Deep Learning Analysis of Viral Genome Evolution in SARS Family and Ebola

Abhishak Raj[1*], Victor Jose M[2]

[1]Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Tamilnadu, India, abhishakraj525@gmail.com

[2]Department of Computer Science and Engineering, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Tamilnadu, India, mvictorjose@gmail.com

Corresponding: abhishakraj525@gmail.com

HOW TO CITE:

Abhishak Raj, Victor Jose M (2026). Deep Learning Analysis of Viral Genome Evolution in SARS Family and Ebola. International Journal of Special Education, 41(11s), 460-483

ABSTRACT:

Infectious diseases remain a hindrance to global health with the pace of anthropogenic, behavioral, and social change fostering emergence and expansion of new pathogens. In this regard, insights into high-impact virus genomic structure and evolutionary processes are important for informed public health interventions. The current study offers an exhaustive comparative examination of the four significant viruses' genomic sequences: COVID-19, SARS, MERS, and Ebola. High-throughput genomic sequencing methods were used in combination with a Gated Recurrent Unit (GRU)-solved deep learning algorithm to decipher the intricate genetic characteristics of the pathogens. The comparative genomics strategy identifies both conserved and shared genetic features between the viruses and highlights their transmission mechanisms, host interactions, and pathogenicity. As SARS-CoV-2 has high sequence identity with SARS-CoV and MERS-CoV, unique spike protein structural differences and receptor-binding domains account for variability in infectivity and host range. Integration of deep learning allows for sophisticated pattern detection in large-scale genomic data, improving the accuracy of evolutionary and functional inference. The GRU model showed excellent performance, with accuracy levels of 99.01% for COVID-19, 98.91% for SARS, 98.35% for MERS, and 98.04% for Ebola, with high recall, precision, and F1-scores, highlighting the model's credibility for viral genome prediction.

These findings support the model's application in differentiating between highly similar viral genomes and extracting principal genetic signatures of importance for diagnostics and therapeutics. This research not only expands our knowledge of emerging viruses' genetic associations and evolutionary pathways but also highlights the promise of machine learning-based genomic analysis as a preparedness and response strategy against future epidemics. The results provide critical information for the design of focused diagnostics, vaccines, and antiviral measures, emphasizing the role of genomic surveillance in global health security.

Keywords: Deep Learning, Gated Recurrent Unit, Genome, Epidemic, Virus.

COPYRIGHT STATEMENT:
Copyright: © 2026 Authors.
Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

Pathogens have a better chance than ever to transmit in a contemporary closely connected and populated world. Averagely, rates of pandemic of infectious diseases have increased and their complexity since 2000 due to the intense international developments or rapids, Urbanization and international traveling. The past few decades have been characterized by numerous viral outbreaks of the global spectrum such as Zika, Ebola, Nipah virus, SARS, MERS, and influenza A (H5N1)[1]. The SARS epidemic of 2003 was the initial 21 st century pandemic to be declared a public health emergency, quickly followed by the H5N1 [2]. Other outbreaks that resulted in the activation of the WHO Public Health Emergency of International Concern (PHEIC) are: the 2009 H1N1 pandemic, the wild poliovirus resurgence in 2014 and the West African Ebola epidemic, the Zika epidemic of 2016, and the COVID-19 pandemic (2020 - present) [3]. Such outbreaks highlight how the globalized world is constantly in danger of emerging infections. Such outbreaks are of the virus Zika, Ebola, Nipah virus, severe acute respiratory Such occurrences confirm the worsening of the risks of contagious diseases in the globalized world caused by other factors, including urbanization, travels, and ecological transformations [4]. Repeated evidence of the declaration of PHEIC shows the necessity of the enhanced worldwide preparedness and integrative response procedures [5]. Figure 1 provides a timeline illustrating the development of these prominent viruses, showcasing the dynamic landscape of global health challenges.

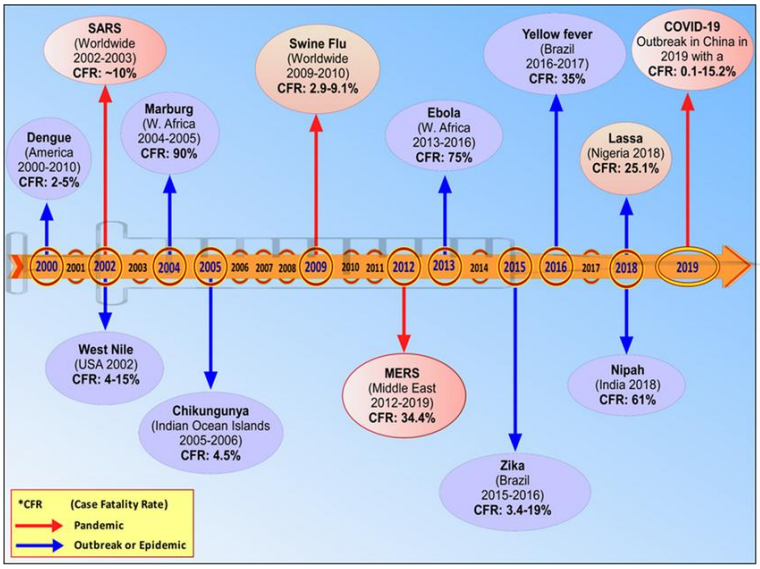
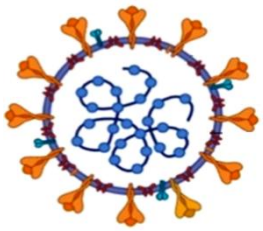
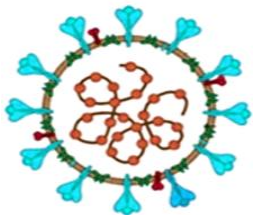
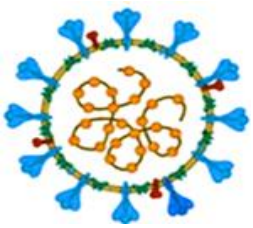
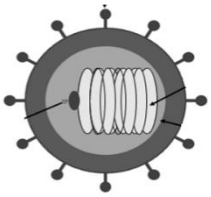


Figure 1 Timeline of the most common viruses that are developing

Even when the agents mentioned above are known, the risks they create remain unpredictable and their contagion is almost impossible to control as the Covid 19 and Ebola outbreaks have shown [6]. Furthermore, considering unfamiliar biodiversity, constant and incredibly quick development and growing human exposure, it is highly likely that many new disease species and strains are still undiscovered [7]. It's harder to evaluate and forecast risks when there are several hosts or vectors, as there are many. For instance, Ebola most likely originated in fruit bats [8], yet it can spread to people and chimpanzees and produce fatal outbreaks. Since genome sequencing is the state-of-the-art method for the open-view discovery of pathogens [9], it is essential to create automated procedures for assessing the infectious potential of sequences that cannot currently be identified. Clinical samples often include less than 100 reads of the pathogenic virus, with host reads and contaminants predominating [10]. Table 1 provides a comprehensive comparison of the taxonomy and epidemiology of different viruses, highlighting key aspects such as family, genus, transmission modes, epidemic status, severe cases and origins.

