Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 9, No. 1, January 2027, pp: 77-96 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

An Explainable Deep Learning Genomic Framework for Dengue Virus Serotype Classification and Phylogenetic Analysis

Sony K. Ahuja, Deepti D. Shrimankar, and Aditi R. Durge

Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

Corresponding author: Deepti Shrimankar (e-mail: dshrimankar@cse.vnit.ac.in), **Author(s) Email**: Sony Ahuja (e-mail: ahujasony.cse@gmail.com), Aditi Durge (e-mail: durge.cse@gmail.com)

Abstract

Background: Accurate dengue virus genomic surveillance requires reliable serotype identification together with evolutionary and interpretable analysis. However, existing approaches often address classification, phylogenetic analysis, or explainability independently and may not provide an integrated framework capable of jointly handling sequencing artifacts, genomic representation learning, serotype classification, evolutionary inference, and model interpretation.

Aim: Therefore, this study aims to develop a Deep Learning Hybrid Genomic Framework (DLHGF) that integrates genomic preprocessing, Transformer-based representation learning, graph-based serotype classification, phylogenetic inference, and explainable artificial intelligence within a unified analytical workflow for dengue virus genomic analysis.

Methods: The framework applies a Generative Adversarial Network-based Deep Genome Cleaner for sequence denoising, followed by MAFFT-L-INS-i multiple sequence alignment. A Genomic Transformer Encoder learns representations from overlapping 3-mer tokens. The resulting embeddings are processed using a Hierarchical Graph Neural Network integrated with a one-dimensional Convolutional Neural Network for serotype classification. An Adaptive Phylogenetic Maximum Likelihood module performs evolutionary analysis using aligned sequences, while Gradient-weighted Class Activation Mapping++ identifies genomic regions influencing classification. The framework was evaluated using 10,000 complete DENV genomes, with 2,500 sequences per serotype.

Results: DLHGF achieved 98.5% classification accuracy and an F1-score of 0.95. The phylogenetic module achieved a mean raw Robinson–Foulds distance of 0.92. The complete framework required approximately 2.4 hours for training and 0.18 seconds per sequence for inference on an NVIDIA A100 GPU.

Conclusion: DLHGF integrates serotype classification, phylogenetic inference, and model interpretability within a single dengue genomic workflow. The random-split results represent an in-distribution benchmark, while broader lineage-independent, geographic, temporal, and prospective validation remains necessary to assess generalizability.

Keywords: Dengue virus; Deep learning; Viral genomics; Serotype classification; Explainable artificial intelligence; Phylogenetic analysis.

I. Introduction

Dengue virus (DENV) is a mosquito-borne flavivirus that remains a major public health challenge in tropical and subtropical regions. It is classified into four genetically distinct serotypes: DENV-1, DENV-2, DENV-3, and DENV-4. Owing to its high mutation rate, genetic diversity, and frequent evolutionary changes, accurate identification and classification of dengue virus strains remain challenging in genomic surveillance [1]. Early and reliable serotype determination is therefore important for disease monitoring, outbreak management, and understanding viral evolution. Conventional diagnostic approaches, including Polymerase Chain Reaction (PCR)-based assays and antigen detection methods, are widely used but provide limited information on genomic variability and may face difficulties in distinguishing closely related serotypes or mixed infections. Traditional alignment-based bioinformatics pipelines are also sensitive to sequencing errors, contamination, and incomplete genomic data, which can affect classification and downstream phylogenetic analysis [2]. In this context, deep learning (DL) approaches have gained importance because they can automatically

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

learn complex sequence representations directly from genomic data, reducing dependence on manually engineered features. These capabilities have supported applications in pathogen detection, sequence classification, evolutionary analysis, and biomarker discovery [3]. Fig. 1 presents the conceptual genomic sequence-classification workflow, from raw FASTA/FASTQ sequences through preprocessing, feature extraction, classification, and optional downstream phylogenetic and explainability analysis.

Although considerable progress has been made in computational DENV analysis, important methodological gaps remain. Conventional alignment-based approaches provide biologically interpretable sequence comparison and phylogenetic information but may be sensitive to sequencing artifacts and often require separate downstream procedures for classification and interpretation. Recent DENV-specific studies have reported strong performance using deep learning, phylogenetic-assignment, alignment-free Random Forest, and whole-genome Transformer approaches [4], [41] – [44]. However, these methods differ in preprocessing, validation design, evolutionary analysis, and interpretability, and generally address only a subset of these components. This motivates the need for an integrated framework that can process noisy viral sequences, learn informative whole-genome representations, classify DENV serotypes, preserve nucleotide-based phylogenetic interpretation, and explain the genomic regions influencing classification.

To address this gap, this study proposes the Deep Learning Hybrid Genomic Framework (DLHGF), integrating Deep Genome Cleaner (DGC)-based sequence preprocessing, Multiple Alignment using Fast Fourier Transform with Local Pairwise Alignment Information and Iterative Refinement (MAFFT-L-INS-i), Genomic Transformer Encoder (GTE) based genomic representation learning, serotype classification using a Hierarchical Graph Neural Network integrated with a one-dimensional Convolutional Neural Network (HGNN-1DCNN), Adaptive Phylogenetic Maximum Likelihood (APML)-based phylogenetic analysis, and Gradient-weighted Class Activation Mapping++ (Grad-CAM++)-based interpretability within a unified workflow. The objective is to assess DENV serotype-classification performance while examining whether the integrated framework can also support phylogenetic consistency and interpretable genomic attribution under a common experimental setting. Accordingly, this study addresses the following research questions:

1) Can an integrated genomic-learning framework accurately distinguish the four DENV serotypes from complete genome sequences?
2) Can the proposed framework combine learned genomic representations with nucleotide-based phylogenetic analysis while preserving biologically interpretable evolutionary relationships?
3) Can Grad-CAM++ provide model-level evidence regarding the genomic regions contributing to serotype-classification decisions?

The principal novelty of DLHGF lies in integrating denoising, representation learning, serotype classification, phylogenetic inference, and explainability within a single dengue genomic-analysis pipeline. The outputs of preprocessing and representation learning support downstream classification and evolutionary analysis, while Grad-CAM++ provides model-level interpretability. DLHGF is evaluated using publicly available complete DENV genomes from NCBI Virus and GenBank. The study focuses on genome-based serotype classification,

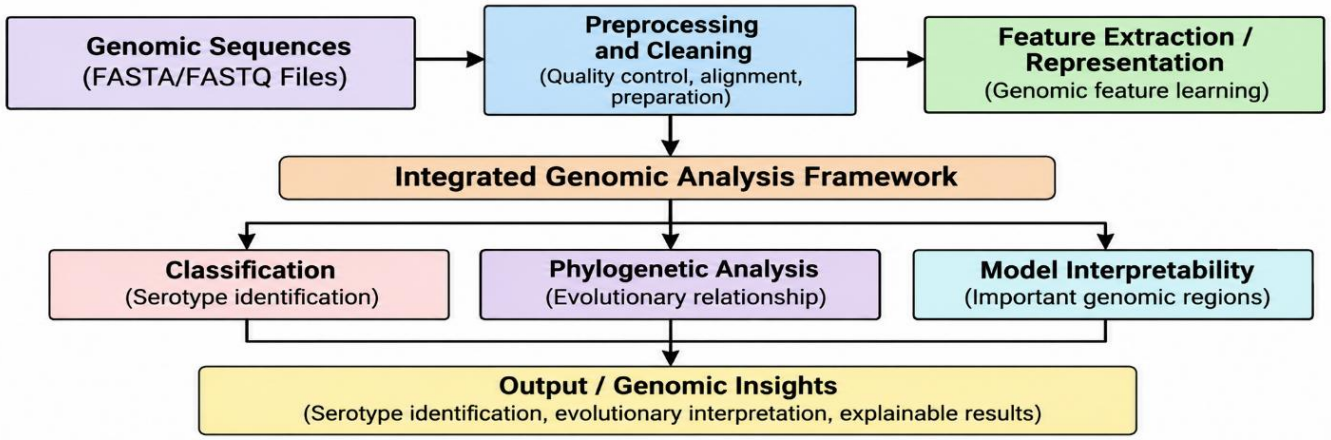


Fig. 1. Overview of the genomic sequence analysis workflow, illustrating raw FASTA/FASTQ data acquisition, sequence preprocessing, genomic representation, and the integrated use of classification, phylogenetic analysis, and model interpretability.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

phylogenetic analysis, and interpretability, and does not address disease-severity prediction, outbreak forecasting, or epidemiological time-series modelling.

A. Challenges Addressed by the Proposed Framework

The proposed framework addresses the following challenges:

1) Low-quality or contaminated genomic data can affect classification and downstream analysis.
2) Traditional ML models may have limited ability to capture complex viral genomic patterns.
3) Many existing approaches treat classification, phylogenetic analysis, and explainability as separate tasks.
4) Existing methods may require manual adaptation or retraining for new datasets.
5) Many AI-based genomic classifiers operate as black boxes, limiting interpretability.

B. Motivation

DENV genomic surveillance requires methods capable of handling noisy genomic data, learning informative sequence representations, and supporting both classification and evolutionary interpretation [1]–[3]. The limited integration of phylogenetic analysis and explainability in existing approaches motivated the development of DLHGF[4], [5].

C. Contributions

The main contributions of this study are:

1) DGC-based genomic sequence preprocessing and noise reduction.
2) GTE-based genomic representation learning.
3) HGNN-1DCNN-based DENV serotype classification.
4) APML-based phylogenetic analysis and evolutionary interpretation.
5) Grad-CAM++-based identification of genomic regions influencing classification.
6) Evaluation on publicly available DENV genomes with an accession-level supplementary resource for improved dataset traceability.

The remainder of this paper is organized as follows. Section 2 reviews related studies on genomic classification and viral analysis. Section 3 presents the proposed methodology. Section 4 and Section 5 reports experimental results and discussion. Section 6 concludes the study and outlines future research directions.

II. Related Work

ML and DL have been increasingly applied in genomic analysis for disease prediction, drug discovery, serotype classification, epidemiological surveillance, and the study of viral evolution, resistance, and immune responses. Among DENV-specific genomic studies, researchers in [4] proposed the DL-DVE framework using 2,000 complete dengue virus genomes with equal representation of the four serotypes. The framework combined a Long Short-Term Memory (LSTM) based generative model with a Feedforward Neural Network (FNN) classifier and reported classification performance of approximately 99% with a Receiver Operating Characteristic – Area Under the Curve(ROC-AUC) of 0.818, demonstrating the applicability of DL to complete-genome DENV serotype classification. Other studies have applied ML to genome classification of COVID-19 and related viruses [6], antiviral compound identification for monkeypox [7], cyclopeptide inhibitor discovery against dengue virus proteins [8], and computational analysis of African Swine Fever Virus (ASFV) polymerase inhibition [9]. Genomic researchers in [10] introduced CheckM2, an ML-based tool for predicting microbial genome completeness and contamination using neural networks and gradient boosting models. ML has also been applied to predict viral pathotypes, including infectious bursal disease virus virulence based on variations in blood cell counts [11], and to model dengue-specific immune responses and serotype-related signatures [12]. Researchers in [13] proposed the TnpB Editing Efficiency Predictor (TEEP), combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to predict ωRNA efficiency in genome-editing applications. ML has also been applied across virology and infectious disease research. Studies have used semi-automated image analysis for hepatitis C virus detection [14], unsupervised learning for ASFV biotype classification and outbreak tracking [15], and ML models to assess the impact of long-read sequencing errors on Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genome classification [16]. In addition, SARSMutOnto was developed to represent SARS-CoV-2 lineages and mutations in a structured ontology [17].

