Kmer-Based DNA Sequence Image Representation for Viral Disease, Translation and Mutated Pattern Prediction

Prasad Sankar[1], *Dhrupad* Sah[1], *Dheeraj* Kodati[2], and *Chandra Mohan* Dasari[1,*]

[1]Indian Institute of Information Technology, Sri City, Andhra Pradesh, India
[2]Apollo Computing Laboratories Pvt. Ltd., Hyderabad, Telangana, India

Abstract. Accurate prediction of viral diseases is crucial for effective public health strategies, as mutations in DNA sequences can lead to various viral infections. The translation rate of these DNA sequences significantly impacts the severity of the disease. DNA sequencing techniques are capable of extracting variable-length sequences associated with these diseases. However, existing computational techniques often struggle to effectively utilize DNA sequence data for predictive modeling. To address this challenge, we propose a generalized Convolutional Neural Networks (CNNs) model trained on DNA sequences for predicting different viral disease classification tasks. In our preprocessing technique, DNA sequences are transformed into image-like structures using 6-mer frequencies. We conducted comprehensive experiments, including realm classification, SARS-CoV2 binary classification, and classification of seven types of coronaviruses (CoVs). Our approach achieved significant improvements in test accuracy: 89.51% for realm (4-class) classification, 99.80% for SARS-CoV2 binary classification, and 90.97% for coronavirus (7-class) classification. Additionally, we identified various mutations and translation rates of different CoVs using CDs. While CNNs demonstrate better performance, they are inherently black boxes. To address this issue, we performed interpretability analyses to extract the relevant features of various CoVs.

1 Introduction

Accurate prediction of viral diseases is a cornerstone for the development and implementation of effective public health strategies. Viral epidemics and pandemics, such as those caused by influenza, HIV, and more recently, SARS-CoV2, have demonstrated a profound impact on global health and economies. Traditional methods for predicting viral diseases often involve epidemiological models, statistical analyses, and basic bioinformatics tools that primarily focus on known risk factors, historical data, and clinical symptoms [1, 2]. While these methods have been valuable, they encounter significant limitations when it comes to harnessing the full potential of genomic data, especially the vast and complex information contained within DNA sequences. Earlier studies in the domain of viral disease prediction have typically employed sequence alignment and phylogenetic analyses, which are

*Corresponding Author e-mail: chandramohan.d@iiits.in

computationally intensive and often lack the sensitivity to detect subtle but critical variations in DNA sequences that may correlate with disease outcomes [3]. Moreover, they often fail to integrate multi-dimensional data efficiently, which is essential for capturing the intricate relationships between genetic variations and phenotypic expressions of diseases. Conventional methods often fail to capture complex relationships in DNA sequences.

One major challenge is the effective representation of DNA sequence data for predictive modeling. Traditional techniques treat DNA sequences as linear strings of characters, which can obscure the complex, multi-layered information encoded in these sequences. Additionally, the high dimensionality of genetic variable length data poses a significant challenge for conventional machine learning algorithms, which can suffer from overfitting and computational inefficiency when dealing with such complex datasets. Virus taxonomy, including realms, seeks to organize viruses based on shared traits such as genome type, replication strategies, and structural proteins, yet significant challenges remain [4].

CoVs are a family of viruses categorized into two groups: human CoVs and zoonotic CoVs. Common human CoVs, such as 229E, NL63, OC43, and HKU1, typically pose a lower risk to humans. In contrast, zoonotic CoVs like SARS-CoV, MERS-CoV, and SARS-CoV2 originated in animals and were transmitted to humans, presenting a significantly higher threat [5]. Identifying mutations in the CoVs is crucial for understanding its spread and developing strategies to combat future pandemics. This information helps not only control current outbreaks but also predict and prevent future ones. Additionally, understanding these mutations is vital for drug discovery, as it allows scientists to develop targeted treatments and vaccines that can effectively neutralize the virus, even as it evolves.

