Heliyon 10 (2024) e32061

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Deep learning guided prediction modeling of dengue virus evolving serotype

Zilwa Mumtaz [a], Zubia Rashid [b], Rashid Saif [c], Muhammad Zubair Yousaf [a,*]

[a] KAM School of Life Sciences, Forman Christian College University, Ferozpur Road, Lahore, Pakistan
[b] Department of Biomedical Engineering, Faculty of Engineering, Science, Technology and Management, Ziauddin University, Karachi, Pakistan
[c] Department of Biotechnology, Qarshi University, Lahore, Pakistan

ARTICLE INFO

Keywords:
Virus forecasting
DL modeling
Virus classification
Dengue evolution
Genome sequence

ABSTRACT

Evolution remains an incessant process in viruses, allowing them to elude the host immune response and induce severe diseases, impacting the diagnostic and vaccine effectiveness. Emerging and re-emerging diseases are among the significant public health concerns globally. The revival of dengue is mainly due to the potential for naturally arising mutations to induce genotypic alterations in serotypes. These transformations could lead to future outbreaks, underscoring the significance of studying DENV evolution in endemic regions. Predicting the emerging Dengue Virus (DENV) genome is crucial as the virus disrupts host cells, leading to fatal outcomes. Deep learning has been applied to predict dengue fever cases; there has been relatively less emphasis on its significance in forecasting emerging DENV serotypes. While Recurrent Neural Networks (RNN) were initially designed for modeling temporal sequences, our proposed DL-DVE generative and classification model, trained on complete genome data of DENV, transcends traditional approaches by learning semantic relationships between nucleotides in a continuous vector space instead of representing the contextual meaning of nucleotide characters. Leveraging 2000 publicly available DENV complete genome sequences, our Long Short-Term Memory (LSTM) based generative and Feedforward Neural Network (FNN) based classification DL-DVE model showcases proficiency in learning intricate patterns and generating sequences for emerging serotype of DENV. The generated sequences were analyzed along with available DENV serotype sequences to find conserved motifs in the genome through MEME Suite (version 5.5.5). The generative model showed an accuracy of 93 %, and the classification model provided insight into the specific serotype label, corroborated by BLAST search verification. Evaluation metrics such as ROC-AUC value 0.818, accuracy, precision, recall and F1 score, all to be around 99.00 %, demonstrating the classification model's reliability. Our model classified the generated sequences as DENV-4, exhibiting 65.99 % similarity to DENV-4 and around 63–65 % similarity with other serotypes, indicating notable distinction from other serotypes. Moreover, the intra-serotype divergence of sequences with a minimum of 90 % similarity underscored their uniqueness.

* Corresponding author.
E-mail addresses: mumtazzilwa@gmail.com (Z. Mumtaz), zubia.rashid2@gmail.com (Z. Rashid), rashid.saif37@gmail.com (R. Saif), mzubairyousaf@fccollege.edu.pk (M.Z. Yousaf).

https://doi.org/10.1016/j.heliyon.2024.e32061
Received 1 March 2024; Received in revised form 28 May 2024; Accepted 28 May 2024
Available online 29 May 2024

Abbreviations

DENV	Dengue Virus
DL:	Deep Learning
ANNs	Artificial Neural Networks
SVMs	Support Vector Machines
RNNs	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
FNN	Feedforward Neural Network
ConV1d	1 Dimensional Convolutional Neural network
GRU	Gated Recurrent Unit
ReLU	Rectified Linear Unit
CGR	Chaos game representation
FCGR	particularly Frequency CGR

1. Introduction

Understanding the evolutionary dynamics of a virus is crucial for discerning its origin, focusing on key characteristics such as structure, classification and evolution. This knowledge plays a pivotal role in unraveling the fundamental biological mechanisms, thereby advancing vaccine and drug development. Despite the ongoing discoveries related to viruses, the potential existence of unidentified viruses remains a constant concern. The rapid and widespread dissemination of viruses has become robust, presenting formidable challenges in controlling and predicting their expansion. Thus, leads to epidemics and pandemics, underscoring the unpredictable risks associated with these agents [1].

Dengue infection is transmitted through the bite of an Aedes mosquito carrying the ∼10 kb genome-size dengue virus (DENV). Clinically, identifying dengue fever has historically been challenging due to the prevalence of other infections with similar syndromes in tropical environments. The ambiguity in distinguishing dengue from various viral diseases, ranging from yellow fever to tropical influenza, has persisted. Additionally, labeling the historical dengue outbreaks as chikungunya, particularly in the 1800s, is complicated due to the inconsistent and conflicting reporting information. Despite these challenges, the global public health system has remained engaged in addressing the persistent threats posed by the Dengue Virus [2].

In the landscape of genomic analysis, Deep Learning (DL) has emerged as a powerful tool, particularly for extracting features and patterns from complex genomic data. Similarly, in the context of infectious diseases like dengue, machine learning applications have gained prominence, leveraging epidemiological and clinical data for predictive modeling. Rachata et al. notably utilized weather data and feature selection algorithms to forecast dengue incidences, employing Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) to predict a number of cases based on weather and gene expression data [3,4]. Despite the success of SVMs using word2vec representation in specific tasks, their efficiency waned when directly applied to nucleic acid sequences. There are multiple methods for viral genome classification, employing both alignment and machine learning approaches. In the alignment-based approach, detection of viral sequences is carried out using tools such as USEARCH, SCUEAL [5] and REGA [6], relying on alignment scores for genome classification. However, these alignment-based methods have limitations, notably their performance dependence on the selection of initial alignments. However, in the machine learning approach, various methods have been proposed for the classification of viral genome sequences. However, these methods face limitations in their ability to detect viral genome contigs and the challenge of extracting useful hidden information. Moreover, those trained exclusively on nucleotide sequence data further constrain their utility in comprehensive genome analysis.