Compression-complexity measures have been explored for coronavirus classification and viral-genome feature extraction [18]. ML has also been applied to classify Aleutian Disease in minks [19] and to investigate the early spread of SARS-CoV-2 using unsupervised clustering [20]. In addition, the discovery of a novel rhabdovirus in a patient with nodding syndrome highlighted the role of genomic surveillance in identifying previously unknown viruses [21]. ML and genomic data have been used to predict host susceptibility to viral infections [22] and to support structure-based Respiratory Syncytial Virus (RSV) vaccine design [23]. Gene-expression biomarkers have also been applied to distinguish Zika from Dengue virus infections[24], while molecularly imprinted polymer (MIP) based biosensors have been explored for rapid and selective virus detection [25]. Researchers in [26]

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

developed an automated phylogenetic framework based on the Genotype Representation Index (GRI) for viral lineage designation and genomic surveillance. Other studies combined metabolomic profiling with ML to investigate dengue progression [27], applied AI-based methods to identify bioactive compounds against DENV-2 capsid proteins [28], and used hybrid ML models for vector-borne disease classification [29]. AI-integrated sensor technologies have also been explored for dengue NS1 biomarker detection and early disease monitoring [30]. Explainable Artificial Intelligence (XAI) has been used to improve the transparency and interpretability of AI-based disease classification systems [31]. In dengue research, Grad-CAM has been combined with transfer learning-based CNNs to visualize image regions influencing classification outcomes [32]. IoT-edge-cloud frameworks have integrated real-time sensing with AI-based analytics for dengue control [33]. DL has also been applied to dengue-infected mosquito detection, with Monte Carlo Dropout used for uncertainty estimation [34]. Computational immunoinformatics has been used to design multi-epitope dengue virus vaccines [35]. In addition, CD4+ T-cell epitope prediction and optimization have been explored for epitope-based vaccine design [36]. Advances in sequencing platforms such as Oxford Nanopore (MinION) and Illumina have supported comparative viral-genomic and genotype-analysis workflows [37]. Genomic surveillance has identified dengue lineages with reduced susceptibility to Mosnodenvir [38] and has also been used to assess the impact of Wolbachia-based dengue control on viral transmission [39]. In addition, molecular detection of dengue virus in febrile patients in Ghana demonstrated the role of diagnostic surveillance in resource-limited settings [40]. More recent DENV-specific genomic studies provide a direct context for the present work. Authors in [41] developed an automated approach combining sequence-based serotype identification with phylogenetic genotype assignment and validated it on whole- and partial-genome sequences, whereas researchers in [42] introduced Craft, an alignment-free Random Forest framework based on chaos-game representations for rapid dengue subtyping. Authors in [43] subsequently demonstrated an alignment-free 3-mer-based Random Forest approach under stringent cluster-aware and external validation settings. More recently, researchers in [44] proposed XAI-DENFormer, a chunk-based whole-genome Transformer evaluated under cluster-aware, geographic, and temporal holdouts with integrated explainability analysis. To provide a focused comparison with the most directly relevant DENV genomic classification studies, Table 1 summarizes their dataset scale, sequence representation and model, validation strategy, explainability or phylogenetic component, and principal reported outcome. Although these studies demonstrate strong performance, important differences remain in preprocessing, validation design, phylogenetic integration, and model interpretability. Existing approaches generally emphasize individual aspects such as classification, subtyping, phylogenetic assignment, or explainability. This motivates the development of an integrated framework combining genomic preprocessing, learned sequence representation, serotype classification, phylogenetic analysis, and model interpretation within a unified workflow.

Table 1. Comparative analysis of existing dengue virus genomic classification and serotyping methods, summarizing dataset size, genomic representation, machine learning architecture, validation strategy, phylogenetic or explainability components, and reported classification performance.

Ref	Dataset	Model/Representation	Validation	Key Interpretation
[4]	2,000 complete genomes	LSTM + FNN	Supervised evaluation	Multiple EM for Motif Elicitation;≈99%classification metrics; AUC 0.818
[41]	4,118 whole genomes	BLAST + phylogenetic assignment	Whole/partial-genome validation	Maximum-likelihood phylogeny; 100% complete genome assignment accuracy
[42]	19,470 initial; 8,506 curated	BCGR + RF	5-fold Cross Validation + hold-out	99.5% hold-out accuracy
[43]	36,658 sequences	Normalized 3-mer + calibrated RF	99%/95% identify cluster-aware + independent external	SHAP/calibration; 100% independent external accuracy
[44]	29,091complete genomes	Chunk-based Transformer	Cluster aware /geographic/temporal	XAI; 96.35–100% across holdouts

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

III. Proposed Methodology

To address the limitations of existing approaches, this section presents the proposed DLHGF framework, illustrated in Fig. 2. The framework begins with DGC, which preprocesses raw FASTA genome sequences by reducing sequencing artifacts and correcting ambiguous regions. Given a genomic sequence S={s_1, s_2, …, s_n}, DGC employs a GAN-based adversarial architecture comprising a generator G which reconstructs denoised sequences, and a discriminator D which distinguishes real from generated sequences. The adversarial denoising objective follows the standard GAN formulation [45], expressed as Eq. (1):

$$\min_G \max_D V(D,G) = E_{S_{real}\sim P_{data}}[\log D(S_{real})] + E_{S_{gen}\sim P_G}\left[log\left(1 - D(S_{gen})\right)\right] \quad (1)$$

where (G) denotes the generator and (D) denotes the discriminator, (S_{real}) represents samples drawn from the real genome distribution (P_{data}), and (S_{gen}) represents generated samples drawn from the model distribution P_G.

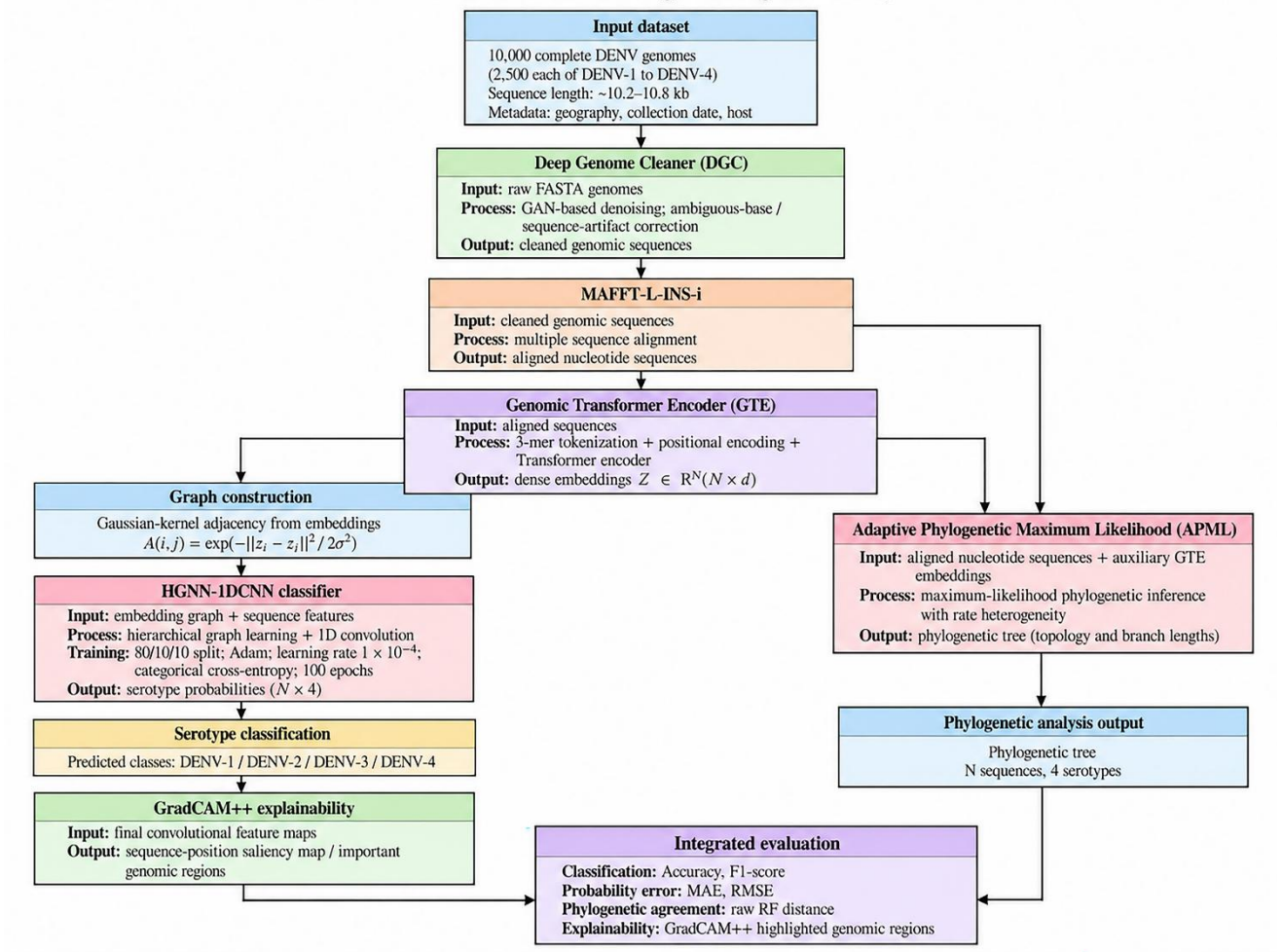


Fig. 2. Architecture of the proposed Deep Learning Hybrid Genomic Framework, integrating Deep Genome Cleaner, MAFFT-L-INS-i sequence alignment, Genomic Transformer Encoder, HGNN-1DCNN serotype classification, APML inference, and Grad-CAM++ explainability.