Motivated by these limitations, our research seeks to advance the field by introducing a novel approach that leverages the power of deep learning, specifically CNNs, to predict viral disease realms and coronavirus classification from DNA sequences. We propose a pioneering method where DNA sequences are transformed into image-like representations, enabling the application of CNNs for genomic data analysis. While CNNs can effectively classify various coronavirus sequences, they do not explicitly reveal the underlying patterns used for differentiation. The Black-box nature of CNN models hampers biological interpretability, reducing their utility for actionable insights. Understanding the divergent patterns extracted by CNNs is crucial, as these patterns help distinguish one coronavirus strain from another. Additionally, it is essential to investigate how mutations in different coronaviruses contribute to the spread of the disease [7]. By uncovering these patterns and studying mutation-driven variations, researchers can gain deeper insights into viral behavior.

This study addresses several key challenges identified in existing works:

- Converting DNA sequences into image-like structures allows the proposed model to capture spatial dependencies and hierarchical features.
- The proposed model efficiently handles variable lengths of large-scale genomic data, enhancing the scalability of predictive models.
- The translation rate of SARS-CoV2 was identified to be higher than that of other coronaviruses by calculating the average proportions of slow codons and slow di-codons.
- The mutations in SARS-CoV2 were found to be fewer than those in other coronaviruses by comparing the codons with the reference sequence.
- To address the lack of a decision-making process in CNNs, we explore relevant features (patterns) that distinguish different types of coronaviruses by utilizing the learned filters.

2 Literature Review

The study in [8] employed deep learning (DL) models such as CNN and LSTM variants with K-mer encoding for DNA sequence classification, aiming at accurate virus detection and drug design. The k-mer approach effectively transforms DNA sequences into numerical formats, aiding analysis with machine learning algorithms [3]. However, limitations include

potential sensitivity to k-mer size selection, scalability issues with larger genomes, and challenges in handling sequence variations like mutations. In a related study [1], the application of CNNs to DNA sequence classification was explored. The research highlights CNNs' adaptation from image processing to genomic sequence analysis, leveraging their hierarchical structure for enhanced classification accuracy. Limitations include potential overfitting due to CNNs' parameter sensitivity and challenges in the interpretability of learned features in biological contexts. DL models show promise in pandemic detection and prediction but are limited by varying performance and methodological gaps in studies [9]. This work [3] proposes various deep learning models for DNA sequence classification, using k-mer and one-hot encoding techniques, and introduces a multi-transformer model with a pairwise features fusion technique. This work [10] introduces an advanced deep learning approach for Hepatitis C virus genotyping using a graphical representation of nucleotide sequences, achieving better classification accuracy. Despite its effectiveness, a major challenge remains in dealing with data scarcity, particularly in computational genomics.

The work in [11] considered the CNN technique for high-accuracy classification of viruses and other organisms from genome sequences without limiting the sequence length. A notable drawback is the extensive parameter tuning required for CNNs, which can be resource-intensive and complex. The EdeepVPP focuses on predicting viral genomes using CNN and CNN-LSTM models that automatically extract classification features and provide model interpretability. Despite achieving increased performance in classification accuracy, a significant drawback is the relatively poor explainability of the DL models. The study [12] introduces an automated genomic image processing approach for detecting COVID-19. It uses genomic graphical mapping techniques to convert genome sequences into grayscale images and applies statistical features from these images to train classifiers, including k- nearest neighbors. The work in [13] proposes a deep neural network classifier based on stacked sparse autoencoders for the SARS-CoV2 genome, using image representations of genome sequences as input. The primary issue is that SARS-CoV2 samples were only used in testing, not training, which may impact the model's adaptability to classify other emerging viruses.

The proposed approach significantly advances earlier works by introducing a pioneering approach that utilizes CNNs trained on DNA sequences represented as images to predict viral disease realms and CoVs classification. Unlike previous methods, which struggled with effectively leveraging DNA sequence data, our novel preprocessing technique transforms DNA sequences into image-like structures using 6-mer frequencies, optimizing CNN performance. This innovative method demonstrates superior accuracy in various classification tasks, in- cluding realm classification, distinguishing SARS-CoV2 from non-SARS-CoV2 sequences, and classifying multiple types of CoVs. In this study, we compared the translation rates of various CoVs by calculating the average proportions of slow codons and slow di-codons. We also identified various mutations at the codon levels of different CoVs. We extracted various patterns that motivate as features to classify viral CoVs by explaining CNNs.