Table.1. A comparison of the virus's taxonomy and epidemiology

Characteristic	SARS-CoV-2	MERS-CoV	SARS-CoV	Ebola Virus
Virus Family	Coronaviridae	Coronaviridae	Coronaviridae	Filoviridae
Virus Genus	Beta coronavirus	Beta coronavirus	Beta coronavirus	Ebolavirus
Natural Hosts	Bats (potentially intermediate hosts)	Bats (potential intermediate hosts)	Bats (potential intermediate hosts)	Bats (natural host, potential others)

Transmission Mode	Human-to-human (respiratory droplets)	Zoonotic (Camels as potential source)	Human-to-human (respiratory droplets)	Animal-to-human (initial zoonotic)
Primary Symptoms	Respiratory illness, Fever, Cough	Respiratory illness, Fever, Cough	Respiratory illness, Fever, Cough	Fever, Severe Haemorrhagic Fever
Incubation Period	2-14 days	2-14 days	2-14 days	2-21 days
Case Fatality Rate	Varied (typically less than 1-2%)	Varied (30% or higher in some areas)	Varied (around 10%)	Varied (25-90% depending on strain)
Pandemic Impact	Global Pandemic	Limited Sporadic Cases	2002-2003 Outbreak	Localized Outbreaks, Occasional Large
Structure				

Comprehending the genetic composition of viruses becomes crucial for deciphering their complex processes of causing disease, spreading and changing over time [11]. Examining a virus's genome is currently an essential tool in the ongoing pursuit of world health, allowing investigators to decode the complex code hidden inside the virus's genetic information [12]. The RNA genome of a virus functions as its biological blueprint, including all the necessary information enabling the virus's ability to reproduce and function within a host tissue [13]. Figure 2 illustrates that the genetic code consists of a solitary strand comprising RNA molecules. Adjacent base pairings are created when four of the RNA-building components, A, U, C, as well as G, are organized in a specific manner. This arrangement encodes directions for duplicating the virus's genome along with producing proteins.

Most RNA viral genomes are essential to have their composition accurately determined to be able to sequence the genome [14]. RT-PCR is a major activity in this procedure and involves the conversion of RNA into complementary DNA (cDNA) so that it can be sequenced. The development of modern DNA sequencing technology has vastly improved the speed and efficiency at which genomics can be sequenced; and as such new research is able to decode at a higher accuracy the genetic make up of RNA viruses.

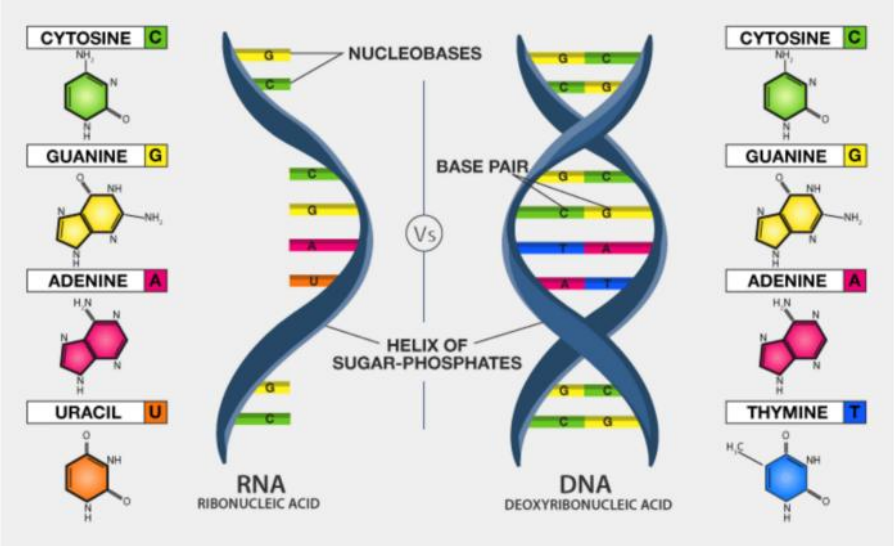


Figure 2. RNA and DNA Composition

The structure, the functioning and the evolution patterns of viruses are should be known to be able to attack them properly [15]. This kind of knowledge assists with the identification of vaccine and antiviral drugs targets. The sequencing of viral genomes has now turned out to be a strategic implementation to such a battle, providing information on pathogenicity, transmission, and drug resistant mutations [16]. Among the most significant objectives of genomic research is the emergence of precision therapeutics, which exploit genetic variations by personalizing medicine against the diseases. Moreover, there are also on-going systematic efforts to connect genomic sequences data with data on population health and behavior, in order to explore more on the relationships between disease and genetics. The method is especially crucial when looking at new viral strains such as SARS, Ebola, COVID-19, and MERS because it assists in monitoring their progress, transmission, and the health hazard that they cause [17]. Machine learning, deep learning, and artificial intelligence enable the new era of analysis of complex genomic data and thus speed up the process of discovery of valuable insights. Nonetheless, AI needs to be perfectly justified to achieve the required outcomes, with both clear goals and critical validation of every step of processing genomic data to gaining its valuable and precise outcomes. Even though the levels of advancement in genomics are currently high, much of the application of AI to this field is still in the research stage. The goals of this paper are based on the above discussion and the current arena of artificial intelligence in genomic sequencing. The most important ones are:

- To create a deep learning architecture of comparative genomics of the highly pathogenic viruses (COVID-19, Ebola, SARS, and MERS)
- To characterize and find evolutionary relationship and functional associations

among viral genome sequences including the examination of the number of amino acids, alignment patterns within genomic sequences and information about DNA similarity.

Computational modeling has changed the direction of viral genomics, as it opened an eye to the world of viral evolution, pathogenicity, and transmission dynamics. One of such innovations became the Gated Recurrent Unit (GRU), which has proved to be a very effective instrument in the analysis of genomic sequences [18]. It is highly suitable to analyze the viral genomes due to its capability to ascertain long-range dependencies and time-dependent relationships in sequential data. Already GRU-based models have come to play a vital role in the study of viruses with a high impact, including COVID-19, SARS, MERS, and Ebola, among other viruses, offering significant insights into the field of public health planning as well as in the invention of therapeutics [19].

2. MATERIALS AND METHODS

The genomic sequence data formed the basis of holistic study of viral genetic structure and evolutionary history. The COVID-19 pandemic has further increased the need to use high-tech approaches to genomic sequencing and analysis: it affects the entire world, which makes the demand even higher. Embedded in these genetic sequences is an important information that can unravel the path of evolution of a virus, its spread and its capacity to cause pathogenesis; these are all-important data in response to the virus in the collective appreciation to the health of the community besides the development of therapies. Understanding the origin, spread and molecular characteristics of the virus is crucial and this abundance of genetic data forms the basis for developing medical and antiviral treatments.