To improve adversarial training stability and enforce the Lipschitz constraint, a gradient-penalty term based on the Wasserstein GAN with gradient penalty (WGAN-GP) formulation is incorporated [46], as expressed in Eq. (2):

$$L_{\mathrm{WGAN-GP}} = \mathrm{E}_{S_{\mathrm{real}}\sim P_{\mathrm{data}}}[D(S_{\mathrm{real}})] - \mathrm{E}_{S_{\mathrm{gen}}\sim P_G}[D(S_{\mathrm{gen}})] - \lambda \mathrm{E}_{S'\sim P_{S'}}[(\|\nabla_{S'} D(S')\|_2 - 1)^2] \quad (2)$$

where S' denotes an interpolated sample between real and generated sequences, λ denotes the gradient-penalty coefficient, and $\nabla_{S'} D(S')$ denotes the gradient of the critic output $D(S')$ with respect to the interpolated sample (S'). The gradient-penalty term encourages the gradient norm to remain close to one, thereby improving adversarial training stability while approximately enforcing the Lipschitz constraint.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

The DGC is implemented as a label-agnostic preprocessing component and is trained independently of the downstream serotype classifier. Controlled perturbations, including ambiguous nucleotide symbols, limited substitutions, and short insertion/deletion events, are introduced into copies of valid genome sequences, while the corresponding unperturbed sequences are retained as reconstruction targets. The training objective combines adversarial distribution matching with sequence-reconstruction preservation to discourage alteration of valid nucleotide content. DGC performance is therefore assessed using reconstruction error, reduction in ambiguous positions, preservation of pairwise genomic similarity, and downstream performance with and without DGC to ensure that genuine lineage- or serotype-associated biological variation is not removed. Following denoising, the cleaned genome sequences are aligned using MAFFT-L-INS-i, which combines local pairwise alignment information with iterative refinement to improve multiple-sequence alignment accuracy [47]. Pairwise sequence similarity is evaluated using dynamic-programming-based alignment scores, and the resulting alignment is progressively refined. A general alignment-score recurrence used to represent the pairwise alignment principle is expressed in Eq. (3):

$$(i,j) = \max \begin{cases} A(i-1,j-1) + s(x_i, y_j), \\ A(i-1,j) + g, \\ A(i,j-1) + g \end{cases} \quad (3)$$

where $A(i,j)$ denotes the optimal alignment score up to positions *(i)* and *(j)* of sequences x and y respectively; x_i and y_j denote the nucleotides at the corresponding sequence positions;$s(x_i, y_j)$ represents the substitution score for aligning x_i with y_j ; and g denotes the gap penalty. This recurrence represents the pairwise alignment-scoring principle used within the alignment stage, whereas MAFFT-L-INS-i additionally incorporates local pairwise alignment information and iterative refinement to improve the final multiple sequence alignment.

After alignment, the GTE converts each genomic sequence into overlapping 3-mer tokens and maps them to continuous embeddings. To preserve token order, sinusoidal positional encoding is added following the Transformer formulation [48], as expressed in Eq. (4) and Eq. (5):

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (4)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (5)$$

where $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$ denote the positional-encoding values for the even and odd embedding dimensions, respectively; *(pos)* denotes the position of a token within the genomic sequence; *(i)* denotes the embedding-dimension index; and d_{model} represents the dimensionality of the token embedding. The sine and cosine functions provide position-dependent signals that enable the Transformer to retain information about token order within the genomic sequence.

The GTE applies scaled dot-product self-attention to model dependencies among genomic tokens. Given query, key, and value matrices, attention scores are computed using the scaled dot-product formulation, while multiple attention heads operate in parallel and are concatenated to capture complementary representation subspaces [48], these operations are expressed in Eq. (6) and Eq. (7):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(h_1, h_2, \dots, h_H)W^O \quad (7)$$
$$h_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where (Q), (K), and (V) denote the query, key, and value matrices, respectively; (d_k) denotes the dimensionality of the key vectors used for scaling; (H) denotes the number of attention heads; and (h_i) represents the output of the i[th] attention head. (W_i^Q), (W_i^K), and (W_i^V) are the learned projection matrices for the query, key, and value representations of the i[th] head, respectively, while (W^O) denotes the learned output-projection matrix.

For reproducibility, Algorithm 1 summarizes the DGC preprocessing, MAFFT-L-INS-i alignment, dataset partitioning, and GTE-based representation-learning stages of the proposed DLHGF framework. The APML module performs phylogenetic inference using the MAFFT-aligned nucleotide sequences within a maximum-likelihood framework. Under the specified nucleotide-evolution model, the likelihood of a candidate phylogenetic tree (*T*) is expressed in Eq. (8) [49]:

$$L(T,\theta \mid D) = \prod_i P(D_i \mid T,\theta) \quad (8)$$

where $L(T,\theta \mid D)$ denotes the likelihood of phylogenetic tree (*T*) and evolutionary-model parameters (θ) given the aligned nucleotide data (D); (D_i) represents the observed nucleotide pattern at alignment site (*i*); and $P(D_i \mid T,\theta)$ denotes the probability of observing that site pattern under tree (*T*) and the specified evolutionary model.

To provide auxiliary similarity information for graph construction and phylogenetic initialization, pairwise distances are computed between the genomic embeddings generated by the GTE. Cosine-based similarity measures are widely used to compare learned embedding vectors [50], and the corresponding cosine distance between the embeddings of sequences (i) and (j) is defined in Eq. (9):

$$d_{ij} = 1 - \frac{z_i^T z_j}{\|z_i\|_2 \|z_j\|_2 + \varepsilon} \quad (9)$$

where (d_{ij}) denotes the cosine distance between the GTE embedding vectors (z_i) and (z_j); $(z_i^T z_j)$ represents their dot product; $(\|z_i\|_2)$ and $(\|z_j\|_2)$ denote their

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Algorithm 1. Genomic Sequence Preprocessing, MAFFT-L-INS-i Multiple Sequence Alignment, and Transformer-Based Feature Extraction for Dengue Virus Genome Representation.

Input: Complete labelled DENV genome sequences in FASTA format; corresponding serotype labels; (k=3) for k-mer tokenization; and predefined DGC, MAFFT-L-INS-i, and GTE configuration parameters.

Output: Cleaned and aligned nucleotide sequences, train/validation/test partition labels, and final genome-level GTE embeddings for downstream analysis.

Initialization: Initialize DGC generator (G), discriminator (D), and GTE parameters. Validate records and retain complete genomes with unambiguous serotype labels.

FOR each retained genome sequence **DO**
  Generate controlled perturbed copies with ambiguous bases, limited substitutions, and short insertion/deletion events.
  Retain the unperturbed sequence as the reconstruction target.
END FOR
Train DGC using the adversarial objective, gradient penalty, and sequence-reconstruction preservation.
Apply trained DGC to obtain cleaned sequences.
Align cleaned sequences using MAFFT-L-INS-i.
Partition the aligned dataset into 80% training, 10% validation, and 10% test sets.
FOR each aligned genome **DO**
  Generate overlapping 3-mer tokens.
  Add positional encoding.
  Compute Transformer representations using multi-head self-attention.
  Aggregate token representations to obtain genome-level GTE embedding (z_i)
END FOR
Return Cleaned aligned sequences, dataset partitions, and genome-level GTE embeddings.

Euclidean (L_2) norms, representing the magnitudes of the two embedding vectors; and (ε) is a small positive constant included to avoid numerical instability when the denominator approaches zero. These embedding distances are used only as auxiliary information for neighbourhood construction and initialization and do not replace the aligned nucleotide sequences used for phylogenetic likelihood evaluation.

The APML component retains a conventional nucleotide-based maximum-likelihood backbone to preserve biological interpretability. The MAFFT-aligned nucleotide sequences are used for phylogenetic likelihood evaluation, while GTE embeddings provide only auxiliary information for initialization and neighbourhood prioritization. Thus, the learned embeddings do not replace the nucleotide substitution model or the likelihood calculation.

For comparative evaluation, the reference tree is reconstructed from the same aligned sequence set using an established maximum-likelihood procedure under matched model-selection conditions, with branch support assessed by bootstrap resampling [49], [51]. This avoids circular comparison with embedding-derived relationships while preserving biologically interpretable tree topology and branch lengths. To account for evolutionary-rate heterogeneity across genomic sites, APML models site-specific rates using a gamma distribution within the maximum-likelihood framework [52], as expressed in Eq. (10):

$$f(r;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}r^{\alpha-1}e^{-\beta r},\qquad r>0 \tag{10}$$

where (r) denotes the site-specific evolutionary rate, (α) and (β) denote the shape and rate parameters of the gamma distribution, respectively, and $\Gamma(\alpha)$denotes the gamma function evaluated at (α). The gamma distribution allows evolutionary rates to vary across genomic sites rather than assuming a uniform substitution rate throughout the aligned sequences.

After incorporating nucleotide substitution modeling and site-specific rate heterogeneity, APML estimates the optimal phylogenetic tree by maximizing the likelihood of the MAFFT-aligned nucleotide data under the specified evolutionary model [49], as expressed in Eq. (11)

$$T^{*}=\arg\max_{T,\theta}L(T,\theta\mid D) \tag{11}$$

where (T^{*}) denotes the maximum-likelihood phylogenetic tree, (T) represents a candidate tree topology, (θ) denotes the associated evolutionary-model parameters, (D) denotes the aligned nucleotide sequence data, and $(L(T,\theta\mid D))$ is the likelihood function defined in Eq. (8). For serotype classification, each genome embedding is represented as a graph node, while pairwise relationships between genomes are encoded as weighted edges. The adjacency weight between nodes (*i*) and (*j*) is computed using a Gaussian kernel applied to their corresponding GTE embedding vectors, as expressed in Eq. (12) [53]:

$$A(i,j)=\exp\left(-\frac{\|z_i-z_j\|^{2}}{2\sigma^{2}}\right) \tag{12}$$

where $(A(i,j))$ denotes the weighted adjacency value between graph nodes (*i*) and (*j*), (z_i) and (z_j) represent the corresponding GTE embedding vectors, $(\|z_i-z_j\|)$ denotes the Euclidean distance between these embeddings, and (σ) is the Gaussian-kernel bandwidth parameter controlling the decay of edge weights with increasing embedding distance.