The deep learning models such as Recurrent Neural Networks (RNNs) have demonstrated their effectiveness in the field of natural language processing. The applications of deep learning in computational biology mainly concentrate on genome analysis and sequencing. However, the RNNs are considered black boxes due to the complexity of the model in interpreting the hidden state of the model. Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) have been used to detect viral genome sequences using training of pattern and frequency of branches. Whereas, the Feedforward Neural Network (FNN), a type of CNN model, is adept at evaluating visual patterns while accommodating the inherent heterogeneity within the data. This network is designed to map fixed-length inputs to a fixed-size output and is trained using a backpropagation algorithm. Consisting of multiple layers, the CNN effectively stores and updates information in filter weights as it learns the intricate relationship between input and output. This study aims to predict the genomic sequences of emerging serotypes to deepen our understanding of Dengue Virus evolution. It also classifies predicted and unknown Dengue Virus genome sequences according to their serotypes. Utilizing RNNs trained on the complete genome, the model focuses on learning patterns in tokens rather than representing character meanings. LSTM, chosen for genomic prediction tasks, excels in handling sequential data and mitigating the vanishing gradient problem inherent in traditional RNNs.

2. Materials and methods

2.1. Data collection and preprocessing

A dataset comprising 2000 complete genome sequences of Dengue Virus were assembled from the NCBI [7] and BV-BRC databases [8]. The datasets incorporated 500 sequences for each of the four serotypes of the Dengue Virus. Before further analysis, a preprocessing step was performed to ensure that the data were in a suitable format. Subsequently, four distinct datasets representing serotypes were built, each containing sequences spanning 10,273 base pairs. All the sequences were concatenated into a single dataset following a labeling process. All nucleotides were selected as features, and DNV1, DNV2, DNV3 and DNV4 as serotype labels.

2.2. DL-DVE architecture and working for sequence generation

The genomic sequences of DENV were extracted from FASTA files using the Biopython library. A tokenization approach was used, treating the sequences at the character level, and an n-gram strategy was employed to generate the input sequences for the model from a FASTA file. To ensure uniformity in the sequence length, the data underwent padding. The generative model, implemented as a sequential model in Tensorflow Keras, comprised an embedding layer [9], two LSTM layers for capturing sequential patterns, and a dense layer for output. The model was compiled using sparse categorical cross-entropy loss and the Adam optimizer. The sequential model is designed for sequence processing with a specific focus on capturing patterns in the sequences related to the classification task. For training, the model utilized prepared predictors and labels over 15 epochs to efficiently capture underlying patterns in the data. Additionally, a function generated a new sequence based on a given seed text and the trained generative model.

$$ft = \sigma(Wf \cdot [ht - 1, xt] + bf) \qquad [10]$$

The sequence generation process involved predicting the next word with a controlled level of randomness to introduce diversity while preserving the patterns in the entire data [10], enhancing the model's generative capabilities. Fig. 1 shows a detailed architecture of the DL-DVE model for sequence classification and generation.

2.2.1. Comparison of generative models

We used 1 1-dimensional convolutional Neural network (ConV1d), a Gated Recurrent Unit (GRU), and a Simple Recurrent Neural Network (RNN) to compare the efficiency of our DL-DVE model [11]. All of them were trained using sparse categorical cross-entropy loss and an Adam optimizer, making them suitable for sequence generation tasks. The performance comparison of models is shown in Table 1.

2.2.2. ConV1d

We used ConV1d, a CNN-based model implemented with an embedding layer, to represent words in a continuous vector space. A

