AI-Powered Classification and Early Detection of Dengue Virus Lineages for Timely Public Health Response

Emmanuel Nyandu Kagarabi [1 2] Dr. Joicymara Xavier [3]

[1]African Institute for Mathematical Sciences (AIMS South Africa) [2]International School for Advanced Studies (SISSA, Italy) [3]Instituto Tecnológico de Aeronáutica, Brazil

Introduction

Dengue is a widespread mosquito-borne virus, especially in tropical and subtropical regions. It is transmitted in a human–mosquito–human cycle (Figure 1).

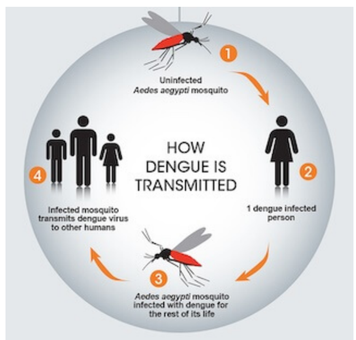

Figure 1. Dengue Transmission Cycle.

Dengue has a complex hierarchical structure : Four serotypes (DENV-1,...,DENV-4), each with genotypes and lineages(major and minor). Genomic surveillance is vital for tracking viral evolution and guide early public health responses. A typical dengue genomic sequence has ~10,700 nucleotides from the DNA alphabet $\Omega = \{A, C, G, T\}$. Traditionally, lineage identification relies on **alignment-based methods**(slow, expensive) while data keeps growing. **Alignment-free methods**: Machine Learning (ML) and Deep Learning (DL) approaches are promising alternatives.

Our main goal is to build ML/DL models to classify DENV sequences into lineages and detect potential new variants to strengthen genomic surveillance.

Methodology

Dataset

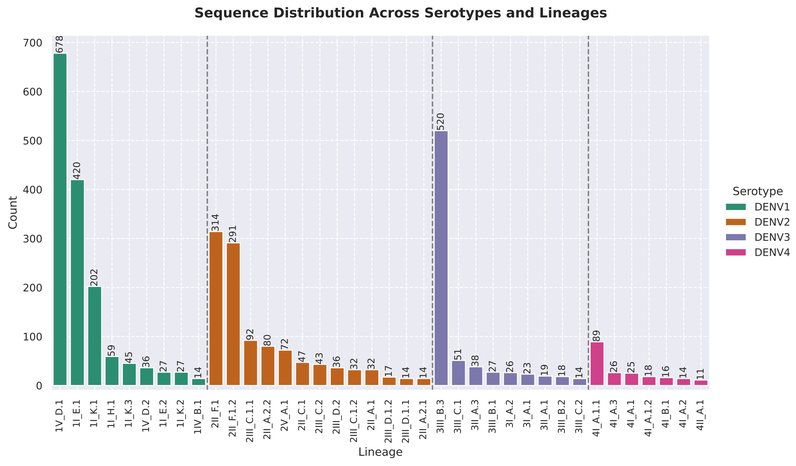

Figure 2. Dataset downloaded from GISAID and cleaned (3,527 samples)

Feature Extraction

One-hot encoding, k-mer encoding, Frequency Chaos Game Representation(FCGR):

$$\begin{cases} X_0 = \left(\frac{1}{2}, \frac{1}{2}\right) \\ X_i = \frac{1}{2}\left[X_{i-1} + \mathcal{C}(S_i)\right] \end{cases}, \mathcal{C}(S_i) = \begin{cases} (0,0), \text{if } S_i = \text{A}, \\ (0,1), \text{if } S_i = \text{C}, \\ (1,1), \text{if } S_i = \text{G}, \\ (1,0), \text{if } S_i = \text{T}. \end{cases}, \forall i \in \{1, 2, \ldots, n\}.$$

$\mathcal{C}(S_i)$ represents the coordinates of the nucleotide $S_i \in \Omega$:

Balancing

Used the Synthetic Minority Oversampling Technique (SMOTE) for class imbalance.

Models

Implemented Random Forest, XGBoost, LightGBM, Hierarchical Classifiers, 1D CNN with Self-Attention, 2D CNN using FCGR.

Metrics

Accuracy, F1-Score, Matthews Correlation Coefficient (MCC), Precision, Recall,...

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \ \text{MCC} \ \in [-1, 1].$$

Variant Detection

Trained a Variational Autoencoder (VAE) to generate synthetic samples (to mimic potential new variants) and suggested a Voting-based detection algorithm : A lineage is accepted if average confidence > 80%; otherwise, flagged as potential new variant.

Results

- Classification : All models achieved $\geq 97\%$ across all metrics.