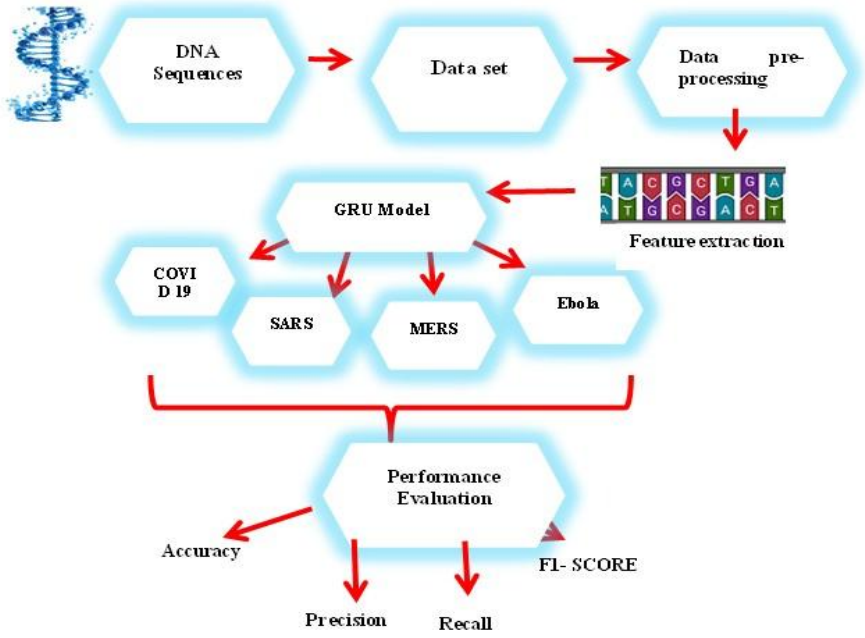


Figure 3. Workflow of the proposed model

The exposure of genomic sequence data has initiated an abundance of computing tools and techniques. One of such methods is the GRU algorithm; it is a deep learning approach that can handle time series exceptionally well, including genomic sequences. By using it to analyze genomics information, the GRU method facilitates the sorting of the viruses based on their genetic sequences, allowing to determine similarities and differences. Figure 3 represents the fine details of this genetic analysis process in a pictorial way.

Algorithm.1. GRU- Algorithm.1: GRU-Based Genome Sequence Analysis

Input: Genomic sequences from four viruses: COVID-19, SARS, MERS, Ebola.

Output: Analytical results on the genome sequences of the four viruses, including relationships and similarities.

Step 1: Data Preparation

- *Transform raw genomic data into numerical form suitable for machine learning.*
- *Convert DNA sequences into RNA format as needed.*
- *Divide the complete dataset into separate training and testing subsets.*

Step 2: Property Engineering

 Extract relevant features from the sequences, such as:

- *GC content*
- *Codon usage bias*
- *Nucleotide sequence patterns*
- *Derived protein sequences*
- *Structural and domain information*
- *Amino acid arrangements*

Step 3: GRU Model Construction

Set up the GRU neural network:

- *Initialize all model parameters and gates (reset, update, candidate, hidden state).*
- *Implement the GRU forward pass for each sequence:*
- *Compute the update gate to determine how much past information to retain.*
- *Generate the candidate activation for the new memory content.*
- *Update the hidden state using the computed gates and candidate activation.*
- *Produce the output for each time step based on the hidden state.*

Step 4: Model Training

- *Define a suitable loss function (such as categorical cross-entropy).*
- *Train the model over multiple epochs, adjusting parameters to minimize the loss.*
- *Use the training data to optimize the GRU's ability to learn sequence patterns.*

Step 5: Model Evaluation

- *Test the trained model on the reserved testing data.*
- *Calculate performance metrics including accuracy, precision, recall, and F1-score.*
- *Interpret the results to assess the model's effectiveness in identifying genetic relationships.*
- *Examine alignment patterns and similarities among the viral genomes.*

Step 6: Biological Interpretation

- *Use the model's output to gain insights into the genetic relationships between the viruses.*
- *Identify significant sequence patterns, evolutionary links, or unique features.*
- *Summarize findings to support further biological or epidemiological research.*

End

2.1 Dataset

Genomic metadata that are used in this study are sourced out of GenBank, the nucleotide sequence overall complete repository of the National Center for Biotechnology Information (NCBI). In addition to offering a huge genetic database, GenBank promotes biological annotation. As seen in Table 2, the sequences in the dataset were carefully sorted according to host details and nucleotide completeness. To ensure a concentrated and uniform dataset, only genome sequences designated as "complete genome" were retrieved for analysis.

Table 2. Metadata for the dataset

Column name	Type
Nucleotide Accession	Number
Species Taxonomy Id	Number
Species Name	String
Virus Genus	String
Virus Family	String
Isolate Name	String
Nucleotide Length	Number
Sequence Type	DNA sequence
Geo_Location	String
Host Taxonomy ID	Number
Collection Date	Date

2.2 Pre-processing

Preprocessing of the genomic data is a very important procedure in readying genomic sequences to undergo analysis using GRU. Raw genetic data are pre-processed to guarantee the best performance of our model through, the normalization of our features scales to create a standard measure, elimination of repetitive and

non coding sequences so as to reduce noise, standardization of lengths to the same dimension of inputs and their conversion into numerical representations to match a GRU architecture. All these measures increase data quality and reduce computational bias, and the capacity of the model to reveal meaningful biological patterns. The pre-processing stage is crucial as it entails converting the data into numerical form and matching it against a genome that serves as a reference, hence facilitating the detection of mutations in genetics and variations.

By eliminating copies of sequences, and any non-coding areas, possible outlining is minimized and the amount of data reduces. Also, the normalization process transforms the values of features to a certain standard range improving the quality of the further deep learning classifier. Since the viruses under analysis are constructed on the basis of RNA type and the information provided as input is in form of FASTA form, which consists of DNA, it is essential that a preprocessing procedure is executed which changes the DNA in form of RNA. The modification that is required is that RNA viruses have uracil (U) rather than thymine (T) to form all the nucleotides of oneself. As an example, the DNA pattern ATTAAAGGTT is used as an input in the process of generating.fna file. Since the virus genomes are RNA, we utilise the transcribe strategy whereby every T is replaced by an U in the sequence. The consequence of this change is the RNA that starts with an amino acid sequence that starts with AUUAAAGGUU. The transcribe function is significant in the conversion of DNA towards mRNA in order to guarantee the correctness of the ensuing studies that also incorporate translation by way of the amino acid chain.

2.3 Feature Extraction

Feature extraction process is the determinant stage in the genomic analysis process, that translates raw data to deserving patterns, that are used as feature on machine learning. This process within our system defines specific genetic determiners - SNPs, deletions, and maintained motifs, defining a wide variety of differences in viral strain with high specificity. We encode the genomic sequences into a bidirectional GRU network representing the sequences as a sequence-based numerical representation that did not lose its biological context. The most important features are such characteristics as GC content, codon usage pattern and nucleotide distribution manner, protein-coding patterns related to specific viral functions or pathogenicity in the current strategy. The biological relevant characteristics in the sequences have tuning capabilities and the noise on the performance of the model can be eliminated by the power of the feature extraction pipeline. Our system seeks out such strain-specific signatures by exploiting the information content in the primary sequence and in the structure, with a high specificity level) not only because we give consideration to primary-sequence information content but also to the derived structured elements (secondary structure and functional protein

domains). This discriminative power is significant to secure the accurate nature of the remaining deep learning models, yet offer interpretations of genomic features that supports the behavior of the virus.

2.4 Deep Learning Classifier -GRU

GRU is a modified version of recurrent neural network (RNN) which endeavored to solve the issue of vanishing gradient among traditional RNNs so that it enhances the ability of the network to construct long-range relationships between the sequential data. Figure 4 shows that GRU cell has two basic gating mechanisms, the update gate (z_t) and reset gate (r_t). These gates evolve so as to dynamically control the information flow; the issue being how to access enough but not too much of historical state to bring on high quality transitions at every timestep. In the case of sequential processing, often the GRU does three important computations at each time the activation of the reset gate, the activation of the update gate, and the calculation of the candidate hidden state ($\tilde{h}_t$). The candidate state is obtained out of the current input and the gated historical state. The last hidden state (h_t) is subsequently calculated as a weighted combination of the prior state and candidate state that allows selective memory to include the significant temporal patterns and adroitly process the inputs sequentially.