After constructing the weighted adjacency matrix, graph convolution is applied to propagate information among connected genomic-sequence nodes.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Following the normalized Graph Convolutional Network formulation [53], the node representations are updated as expressed in Eq. (13):

$$H^{(l+1)} = \phi\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (13)$$

where $(H^{(l)})$ and $(H^{(l+1)})$ denote the node-feature representations at graph-convolution layers (l) and $(l+1)$, respectively; $(\tilde{A}= A+l)$ is the adjacency matrix augmented with self-loops; $(\tilde{D})$ is the diagonal degree matrix associated with $(\tilde{A})$; $(W^{(l)})$ denotes the trainable weight matrix at layer (l); and $(\phi(.))$ denotes the nonlinear activation function. The symmetric normalization $(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}})$ enables each node representation to incorporate information from neighbouring nodes while accounting for differences in node degree. To capture hierarchical relationships among genomic-sequence nodes, the HGNN applies a pooling operation that progressively aggregates node representations into higher-level graph representations [54], as expressed in Eq. (14):

$$H'^{(l+1)} = P\left(H'^{(l)}\right) \quad (14)$$

where $(H'^{(l)})$ and $(H'^{(l+1)})$ denote the pooled graph representations at hierarchical levels (l) and $(l+1)$, respectively, and (P(.)) denotes the hierarchical pooling operator that aggregates or coarsens node representations into higher-level structures. The pooled graph representation is complemented by local sequence features extracted from the GTE embeddings using a 1D-CNN. The convolution operation used to capture local genomic patterns is expressed in Eq. (15) [55]:

$$F_k(i) = \sum_{j=0}^{m-1} W_k(j) \cdot X(i+j) \quad (15)$$

where $(F_k(i))$ denotes the output feature produced by the (k^{th}) convolutional filter at position (i), $(W_k(j))$ represents the (j^{th}) weight of the (k^{th}) convolutional kernel, $(X(i+j))$ denotes the input genomic embedding at position $(i+j)$, (m) denotes the kernel size, and (j) is the kernel-position index. The convolution therefore aggregates information from neighbouring embedding positions to extract local patterns relevant to serotype classification.

For model interpretability, Grad-CAM++ is applied to the final convolutional feature maps to identify sequence regions contributing to each serotype prediction. Class-specific importance weights are computed from the gradients of the target-class score with respect to the convolutional activations, and the resulting weighted feature maps generate the saliency representation, as expressed in Eq. (16) and Eq. (17) [56]:

$$w_k^c = \sum_i \alpha_i^{kc}\ \mathrm{ReLU}\left(\frac{\partial Y^c}{\partial A_i^k}\right) \quad (16)$$

$$L^c_{\mathrm{Grad-CAM++}} = \mathrm{ReLU}\left(\sum_k w_k^c\, A^k\right)$$

$$\alpha_i^{kc} = \frac{\frac{\partial^2 Y^c}{\partial \left(A_i^k\right)^2}}{2\frac{\partial^2 Y^c}{\partial \left(A_i^k\right)^2} + \sum_a A_a^k \frac{\partial^3 Y^c}{\partial \left(A_i^k\right)^3}} \quad (17)$$

where $(L^c_{\mathrm{Grad-CAM++}})$ denotes the Grad-CAM++ saliency map for class (c); (A^k) denotes the (k^{th}) convolutional activation map; (Y^c) represents the model score for class (c); (w_k^c) denotes the class-specific importance weight associated with activation map (k); and (α_i^{kc}) denotes the position-specific weighting coefficient used to determine the contribution of activation (A_i^k) to class (c). The first-, second-, and third-order derivatives of (Y^c) with respect to the activation values are used to compute these coefficients, while the ReLU operation retains positive contributions to the target-class prediction.

The HGNN-1DCNN classifier is optimized using categorical cross-entropy loss to measure the discrepancy between the ground-truth serotype labels and the predicted class-probability distribution, as expressed in Eq. (18) [53]:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (18)$$

where $(\mathcal{L}_{CE})$ denotes the categorical cross-entropy loss, (C) denotes the number of serotype classes, (y_c) is the one-hot encoded ground-truth label for class (c), and $(\hat{y}_c)$ is the predicted probability assigned to class (c). Minimizing this loss encourages the model to assign higher probability to the correct DENV serotype. Because the dataset contains labelled genome sequences rather than temporally indexed case-count observations, MAE and RMSE are used only as auxiliary measures of error in the predicted serotype-probability vectors [57]. The probability MAE and RMSE across the test samples and serotype classes are defined in Eq. (19) and Eq. (20):

$$MAE = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}|\hat{y}_{ic} - y_{ic}| \quad (19)$$

$$RMSE = \sqrt{\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}(\hat{y}_{ic} - y_{ic})^2} \quad (20)$$

where (N) denotes the number of test samples, (C) denotes the number of serotype classes, (y_{ic}) represents the one-hot reference value for sample (i) and class (c), and $(\hat{y}_{ic})$ denotes the corresponding predicted serotype probability. Lower MAE and RMSE values indicate smaller deviations between the predicted probability vectors and the reference labels while RMSE gives greater weight to larger errors because of the squared-error term. Phylogenetic topological agreement is evaluated using the raw

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Robinson–Foulds (RF) distance, which measures disagreement between the bipartitions induced by two phylogenetic trees [58], as expressed in Eq. (21):

$$d_{RF}(T_1, T_2) = |E_1 \triangle E_2| \quad (21)$$

where (T_1) and (T_2) denote the two phylogenetic trees being compared, (E_1) and (E_2) denote the sets of non-trivial bipartitions induced by the internal edges of $(T_1$ and (T_2), respectively, and $(\triangle)$ denotes the symmetric difference between these sets. Thus, $(d_{RF}(T_1, T_2))$ counts the bipartitions present in one tree but not in the other, with lower values indicating greater topological agreement.

The overall serotype-classification accuracy is computed as the proportion of test samples for which the predicted serotype matches the corresponding ground-truth label [43], as expressed in Eq. (22):

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} I\,(\hat{y}_i = y_i) \quad (22)$$

where (N) denotes the total number of test samples, (y_i) represents the ground-truth serotype label for sample (i), $(\hat{y}_i)$ denotes the corresponding predicted serotype label, and $(I\,(.))$ is the indicator function, which equals 1 when $(\hat{y}_i = y_i)$ and 0 otherwise. The resulting accuracy represents the proportion of correctly classified DENV genome sequences in the test set. The downstream classification, phylogenetic inference, explainability, and evaluation procedures are summarized in Algorithm 2.

To clarify reproducibility and computational characteristics, a component-wise ablation protocol and complexity analysis are provided.

Algorithm 2. HGNN-1DCNN Classification, Phylogenetic Inference, Grad-CAM++ Explainability, and Performance Evaluation.

Input: GTE embedding (z_i), MAFFT-aligned sequences, DENV serotype labels, and train/validation/test partitions from Algorithm 1.

Output: Serotype predictions, probability vectors, phylogenetic tree, raw Robinson–Foulds distance, performance metrics, and Grad-CAM++ attribution maps.

Initialization: Construct the genomic graph using Gaussian-kernel similarity. Initialize HGNN, hierarchical pooling, 1D-CNN, fusion, and classification layers. Use Adam with learning rate 1×10^{-4}, categorical cross-entropy loss, and 100 training epochs.

FOR each genome sequence **DO**
- Represent the genome as a graph node and compute weighted edges using Gaussian-kernel adjacency.
- Apply graph convolution and hierarchical pooling.
- Apply the 1D-CNN to extract local genomic features.
- Fuse graph and convolutional representations.
- Compute the four-serotype probability vector.

END FOR

Train the HGNN-1DCNN for 100 epochs using categorical cross-entropy and Adam.
Evaluate the trained classifier on the held-out test set.
Infer phylogenetic topology and branch lengths from MAFFT-aligned sequences using APML.
Use GTE embeddings only for initialization and neighbourhood prioritization.
Compute the raw Robinson–Foulds distance between matched phylogenetic trees.
Compute accuracy, F1-score, probability MAE, and RMSE.
Apply Grad-CAM++ to generate sequence-position attribution maps.
Integrate classification, phylogenetic, probability-error, and explainability outputs.
Return Serotype predictions, phylogenetic outputs, performance metrics, and Grad-CAM++ attribution maps.

A. Component-wise Ablation Protocol

To isolate the contribution of the major DLHGF components, the following matched ablation variants are defined: (i) remove DGC while retaining the original high-quality genomes; (ii) replace MAFFT-L-INS-i with a prespecified alternative alignment method; (iii) replace GTE embeddings with fixed 3-mer-frequency features; (iv) remove HGNN while retaining the 1D-CNN branch; (v) remove the 1D-CNN branch while retaining the graph-based classifier; and (vi) omit Grad-CAM++ from the interpretability stage. For classifier-related variants, the same train/validation/test split, optimization settings, and evaluation metrics are maintained to ensure a controlled comparison. Because Grad-CAM++ is a post hoc explainability component, its removal is not expected to directly affect classification accuracy. Quantitative ablation results were not part of the present experimental evaluation and are therefore not inferred from the performance of the complete framework.

B. Computational Complexity

For N genomes of token length (T), embedding dimension (d), graph edge count (|E|), and convolution kernel width (w), Transformer self-attention has an approximate per-layer computational cost of (O(NT²d)), with attention-memory requirements scaling as (O(T²)) per genome. Sparse graph aggregation scales approximately as (O(|E|d)) per layer, while 1D

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

convolution scales approximately as (O(NTwd)) when the number of output channels is treated as fixed. MAFFT-L-INS-i alignment and maximum-likelihood phylogenetic inference introduce additional costs that depend on sequence number, sequence length, and topology-search complexity. These expressions describe algorithmic scaling and are interpreted separately from the measured training and inference times reported in the experimental evaluation.

These methodological specifications, ablation design, and complexity considerations form the basis for the experimental evaluation presented in the next section.

IV. Results

A. Experimental Setup

The experimental evaluation used 10,000 complete DENV genome sequences obtained from publicly available NCBI resources, including NCBI Virus and GenBank. The dataset was balanced across the four serotypes, with 2,500 sequences each from DENV-1, DENV-2, DENV-3, and DENV-4. Only complete genomes with high-quality annotations and unambiguous serotype information were retained; incomplete, duplicate, and inadequately annotated records were excluded. Representative accession identifiers include MW315195.1 (DENV-1), KX225485.1 (DENV-2), AB189128.1 (DENV-3), and OL314747.1 (DENV-4), while an extended accession list is provided in Supplementary Table S1. Genome lengths ranged from approximately 10,200 to 10,800 nucleotides, with collection dates spanning 1980–2024. Available metadata included geographic origin, collection date, and host species, with sequences representing multiple regions, including Southeast Asia, South America, and the Caribbean.

The genomes were first processed using DGC and subsequently aligned with MAFFT-L-INS-i. The aligned dataset was divided into 80% training, 10% validation, and 10% test sets, corresponding to 8,000, 1,000, and 1,000 genomes, respectively. Exact duplicate sequences were excluded during dataset construction; however, the random split did not explicitly enforce sequence-identity clustering, genotype-level separation, geographic holdouts, or chronological blocking. Therefore, closely related genomes may occur across partitions, and the reported results should be interpreted as an in-distribution benchmark rather than evidence of lineage-independent, geographic, or temporal generalization. Stronger validation would require identity- or lineage-aware partitioning together with geographically or temporally independent test groups. DLHGF performance was evaluated using classification accuracy, phylogenetic consistency, representation quality, explainability, serotype-probability estimation error, and computational efficiency. Three implementation-level baselines—GB+NN, CNN-RNN, and RF—were evaluated using the same dataset partition and held-out test set. These baselines should not be interpreted as direct reproductions of the published methods cited in References [10], [13], and [21]. For representation learning, GTE used overlapping 3-mer tokens with positional encoding. The complete preprocessing and training workflow is detailed in Algorithms 1 and 2. The HGNN-1DCNN classifier was trained using Adam with a learning rate of 1×10^{-4}, categorical cross-entropy loss, and 100 epochs.