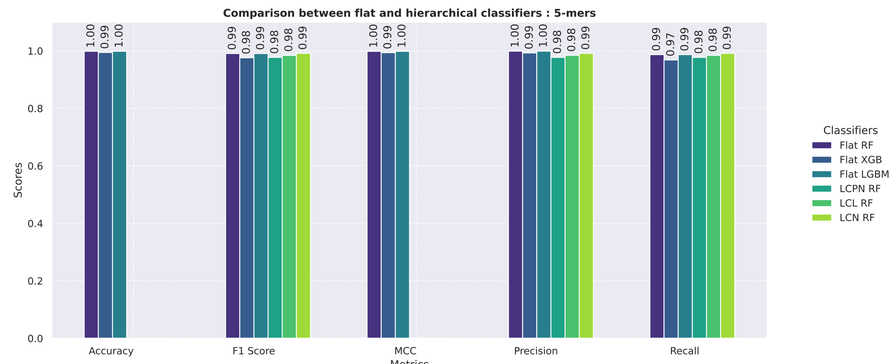

Figure 3. ML Models Results using 5-mers encoding as feature extraction.

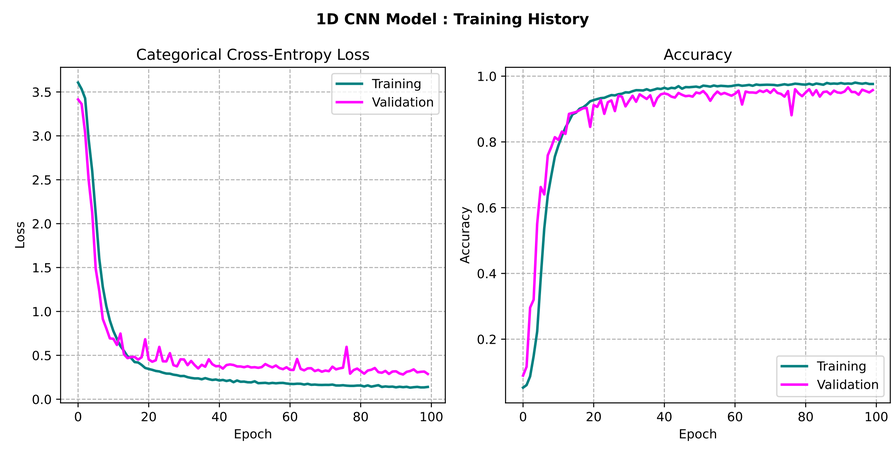

Figure 4. 1D CNN with Self-Attention using one-hot encoding as feature extraction.

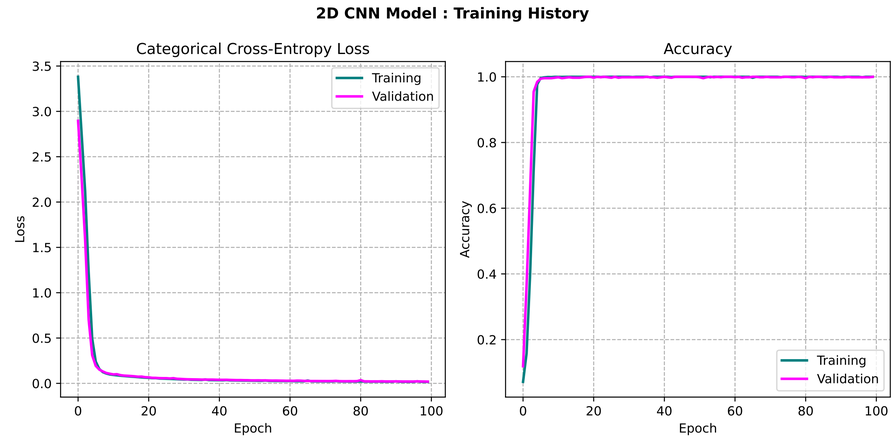

Figure 5. 2D CNN using FCGR (Resolution=32) as feature extraction.

- The VAE-generated sequences showed high matching scores with known Dengue lineages (validated through Nextclade).
- The outputs from our suggested Variant Detection Algorithm (voting from LightGBM, 1D CNN, and 2D CNN models) aligned with the Genome Detective tool, and did so in a shorter time.

Discussion

The proposed pipeline offers rapid and (reliable) classification of DENV sequences and timely detection of emerging variants. It complements traditional phylogenetic tools and might accelerates timely public health response.

In the future, Integrating interpretable and explainable AI techniques, alongside relevant biological knowledge of the pathogen, may enhance the reliability and credibility of both the models and their predictions.

References

[1] Verity Hill, Sara Cleemput, James Siqueira Pereira, Robert J Gifford, Vagner Fonseca, Houriiyah Tegally, Anderson F Brito, Gabriela Ribeiro, Vinicius Carius de Souza, Isabela Carvalho Brcko, et al.
A new lineage nomenclature to aid genomic surveillance of dengue virus.
PLoS biology, 22(9):e3002834, 2024.

[2] Emmanuel Nyandu Kagarabi.
Ai-powered classification and early detection of dengue virus lineages for timely public health response.
Master's thesis, African Institute for Mathematical Sciences (AIMS), Cape Town, South Africa, October 2024.
Full thesis available here.

DEEP LEARNING INDABA Kigali 2025 GitHub emmanuelnyandukagarabi LinkedIn Gmail emmanuelnk@aims.ac.za