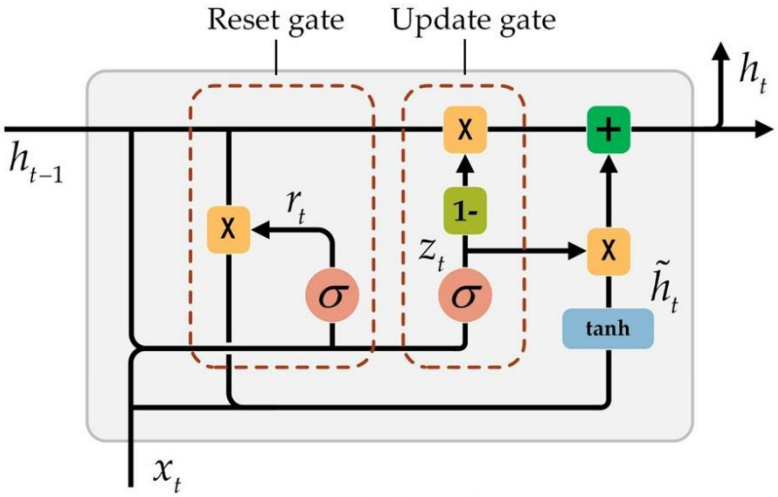


Figure 4. GRU Architecture

Figure 4 shows the current time step to be x_t, and the update gate and reset gate are Z_t and r_t, respectively. The current and the previous time step outputs are h_t and h_{t-1}. The GRU parameters are calculated in the following way.

The current time-step input is defined as x t, the update and reset gates are Z_t and r_t, respectively as shown in the Figure 4. h_t is the previous state output and h_{t-1} is current state output. Parameters of the GRU are calculated using the following expressions Equation (1) to Equation (6):

$$Z_t=\sigma\left(W_{z}\cdot[h_{t-1},x_t]\right) \quad (1)$$

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]) \quad (2)$$

$$\widetilde{h}_t = tanh(W.[r_t * h_{t-1}, x_t]) \quad (3)$$

$$h_t = (1-Z_t) * h_{t-1} + Z_t * \widetilde{h}_t \quad (4)$$

These matrices of parameters are W_z, W and W_r $\widetilde{h}_t$ represents the candidate hidden state. In this architecture, the sigmoid ($\sigma(x)$ and hyperbolic tangent $tanh(x)$) activations are used.

$$sigmoid = \frac{1}{1+e^{-x}} \quad (5)$$

$$tanh = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (6)$$

GRUs have emerged as a strong structure of sequence modeling, especially since they have demonstrated a strong ability to exploit long-range dependencies with the vanishing gradient issue. During our viral genomes analysis framework, we modulated the GRUs as a way to read and interpret genomics sequences in order to predict behavior and evolutionary paths. Implementation This is achieved by transliting viral RNA/DNA sequences into numerical form by one-hot encoding: each nucleotide (A, C, G, T/U) is encoded by a separate binary vector. The GRU network proceeds to process these coded sequences in a sequential manner and, as a result, the model learns complicated genomic relationships and contextual patterns.

The architecture of GRU provides peculiar framing of genomic analysis tasks by means of advanced information control that is implemented through gating methods supportive reset gates, update gates. It permits both local inputs of sequence feature and global patterns of evolution to be modeled - essential in the detection of protein-coding regions, regulatory elements and phylogenetic limit. Figure 5 shows the three abilities of our proposed model architecture, which can be applied to retain appropriate genomic surrounding and selectively revise such internal states according to the related sequence input and prior network states.

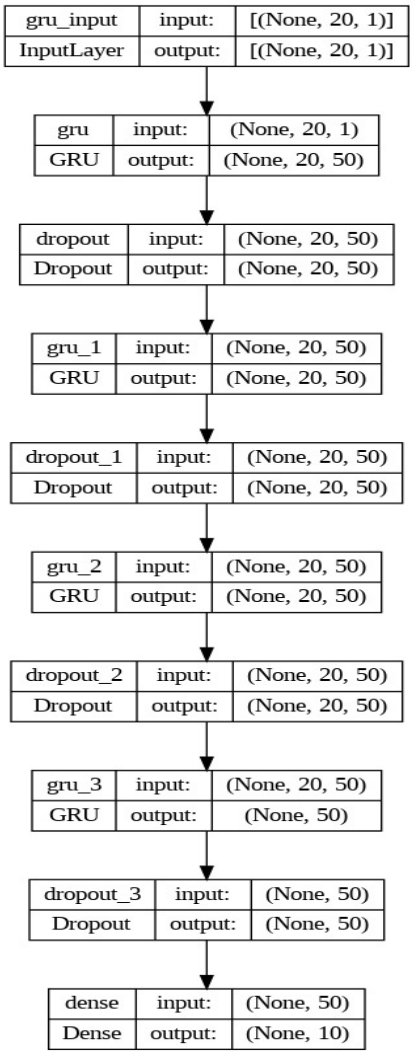


Figure 5. Model Architecture of the proposed model

2.5 Model optimization and hyper-parameter

The hyperparameters of a Deep neural network are determined empirically and make a profound difference to learning. Consequently, a huge number of variables are exhaustively conducted to determine the model that provides the best classification accuracy which is tabulated in Table 3.

Table 3. Hyperparameter specifications

Hyperparameters	**Values**
Learning Rate	0.001
Epochs	50
Activation function	ReLU
Dropout	0.2

Batch size	32
Loss Function	Categorical cross-entropy
Optimizer	ADAM

2.6 Evaluation Metrics

Measures of evaluation allow measuring the performance of models systematically by giving quantitative measures. The metrics result in a possibility to compare and justify the effectiveness of the proposed methodology in an objective way. The table 4 shows the main performance indicators that we applied in our evaluation framework.

Table 4. Evaluation Metrics

Performance Metrics	**Formulas**
Precision	TP/ (TP + FP)
F1-score	$\frac{2*precision*recall}{precision+recall}$
Recall	TP/(TP+FN)
Accuracy	(TP+TN)/ (TP + TN +FP + FN)

TP stands for the true positives, TN stands for true negative, FP stands for false positive and FN stands for false negative

3 . RESULTS AND DISCUSSIONS

The coding explorations were also carried out using the Google colab notebook tool, which deployed the Bio-Python as a smooth entry point programming language. The suggested model used only functionalities offered by sci-kit-learn and Keras to make data pre-processing effective. The platform on which the algorithm was developed had 8 gigabytes of RAM & 1.6 gigahertz Intel core i5 processor. The backpropagation computational framework enhanced learning and model optimization by resilient training. The effective use of Google Colab, Biopython, Scikit-learn, and Keras allowed the researchers to reach the goals effectively. This team work enabled an easier and powerful approach in discussing the genetic data.

Figure 6 displays the conventional codon table, demonstrating the correspondence between RNA codons with certain amino acids in their simpler forms. Codons, which are composed of three-

letter sequences of events, function as the core units of proteins. The structure of protein culminates during transformation indicated by the emergence of three-stop codons, specifically UAG, UAA, as well as UGA.