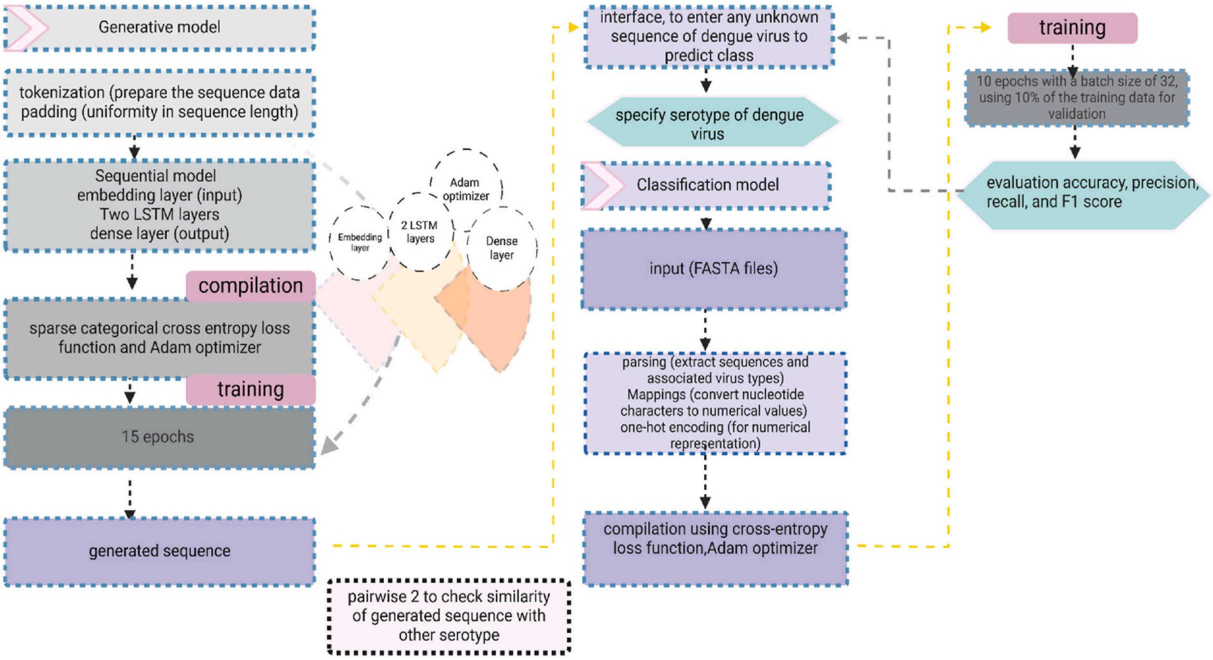
Generative model
tokenization (prepare the sequence data padding (uniformity in sequence length)
Sequential model embedding layer (input) Two LSTM layers dense layer (output)
compilation
sparse categorical cross entropy loss function and Adam optimizer
training
15 epochs
generated sequence
Adam optimizer
Embedding layer
2 LSTM layers
Dense layer
pairwise 2 to check similarity of generated sequence with other serotype
interface, to enter any unknown sequence of dengue virus to predict class
specify serotype of dengue virus
Classification model
input (FASTA files)
parsing (extract sequences and associated virus types) Mappings (convert nucleotide characters to numerical values) one-hot encoding (for numerical representation)
compilation using cross-entropy loss function,Adam optimizer
training
10 epochs with a batch size of 32, using 10% of the training data for validation
evaluation accuracy, precision, recall, and F1 score

Fig. 1. Workflow and architecture of DL-DVE for Dengue Virus genome classification and prediction.

Table 1
Performance comparison of different models for emerging sequence generation.

Sr. no.	Model	Accuracy (%)	Epoch	Sequence generated (length)
1	ConV1d	29	10	<10 kb
2	GRU	35	10	<10 kb
3	Simple RNN	30	10	<10 kb
4	Our proposed LSTM model	37	10	>10 kb
		93	15	>10 kb

single Conv1D layer with 128 filters and a kernel size of 5 captured local patterns. GlobalMaxPooling1D reduced the dimensionality, and a dense layer with a softmax activation generated a probability distribution over the entire network.

2.2.3. GRU

The model consisted of an embedding layer for word representation, followed by two GRU layers with 100 units, capturing the sequential patterns. The final dense layer outputs a probability distribution.

2.2.4. Simple RNN

A simple RNN model was implemented using an embedding layer converting nucleotides into 50-dimensional vectors, followed by two simple RNN layers with 100 units each. The first layer returned sequences, capturing the temporal patterns, while the second provided a condensed representation. The final dense layer outputs probabilities, making it suitable for our sequence generation task.

We further analyzed the generated sequences using BLAST search and Pairwise2 algorithm to assess the most similar type of sequence and similarity scores with other serotypes using Biopython. Furthermore, we used MEME Suite (version 5.5.5) to enrich our understanding of the conserved patterns and motifs in the sequences [12].

2.3. DL-DVE architecture and working for sequence classification

The Biopython library was used to extract genomic sequences of DENV from FASTA files. Sequence parsing facilitated the extraction of sequences and associated virus types. The unique nucleotide characters within the sequences were identified to assess dataset diversity. To enable the deep learning model utilization, nucleotide characters were mapped to numerical values and vice versa. One-hot encoding transformed the sequences into numerical representations [13]. Dengue Virus types underwent conversion to numerical labels using Scikit-learn's LabelEncoder for multi-class classification modeling [14]. The dataset was divided into training and testing sets, with 80 % of the data allocated for training and 20 % for testing, ensuring the model's ability to generalize to the unseen data.

An FNN model was deployed using the Tensorflow Keras library. The model consists of a flattening layer, a dense hidden layer employing Rectified Linear Unit (ReLU) activation, and an output layer utilizing softmax activation for effective multi-class classification. The pivotal role of the non-linear activation function is highlighted post-convolution, particularly in comprehending CNN dynamics. Among the commonly employed activation functions, namely ReLU, Sigmoid, and Tanh, ReLU demonstrates accelerated learning. The outer layer employed Softmax activation, enabling the assessment of class probabilities in prediction scenarios. The FNN integrated multiple filters traversing a one-hot encoded binary vector representing the sequence.

The FNN model underwent compilation utilizing the categorical cross-entropy loss function and the Adam optimizer. During training, the Adam optimizer dynamically adjusted weights and biases, while the sparse categorical cross-entropy loss function quantified the dissimilarity between predicted probabilities and proper labels. The training spanned ten epochs with a batch size of 32, incorporating 10 % of the training data for validation purposes. Evaluation of the testing set gauged the FNN's accuracy in predicting virus type, with predictions made on a subset of testing data compared against actual virus types to assess performance. To comprehensively evaluate the model's effectiveness, various classification metrics, including accuracy, precision, recall, and F1 score, were computed. A user-friendly interface was developed to facilitate the input of new viral sequences. The provided sequence underwent preprocessing and was fed to the trained model to predict the associated virus type. The workflow and architecture of the classification and generative model are shown in Fig. 1.