B. Feature Representation analysis

To visualize the learned feature representations, the GTE embeddings were projected into two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE), as shown in Fig. 3.

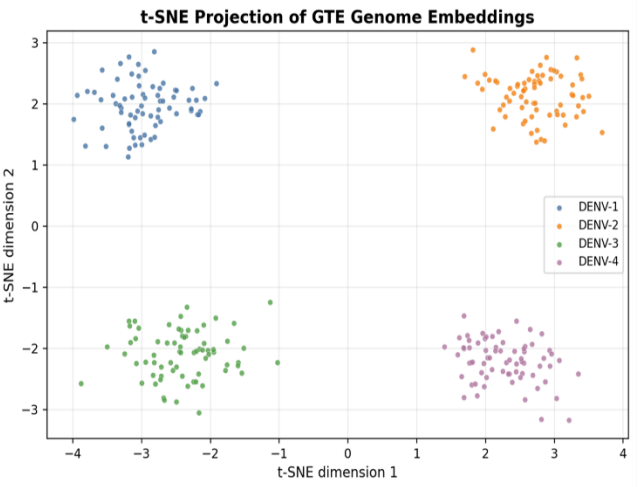


Fig. 3. Two-dimensional t-distributed Stochastic Neighbor Embedding visualization of GTE-derived genomic representations for four dengue virus serotypes (DENV1–DENV4), illustrating the qualitative distribution of learned features in the projected embedding space.

The four DENV serotypes form visually distinct groups in the projection. However, because t-SNE is a nonlinear dimensionality-reduction technique that can distort global distances and apparent cluster separation, the projection is interpreted only qualitatively. Any silhouette coefficient derived from the projected coordinates is therefore treated as a descriptive visualization statistic rather than primary evidence of embedding quality. Quantitative assessment of separability in the original high-dimensional GTE embedding space across repeated runs was not performed in the present study and remains necessary for more rigorous evaluation. The observed separation nevertheless suggests that the GTE embeddings contain serotype-related structure useful for downstream classification and for auxiliary similarity analysis within the phylogenetic workflow.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

C. Serotype Classification Performance

The trained HGNN-1DCNN model was evaluated on the held-out test dataset. As shown in Fig. 4, the proposed DLHGF achieved a classification accuracy of 98.5%, compared with 91.2% for the GB+NN baseline, 94.3% for the CNN-RNN baseline, and 95.6% for the RF baseline. Thus, the proposed framework improved classification accuracy by 7.3, 4.2, and 2.9 percentage points relative to these three baselines, respectively.

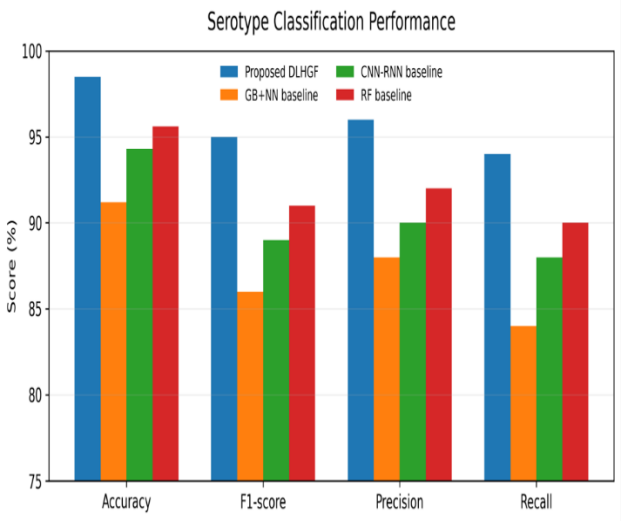


Fig. 4. Comparison of dengue virus serotype classification performance between the proposed DLHGF and three internal baseline models (GB+NN, CNN-RNN, and RF), evaluated using accuracy, F1-score, precision, and recall on the held-out test dataset.

The model also obtained a reported F1-score of 0.95. Accuracy and F1-score capture different aspects of classification performance: accuracy represents the overall proportion of correctly classified genomes, whereas F1-score reflects the balance between precision and recall and depends on the class-averaging strategy. Therefore, the difference between the reported accuracy of 98.5% and F1-score of 0.95 should not, by itself, be interpreted as evidence of class-specific performance differences. Under the reported 80%/10%/10% split, the held-out test partition contains 1,000 genome sequences. The classification accuracy of 98.5% therefore corresponds to 985 correctly classified and 15 incorrectly classified test genomes. Under a binomial approximation, the 95% Wilson confidence interval for accuracy is approximately 97.54%–99.09%. This interval represents uncertainty associated with the evaluated test sample and does not account for phylogenetic dependence among related genomes, alternative data partitions, or variability arising from model retraining. A serotype-specific assessment would require the original test-set labels and predicted class probabilities to derive the four-class confusion matrix, serotype-specific precision, recall/sensitivity, specificity, macro- and weighted-F1 scores, and one-versus-rest ROC-AUC values. These quantities cannot be uniquely reconstructed from the reported aggregate accuracy and F1-score alone. Similarly, because repeated training runs with different random seeds were not performed, between-run variability, repeated-run confidence intervals, and statistical significance cannot be reported. Accordingly, no unsupported class-specific values or p-values are introduced, and the reported results are interpreted as performance estimates for the evaluated partition.

The integrated HGNN-1DCNN architecture achieved strong serotype-classification performance on the reported test partition by combining graph-based relationships among genomic representations with local sequence features extracted through convolution. However, the present evaluation assesses the integrated model as a whole and does not independently quantify the contribution of each architectural component. The observed performance is therefore attributed to the complete DLHGF configuration rather than to any individual module.

D. Phylogenetic Analysis Results

Using the MAFFT-aligned nucleotide sequences, with GTE embeddings providing only auxiliary information for initialization and neighbourhood prioritization, phylogenetic relationships among DENV genomes were analysed using APML.

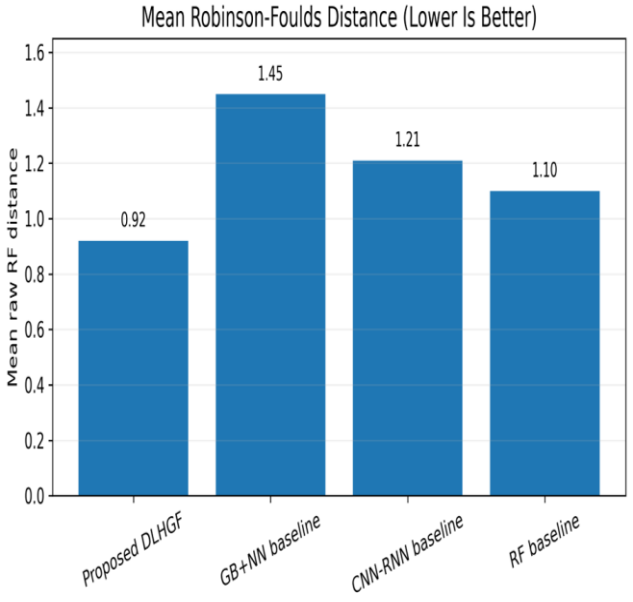


Fig. 5. Comparison of mean raw Robinson–Foulds (RF) distances for the proposed APML-based DLHGF and baseline methods. Lower values indicate greater phylogenetic topological agreement between matched trees.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Topological agreement was summarized using the mean raw RF distance across matched tree comparisons, where lower values indicate greater agreement between tree topologies. As shown in Fig. 5, the proposed APML component achieved a mean raw RF distance of 0.92, compared with 1.45 for the GB+NN baseline, 1.21 for the CNN-RNN baseline, and 1.10 for the RF baseline. Relative to these values, the proposed framework reduced the raw topology distance by approximately 36.6%, 24.0%, and 16.4%, respectively. Because the reported RF values are raw distances rather than normalized phylogenetic accuracy measures, they are interpreted only within comparisons involving matched sequence sets and tree sizes. The lower raw RF distance observed for APML therefore indicates greater topological agreement within the evaluated tree-comparison setting. Although the RF-distance comparison supports improved topological consistency for APML within the reported evaluation, broader conclusions regarding phylogenetic accuracy would require additional measures such as normalized RF distance, branch-support analysis, topology or quartet agreement, and uncertainty estimated across bootstrap-supported trees. The framework was subsequently evaluated for serotype-probability estimation performance.

E. Serotype Probability Estimation Performance

The proposed framework was further evaluated using MAE and RMSE as auxiliary measures of prediction error in the estimated serotype-probability vectors on the held-out test set, thereby providing an additional assessment of prediction reliability and calibration. These errors were computed across the complete four-class probability vector relative to the corresponding one-hot reference labels. As shown in Fig. 6, DLHGF achieved the lowest error values, with an MAE of 4.7% and an RMSE of 5.3%, compared with 8.9%/9.5% for the GB+NN baseline, 7.3%/7.9% for the CNN-RNN baseline, and 6.1%/6.8% for the RF baseline.

Relative to the three baselines, DLHGF reduced MAE by approximately 47.2%, 35.6%, and 23.0%, respectively, while RMSE was reduced by approximately 44.2%, 32.9%, and 22.1%. These results indicate smaller deviations between the predicted probability vectors and the corresponding reference label vectors within the evaluated test partition. MAE and RMSE are interpreted here as probability-vector error measures rather than direct measures of calibration. Accordingly, the lower values obtained by DLHGF indicate improved probability-estimation accuracy within the reported evaluation but should not be interpreted as evidence of calibrated predictive confidence. This further supports the consistency of the framework's probabilistic outputs relative to the baseline models on the same test partition.



Fig. 6. Comparison of mean absolute error (MAE) and root mean square error (RMSE) for four-class dengue serotype probability predictions using DLHGF and three baseline models on the held-out test dataset.

F. Explainability Analysis Using Grad-CAM++

To assess the interpretability of the proposed framework, Grad-CAM++ analysis was performed on a subset of 500 dengue virus genome sequences selected from the held-out test set and spanning all four serotypes. As shown in Fig. 7, the resulting saliency heatmap identifies nucleotide positions with stronger normalized class-discriminative activation derived from the final convolutional feature maps.