	U	C	A	G	
U	UUU F	UCU S	UAU Y	UGU C	U
U	UUC F	UCC S	UAC Y	UGC C	C
U	UUA L	UCA S	UAA Stop	UGA Stop	A
U	UUG L(s)	UCG S	UAG Stop	UGG W	G
C	CUU L	CCU P	CAU H	CGU R	U
C	CUC L	CCC P	CAC H	CGC R	C
C	CUA L	CCA P	CAA Q	CGA R	A
C	CUG L(s)	CCG P	CAG Q	CGG R	G
A	AUU I	ACU T	AAU N	AGU S	U
A	AUC I	ACC T	AAC N	AGC S	C
A	AUA I	ACA T	AAA K	AGA R	A
A	AUG M(s)	ACG T	AAG K	AGG R	G
G	GUU V	GCU A	GAU D	GGU G	U
G	GUC V	GCC A	GAC D	GGC G	C
G	GUA V	GCA A	GAA E	GGA G	A
G	GUG V	GCG A	GAG E	GGG G	G

Figure 6. RNA standard codon table

To have a comprehensive understanding of the genetic makeup of the viruses being investigated, it is important to analyze the DNA patterns' length, as depicted throughout Table 5. The overall length is obtained using a precise enumeration of the initial base pairs within the series, which is subsequently divided by the amount of space among each succeeding base pair. It allows the utilization of precise assessments in the analysis of the genetic segment.

Table 5. DNA sequence length of each sequence in investigation

Virus	DNA Sequence Length (Base Pairs)
SARS (SARS COV-1)	29,751
COVID-19 (SARS COV-2)	29,903

Ebola	18,959
MERS	30,119

The unique biological properties of different viruses are determined by a unique composition and sequence of nucleotides. Information on these genetic patterns is essential in interpretation of viral-host interactions as well as the dynamics in the spread of the virus among human populations. Genetic algorithm was accurately used to consider the structural complexity of the viral genes in terms of their based DNA structure and the respective amino acids sequences. Genomic data also involves the mapping of the complete genomic sequences of the pathogens, together with indication of specific genetic signatures, as per Figure 7. Such visualization not just explains the morphology in the interaction of viral proteins with the host systems but also underlines the biochemical mechanism that explains the viral behaviour.

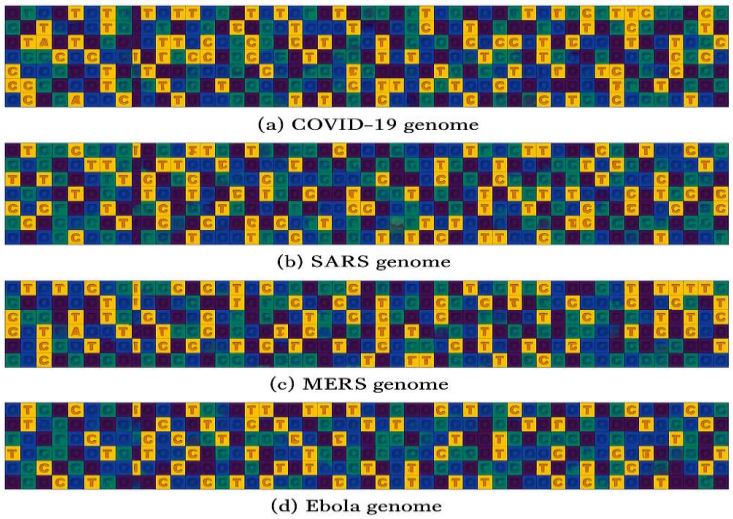


Figure 7. Illustration of virus DNA chains of nucleotides

Figure 8 presents a graphical representation of the nucleotide compositions of various genome sequences, highlighting distinct patterns in the abundance of adenine (A), thymine (T), cytosine (C) and guanine (G) across different sequences. This disparity results from the normal nucleotide pairing in DNA, which matches adenine with thymine and cytosine with guanine. Consequently, the prevalence of one nucleotide often inversely correlates with its pairing partner. Despite this, it's evident that both adenine and thymine remain the most prevalent nucleotides across all genome sequences, highlighting their fundamental role in the genetic makeup. The variable abundance of other nucleotides underscores the unique genomic signatures of each virus.

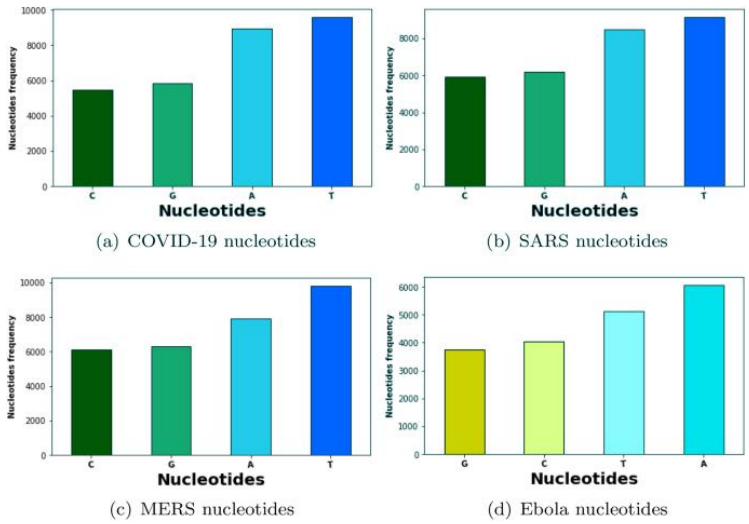


Figure 8. Nucleotide frequency in viral DNA sequences

Figure 9 illustrates the GC content of the aforementioned viruses, providing a key insight into the stability of their DNA structures. GC base pairs, characterized by three hydrogen bonds, exhibit a stronger bonding than AT base pairs, which have two hydrogen bonds. Consequently, double-stranded DNA with a higher prevalence of GC base pairs tends to be more robust and stable and boasts a higher melting temperature. The thermal stability of DNA is a crucial factor influencing various molecular processes. As observed in the table, all the viruses exhibit higher thermal stability compared to the COVID-19 virus. This implies that the DNA of these viruses will have a higher melting temperature, indicating increased resistance to separation into single strands. The lower thermal stability of the COVID-19 virus suggests that its DNA will require a comparatively lower annealing temperature for processes such as PCR, providing valuable insights for molecular biology and diagnostics.