3. Results and discussion

With the emergence of next generation sequencing technology, it has become possible to make predictions using comprehensive data of genomic sequences [15]. This capability has been shown in the classification of COVID-19 variants and other viruses through deep-learning approaches [11,16,17]. Additionally, CNN and LSTM models have been increasingly used for predicting dengue cases [3,18–24]. The CNN and LSTM models underscore the importance of integrated neural networks with genomic sequences as an innovative method capable of revolutionizing virus studies [25–27].

Machine learning models go beyond human reasoning and build prediction models from several complex combinations. The DL models such as LSTM, GRU and CNN have been used for sequence classification and generative tasks [28]. RNNs, which are CNN-based models, are effective in natural language processing and genome analysis but can be challenging to interpret due to their complexity. LSTMs and CNNs excel in genome sequence detection by learning patterns, while FNNs are adept at recognizing patterns and handling data heterogeneity. They map fixed-length inputs to fixed–size outputs and are trained using backpropagation. The information is stored and updated effectively in filter weights, enabling learning of complex input and output relationships. Studies have been performed to utilize these models for the detection of HCV variants from complete genome sequence data that showed resistance to direct-acting antivirals. The identified variants were incorporated into machine learning algorithms for assessment of the effectiveness of the predictive model [29]. The "Long Short-Term Memory" model is considered effective in capturing complex patterns in data and multiple features to make accurate predictions [30]. Being a subtype of RNNs, these models possess an enhanced capability to learn information from distant points in time. Traditional RNNs encounter the vanishing gradient problem, impeding their ability to capture changes that occurred in data long ago. LSTMs overcome this challenge through a gating mechanism where the gates open and close based on values learned from each input. This mechanism enables LSTMs to accumulate information over an extended period by dynamically learning to forget certain aspects of information. Complete genome sequence data of DENV from the four existing serotypes was employed in our LSTM model with the aim of generating sequences that exhibit the probability of emerging as a new serotype.

Initially, we employed a range of models including ConV1d, GRU, Simple RNN and our proposed LSTM model, aiming to ascertain the most effective approach. Notably, at ten epochs, our LSTM model demonstrated 37 % accuracy, outperforming ConV1d, GRU and Simple RNN. Moreover, when trained at ten epochs, our LSTM model generated sequences exceeding 10 kb in size, which distinguished it from the other models that generated sequences of less than 10 kb. This aspect of sequence length carried significant implications for the model's predictive capacity and biological relevance.

Recognizing the potential for further enhancement, we extended the training duration of our LSTM model to 15 epochs. This adjustment yielded a substantial increase in accuracy, reaching 93 %. This improvement underscored the efficacy of prolonged training in refining the model's predictive capabilities, as shown in Table 1.

To further analyze the origin of generated sequences, we employed a multi-class classification approach using the FNN model for genomic sequence classification. In the realm of biological sequence analysis, machine learning and deep learning using CNNs have demonstrated high precision for binary or multi-class classification [31]. Nucleotide sequence-based studies have typically employed one-hot encoding vectors to represent each nucleotide, with all unknown nucleotides represented as all zero vectors. Chaos game representation (CGR), particularly Frequency CGR (FCGR), has shown promise in encoding sequences in image format and has been applied to predict drug resistance [32]. Our study contributes to the landscape of genomic analysis, employing one-hot encoding, a proven method for next-generation sequencing reads and phenotype label abstraction.

The identification of novel genomic regions in viral pathogens using CNNs and LSTMs has emerged as a compelling area of exploration among researchers [1]. In the CNN framework, the initialization of filter weights involves random uniformness, and these weights are subsequently refined through the backpropagation process to minimize the loss or cost function. The iterative learning process allowed the network to adapt and optimize its performance, enhancing its ability to discern meaningful patterns and features within the given data. Our CNN based FFN classification model trained on DENV complete genome data achieved an accuracy of approximately 99.00 %. Across the ten training epochs, the model consistently improved, achieving a final validation accuracy of 98.12 % (Fig. 2).

The model demonstrated reliable results in achieving high-quality predictions across multiple evaluation metrics with overall accuracy, precision, recall, and an F1 score of 99.00 %, as shown in Fig. 3.

The model's performance was further assessed through the ROC-AUC curve. The loss function appeared to decrease significantly and the area under the ROC-AUC on the validation set showed fluctuations, ultimately stabilizing around 0.818. This robust accuracy, coupled with the low training loss, suggests the model's effectiveness in accurately classifying instances, as shown in Fig. 4.

We further investigated the model's performance by inserting unknown sequences of the Dengue Virus genome. The model defined