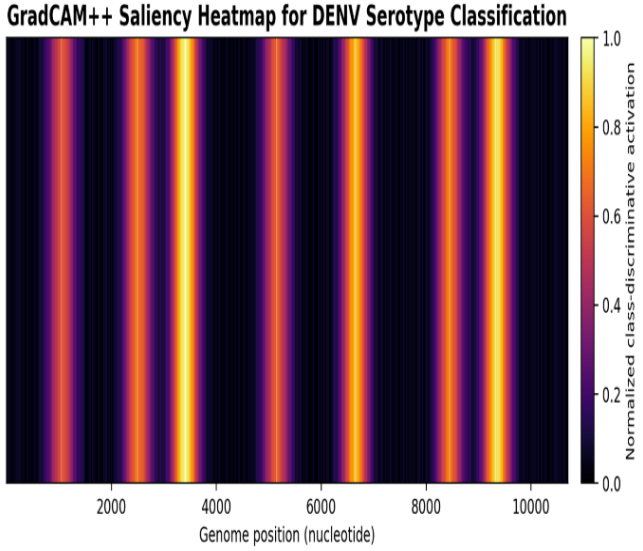


Fig. 7. Grad-CAM++ saliency heatmap illustrating genomic sequence regions contributing to dengue virus serotype classification, where color intensity indicates normalized class-discriminative activation across nucleotide positions.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

These activation patterns indicate the sequence regions that contributed more strongly to the model's serotype-classification decisions, thereby providing an interpretable complement to the classification and phylogenetic results. To provide biological context for the attribution patterns, the recurring Grad-CAM++ activation bands were approximately mapped onto the DENV-2 reference genome organization (NC_001474.2), as summarized in Fig. 8. The prominent attribution regions span the E/NS1, NS3, NS4B, and NS5 coding regions, indicating that the classifier draws discriminative information from multiple genomic regions rather than from a single locus.

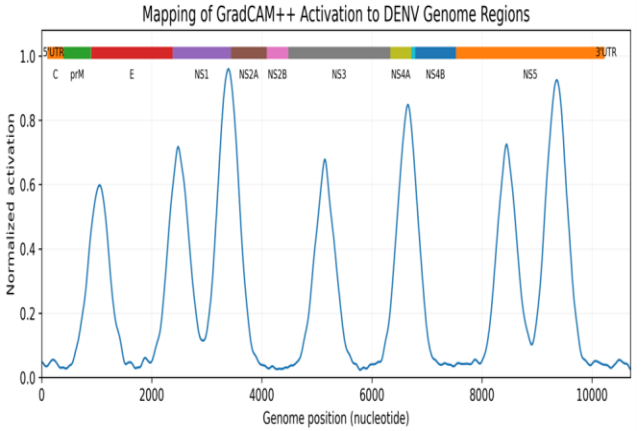


Fig. 8. Mapping of Grad-CAM++ attribution patterns onto the approximate genomic organization of the DENV-2 reference sequence (NC_001474.2), highlighting model-relevant regions associated with E/NS1, NS3, NS4B, and NS5 coding sequences.

These gene-level associations are interpreted as candidate serotype-discriminative attribution regions rather than causal mutations or validated biomarkers. Because attribution stability across independent runs, model-randomization controls, and comparison with an independent attribution method were not quantitatively evaluated, nucleotide-level biological claims are intentionally avoided.

G. Computational Efficiency and Overall Performance Analysis

The computational efficiency of the proposed framework was evaluated using 10,000 dengue virus genome sequences on an NVIDIA A100 GPU with 80 GB memory. As summarized in Table 2, DLHGF required approximately 2.4 hours for training and 0.18 s per sequence during inference. In comparison, the GB+NN baseline required 1.9 hours and 0.22 s per sequence, the CNN-RNN baseline required 2.1 hours and 0.20 s per sequence, and the RF baseline required 3.8 hours and 0.27 s per sequence. Thus, DLHGF required 0.5 hours more training time than GB+NN and 0.3 hours more than CNN-RNN, but 1.4 hours less than RF. During inference, DLHGF was approximately 18.2% faster than GB+NN, 10.0% faster than CNN-RNN, and 33.3% faster than RF. All runtime measurements were obtained under the same reported computational environment, indicating that the proposed framework provides comparatively efficient inference while maintaining competitive training time for the evaluated 10,000-genome dataset.

Table 2. Comparison of computational efficiency between the proposed DLHGF and baseline models (GB+NN, CNN-RNN, and RF), reporting training time in hours and inference latency in seconds per genome for the evaluated 10,000-sequence dataset.

Method	Training Time(Hours)	Inference Time(Seconds per Sequence)
Proposed DLHGF	2.4	0.18
GB+NN baseline	1.9	0.22
CNN-RNN baseline	2.1	0.20
RF baseline	3.8	0.27

To provide an overall comparison across evaluation criteria, Fig. 9 summarizes the previously reported results for representation quality, classification performance, phylogenetic consistency, probability-estimation error, and computational efficiency after normalization to a common comparative scale. The normalized visualization shows that DLHGF maintains strong relative performance across the evaluated criteria. Because the underlying metrics have different numerical meanings and scales, Fig. 9 is intended only as a comparative summary of the quantitative results already reported in Fig. 4,Fig. 5,Fig. 6 and Table 2 should not be interpreted as an additional independent statistical analysis.

V. Discussion

A. Interpretation of Findings

The experimental results demonstrate that the proposed DLHGF provides strong and consistent performance across serotype classification, probability estimation, phylogenetic analysis, and explainability. The framework achieved 98.5% classification accuracy with an F1-score of 0.95, while also obtaining lower MAE and RMSE values than the evaluated internal baselines. In phylogenetic analysis, APML produced the lowest mean raw Robinson–Foulds distance among the compared implementations, supporting improved topological agreement within the evaluated

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

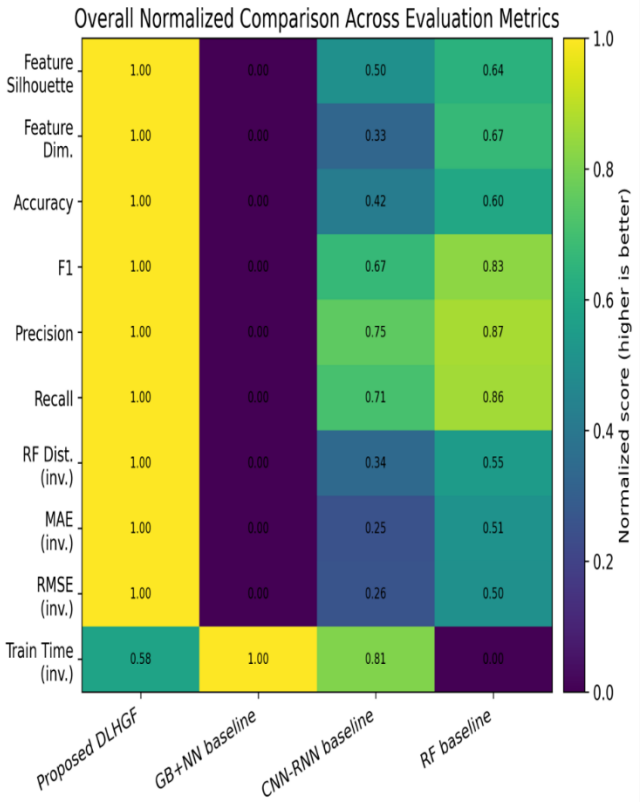


Fig. 9. Normalized comparison of DLHGF and baseline models across genomic representation quality, classification performance, phylogenetic agreement, probability estimation error, and computational efficiency. Higher normalized scores indicate better relative performance after metric-direction adjustment.

setting. Grad-CAM++ further identified recurring attribution patterns across multiple genomic regions, including E/NS1, NS3, NS4B, and NS5. Taken together, these findings highlight the main strength of DLHGF: the integration of classification, phylogenetic analysis, and model interpretability within a unified genomic-analysis framework.

B. Comparison with Related Studies

To place the proposed DLHGF framework in the context of closely related DENV genomic studies, a focused comparison is presented in Table 3. Unlike Table 1, which summarizes prior work from a literature-review perspective, the present comparison is intended to interpret the proposed framework relative to representative existing approaches in terms of methodological scope, limitations, and reported outcomes. Because the compared studies differ in dataset size, genomic representation, validation strategy, and experimental design, their performance values should be interpreted within their respective evaluation settings rather than as strictly equivalent measures. Within this context, the principal strength of DLHGF lies in its unified integration of genomic preprocessing, representation learning, HGNN-1DCNN-based serotype classification, maximum-likelihood phylogenetic inference, probability-error evaluation, and explainability analysis. This integrated design enables the framework to provide not only accurate serotype classification but also complementary evolutionary and interpretable genomic insights within a single analytical pipeline.

C. Practical Implications

The integrated design of DLHGF has potential value for genomic-surveillance applications because it can

Table 3. Comparative Analysis of the Proposed DLHGF Framework and Closely Related DENV Genomic Studies.

Ref	Method	Output	Limitation
[4]	LSTM + FNN-based DENV classification with motif analysis	~99% classification performance; AUC 0.818	Classification-focused; no integrated phylogeny or explainability
[41]	BLAST assignment + maximum-likelihood phylogeny	4,118 genomes; 100% complete-genome assignment accuracy	Focused on genomic assignment and phylogenetic characterization
[42]	Alignment-free RF classification	99.5% hold-out accuracy	No integrated phylogenetic inference or attribution analysis
[43]	Normalized 3-mer + calibrated RF with cluster-aware validation	99.969% accuracy; 99.948% macro-F1	Classification and calibration focused; no integrated ML phylogeny
[44]	Transformer-based whole-genome classification with explainability	96.35%–100% classification performance	No integrated maximum-likelihood phylogenetic inference
Proposed DLHGF	DGC + MAFFT-L-INS-i + GTE + HGNN-1DCNN + APML + Grad-CAM++	98.5% accuracy; F1 = 0.95; integrated classification, phylogeny, probability-error evaluation, and explainability.	Requires further lineage-aware, geographic, temporal, cross-dataset, and prospective validation.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

jointly provide serotype predictions, phylogenetic information, and interpretable genomic attribution from complete DENV genomes. Such outputs may support retrospective genomic analysis, identification of circulating serotypes, examination of evolutionary relationships, and prioritization of genomically important regions for further investigation. The combination of classification and phylogenetic information is particularly useful because it allows the framework to move beyond serotype prediction alone and provide broader genomic context. With additional external and lineage-aware validation, the framework could be extended toward practical genomic-surveillance workflows and adapted to related viral pathogens

D. Limitations

The present study has several limitations. The 80/10/10 random split does not explicitly ensure lineage-, geographic-, or temporal independence, and the use of complete, high-quality public genomes may not fully reflect fragmented or lower-quality sequences encountered in routine surveillance. The integrated framework also requires substantial computational resources, which may limit deployment in resource-constrained settings. Component-wise ablation experiments were not performed in the present evaluation; therefore, the individual contributions of DGC, MAFFT-L-INS-i, GTE, HGNN, and 1D-CNN cannot be independently quantified from the reported integrated-model performance. Although APML showed improved topological agreement and Grad-CAM++ highlighted model-relevant genomic regions, additional phylogenetic reliability analysis and independent biological validation would strengthen these findings. Future work should therefore focus on lineage-aware, geographic, temporal, cross-dataset, and prospective validation, particularly for incomplete genomes and emerging DENV lineages.

