates stable performance. This is followed by DCT-III with an average accuracy of

Table 4. : Signal Processing Techniques Analysis for PCA and LDA

Technique	Dimensionality Reduction	Avg. Acc.	Std Dev	Worst Acc.	Best Acc.
DCT II	PCA	75.7	0.1594	56	96.2
DCT III	PCA	79.44	0.152	59	97.19
FFT	PCA	89.35	0.078013	77	98.65
Haar Wavelet	PCA	72.67	0.1826	50	99.21
Coiflet Wavelet	PCA	71.64	0.15249	52	95.7
DCT II	LDA	98.75	0.007668	97	99.8
DCT III	LDA	98.79	0.0061911	97.53	99.8
FFT	LDA	98.90	0.00738	97.97	99.8
Haar Wavelet	LDA	98.72	0.009345	96.74	99.8
Coiflet Wavelet	LDA	98.85	0.00626	97.86	99.6

79.44% and a standard deviation of 0.152 with an accuracy range from 59% to 97.19%. DCT-II comes next with an average accuracy of 75.7% with a standard deviation of 0.159 and an accuracy range between 56% and 96.2%. Haar Wavelet has the largest variability (standard deviation of 0.183), with an average accuracy of 72.67% and an accuracy ranging from 50% to 99.21%. Coiflet Wavelet has an average accuracy of 71.64%, a standard deviation of 0.152, and an accuracy range of 52% to 95.7% Under LDA, all methods show significant improvement. FFT maintains excellent performance with an average accuracy of 98.90%, a low standard deviation of 0.007, and accuracy ranging from 97.97 % to 99.8%. DCT-II and DCT-III also have good performances with average accuracies of 98.75% and 98.79%, respectively, and narrow accuracy ranges. The Haar Wavelet has an improved average accuracy of 98.72% with a standard deviation of 0.009 and thus varies between 96.74% and 99.8% in accuracy. On the other hand, the Coiflet Wavelet has an average accuracy of 98.85% with a standard deviation of 0.006 and accuracy range of 97.86%.

Table 5. : ML Techniques Analysis for PCA and LDA

Technique	Dimensionality Reduction	Avg. Acc.	Std Dev	Worst Acc.	Best Acc.
Voting Classifier	PCA	93.9	0.06307	80	99.21
Decision Tree	PCA	70	0.096693	57	88
SVC	PCA	71	0.163867	50	95
Voting Classifier	LDA	99.57	0.002059	99.3	99.81
Decision Tree	LDA	99	0	99	99
SVC	LDA	97.93	0.00533	97.19	98.65

Table 5 presents an overview of the various machine learning algorithms' performances under PCA and LDA, including Voting Classifier, Decision Tree, and SVC. The Voting Classifier, under PCA, has the maximum average accuracy of 93.9%, ranging from 80% to 99.21%, with a standard deviation of 0.06307. With greater variability, Decision Tree and SVC exhibit lower average accuracies of 70% and 71%, respectively. SVC is between 50% and 95%, and Decision Tree is between 57% and 88%. All strategies demonstrate a notable improvement under LDA. With an extremely low standard deviation of 0.002059, the Voting Classifier achieves an average accuracy of 99.57%, ranging from 99.3% to 99.81%. With no fluctuation, Decision Tree continuously maintains 99% accuracy. With a standard deviation of 0.00533, SVC increases to an average accuracy of 97.93%, ranging from 97.19% to 98.6%. Overall, it can be seen that LDA enhances the accuracy and consistency of all algorithms, with the Voting Classifier performing the best.

Table 6. : Comparison of Dimensionality Reduction Techniques

Technique	Avg. Acc.	Std Dev	Worst Acc.	Best Acc.
PCA	78%	0.1577	52%	99.21%
LDA	96.74%	0.0075	96.74%	99.80%

A thorough comparison of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with an emphasis on important performance indicators is provided in Table 6. The accuracy achieved using PCA is as follows: 52% at lowest, 99.21% at maximum, 78.17% at mean, and 0.1577 at standard deviation. With a mean accuracy of 96.74%, a maximum accuracy of 99.80%, a minimum accuracy of 96.74%, and a substantially smaller standard deviation of 0.00755, LDA, on the other hand, performs noticeably better than PCA. The better performance and consistency of LDA are demonstrated by this data. LDA is the best option for data processing tasks needing great precision and stability, as Table 6 amply illustrates its benefits. LDA assures higher dependability by reducing variability and improving average accuracies in complicated data contexts where precision and reliability are critical. This essentially makes LDA a very important tool for multifarious data analysis projects, offering improved performance and uniformity compared to PCA.

5. CONCLUSION

In this study, the interaction of different coding schemes and transformation methods was extensively studied to enhance the performance of machine learning with specific focus on PCA and LDA as dimension reduction techniques. With the assessment of EIIP, Numeric, and Complex coding systems with transformations such as FFT, DCT II, DCT III, and wavelet approaches like Haar and Coiflet, it was found through the study that FFT with Voting Classifier and PCA performed optimally to detect strong frequency-domain characteristics in EIIP and Numeric codings. Although Haar and Coiflet wavelets opened the prospect to detect local as well as global patterns, their performance depended more on the context and the type of data and hence they needed to be very carefully selected based on the data nature. However, LDA consistently delivered a very discriminative feature space regardless of coding and greatly enhanced class separability as well as accuracy of the classifier.