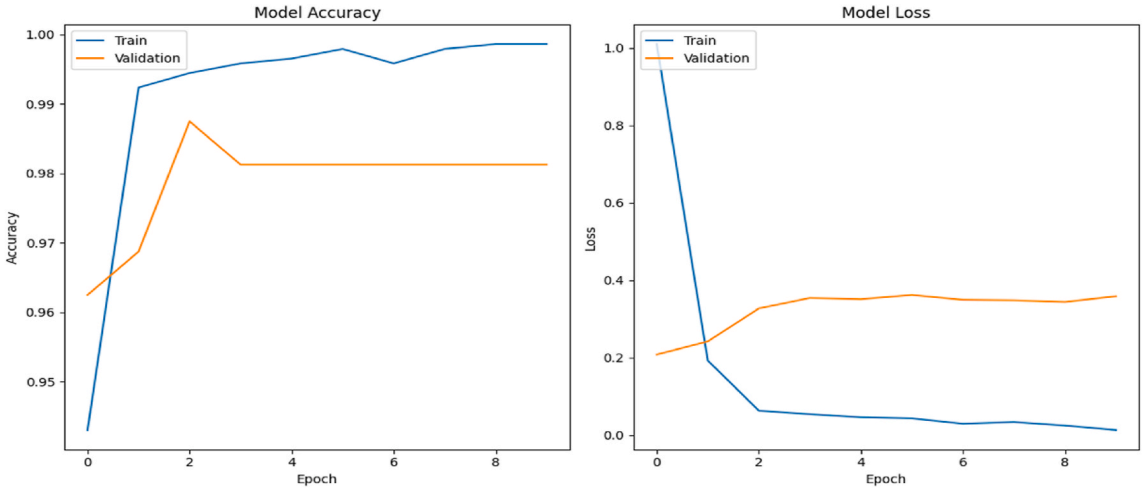

Fig. 2. The validation accuracy of FFNN model over 10 epochs.

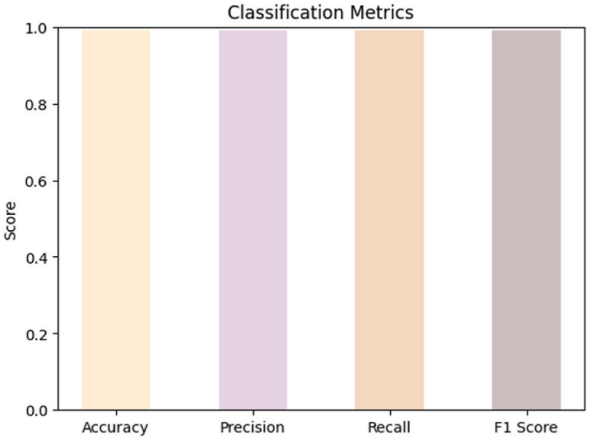

Fig. 3. Classification metrics illustrating accuracy, precision, recall and F1 score.

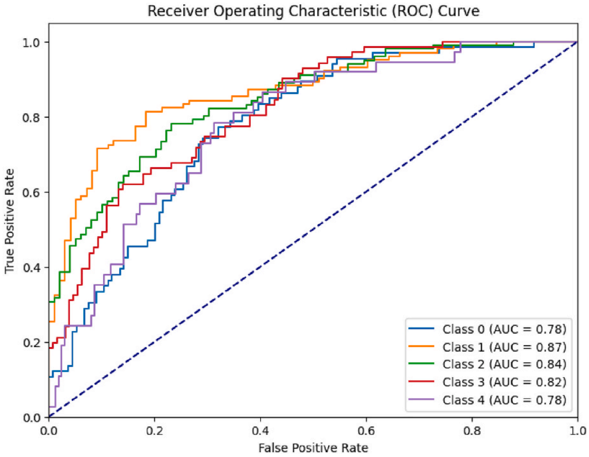

Fig. 4. ROC-AUC for FFNN model.

a serotype for that sequence. We further cross-checked the predicted serotype with the actual serotype confirmed from its source. The FNN model worked efficiently on the unknown DENV sequence data and predicted the actual classes or serotypes of the Dengue Virus (Fig. 5).

Using our DL-DVE generative and classification model, the generated sequences were successfully classified as belonging to the DENV-4 serotype. We assessed the similarity between the sequences generated by our LSTM model and the DENV-4 serotype. Surprisingly, our investigation revealed a similarity of approximately 66 %, suggesting that the generated sequences did not align closely with the DENV-4 serotype. We further checked the similarity of generated sequences with other serotypes. The approximate similarity observed was 63–66 %, suggesting that the generated sequences did not align closely with any known serotype, as shown in Fig. 6. The generated sequences meet the specified sequence length requirement and emphasize the effectiveness of our model in accurately predicting and classifying DENV serotypes. An approximate sequence similarity of 65 % among DENV serotypes has been demonstrated in multiple studies [33–35].

The exploration of motifs with statistical significance was integral in our study of complete genome sequence data encompassing DENV 1 to 4 serotypes, juxtaposed with the generated sequences [36]. The MEME Suite analysis contributed to an enhanced comprehension of conserved patterns shedding light on some motifs within all serotypes and the emerging serotype. Fig. 7 illustrates the conserved motif in DENV 1–3 alongside the generated sequence, replacing DENV-4. This replacement was based on pairwise2 function and BLAST search, showing that the generated sequences closely match DENV-4.

While our study provides valuable insights as it utilizes RNN and LSTM models to accurately classify and generate sequences of DENV that can be valuable in terms of designing potential vaccine candidates, it is imperative to acknowledge certain limitations that accompany such findings. Firstly, the capabilities of our trained model are confined to DENV sequences of similar size, which may restrict the generalizability of our findings to larger or smaller genomes. Moreover, the predictive accuracy may vary when applied to genomes with significant sequence or structural variations. As with any predictive model, the potential for unforeseen mutations could affect the reliability of our predictions. Despite these constraints, our study highlights the potential utility of predictive modeling in