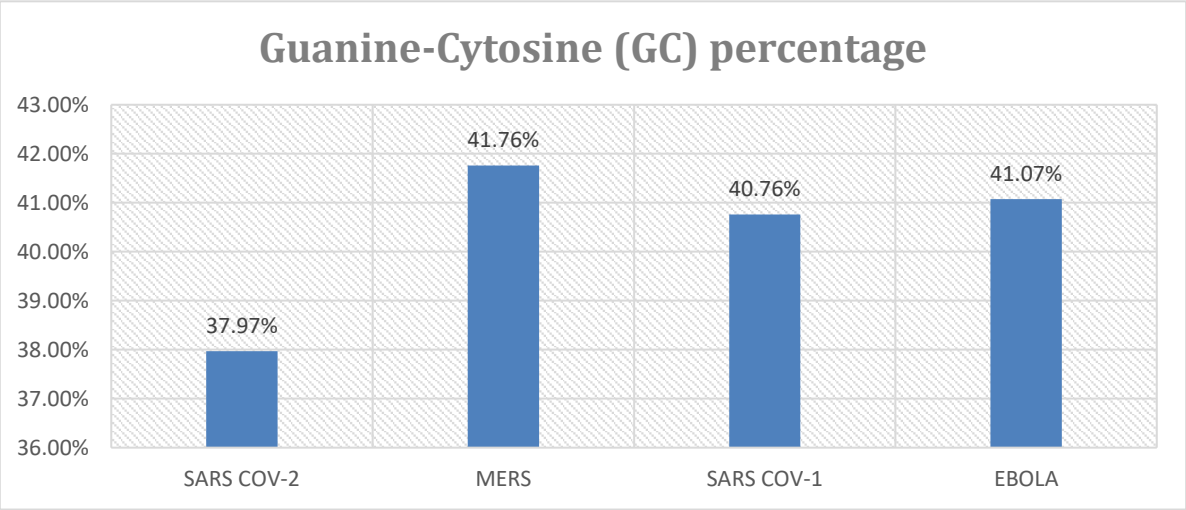
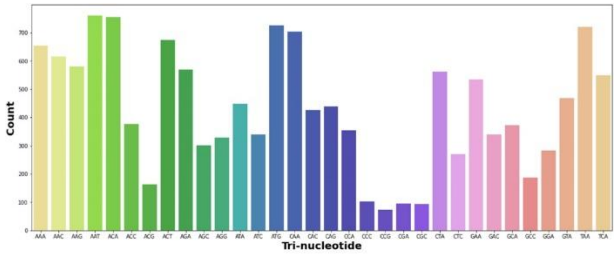


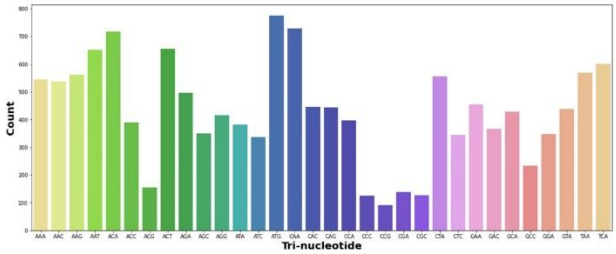
Figure 9. GC content associated with each virus

Figure 10 shows genomic basics and their variation between species. This difference greatly affects DNA sequences' tri-nucleotide frequency, which is a particular arrangement of three different nucleotides called a codon. An analysis of the COVID-19 genome's tri-nucleotide frequency shows the AAT codon's frequency. Due to selection pressure, the virus holds onto codons that improve its ability to reproduce and transcription in human hosts.

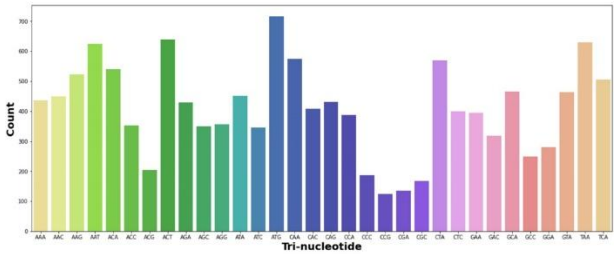
In contrast, SARS and MERS viruses favor the trinucleotide ATG. Translation begins with this trinucleotide, a critical protein synthesis step. Ebola's CAA trinucleotide synthesizes glutamine. The infectious agent may have evolved to improve cell communication, according to the research. A complex analysis of trinucleotide frequencies reveals viruses' strategies for thriving in their host species.



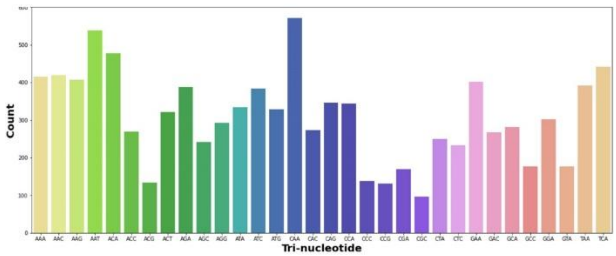
(a) COVID-19 Tri-Nucleotides



(b) SARS Tri-Nucleotides



(c) MERS Tri-Nucleotides



(d) Ebola Tri-Nucleotides

Figure 10 . DNA's tri-nucleotide frequency

Basic amino acid proteins are important to infections. Figure 11 comparison table shows how viral proteins differ. Though they share 20 essential amino acids, virus genomes differ in their peptide amino acid makeup.

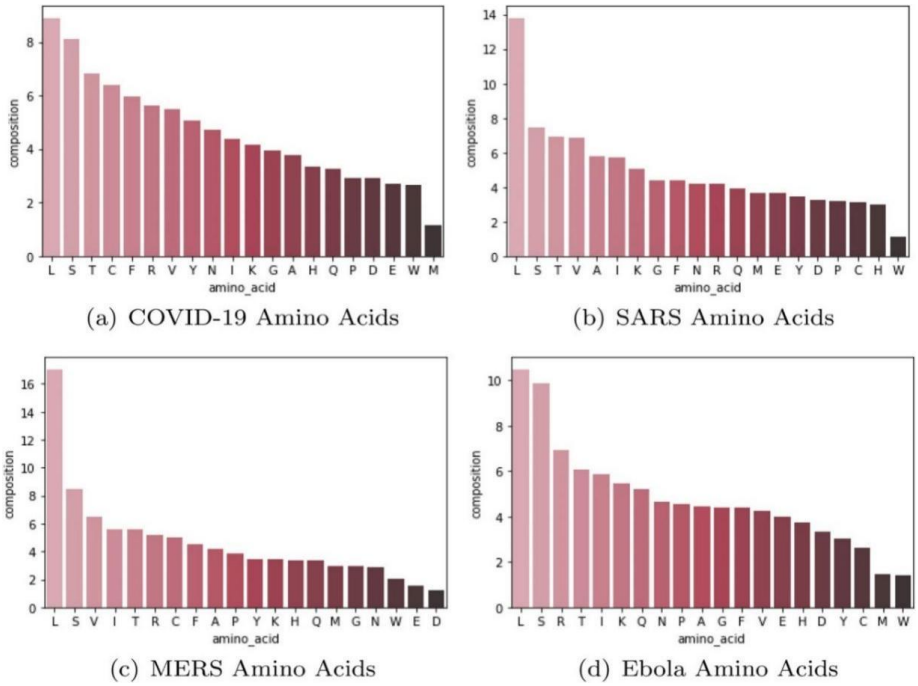


Figure 11. Composition of amino acids

In particular, serine (S) and leucine (L), which are essential amino acids for the infection process when the spike protein interacts with host cells, are highly expressed in the spike proteins of SARS, COVID-19 and MERS. These viruses may have some similar amino acids, but their proteins show different combinations of different amino acids. As opposed to coronaviruses, Ebola viruses have a unique protein structure, which makes periodic revisions to Ebola immunizations necessary because various Ebola virus strains have different amino acid changes. The leucine and valine concentration in the proteins of the Ebola virus is a result of its distinct features, which include fast mutation and immune system evasion. The complex interactions between amino acids highlight the various methods viruses use to tailor their protein structures for both infection and survival.

The number of protein strands connected to the genomes of various viruses is depicted in Figure 12. The numbers that are acquired represent the number of unique protein strands that are encoded by the genome of each virus. The variations in these figures correspond to the distinct genetic features and intricacies of every virus.