3 Methodology

We propose a generalized CNN model for diverse classification tasks by transforming variable-length DNA sequences into fixed-size images.

3.1 Model Architecture

The proposed CNN based model comprises an input layer, three 2D convolutional layers, two fully connected layers, and an output layer as shown in **Figure 1**.

The CNN based architecture is specifically designed for robust classification tasks. For realm classification and binary classification of SARS-CoV2, the model operates on grayscale images of size 64 × 64 pixels, while for the 7-class classification of CoVs, images are resized to 32 × 32 pixels. The model architecture begins with an input layer **X**, representing the input image tensor. Following the input layer, a rescaling layer normalizes the pixel values to the range [0, 1]. The core of the CNN architecture consists of three 2D convolutional blocks. Each block starts with a convolutional layer Conv2D, which convolves the input tensor $\mathbf{X}_{\mathbf{norm}}$ with a set of learnable filters W. The ReLU activation function is used to introduce non-linearity.

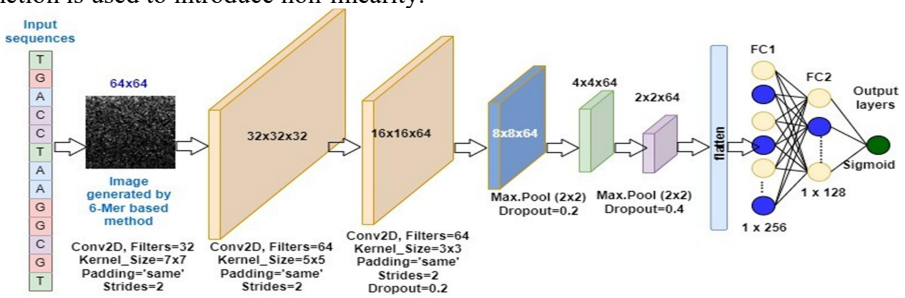


Fig.1. Generalized CNN Model Architecture

To maintain spatial dimensions and reduce computational load, 'same' padding is applied, preserving the spatial dimensions. A stride of two is used to down-sample the feature maps. Dropout layers **Dropout** are inserted after each convolutional layer to prevent over-fitting. Interleaved with the convolutional blocks, Max Pooling layers **Pool** perform spatial down-sampling. After convolutional and pooling layers, a Flatten operation reshapes the tensor into a vector. The fully connected layers consist of dense layers **Dense** that progressively reduce dimensionality. The final dense layer with c neurons, where c is the number of output classes, computes logits. The proposed generalized model is used for three various tasks by changing the number of outputs, *four* for Realm, *two* for SARS-CoV2, and *seven* for coronavirus classifications. This architecture optimizes feature extraction from DNA sequences, enabling accurate classification across diverse datasets.

In the proposed approach, DNA sequences are transformed into image-like structures using a 6-mer based method. A 6-mer refers to a sequence of six nucleotides, which are computed from the input DNA sequence (composed of nucleotides A, C, G, and T) to generate frequency matrices. These matrices represent the frequency of different 6-mers, which are then reshaped into a 64x64 pixel image. This transformation is essential because it lever- ages CNNs' ability to process image data, allowing the model to learn patterns and structures within the DNA sequences in a more efficient manner. By representing genomic data as images, the model can identify critical motifs or structural features that correlate with specific viral disease realms or strains, enabling it to distinguish between various viral families with high accuracy.

4 Dataset Collection and Image Representation

This section discusses the collection of various datasets for different classification tasks and a novel technique to convert variable length genome sequences into fixed length grayscale images.

