Coronaviridae Classifier & Mutation Descriptor: An Explainable BiLSTM-Based Approach for Variant Classification and Mutation Analysis Without Reference Genomes

Problem Statement and Motivation:

The rapid evolution of RNA viruses, particularly within the Coronaviridae family, poses significant challenges for accurate classification and mutation detection. Traditional methods often rely on reference genomes, which may not be available or up-to-date for emerging variants. This project aims to develop an explainable deep learning model that can classify viral variants and identify mutation hotspots without the need for reference genomes, thereby facilitating timely and accurate viral surveillance.

Introduction:

This project presents a novel approach to classify variants within the Coronaviridae family using a Bidirectional Long Short-Term Memory (BiLSTM) model. By leveraging deep learning techniques, the model not only classifies viral sequences but also identifies mutation hotspots through pattern-based inference. The incorporation of explainability methods ensures that the model's predictions are interpretable, enhancing trust and facilitating further biological insights.

Base Paper and Related Works:

- **Preprocessing:**
  Gene Sequence to 2D Vector Transformation for Virus Classification
  Link

- **Model Architecture:**
  BiLSTM-5mC: A Bidirectional Long Short-Term Memory-Based Approach for Predicting 5-Methylcytosine Sites in Genome-Wide DNA Promoters
  Link

- **Explainability:**
  Explainable Deep Neural Networks for Novel Viral Genome Prediction
  Link
- **Mutation Detection:**
  SPLASH: A Statistical, Reference-Free Genomic Algorithm Unifies Mutation Detection
  Link

- **Viral Genome Classification:**
  Classification of Highly Divergent Viruses from DNA/RNA Sequence Using Transformer-Based Models
  Link

Datasets:

- **Original Dataset:**
  Collected 1,900+ viral genomic sequences from the NCBI Virus database, encompassing various Coronaviridae variants such as MERS-CoV, SARS-CoV-2, HCoV-NL63, HCoV-HKU1, HCoV-OC43, HCoV-229E, and SARS-CoV.

- **Preprocessed Dataset:**
  Sequences were encoded using the mapping {"A": 1, "C": 2, "G": -1, "T": -2} and stored in `.npz` format for efficient loading and processing.

- **Source Link:**
  NCBI Virus Database

Model Architecture:

The model employs a Sequential architecture with the following layers:

1. **Embedding Layer:**
   Transforms input sequences into dense vector representations.

2. **Bidirectional LSTM Layers:**
   Captures contextual information from both forward and backward directions, enhancing the model's understanding of sequence dependencies.

3. **Dropout Layers:**
   Prevents overfitting by randomly deactivating neurons during training.

4. **Dense Layers:**
   Performs classification based on the features extracted by the LSTM layers.

5. **Output Layer:**
   Utilizes a softmax activation function to output probabilities for each virus class.

The model is compiled with the Adam optimizer and trained using the sparse categorical cross-entropy loss function. Mixed precision training and XLA (Accelerated Linear Algebra) compilation are employed to optimize performance on compatible hardware.

Flow of our model:

[Data Input]

↓

[Embedding Layer]

↓

[Bidirectional LSTM Layer 1 (128 units)]

↓

[Dropout Layer (0.3)]

↓

[Bidirectional LSTM Layer 2 (64 units)]

↓

[Dense Layer (64 units, ReLU)]

↓

[Dropout Layer (0.3)]

↓

[Dense Output Layer (lenVirusList units, Softmax)]