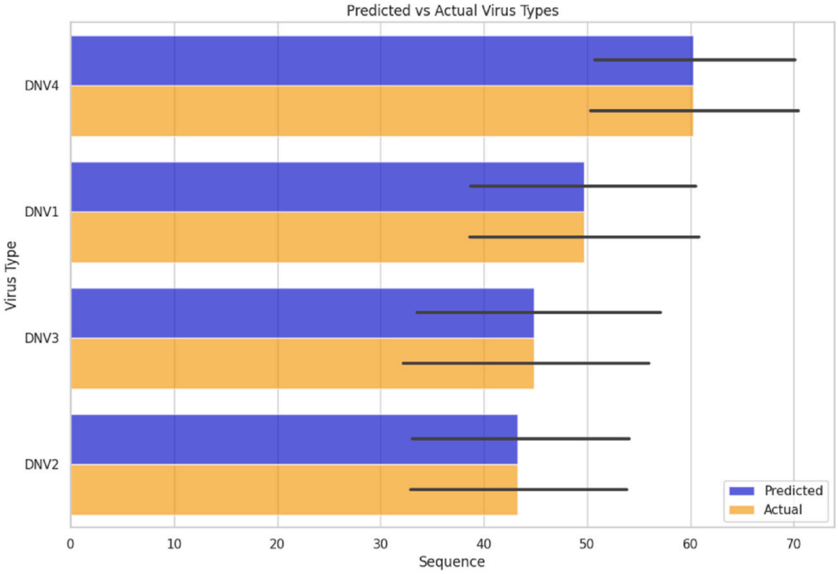

Fig. 5. FFNN model accuracy through predicted vs actual virus types.

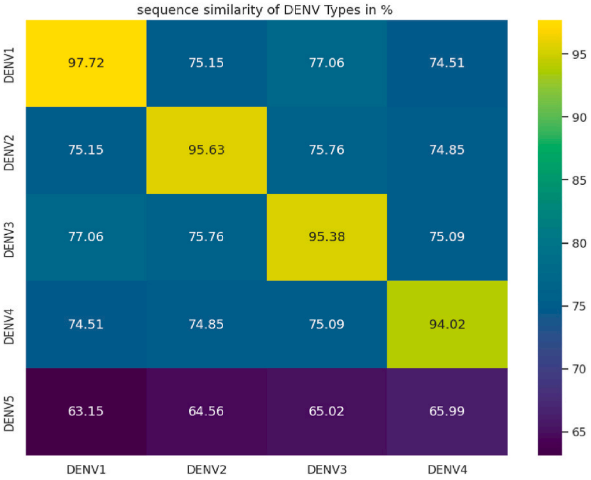

Fig. 6. Sequence similarity of intra and inter DENV serotypes and comparison with generated sequences.

informing vaccine development and diagnostic strategies.

4. Conclusion

Our study demonstrates the effectiveness of utilizing sequential models for classifying and generating DENV genomic sequences resulting in the generation of sequences resembling a potential emerging serotype. Our predictive model classified the generated sequences serotypes, showing a close resemblance to a specific serotype while diverging significantly from existing serotypes. The intra-serotype divergence affirms the distinctiveness of generated sequences within the DENV serotype landscape; further exploration includes the analysis of conserved motifs. Considering the complete genome sequence size of the DENV genome is approximately 10 kb, our trained model is limited to predicting DENV sequences of similar size. Nonetheless, in future directions, predictive modeling applied well in advance of mutations holds promise for informing vaccine design and the development of diagnostic kits.

Data availability statement

All data and code used for running experiments, model fitting, and plotting is available on a GitHub repository at https://github.com/Ziloeuvre/DL-DVE.git.We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.10,988,910.

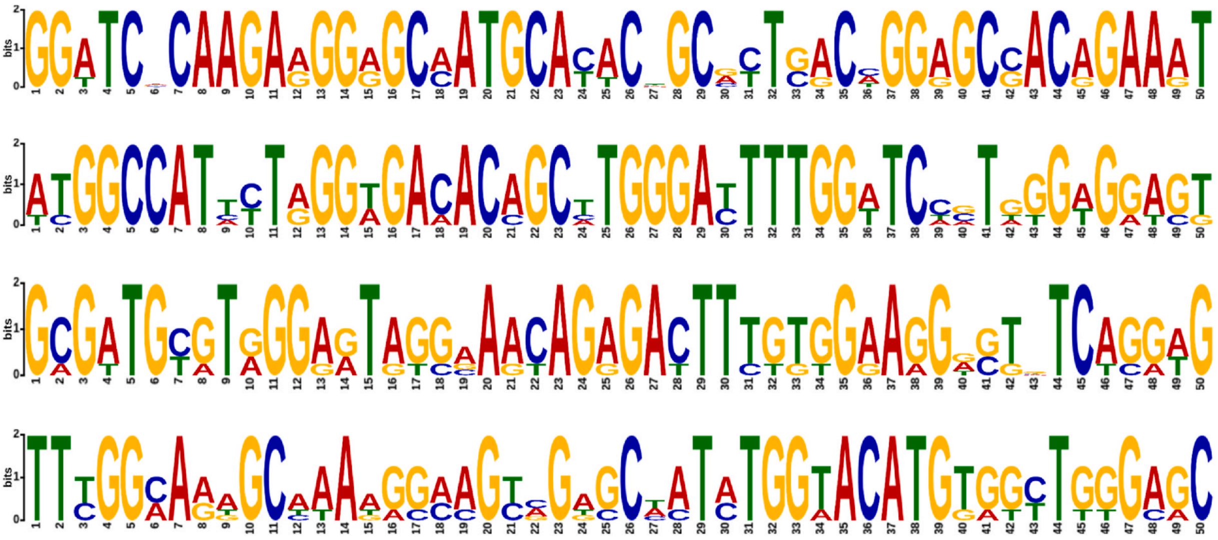

Fig. 7. Sequence Logo Analysis of generated sequence with DENV-4 sequences: A visual representation of conserved motifs in the nucleotide sequences, where the y-axis depicts information content in bits (0, 1, 2), and the x-axis represents nucleotides. E-values highlight the statistical significance of motifs, with the first motif at 1.0e-028, the second at 3.8e-026, the third at 1.8e-019, and the fourth at 3.7e-018. The sequence logo provides insights into the nucleotide composition and conservation within the analyzed motifs.