4.1 Dataset Collection

We collected three types of genome sequence datasets for three different tasks, such as Realm (4-class), SARS-CoV2 (binary), and Coronavirus (7-class) classifications. The first dataset

contains 14,684 DNA sequences of four different Realms: Duplodnaviria, Monodnaviria, Riboviria, and Varidnaviria collected from Virus-Host DB. The second dataset comprises 1,553 SARS-CoV2 RNA samples, downloaded from the National Center for Biotechnology Information (NCBI). In the third dataset, we collected and analyzed 4,143 coding sequences (CDs) from seven types of CoVs capable of infecting human hosts, sourced from the Virus Pathogen Resource (ViPR) [14]. A CDs refers to the segment of a gene, either DNA or RNA, that codes for proteins. These CDs represent host-specific viral-infected mRNAs. The three datasets' details like the number of sequences or CDS and length in base pairs (bps) are outlined in **Table 1**. Additionally, reference genomes and test sequences from various CoVs were obtained from NCBI to assess mutational bias.

Table 1. Number of sequences and its ranges of three datasets.

Dataset	Type	# Sequences	Range (bps)	Dataset	Type	# Sequences	Range (bps)
Realm	Duplodnaviria	3671	174-497513	Corona Virus	229E	113	51-20292
	Monodnaviria	2973	618-6334		OC43	1130	239-21288
	Riboviria	2703	229-124279		NL63	270	51-20190
	Varidnaviria	5337	279-1473573		HKU1	238	249-21654
SARS-CoV2	SARS-CoV2	1553	29432-29945		MERS-CoV	27	42-21237
	Non-SARS-CoV2	1702	174-497513		SARS-CoV	1863	80-21291
					SARS-CoV2	502	107-21291

4.2 Image representation based on K-mer Frequency

The collected dataset shown in **Table 1**, indicates significant variation in the number of base pairs across sequences. Before training the model, each sequence is converted into an image. Since CNNs require fixed-length input, converting these sequences to a fixed length results in a loss of information. Therefore, it is essential to encode this data effectively to retain important information and minimize information loss. Our approach utilizes a novel technique that converts DNA sequences into images based on k-mer frequencies. Specifically, we focus on 6-mers, which are subsequences of six nucleotides, as this has previously been proven [15] to provide informative encoding.

The DNA/RNA sequence contains four bases {A, C, G, T}. The number of possible 6-mers with 4 bases is 46, which is converted into 64x64 images. Each pixel in the generated image corresponds to the frequency of a specific 6-mer, with the Pixel Intensity (PI) reflecting its prevalence. The frequency of each k-mer ki j in the sequence is extracted and then Image PI is calculated as follows:

$$PI_{ij} = \frac{\text{Frequency of } k_{ij} \text{ in sequence } S}{\text{Maximum frequency of any k-mer in } S} \times 255 \quad (1)$$

Each sequence is represented as a 64x64 grayscale image and each pixel value is PI_{ij}. This k-mer based image representation preserves the sequential information of nucleotides, enabling the CNN to extract spatial features effectively. By converting the DNA sequences into a format that leverages CNN's strengths in image processing, we enhance the model's ability to identify patterns and relationships within the genetic data, ultimately improving the accuracy and reliability of the classification task. Our approach ensures that critical sequence information is maintained and leverages the powerful feature extraction capabilities of CNNs, providing a robust method for genome sequence analysis across varying sequence lengths.

5 Results and Discussion

The proposed generalized CNN based model is applied to k-mer based DNA sequence image representation approach and was evaluated on three different classification tasks:

Realm(4-class), SARS-CoV2(binary), and coronavirus (7-class) classifications. The training was conducted on Google Colab with GPU support and implemented by using Keras. For local development, Jupyter notebooks on a system with 8GB RAM, an Intel i7 processor, and 1 TB SSD storage ensured efficient experimentation. Evaluation metrics included accuracy and loss function with hyperparameters optimized via random search.

5.1 Realm (4-Class) Classification

This task involves classifying DNA sequences into four different realms: Duplodnaviria, Monodnaviria, Riboviria, and Varidnaviria. The proposed CNN model achieved a test accuracy of 89.51%. This result demonstrates the effectiveness of the k-mer based image repre- sentation in capturing discriminative features for distinguishing between different realms of life. The realm classification task is particularly challenging due to the diverse nature of the four realms each with its unique characteristics and evolutionary lineages.

5.2 SARS-CoV2 Binary Classification

In this task, the goal is to identify whether a given DNA sequence belongs to the SARS-CoV2 virus or not. The accurate identification of SARS-CoV2 sequences is crucial for effective disease monitoring, contact tracing, and therapeutic development. The proposed approach achieved an impressive test accuracy of 99.80%. This high accuracy showcases the ability of the k-mer based image representation to accurately classify SARS-CoV2 sequences.