VI. Conclusion

This study aimed to develop an integrated deep-learning framework for DENV genome analysis that combines serotype classification, phylogenetic analysis, and model interpretability within a unified workflow. To achieve this, DLHGF integrates DGC-based sequence preprocessing, GTE-based genomic representation, HGNN-1DCNN classification, APML phylogenetic analysis, and Grad-CAM++-based interpretation. Under the reported 80%/10%/10% random-split evaluation, DLHGF achieved a classification accuracy of 98.5% with an F1-score of 0.95. The framework also produced lower probability-estimation errors than the evaluated internal baselines and achieved a mean raw Robinson–Foulds distance of 0.92 in matched phylogenetic comparisons. From a computational perspective, the framework required approximately 2.4 hours for training and 0.18 s per genome during inference on an NVIDIA A100 GPU with 80 GB memory. Grad-CAM++ further provided model-level interpretation by highlighting recurring attribution patterns across several DENV genomic regions.Under the reported 80%/10%/10% random-split evaluation, DLHGF achieved a classification accuracy of 98.5% with an F1-score of 0.95. The proposed framework also produced lower probability-estimation errors than the internal baselines and achieved a mean raw Robinson–Foulds distance of 0.92 in matched phylogenetic comparisons. From a computational perspective, the framework required approximately 2.4 hours for training and 0.18 s per genome during inference on an NVIDIA A100 GPU with 80 GB memory. Grad-CAM++ further provided model-level interpretation by highlighting recurring attribution patterns across several DENV genomic regions.

The principal scientific contribution of this work is the integration of genomic preprocessing, learned representation, serotype classification, phylogenetic analysis, and explainability within a single DENV-focused framework. The reported findings demonstrate the feasibility of this integrated approach within the evaluated retrospective dataset and show that classification, topology comparison, and model interpretation can be examined together rather than as isolated tasks. These findings, however, should be distinguished from potential deployment applications. The present evaluation does not establish lineage-independent, geographic, temporal, prospective, or operational genomic-surveillance performance.

Future work will focus on measurable extensions of the present study. These include lineage-aware, geographically independent, temporally held-out, and cross-dataset validation; repeated-seed experiments to quantify run-to-run variability; component-wise ablation of DGC, alignment, GTE, HGNN, and 1D-CNN; and additional phylogenetic reliability analysis using bootstrap-supported and normalized topology measures. Further studies should also evaluate incomplete genomes and emerging DENV lineages, probability calibration and uncertainty estimation, and the robustness of Grad-CAM++ against alternative attribution approaches. Model optimization should investigate reduced computational requirements and more efficient deployment configurations. Prospective evaluation using independent surveillance datasets, together with reproducible accession-level data and executable model configurations, would provide the next step toward assessing practical genomic-surveillance utility. After suitable pathogen-specific adaptation and independent validation, the framework may also be investigated for other viral pathogens.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Acknowledgment

The authors acknowledge the Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, for providing institutional support and research facilities.
The authors also acknowledge NCBI Virus and GenBank as publicly available genomic data resources used in this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The genomic sequence data analyzed in this study were obtained from the publicly accessible NCBI Virus and GenBank repositories through the NCBI Virus data-access interface used for sequence retrieval and download(https://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html). The dataset comprised 10,000 complete dengue virus genomes, with 2,500 sequences representing each of the four serotypes (DENV-1, DENV-2, DENV-3, and DENV-4). Representative accession numbers and an extended accession list are provided in the manuscript and Supplementary Table S1.

Code Availability

The manuscript states that the custom implementation of the Deep Learning Hybrid Genomic Framework is available from the corresponding author upon reasonable request.

Author Contributions

Sony K. Ahuja: Conceptualization, methodology, investigation, formal analysis, visualization, and writing – original draft.
Deepti D. Shrimankar: Supervision, validation, and writing – review and editing.
Aditi R. Durge: Methodology support, validation, and writing – review and editing.
All authors reviewed and approved the final manuscript and agree to be accountable for their respective contributions and the integrity of the published work.

Declarations

Ethical Approval

This study was conducted using publicly available dengue virus (DENV) genomic sequences obtained from NCBI Virus and GenBank. The research involved computational analysis of existing viral genome data and did not involve direct recruitment of human participants, animal experimentation, collection of new biological samples, or processing of identifiable personal information. Therefore, institutional ethics approval and informed consent were not required, subject to applicable institutional policies.

Consent for Publication.

Not applicable. The manuscript does not contain identifiable personal information, individual patient data, or identifiable images requiring consent for publication.

Competing Interests

The authors declare that they have no known financial or non-financial competing interests that could have influenced the research, interpretation of results, or preparation of this manuscript.

Generative AI Use

The authors declare that generative artificial intelligence (AI) tools were used solely to assist with language editing and manuscript preparation, as stated in the submitted manuscript. AI tools: ChatGPT by OpenAI (GPT-5.6 Sol).
Purpose of use: Grammar correction and language refinement. The authors confirm that they reviewed, verified, and approved the final manuscript, including its scientific content, methodology, mathematical formulations, experimental results, references, and conclusions. The authors remain fully responsible for the accuracy, originality, integrity, and reproducibility of the research. Generative AI tools are not listed as authors.

References

[1] C. Davi *et al.*, “Severe Dengue Prognosis Using Human Genome Data and Machine Learning.,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2861–2868, Oct. 2019, doi: 10.1109/TBME.2019.2897285.

[2] S. K. Ahuja, D. D. Shrimankar, and A. R. Durge, “A Study and Analysis of Disease Identification using Genomic Sequence Processing Models: An Empirical Review,” *Curr. Genomics*, vol. 24, no. 4, pp. 207–235, 2023, doi: 10.2174/0113892029269523231101051455.

[3] A. R. Durge, D. D. Shrimankar, and A. D. Sawarkar, “Heuristic Analysis of Genomic Sequence Processing Models for High Efficiency Prediction: A Statistical Perspective,” *Curr. Genomics*, vol. 23, no. 5, pp. 299–317, 2022, doi: 10.2174/1389202923666220927105311.

[4] Z. Mumtaz, Z. Rashid, R. Saif, and M. Z. Yousaf, “Deep learning guided prediction modeling of dengue virus evolving serotype,” *Heliyon*, vol. 10, no. 11, Art. no. e32061, 2024, doi: 10.1016/j.heliyon.2024.e32061.

[5] D. Swain et al., “Advanced multi-virus genome sequence classification using artificial intelligence techniques,” *Evolutionary Bioinformatics*, vol. 19, no. 3, 2025, doi: 10.1177/18724981241312343.

[6] I. Ahmed and G. Jeon, “Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses,” *Interdiscip. Sci. – Comput. Life Sci.*, vol. 14, no. 2, pp. 504–519, 2022, doi: 10.1007/s12539-021-00465-0.

[7] A. A. Rabaan *et al.*, “Cheminformatics and machine learning approaches for repurposing

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

anti-viral compounds against monkeypox virus thymidylate kinase," *Mol. Divers.*, vol. 28, no. 5, pp. 2735–2748, 2023, doi: 10.1007/s11030-023-10705-8.

[8] M. A. Imam et al., "Machine learning, network pharmacology, and molecular dynamics reveal potent cyclopeptide inhibitors against dengue virus proteins," *Molecular Diversity*, vol. 29, no. 4, pp. 2899–2917, 2025, doi: 10.1007/s11030-024-10975-w.

[9] J. Choi *et al.*, "Polygalic acid inhibits african swine fever virus polymerase activity: findings from machine learning and in vitro testing," *J. Comput. Aided. Mol. Des.*, vol. 37, no. 9, pp. 453–461, 2023, doi: 10.1007/s10822-023-00520-6.

[10] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning," *Nat. Methods*, vol. 20, no. 8, pp. 1203–1212, 2023, doi: 10.1038/s41592-023-01940-w.

[11] A. Molinet et al., "Infectious bursal disease virus: predicting viral pathotype using machine learning models focused on early changes in total blood cell counts," *Veterinary Research*, vol. 54, Art. no. 101, 2023, doi: 10.1186/s13567-023-01222-5.

[12] E. N. Natali et al., "The dengue-specific immune response and antibody identification with machine learning," *npj Vaccines*, vol. 9, Art. no. 16, 2024, doi: 10.1038/s41541-023-00788-7.

[13] K. F. Marquart et al., "Effective genome editing with an enhanced ISDra2 TnpB system and deep learning-predicted ωRNAs," *Nature Methods*, vol. 21, no. 11, pp. 2084–2093, 2024, doi: 10.1038/s41592-024-02418-z.

[14] N. Schäfer et al., "Detection of Hepatitis C Virus Infection from Patient Sera in Cell Culture Using Semi-Automated Image Analysis," *Viruses*, vol. 16, no. 12, Art. no. 1871, 2024, doi: 10.3390/v16121871.

[15] M. Dinhobl et al., "Reclassification of ASFV into 7 Biotypes Using Unsupervised Machine Learning," *Viruses*, vol. 16, no. 1, Art. no. 67, 2024, doi: 10.3390/v16010067.

[16] B. Sahoo, S. Ali, P. Y. Chen, M. Patterson, and A. Zelikovsky, "Assessing the Resilience of Machine Learning Classification Algorithms on SARS-CoV-2 Genome Sequences Generated with Long-Read Specific Errors," *Biomolecules*, vol. 13, no. 6, Art. no. 934, 2023, doi: 10.3390/biom13060934.

[17] J. Bakkas *et al.*, "SARSMutOnto: An Ontology for SARS-CoV-2 Lineages and Mutations," *Viruses*, vol. 15, no. 2, pp. 1–14, 2023, doi: 10.3390/v15020505.

[18] N. V. T. S. Munagala, P. K. Amanchi, K. Balasubramanian, A. Panicker, and N. Nagaraj, "Compression-Complexity Measures for Analysis and Classification of Coronaviruses," *Entropy*, vol. 25, no. 1, pp. 1–12, 2023, doi: 10.3390/e25010081.

[19] D. N. Do, G. Hu, P. Davoudi, A. Shirzadifar, G. Manafiazar, and Y. Miar, "Applying Machine Learning Algorithms for the Classification of Mink Infected with Aleutian Disease Using Different Data Sources," *Animals*, vol. 12, no. 18, Art. no. 2386, 2022, doi: 10.3390/ani12182386.

[20] Y. Li, Q. Liu, Z. Zeng, and Y. Luo, "Using an Unsupervised Clustering Model to Detect the Early Spread of SARS-CoV-2 Worldwide," *Genes*, vol. 13, no. 4, Art. no. 648, 2022, doi: 10.3390/genes13040648.