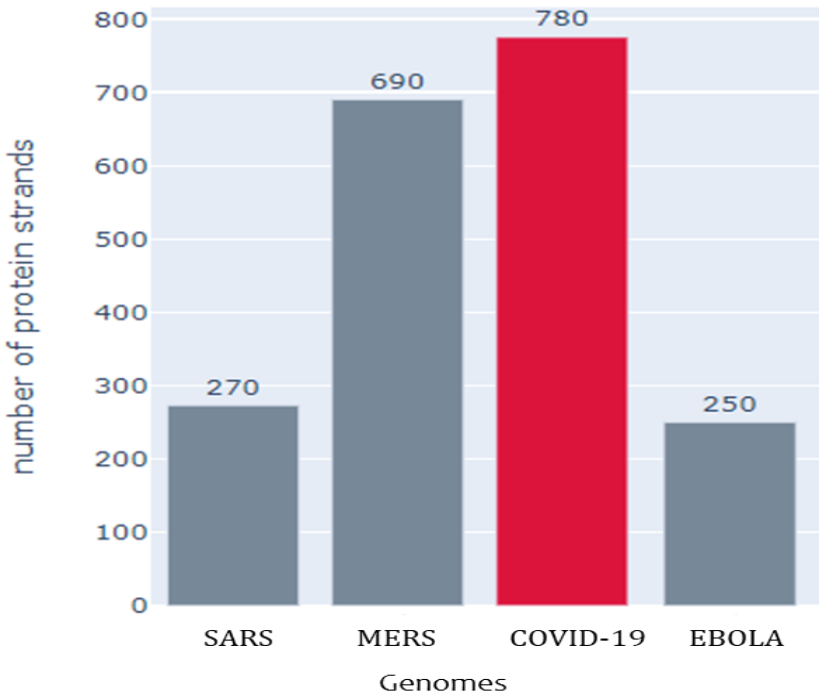


Figure 12. Number of protein strands

The COVID-19 has the greatest count (780), suggesting a relatively complex genetic makeup. MERS is closely followed by 690 protein strands, indicating a significant coding ability. While Ebola has a unique profile with 205 protein strands, SARS has a lower count of 270. Understanding the possible diversity and usefulness of viral proteins—which influence aspects like virulence, transmission and host interactions—requires an understanding of the number of protein strands.

Figure 13 provides an explanation of a sequence align methodology, a crucial method for analyzing protein, RNA and DNA structures. Sequence alignment refers to the process of organizing such sequences by trying to identify comparable parts that may arise from fundamental, functioning, or genetic connections between the nucleotides. This method is valuable for determining sequence homology, clarifying evolutionary connections and constructing homology modeling involving protein architectures. The fundamental assumption of alignment of genes is the belief that variants that have an ancestral relationship will undergo some level of evolutionary shift as time goes on. By employing sequence symmetry, investigators may detect conserved regions, understand the functional significance of certain elements and investigate the molecular progression of creatures that exist. Aligned sequencing is an invaluable technique that has significantly advanced the areas of genetics as well as bioinformatics. It aids in understanding genetic connections and establishes the basis for elucidating the structural as well as functional attributes of macromolecules.

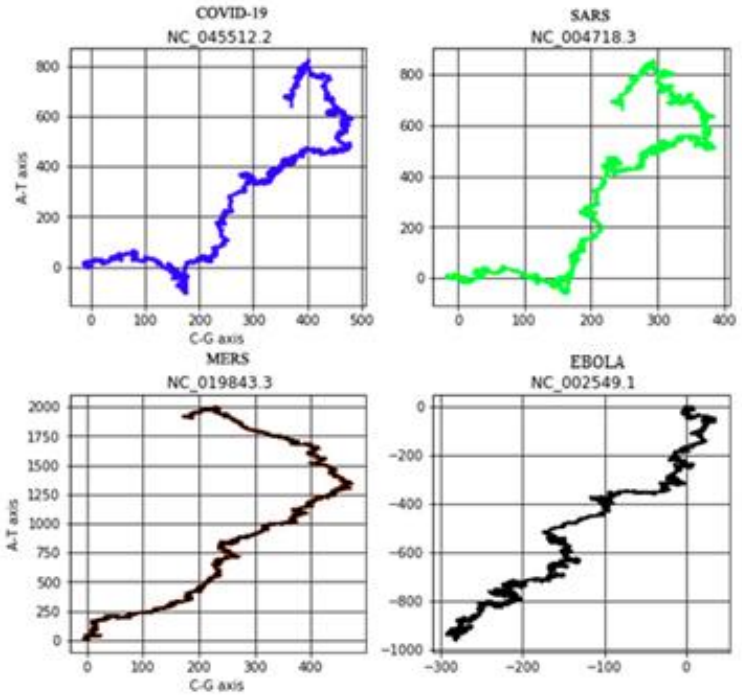


Figure 13. Sequence Alignment

By analyzing the genomic sequence of MERS, COVID 19, SARS, and Ebola allow visualizing clear indicators of common genetic organization, as well as pathogen specific patterns, as in Table 6. Such comparative genomic information plays a pivotal part in terms of comprehending evolutionary relationships, transmission mechanisms, and virulence factors in cases of new outbreak pathogens. All the above results indicate the effectiveness and reliability of the GRU model in correct categorization of viral genes.

Table 6. Comparison of the proposed methodology performance with others

Sl.No:	Infection	Dataset	Methodology	Accuracy(%)	References
1	Ebola	GRU	GRU	98.04%	[6]
2	Ebola	GenBank	XML	98%	[7]
3	SARS	BLAST	CNN	99%	[8]
4	SARS-s	BLAST	CNN	99%	[9]
5	COVID-19	Genebank	XGBoost	86%	[10]
6	COVID-19	NCBI	MD and SVM	75%	[11]

7	MERS	NCBI-GEO	PERL	69.6%	[12]
8	MERS	FASTA	CNN	100%	[13]
9	COVID-19	GenBank	GRU	99.01 %	Proposed
10	MERS	GenBank	GRU	98.35	Proposed
11	SARS	GenBank	GRU	98.91	Proposed
12	EBOLA	GenBank	GRU	98.04	Proposed

The use of GRU networks in phylogenetic analysis allows fine examination of the molecular properties of viruses, as well as their evolutionary dynamics, and especially the discovery of mutations that influence virulence, host tropism, and transmission efficacy. This has greatly led to more knowledge in SARS and MERS epidemics that led to development of better diagnostic tests and treatment methods that can curb worldwide outbreaks. Analysis of the genomes of the Ebola virus using GRU networks has been especially useful to the researchers to track the transmission routes between the animal reservoirs and humans in the localities and discover the hotspots in the region that have an active existence of the virus. In addition, predictive modeling of GRU has made it possible to forecast the epidemic earlier on thus reducing the cases of disease outbreak and death through enhanced containment strategies. Viruses are very successful and as such, inflict a lot of harm on human lives and our understanding on viral evolution, its transmission and interaction with the host has been revolutionized by the introduction of GRU modeling in viral genomics. The implementation capabilities of the GRU networks on temporal patterns of large-scale genomic data can incorporate actionable knowledge that can have immediate implications to public health actions and interventions in disease control processes. With the ever-evolving threats posed by new viral pathogens, there is an urgent need to have a synergistic approach to the response and preparations to outbreaks on the global scale through GRU modeling and experimental virology.