5.3 Coronavirus (7-Class) Classification

The task involves classifying DNA sequences into seven types of CoVs. The proposed generalized model achieved a test accuracy of 90.97%. This result demonstrates the potential of the k-mer based image representation in identifying and classifying coronavirus sequences. The results of the classification are summarized in **Table 2**. The Realm dataset, comprising 126,340 parameters with a 64x64 image size, achieved an accuracy of 89.51% and a test loss of 0.242. The classification of viral realms is a complex and evolving task due to the distinct characteristics and evolutionary pathways of each realm. In contrast, the model excelled on the SARS-CoV2 dataset, with only 17,585 parameters and the same image size, achieving an impressive accuracy of 99.80% and a minimal loss of 0.001, indicating enhanced feature discernibility. The Coronavirus dataset, using a smaller 32x32 image size, recorded a test accuracy of 90.97% and a loss of 0.218 with 17,606 parameters, suggesting commendable but less optimal performance compared to SARS-CoV2. Overall, this evaluation underscores the critical role of dataset characteristics and model parameterization in achieving high classification accuracy, guiding future optimization strategies in CNN architectures for viral classification tasks.

Table 2. Performance evaluation of generalized CNN model on three datasets

Classification	Model Parameters	Image Size	Test Loss	Test Accuracy (%)
Realm	126,340	64x64	0.242	89.51
SARS-CoV2	17,585	64x64	0.001	99.80
Coronaviruses	17,606	32x32	0.218	90.97

The proposed model is also used for 7-class coronavirus classification by converting the sequences into one-hot encoding vectors. We prepare the fixed length sequences from variable length by adding ‘N’ at the end of small sequences. In this model, Conv1D’s are used instead of Conv2D’s. The 10-fold classification AUC-ROC and AUC-PR curves are shown in Figure 2 and the average accuracy of the model is 84.26%. The model gives better accuracy if we pass viral sequences as images than one-hot encoding vectors.

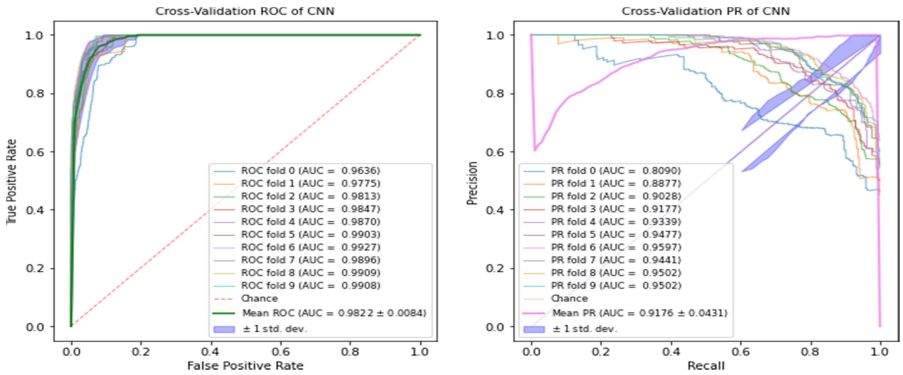


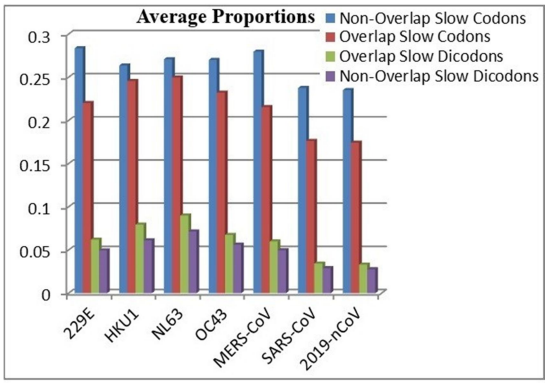
Fig. 2. The AUC-ROC and AUC-PR curves of corona virus classification