6. REFERENCES

[1] NIH: National Institute of Allergy and Infectious Diseases. Coronaviruses, March 2022.

[2] J. Emrani. Sars-cov-2, infection, transmission, transcription, translation, proteins, and treatment: A review. *International Journal of Biological Macromolecules*, 193:1249–1273, December 2021.

[3] S. Amin, A. Alharbi, M. I. Uddin, and H. Alyami. Adapting recurrent neural networks for classifying public discourse on covid-19 symptoms in twitter content. *Soft Computing*, 26(20):11077–11089, August 2022.

[4] A. Aleem, A. B. A. Samad, and S. Vaqar. Emerging variants of sars-cov-2 and novel therapeutics against coronavirus (covid-19), May 2023.

[5] W. Hariri and A. Narin. Deep neural networks for covid-19 detection and diagnosis using images and acoustic-based techniques: a recent review. *Soft Computing*, 25(24):15345–15362, August 2021.

[6] A. Khodaei, P. Shams, H. Sharifi, and B. M. Tazehkand. Identification and classification of coronavirus genomic signals based on linear predictive coding and machine learning methods. *Biomedical Signal Processing and Control*, 80:104192, February 2023.

[7] S. M. Naeem, M. S. Mabrouk, S. Y. Marzouk, and M. A. Eldosoky. A diagnostic genomic signal processing (gsp)-based system for automatic feature analysis and detection of covid-19. *Briefings in Bioinformatics*, 22(2):1197–1205, August 2020.

[8] K. Patel, V. Shah, N. Patel, and Y. Mehta. An non-invasive approach of corona genome detection. In *2020 International Conference on Advances in Computing, Communication and Materials (ICACCM)*, pages 154–157, 2020.

[9] T. Meng, A. Soliman, M. Shyu, Y. Yang, S. Chen, and S. Iyengar. Wavelet analysis in current cancer genome research: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(6):1442–1459, December 2013.

[10] Y. Yadav, S. N. Sharma, and D. K. Shakya. Detection of tandem repeats in dna sequences using short-time ramanujan fourier transform. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3):1583–1591, June 2022.

[11] J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr. Identification of protein coding regions using the modified gabor-wavelet transform. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):198–207, June 2008.

[12] G. S. Randhawa, M. P. M. Soltysiak, H. E. Roz, C. P. E. De Souza, K. A. Hill, and L. Kari. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *PLOS ONE*, 15(4):e0232391, April 2020.

[13] I. Muhammad, I. Mukhlash, M. Jamhuri, M. Iqbal, and M. I. Irawan. Classification of covid-19 variants using boosting algorithm. In *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 29–34, 2022.

[14] M. S. Hammad, V. F. Ghoneim, M. S. Mabrouk, and W. Al-Atabany. A hybrid deep learning approach for covid-19 detection based on genomic image processing techniques. *Scientific Reports*, 13(1), March 2023.

[15] I. Saha, N. Ghosh, D. Maity, A. Seal, and D. Plewczynski. Covid-deeppredictor: Recurrent neural network to predict sars-cov-2 and other pathogenic viruses. *Frontiers in Genetics*, 12, February 2021.

[16] M. A. El-Dosuky, M. Soliman, and A. E. Hassanien. Covid-19 vs influenza viruses: A cockroach optimized deep neural network classification approach. *International Journal of Imaging Systems and Technology*, 31(2):472–482, February 2021.

[17] NCBI Virus. Ncbi virus: Sequences for discovery, 2024.

[18] S. S. Sahu and G. Panda. Identification of protein-coding regions in dna sequences using a time-frequency filtering approach. *Genomics, Proteomics and Bioinformatics*, 9(1–2):45–55, April 2011.

[19] M. Akhtar, J. Epps, and E. Ambikairajah. Signal processing in sequence analysis: Advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):310–321, June 2008.

[20] M. Akhtar, J. Epps, and E. Ambikairajah. On dna numeric representations for period-3 based exon prediction. In *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, pages 1–4, 2007.

[21] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, January 1974.

[22] M. Kumar and T. K. Rawat. Design of fractional order differentiator using type-iii and type-iv discrete cosine transform. *Engineering Science and Technology, an International Journal*, 20(1):51–58, February 2017.

[23] M. Hassanzadeh and B. Shahrrava. Linear version of parseval's theorem. *IEEE Access*, 10:27230–27241, 2022.

[24] Patrick J. Van Fleet. The haar wavelet transformation. In *Discrete Wavelet Transformations: An Elementary Approach with Applications*, pages 125–181. Wiley, 2019.

[25] Shyh-Jier Huang and Cheng-Tao Hsieh. Coiflet wavelet transform applied to inspect power system disturbance-generated signals. *IEEE Transactions on Aerospace and Electronic Systems*, 38(1):204–210, January 2002.

[26] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016.

[27] N. Zhao, W. Mio, and X. Liu. A hybrid pca-lda model for dimension reduction. In *The 2011 International Joint Conference on Neural Networks*, pages 2184–2190, 2011.

[28] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee. Ai-based smart prediction of clinical disease using random forest classifier and naive bayes. *The Journal of Supercomputing*, November 2020.

[29] J. Hatwell, M. M. Gaber, and R. M. Atif Azad. Ada-whips: explaining adaboost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making*, 20(1), October 2020.

[30] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986.

[31] T. M. T. A. Hamid, R. Sallehuddin, Z. M. Yunos, and A. Ali. Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. *Machine Learning with Applications*, 4, March 2021.