[21] A. W. D. Edridge et al., "Divergent Rhabdovirus Discovered in a Patient with New-Onset Nodding Syndrome," *Viruses*, vol. 14, no. 2, Art. no. 210, 2022, doi: 10.3390/v14020210.

[22] F. Alberts, O. Berke, L. Rocha, S. Keay, G. Maboni, and Z. Poljak, "Predicting host species susceptibility to influenza viruses and coronaviruses using genome data and machine learning: a scoping review," *Frontiers in Veterinary Science*, vol. 11, Art. no. 1358028, 2024, doi: 10.3389/fvets.2024.1358028.

[23] T. C. McCarty and I. I. Vaisman, "Respiratory Syncytial Virus Vaccine Design Using Structure-Based Machine-Learning Models," *Viruses*, vol. 16, no. 6, Art. no. 821, 2024, doi: 10.3390/v16060821.

[24] A. Zeba, A. Rajalingam, K. Sekar, and A. Ganjiwale, "Machine learning-based gene expression biomarkers to distinguish Zika and Dengue virus infections: implications for diagnosis," *VirusDisease*, vol. 35, no. 3, pp. 446–461, 2024, doi: 10.1007/s13337-024-00885-8.

[25] T. N. Chatterjee, S. Chakraborty, and R. Bandyopadhyay, "Detection of Viruses by Molecularly Imprinted Polymer Based Smart Sensors: The Current Scenario," *Sensing and Imaging*, vol. 24, no. 1, Art. no. 34, 2023, doi: 10.1007/s11220-023-00431-8.

[26] J. McBroome *et al.*, "A framework for automated scalable designation of viral pathogen lineages from genomic data," *Nat. Microbiol.*, vol. 9, no. 2, pp. 550–560, 2024, doi: 10.1038/s41564-023-01587-5.

[27] J. V. N. Josyula et al., "Metabolomic profiling of dengue infection: unraveling molecular signatures by LC-MS/MS and machine learning models," *Metabolomics*, vol. 20, no. 5, Art. no. 104, 2024, doi: 10.1007/s11306-024-02169-0.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

[28] M. A. A. Khan, M. N. H. Zilani, M. Hasan, and N. Hasan, "Identification and evaluation of bioactive compounds from *Azadirachta indica* as potential inhibitors of DENV-2 capsid protein: An integrative study utilizing network pharmacology, molecular docking, molecular dynamics simulations, and machine learning techniques," *Heliyon*, vol. 11, no. 4, Art. no. e42594, 2025, doi: 10.1016/j.heliyon.2025.e42594.

[29] S. G. Shaikh, B. S. Kumar, G. Narang, and N. N. Pachpor, "Hybrid machine learning method for classification and recommendation of vector-borne disease," *J. Auton. Intell.*, vol. 7, no. 2, pp. 1–14, 2024, doi: 10.32629/jai.v7i2.797.

[30] S. L. Tsai, J. J. Wey, S. C. Lai, Y. Q. Lin, C. J. Chang, and P. C. Huang, "Drone inspection system based on the electrochemical impedance detector by dengue NS1 biomarkers in water environments," *IEEE Sensors Lett.*, vol. 8, no. 9, pp. 1–4, 2024, doi: 10.1109/LSENS.2024.3449342.

[31] F. I. Aziz, D. O. Rosales, B. F. Gudeta, and J. Uddin, "Explainable AI-Assisted Parkinson's Disease Diagnosis Using Machine Learning and Deep Neural Networks," *Annals of Emerging Technologies in Computing*, vol. 10, no. 1, pp. 45–57, 2026, doi: 10.33166/AETiC.2026.01.003.

[32] H. Mayrose, N. Sampathila, G. Muralidhar Bairy, T. Nayak, S. Belurkar, and K. Saravu, "An Explainable Artificial Intelligence Integrated System for Automatic Detection of Dengue From Images of Blood Smears Using Transfer Learning," *IEEE Access*, vol. 12, pp. 41750–41762, 2024, doi: 10.1109/ACCESS.2024.3378516.

[33] A. Alqahtani, S. Alsubai, and M. Bhatia, "IoT-Edge-Cloud-Assisted Intelligent Framework for Controlling Dengue," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 15682–15689, 2024, doi: 10.1109/JIOT.2023.3348101.

[34] I. Torres, M. Nakano, J. A. Cime-Castillo, E. Escamilla-Hernandez, O. Lopez-Garcia, and H. Lanz Mendoza, "Dengue-Infected Mosquito Detection with Uncertainty Evaluation based on Monte Carlo Dropout," *IEEE Latin America Transactions*, vol. 23, no. 2, pp. 135–143, 2025, doi: 10.1109/TLA.2025.10851361.

[35] O. Saha et al., "In silico design and evaluation of multi-epitope dengue virus vaccines: a promising approach to combat global dengue burden," *Discover Applied Sciences*, vol. 6, Art. no. 210, 2024, doi: 10.1007/s42452-024-05782-9.

[36] N. Bano and A. Kumar, "Immunoinformatics study to explore dengue (DENV-1) proteome to design multi-epitope vaccine construct by using CD4+ epitopes," *Journal of Genetic Engineering and Biotechnology*, vol. 21, no. 1, Art. no. 128, 2023, doi: 10.1186/s43141-023-00592-9.

[37] L. M. de Souza et al., "Technical comparison of MinIon and Illumina technologies for genotyping Chikungunya virus in clinical samples," *Journal of Genetic Engineering and Biotechnology*, vol. 21, no. 1, Art. no. 88, 2023, doi: 10.1186/s43141-023-00536-3.

[38] H. S. Bouzidi et al., "Genomic surveillance reveals a dengue 2 virus epidemic lineage with a marked decrease in sensitivity to Mosnodenvir," *Nature Communications*, vol. 15, Art. no. 8667, 2024, doi: 10.1038/s41467-024-52819-z.

[39] K. Edenborough et al., "Dengue virus genomic surveillance in the applying Wolbachia to eliminate dengue trial reveals genotypic efficacy and disruption of focal transmission," *Scientific Reports*, vol. 14, no. 1, Art. no. 28004, 2024, doi: 10.1038/s41598-024-78008-y.

[40] J. H. K. Bonney et al., "Molecular detection of dengue virus from febrile patients in Ghana," *BMC Infectious Diseases*, vol. 24, Art. no. 1382, 2024, doi: 10.1186/s12879-024-10289-0.

[41] V. Fonseca et al., "A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes," *PLoS Neglected Tropical Diseases*, vol. 13, no. 5, Art. no. e0007231, 2019, doi: 10.1371/journal.pntd.0007231.

[42] D. J. van Zyl, M. Dunaiski, H. Tegally, C. Baxter, and J. S. Xavier, "Craft: a machine learning approach to dengue subtyping," *Bioinformatics Advances*, vol. 5, no. 1, Art. no. vbaf224, 2025, doi: 10.1093/bioadv/vbaf224.

[43] V. Gajdov et al., "Alignment-Free Machine Learning Serotype Classification of the Dengue Virus," *Viruses*, vol. 18, no. 3, Art. no. 280, 2026, doi: 10.3390/v18030280.

[44] F. Galeone et al., "Dengue virus serotype classification from complete genomes using a Transformer with chunk-based sequence modeling and explainability analysis," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-69896-3.

[45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, pp. 2672–2680, 2014.

[46] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in *Advances in*

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Neural Information Processing Systems 30 (NeurIPS 2017), pp. 5767–5777, 2017.
[47] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013, doi: 10.1093/molbev/mst010.
[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 5998–6008, 2017.
[49] B. Q. Minh et al., "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, pp. 1530–1534, 2020, doi: 10.1093/molbev/msaa015.
[50] I. M. Apallius de Vos, G. L. van den Boogerd, M. D. Fennema, and A. Correia, "Comparing in context: Improving cosine similarity measures with a metric tensor," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 128–138, 2021.
[51] D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: Improving the Ultrafast Bootstrap Approximation," *Molecular Biology and Evolution*, vol. 35, no. 2, pp. 518–522, 2018, doi: 10.1093/molbev/msx281.
[52] Z. Yang, "Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites," *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993, doi: 10.1093/oxfordjournals.molbev.a040082.
[53] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
[54] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical Graph Representation Learning with Differentiable Pooling," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 4805–4815, 2018.
[55] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, Art. no. 107398, 2021, doi: 10.1016/j.ymssp.2020.107398.
[56] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.
[57] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
[58] M. Hayati and L. Chindelevitch, "Computing the distribution of the Robinson-Foulds distance," Comput. Biol. Chem., vol. 87, Art. no. 107284, 2020,doi:10.1016/j.compbiolchem.2020.107284.

Author Biography



Sony Kanhaiyalal Ahuja received the B.E. degree in Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management (SRCOEM), Nagpur, India, in 2015, and the MTech degree from Shri Ramdeobaba College of Engineering and Management, Nagpur, India, in 2019. She is currently pursuing the Ph.D. degree in Computer Science and Engineering at Visvesvaraya National Institute of Technology (VNIT), Nagpur, India. Her research interests include Bioinformatics, Artificial Intelligence, Machine Learning, and genomic sequence analysis. Her current research focuses on the application of advanced machine-learning and deep-learning techniques for biomedical and genomic data analysis, including disease classification, viral genome analysis, phylogenetic interpretation, and explainable artificial intelligence. She is particularly interested in developing computational frameworks that integrate genomic information with intelligent learning models for accurate and interpretable biomedical prediction.



Deepti Deepak Shrimankar received the B.E. degree in Computer Technology from Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, India, in 1997, the MTech degree from G. H. Raisoni College of Engineering, Nagpur, India, in 2007, and the Ph.D. degree in Computer Science and Engineering from Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 2013. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Visvesvaraya National Institute of Technology (VNIT), Nagpur, India. She has several publications in international journals and conferences and has also served as a reviewer for many international journals

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

and conferences. Her research interests include bioinformatics, Internet of Things (IoT) and embedded systems, high-performance computing, machine and deep learning, computer and wireless networks, embedded and real-time systems, and image and video processing.



Aditi Rajratna Durge received the B.E. degree in Computer Science and Engineering from G. H. Raisoni Institute of Engineering and Technology (GHRIET), Nagpur, India, in 2015, and the MTech degree from Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 2018. She is currently pursuing the Ph.D. degree in Computer Science and Engineering at Visvesvaraya National Institute of Technology (VNIT), Nagpur, India. Her research interests include Bioinformatics, Artificial Intelligence, and Machine Learning. Her academic and research work focuses on applying computational and intelligent techniques to problems in genomic and biomedical data analysis. She is particularly interested in exploring machine-learning-based approaches for the analysis, interpretation, and classification of biological and genomic data, with an emphasis on developing efficient and reliable computational models for research applications.

Manuscript received 17 September 2026; Accepted 4 November 2026; Available online 20 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2105
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).