5.4 Translation Rate Prediction

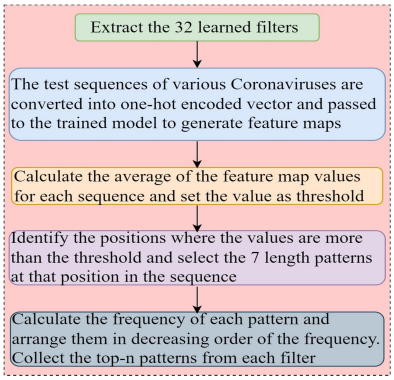
In humans, there are 13 recognized slow codons (ACC, AGT, CAT, CCC, CGC, CTC, GAT, GCC, GGT, GTC, TCC, TGT, TTT) [16]. From these slow codons, 169 slow di-codons were generated by forming pairs of two consecutive codons. The presence of two consecutive slow codons can significantly reduce the translation rate. The frequency of each codon type is measured by using a codon usage package of sequence manipulation suite [17]. The average proportions of human overlapping and non-overlapping slow codons and slow di-codons are extracted from CDs of seven CoVs by using approach [5] and depicted in **Figure 3a**. The number of slow codons and slow di-codons is inversely proportional to the translation rate. The figure clearly illustrates that the translation rate in SARS-CoV2 is higher than the other coronaviruses.

5.5 Mutation Rate Prediction

During replication, coronaviruses can accumulate mutations that alter their codons. DNA mutations are alterations in the genetic sequence that can lead to various diseases. Some mutations might change a slow codon to a fast one or vice versa. These mutations can occur in different forms: Transition, Transversion, Silent, Missense, and Nonsense Mutations. To identify mutations across various coronavirus sequences, we perform codon-based alignments. We analyzed the CDs of three coronaviruses at the codon level, comparing them with their corresponding reference sequences to identify various mutations. The number of distinct mutations in certain CDs of NL63, MERS-CoV, and SARS-CoV2 are provided in **Table 3**. The number of Missense mutations is higher in all coronaviruses. The number of various mutations of SARS-COV2 is much lesser than NL63 and less than MERS-CoV. Notably, NL63 strains, such as KU521535/UF-2/2015, exhibit higher mutation rates, with 70 transitions and 49 missense mutations, suggesting significant sequence variation that could affect protein-coding regions. In contrast, MERS-CoV strains, like MH454272/HCoV- EMC, display similarly high mutation counts, with 57 transitions and 56 missense mutations, reflecting the virus's genetic diversity and its impact on sporadic human transmission. SARS-CoV2 strains, however, show relatively lower mutation frequencies, as observed in MT163716/SARS-CoV2/human/USA with only 2 transitions and 5 missense mutations, suggesting a more conserved genome, possibly aiding its rapid transmission and adaptation. Nonsense mutations are entirely absent, while silent mutations are occurring in SARS-CoV-2 strains. According to the literature, there are no existing models for Realm and COVs classifications. So, we are not comparing our results with state-of-the-art models. Due to variable length sequences, the model gives better results for k-mer based image representation than the one-hot encoding technique.



(a)



(b)

Fig. 3. (a) The average proportion of slow codons and di-codons in both approaches.
(b) Steps for CNN interpretability to extract major features from CoV sequences

Table 3. The number of distinct mutations in CoV disease sequences from various locations

Type	Accession Number/Strain	Transitions	Transversions	Silent	Missence	Nonsense
NL63	CS124012/UNKNOWN	21	46	3	64	0
	DJ009246/UNKNOWN	21	46	3	64	0
	JX504050/RPTEC/2004	3	0	1	2	0
	KT266906/Haiti-1/2015	48	6	27	26	1
	KT381875/UF-1/2015	67	12	31	46	2
	KU521535/UF-2/2015	70	12	31	49	2
	KX179500/UF-2/2015	68	12	31	47	2
MERS-COV	MH306207/HCoV-EMC	46	10	11	44	1
	MH013216/HCoV-EMC	43	5	6	42	0
	MH454272/HCoV-EMC	57	14	12	56	3
	MH432120/2366	57	14	11	58	2
	MH395139/2363	57	14	12	56	3
SARS-CoV2	MT049951/SARS-CoV2/ human/CHN/Yunnan-01	1	1	0	2	0
	MT163716/SARS-CoV2/human/USA/WA3-UW1	2	4	1	5	0
	MT328032/SARS-CoV2/human/GRC/10	3	0	1	2	0
	MT350282/SARS-CoV2/human/BRA/SP02cc	2	2	2	2	0
	MT358637/SARS-CoV2/human/IND/GBRC1	1	0	0	1	0
	MT371048/SARS-CoV2/human/LKA/COV53	1	0	0	1	0
	MT371048/SARS-CoV2/LC542809/TKYE6947	2	1	0	3	0
	MT374104/SARS-CoV2/ human/TWN/CGMH-CGU08	2	1	0	3	0