Funding

This research was funded by ORIC-FCCU with grant number IRIF-17.

CRediT authorship contribution statement

Zilwa Mumtaz: Writing – original draft, Visualization, Validation, Project administration, Methodology, Conceptualization. **Zubia Rashid:** Writing – review & editing, Validation, Resources, Formal analysis. **Rashid Saif:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis. **Muhammad Zubair Yousaf:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.M. Bartoszewicz, A. Seidel, B.Y. Renard, Interpretable detection of novel human viruses from genome sequencing data, NAR genomics and bioinformatics 3 (1) (2021 Mar) lqab004, https://doi.org/10.1093/nargab/lqab004.

[2] G. Kuno, A re-examination of the history of etiologic confusion between dengue and chikungunya, PLoS Neglected Trop. Dis. 9 (11) (2015 Nov 12) e0004101, https://doi.org/10.1371/journal.pntd.0004101.

[3] J.D. Mello-Román, J.C. Mello-Román, S. Gomez-Guerrero, M. García-Torres, Predictive models for the medical diagnosis of dengue: a case study in Paraguay, Comput. Math. Methods Med. (2019 Jul 29;2019), https://doi.org/10.1155/2019/7307803.

[4] Rachata N, Charoenkwan P, Yooyativong T, Chamnongthal K, Lursinsap C, Higuchi K. Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network. In2008 International Symposium on Communications and Information Technologies 2008 Oct 21 (pp. 210-214). IEEE. DOI: 10.1109/ISCIT.2008.4700184.

[5] S.L. Kosakovsky Pond, D. Posada, E. Stawiski, C. Chappey, A.F. Poon, G. Hughes, E. Fearnhill, M.B. Gravenor, A.J. Leigh Brown, S.D. Frost, An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1, PLoS Comput. Biol. 5 (11) (2009 Nov 26) e1000581, https://doi.org/10.1371/journal.pcbi.1000581.

[6] A.C. Pineda-Peña, N.R. Faria, S. Imbrechts, P. Libin, A.B. Abecasis, K. Deforche, A. Gómez-López, R.J. Camacho, T. De Oliveira, A.M. Vandamme, Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools, Infect. Genet. Evol. 19 (2013 Oct 1) 337–348, https://doi.org/10.1016/j.meegid.2013.04.032.

[7] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990 Oct 5) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.

[8] R.D. Olson, R. Assaf, T. Brettin, N. Conrad, C. Cucinell, J.J. Davis, D.M. Dempsey, A. Dickerman, E.M. Dietrich, R.W. Kenyon, M. Kuscuoglu, Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR, Nucleic Acids Res. 51 (D1) (2023 Jan 6) D678–D689, https://doi.org/10.1093/nar/gkac1003.

[9] F. Cui, Z. Zhang, Q. Zou, Sequence representation approaches for sequence-based protein prediction tasks that use deep learning, Briefings in Functional Genomics 20 (1) (2021 Jan) 61–73, https://doi.org/10.1093/bfgp/elaa030.

[10] C.M. Dasari, R. Bhukya, Explainable deep neural networks for novel viral genome prediction, Appl. Intell. 52 (3) (2022 Feb) 3002–3017, https://doi.org/10.1007/s10489-021-02572-3.

[11] S. Ali, B. Sahoo, A. Zelikovsky, P.Y. Chen, M. Patterson, Benchmarking machine learning robustness in COVID-19 genome sequence classification, Sci. Rep. 13 (1) (2023 Mar 13) 4154, https://doi.org/10.1038/s41598-023-31368-3.

[12] Bailey TL, Elkan C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers.

[13] A.C. Choong, N.K. Lee, Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method, in: In2017 International Conference on Computer and Drone Applications (IConDA), IEEE, 2017 Nov 9, pp. 60–65, https://doi.org/10.1109/ICONDA.2017.8270400.

[14] J. Langton, K. Srihasam, J. Jiang, Comparison of machine learning methods for multi-label classification of nursing education and licensure exam questions, InProceedings of the 3rd Clinical Natural Language Processing Workshop (2020 Nov) 85–93, https://doi.org/10.18653/v1/2020.clinicalnlp-1.10.

[15] H. Shim, Futuristic methods in virus genome evolution using the Third-Generation DNA sequencing and artificial neural networks. Global Virology III: Virology in the 21st Century, 2019, pp. 485–513, https://doi.org/10.1007/978-3-030-29022-1_17.

[16] S. Basu, R.H. Campbell, Classifying COVID-19 variants based on genetic sequences using deep learning models. InSystem Dependability and Analytics: Approaching System Dependability from Data, System and Analytics Perspectives, Springer International Publishing, Cham, 2022 Jul 26, pp. 347–360, https://doi.org/10.1007/978-3-031-02063-6_19.

[17] L.C. de Souza, K.S. Azevedo, J.G. de Souza, R.D. Barbosa, M.A. Fernandes, New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning, BMC Bioinf. 24 (1) (2023 Dec) 1–9, https://doi.org/10.1186/s12859-023-05188-1.

[18] S.N. Manoharan, K.M. Kumar, N. Vadivelan, A novel CNN-TLSTM approach for dengue disease identification and prevention using IoT-Fog cloud architecture, Neural Process. Lett. 55 (2) (2023 Apr) 1951–1973, https://doi.org/10.1007/s11063-022-10971-x.