4 . CONCLUSION

Throughout history, humanity has faced various novel and deadly diseases, prompting the essential use of genome sequencing for their detection and effective management. In this study, an exhaustive analysis of the genomes of SARS, COVID-19, Ebola and MERS was conducted. This study utilized advanced techniques including sequence alignment, dot plot analysis and the GRU methodology to deepen understanding of the structure, connections and potential functions of these viruses. These methodologies successfully distinguish between viral genomes with high levels

of recall, accuracy, F1-score and precision in classification, with the COVID-19 classification algorithm achieving an impressive accuracy rate of 99.01%. The distinct characteristics identified within the viral genomes through this research have profound implications for understanding of viral pathogenesis, transmission dynamics and potential targets for therapeutic intervention. By elucidating the genomic signatures associated with virulence, transmission and drug resistance, this study paves the way for the development of customized medications, reliable diagnostic tools and more efficient therapeutic approaches. Such advancements hold the promise of mitigating the impact of emerging infectious illnesses by facilitating timely and targeted interventions. Furthermore, the findings of this research underscore the pivotal role of genome sequencing in comprehending, predicting and mitigating the impact of infectious diseases, particularly in the context of novel and emerging threats. By continuously monitoring and analyzing viral genomes, researchers can anticipate evolutionary changes, track the emergence of new variants and inform public health strategies accordingly. Harnessing the power of advanced computational techniques and interdisciplinary collaborations enables the ongoing advancement of understanding in viral genomics. These insights can be translated into tangible solutions for the prevention and control of infectious diseases on a global scale.

The future scope of this work extends towards several promising directions. Firstly, ongoing research can delve deeper into the molecular mechanisms underlying viral pathogenesis and host interactions, leveraging the insights gained from genome analysis. This could lead to the identification of novel therapeutic targets and the development of more effective treatments for viral infections. Furthermore, the application of machine learning and computational modeling techniques to viral genomics holds great potential. Continued refinement of classification algorithms and predictive models could enhance this study ability to rapidly identify and characterize emerging viral strains, facilitating proactive measures to prevent outbreaks and mitigate their impact. Additionally, the integration of genomic data with other omics datasets, like transcriptomics and proteomics, could provide a more comprehensive understanding of viral infection dynamics and host responses. This multidimensional approach could unveil new biomarkers for disease diagnosis and prognosis, as well as inform the development of personalized treatment strategies tailored to individual patients.

AKNOWLEDGEMENT

The authors gratefully acknowledge the support of the Department of Computer Applications, Noorul Islam Centre for Higher Education, Tamil Nadu, for this research

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

R. Yin, Z. Luo, and C. K. Kwoh, "Exploring the lethality of human-adapted coronavirus through alignment-free machine learning approaches using genomic sequences," Current Genomics, vol. 22, no. 8, pp. 583–595, 2021.

G. Andrade, G. Matias, L. Chrisóstomo, J. da Costa-Neto, J. Sampaio, A. Silva, and I. Cansanção, "Genetic analysis and epitope prediction of SARS-CoV-2 genome in Bahia, Brazil: An in silico analysis of first and second wave genomics diversity," COVID, vol. 3, no. 5, pp. 655–663, 2023.

D. de Assis Paiva, K. S. C. Yotoko, C. A. Zarzar, J. T. de Paiva, D. P. Siqueira, G. Rodrigues, and T. Sáfadi, "Advancing viral genome classification: Assessing the efficiency and accuracy of the alignment-free k-mer method in emerging pandemics," Caderno Pedagógico, vol. 22, no. 8, pp. e17343–e17343, 2025.

J. J. Park and S. Chen, "Metaviromic identification of discriminative genomic features in SARS-CoV-2 using machine learning," Patterns, vol. 3, no. 2, 2022.

R. L. Gonçalves, M. P. et al., "Integrative transcriptome analysis of SARS-CoV-2 human-infected cells combined with deep learning algorithms identifies two potential cellular targets for the treatment of coronavirus disease," Brazilian Journal of Microbiology, vol. 54, no. 1, pp. 53–68, 2023.

X. Su, Z. You, L. Wang, L. Hu, L. Wong, B. Ji, and B. Zhao, "SANE: A sequence combined attentive network embedding model for COVID-19 drug repositioning," Applied Soft Computing, vol. 111, pp. 107831, 2021.

J. Ivanov, D. Polshakov, J. Kato-Weinstein, Q. Zhou, Y. Li, R. Granet, and C. Aultman, "Quantitative structure–activity relationship machine learning models and their applications for identifying viral 3CLpro- and RdRp-targeting compounds as potential therapeutics for COVID-19 and related viral infections," ACS Omega, vol. 5, no. 42, pp. 27344–27358, 2020.

K. S. Azevedo, A. L. et al., "Deepvirusclassifier: A deep learning tool for classifying SARS-CoV-2 based on viral subtypes within the Coronaviridae family," BMC Bioinformatics, vol. 25, no. 1, pp. 231, 2024.

G. S. Dlamini, D. et al., "Classification of COVID-19 and other pathogenic sequences: A dinucleotide frequency and machine learning approach," IEEE Access, vol. 8, pp. 195263–195273, 2020.

Q. Tang, Y. et al., "Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition," Scientific Reports, vol. 5, no. 1, pp. 17155, 2015.

A. M. Alsamman and H. Zayed, "The transcriptomic profiling of SARS-CoV-2 compared to SARS, MERS, EBOV, and H1N1," PLoS ONE, vol. 15, no. 12, pp. e0243270, 2020.

R. Ahsan, K. et al., "Image processing unravels the evolutionary pattern of SARS-CoV-2 against SARS and MERS through position-based pattern recognition," Computers in Biology and Medicine, vol. 134, pp. 104471, 2021.

I. Y. Kawashima, M. et al., "SARS-CoV-2 host prediction based on virus-host genetic features," Scientific Reports, vol. 12, no. 1, pp. 4576, 2022.

G. B. Câmara, P. et al., "Convolutional neural network applied to SARS-CoV-2 sequence classification," Sensors, vol. 22, no. 15, pp. 5730, 2022.

V. B. Franceschi, L. et al., "Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil," BMC Genomics, vol. 22, no. 1, pp. 371, 2021.

D. G. Teixeira, C. et al., "Understanding SARS-CoV-2 spike glycoprotein clusters and their impact on immunity of the population from Rio Grande do Norte, Brazil," Infection, Genetics and Evolution, vol. 118, pp. 105556, 2024.

S. N. Slavov, L. et al., "Dynamics of SARS-CoV-2 variants of concern in vaccination model city in the State of Sao Paulo, Brazil," Viruses, vol. 14, no. 10, pp. 2148, 2022.

F. Zhao, Z. Yang, N. Wang, K. Jin, and Y. Luo, "Traditional Chinese medicine and Western medicine share similar philosophical approaches to fight COVID-19," Aging and Disease, vol. 12, no. 5, pp. 1162, 2021.

M. R. Straus, M. K. Bidon, T. Tang, J. A. Jaimes, G. R. Whittaker, and S. Daniel, "Inhibitors of L-type calcium channels show therapeutic potential for treating SARS-CoV-2 infections by preventing virus entry and spread," ACS Infectious Diseases, vol. 7, no. 10, pp. 2807–2815, 2021.

J. D. Ramírez, M. Muñoz, C. Hernández, C. Flórez, S. Gomez, A. Rico, and A. E. Paniz-Mondolfi, "Genetic diversity among SARS-CoV-2 strains in South America may impact performance of molecular detection," Pathogens, vol. 9, no. 7, pp. 580, 2020.