5.6 Interpretability for Mutated Pattern Prediction

To identify the underlying patterns that drive the prediction of various coronaviruses, we employed a computational approach similar to that used in [6]. The trained CNN model contains the learned filters. These filters are used to extract more important patterns from various corona viruses. **Figure 3b** illustrates the process for interpreting a CNN model to extract key features (patterns) from coronavirus sequences. We extracted the top five patterns, along with their filter IDs and frequencies, from each coronavirus, as shown in **Table 4**. It presents a comprehensive analysis of sequence motifs and their frequencies across six coronavirus

Table 4. Patterns that motivate various coronaviruses

229E			**HKU1**			**NL63**		
Filter Id	Pattern	Frequency	Filter Id	Pattern	Frequency	Filter Id	Pattern	Frequency
28	TGTTGTT	9089	32	TTAATAA	24414	32	TAAAAAT	34830
32	TTACAAA	8966	32	TAATAAA	23539	28	TGTTGTT	31274
32	TGAAAAT	8566	31	TAATAAT	21495	28	TTTTGTT	27449
32	TTAAACA	8259	28	TGTTGTT	20456	25	TAAAAAT	26463
28	TGGTGTT	7811	32	TGATAAA	18636	28	TAATGGT	24988
MERS-CoV			**SARS-CoV**			**SARS-CoV2**		
Filter Id	Pattern	Frequency	Filter Id	Pattern	Frequency	Filter Id	Pattern	Frequency
32	AACAAGT	15616	31	GAAGAAG	110057	32	AAAAAAA	53899
32	TTAAAGA	13539	32	TTGTTAA	98937	32	GAAGAAG	29201
32	TTAACAA	12519	31	TTACAAA	97783	32	TTAAAAA	28427
32	ACTACAA	12488	31	AACAATG	97231	32	AAAAGAA	28381
32	ATTACAG	12167	32	AAAAGAA	95926	32	TGAAGAA	28289

strains-229E, HKU1, NL63, MERS-CoV, SARS-CoV, and SARS-CoV2 highlighting critical patterns that may play pivotal roles in viral replication, evolution, and host interactions. A key observation is the recurrence of the TGTTGTT motif across multiple strains, including 229E (9089 occurrences), NL63 (31274 occurrences), and HKU1 (20456 occurrences), suggesting a conserved regulatory or structural function. Additionally, the GAAGAAG motif is highly prevalent in both SARS-CoV (110057 occurrences) and SARS-CoV2 (29201 occurrences), indicating a shared functional significance in these closely related viruses, potentially linked to replication or immune evasion mechanisms. Specific motifs such as TAAAAAT in NL63 (34830 occurrences) and TGTTGTT in HKU1 further emphasize the evolutionary conservation of essential genomic elements. The filter 32 is the most significant of all the filters. These common and strain-specific patterns offer critical insights into the conserved and diver- gent genomic features across coronaviruses, suggesting potential targets for antiviral strategies or vaccine design by exploiting these conserved motifs critical for viral pathogenicity and transmission.