[19] M.A. Majeed, H.Z. Shafri, Z. Zulkafli, A. Wayayok, A deep learning approach for dengue fever prediction in Malaysia using LSTM with spatial attention, Int. J. Environ. Res. Publ. Health 20 (5) (2023 Feb 25) 4130, https://doi.org/10.3390/ijerph20054130.

[20] V.H. Nguyen, T.T. Tuyet-Hanh, J. Mulhall, H.V. Minh, T.Q. Duong, N.V. Chien, N.T. Nhung, V.H. Lan, H.B. Minh, D. Cuong, N.N. Bich, Deep learning models for forecasting dengue fever based on climate data in Vietnam, PLoS Neglected Trop. Dis. 16 (6) (2022 Jun 13) e0010509, https://doi.org/10.1371/journal.pntd.0010509.

[21] W. Nadda, W. Boonchieng, E. Boonchieng, Influenza, dengue and common cold detection using LSTM with fully connected neural network and keywords selection, BioData Min. 15 (1) (2022 Feb 14) 5, https://doi.org/10.1186/s13040-022-00288-9.

[22] A.R. Doni, T. Sasipraba, LSTM-RNN based approach for prediction of dengue cases in India, Ingénierie Des. Systèmes Inf. 25 (3) (2020 Jun 1), https://doi.org/10.18280/isi.250306.

[23] X. Zhao, K. Li, C.K. Ang, K.H. Cheong, A deep learning based hybrid architecture for weekly dengue incidences forecasting, Chaos, Solit. Fractals 168 (2023 Mar 1) 113170, https://doi.org/10.1016/j.chaos.2023.113170.

[24] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, C. Venkatesan, C. Suresh Gnana Dhas, Analysis of DNA sequence classification using CNN and hybrid models, Comput. Math. Methods Med. 15 (2021 Jul), https://doi.org/10.1155/2021/1835056, 2021.

[25] M.A. Helaly, S. Rady, M.M. Aref, Convolutional neural networks for biological sequence taxonomic classification: a comparative study, in: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019, Springer International Publishing, 2020, pp. 523–533, https://doi.org/10.1007/978-3-030-31129-2_48.

[26] C. Ao, S. Jiao, Y. Wang, L. Yu, Q. Zou, Biological sequence classification: a review on data and general methods, Research 19 (2022 Dec) 11, https://doi.org/10.34133/research.0011, 2022.

[27] M. Pérez-Enciso, L.M. Zingaretti, A guide on deep learning for complex trait genomic prediction, Genes 10 (7) (2019 Jul 20) 553, https://doi.org/10.3390/genes10070553.

[28] S.T. Tsai, E.J. Kuo, P. Tiwary, Learning molecular dynamics with simple language model built upon long short-term memory neural network, Nat. Commun. 11 (1) (2020 Oct 9) 5115, https://doi.org/10.1038/s41467-020-18959-8.

[29] H. Haga, H. Sato, A. Koseki, T. Saito, K. Okumoto, K. Hoshikawa, T. Katsumi, K. Mizuno, T. Nishina, Y. Ueno, A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus, PLoS One 15 (11) (2020 Nov 5) e0242028, https://doi.org/10.1371/journal.pone.0242028.

[30] Z. Lv, C. Ao, Q. Zou, Protein function prediction: from traditional classifier to deep learning, Proteomics 19 (14) (2019 Jul) 1900119, https://doi.org/10.1002/pmic.201900119.

[31] I. Ahmad, M.J. Iqbal, M. Basheri, Biological data classification and analysis using convolutional neural network, J. Med. Imaging Health Inform. 10 (10) (2020 Oct 1) 2459–2465, https://doi.org/10.1166/jmihi.2020.3179.

[32] T. Murad, S. Ali, I. Khan, M. Patterson, Spike2CGR: an efficient method for spike sequence classification using chaos game representation, Mach. Learn. 112 (10) (2023 Oct) 3633–3658, https://doi.org/10.1007/s10994-023-06371-4.

[33] I. Dieng, M. dos Passos Cunha, M.M. Diagne, P.M. Sembène, P.M. de Andrade Zanotto, O. Faye, O. Faye, A.A. Sall, Origin and spread of the dengue virus type 1, genotype V in Senegal, 2015–2019, Viruses 13 (1) (2021 Jan 4) 57, https://doi.org/10.3390/v13010057.

[34] G. Sánchez-González, Z.R. Belak, L. Lozano, R. Condé, Probability of consolidation constrains novel serotype emergence in dengue fever virus, PLoS One 16 (4) (2021 Apr 5) e0248765, https://doi.org/10.1371/journal.pone.0248765.

[35] L.C. Katzelnick, J.M. Fonville, G.D. Gromowski, J.B. Arriaga, A. Green, S.L. James, L. Lau, M. Montoya, C. Wang, L.A. VanBlargan, C.A. Russell, Dengue viruses cluster antigenically but not as discrete serotypes, Science 349 (6254) (2015 Sep 18) 1338–1343, https://doi.org/10.1126/science.aac5017.

[36] N. Srionrod, P. Nooroong, N. Poolsawat, S. Minsakorn, A. Watthanadirek, W. Junsiri, S. Sangchuai, R. Chawengkirttikul, P. Anuracpreeda, Molecular characterization and genetic diversity of Babesia bovis and Babesia bigemina of cattle in Thailand, Front. Cell. Infect. Microbiol. 12 (2022 Nov 29) 1065963.