6 Conclusion

This study introduces a generalized CNN model that effectively addresses the challenge of utilizing DNA sequence data for viral disease prediction by transforming sequences into image-like structures using 6-mer frequencies. Through comprehensive experiments, we demonstrated significant improvements in classification accuracy across multiple tasks, including realm classification, SARS-CoV2 binary classification, and classification of seven types of coronaviruses. The strength of our method lies in the synergy between k-mer based image representation, which captures local sequence patterns, and CNNs' capability to learn hierarchical features from image data. Additionally, our approach identified critical mutations and translation rates associated with various coronaviruses. Although CNNs are inherently black boxes, our interpretability analyses provided insights into the key features that differentiate these viral sequences. This work paves the way for more accurate predictive modeling and better understanding of viral mutations, contributing to improved public health strategies. Future work will focus on studying mutations and translation rates in SARS-CoV2 sequences from different geographical locations.

Acknowledgements: This research is supported by the Science and Engineering Re- search Board (SERB) (Sanction order File No. SRG/2023/000941), Government of India.

References

1. Qayyum, Abdul and Abdesslam, and others, Assessment and classification of covid-19dna sequence using pairwise features concatenation from multi-transformer and deep features with machine learning models, SLAS Technology 29, 100147 (2024).

2. S. Ajagbe, M. Adigun, Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review, Multimedia Tools and Applications 83, 1 (2023). 10.1007/s11042-23-15805-z
3. Sarkar, Bimal and Sharma et.al.,, Determination of k-mer density in a dna sequence and subsequent cluster formation algorithm based on the application of electronic filter,
4. G. Caetano-Anollés, J.M. Claverie, A. Nasir, A critical analysis of the current state of virus taxonomy, Frontiers in Microbiology 14, 1240993 (2023).
5. C.M. Dasari, R. Bhukya, Comparative analysis of protein synthesis rate in covid-19 with other human coronaviruses, Infection, Genetics and Evolution 85, 104432 (2020).
6. C.M. Dasari, R. Bhukya, Intersspp: Investigating patterns through interpretable deep neural networks for accurate splice signal prediction, Chemometrics and Intelligent Laboratory Systems 206, 104144 (2020).
7. Z. Almubaid, H. Al-Mubaid, Analysis and comparison of genetic variants and mutations of the novel coronavirus sars-cov-2, Gene reports 23, 101064 (2021).
8. Gunasekaran, Hemalatha and Ramalakshmi, et.al.,, Analysis of dna sequence classification using cnn and hybrid models, Computational and Mathematical Methods in Medicine 2021, 1835056 (2021).
9. K. Thakur, M. Kaur, Y. Kumar, A comprehensive analysis of deep learning-based approaches for prediction and prognosis of infectious diseases, Archives of Computational Methods in Engineering 30 (2023). 10.1007/s11831-023-09952-7
10. Fahmy, Ahmed and Hammad, Muhammed and Mabrouk, Mai and Al-Atabany, Walid, On leveraging self-supervised learning for accurate hcv genotyping, Scientific Reports 14 (2024).
11. Câmara, Gabriel B. M. et.al.,, Convolutional neural network applied to sars-cov-2 sequence classification, Sensors 22 (2022).
12. M. Hammad, M. Mabrouk, W. Al-Atabany, V. Ghoneim, Genomic image representation of human coronavirus sequences for covid-19 detection, AEJ - Alexandria Engineering Journal (2022). 10.1016/j.aej.2022.08.023
13. M. Coutinho, G. Câmara, R. De Melo Barbosa, M. Fernandes, Sars-cov-2 virus classifi- cation based on stacked sparse autoencoder, Computational and Structural Biotechnol- ogy Journal 21 (2022). 10.1016/j.csbj.2022.12.007
14. Pickett, Brett E and Sadat, and others, Vipr: an open bioinformatics database and analysis resource for virology research, Nucleic acids research 40, D593 (2012).
15. M. Sanabria, J. Hirsch, P.M. Joubert, A.R. Poetsch, Dna language model grover learns sequence context in the human genome, Nature Machine Intelligence pp. 1–13 (2024).
16. C.W. Yang, M.F. Chen, Composition of human-specific slow codons and slow dicodons in sars-cov and 2019-ncov are lower than other coronaviruses suggesting a faster protein synthesis rate of sars-cov and 2019-ncov, Journal of Microbiology, Immunology and Infection 53, 419 (2020).
17. P. Stothard, The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences, Biotechniques 28, 1102 (2000).