↓

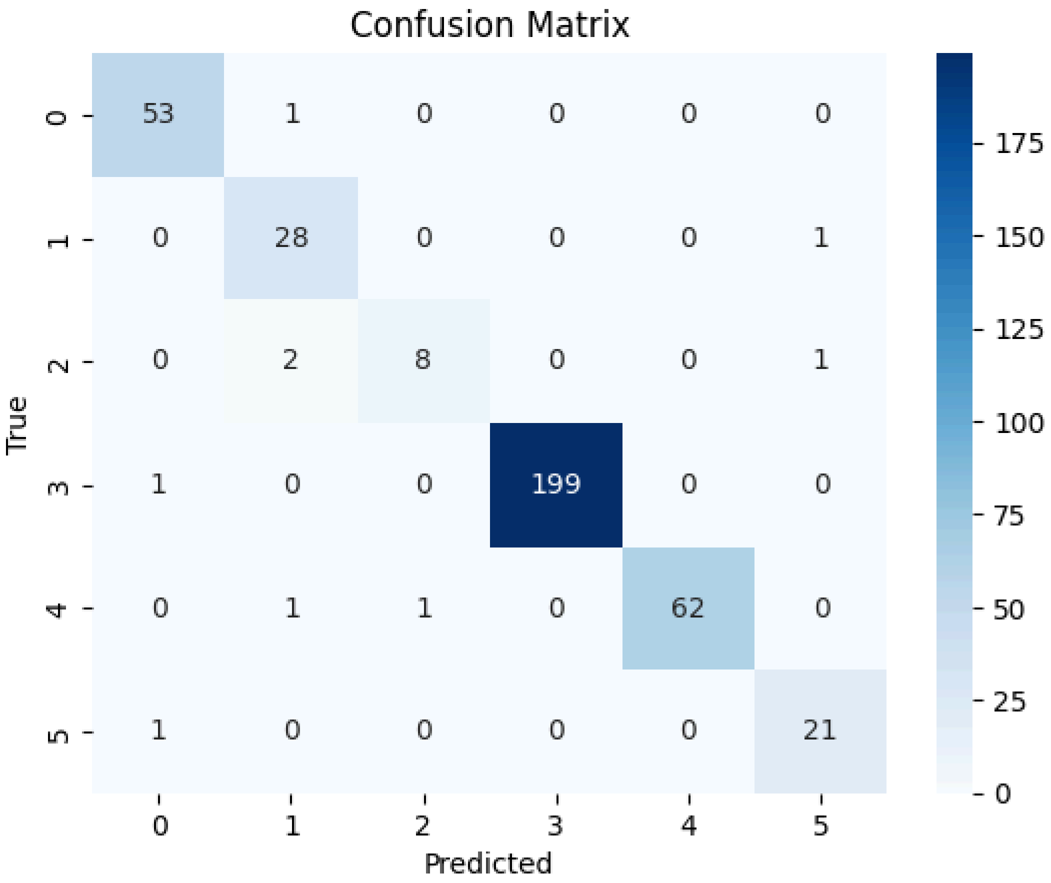
[Prediction: Virus Class]

Results and Significance:

The model achieved high accuracy in classifying various Coronaviridae variants, demonstrating its effectiveness in handling diverse viral sequences. The integration of explainability techniques provided insights into the model's decision-making process, highlighting the most influential genomic positions contributing to each classification.

We achieved a **Test Accuracy** of **97.63%** after 10 epochs.

Below is the **Confusion Matrix** we plotted.



Mutation Detection Without Reference Genome (Pattern-Based Inference):

By analyzing nucleotide variations across sequences within the same class, the model identified mutation hotspots without relying on a reference genome. For instance, in Virus Class 0, position 48 exhibited significant variability, indicating a potential mutation hotspot.

We found top **5 mutation hotspots** for every class.

For example:

- Virus Class 0: Top Mutation Positions-

Position 0: {np.int32(-2): 6, np.int32(2): 8, np.int32(1): 19, np.int32(-1): 22}

Position 7: {np.int32(-2): 27, np.int32(1): 3, np.int32(-1): 22, np.int32(2): 3}

… So on for top 5 positions in every class.

Explainability: Why the Model Predicts a Specific Virus:

Using gradient-based methods, the model highlighted the top influential positions in the genome that contributed to its predictions. This transparency aids in understanding the biological relevance of specific genomic regions in virus classification.

We found -

Top influential positions: [31103 31102 0 31101 1 31100 2 3 31099 4]

Comparison Graphs and Tables:

- **Confusion Matrix:**
  Illustrated the model's performance across different classes, highlighting areas of high accuracy and potential misclassifications.

- **Precision, Recall, and F1-Score:**
  Provided a comprehensive evaluation of the model's classification capabilities, ensuring balanced performance across all classes.

- **Mutation Frequency Charts:**
  Visualized the frequency of mutations at each genomic position, facilitating the identification of mutation hotspots.

Merits and De-merits:

Merits:

Reference-Free Mutation Detection:
Enables the identification of mutation hotspots without the need for a reference genome.

- **Explainability:**
  Enhances trust in the model's predictions by providing insights into its decision-making process.

- **Scalability:**
  Efficiently handles large datasets, making it suitable for real-time viral surveillance.

De-merits:

- **Limited Data for Certain Classes:**
  Some virus classes had fewer samples, potentially affecting the model's performance for those classes.

- **Encoding Limitations:**
  The chosen nucleotide encoding may not capture all the complexities of genomic sequences.

Full Code and Execution Procedure:

1. **Data Collection:**
   Use NCBI accession numbers to fetch viral sequences.

2. **Data Preprocessing:**
   Encode sequences using the specified mapping and store them in `.npz` format.

3. **Model Training:**
   Train the BiLSTM model using the preprocessed data.

4. **Evaluation:**
   Assess the model's performance using appropriate metrics.

5. **Mutation Detection:**
   Analyze sequences within each class to identify mutation hotspots.

6. **Explainability Analysis:**
   Utilize gradient-based methods to determine influential genomic positions.

Conclusion:

This project presents a comprehensive approach to classifying Coronaviridae variants and

identifying mutation hotspots without relying on reference genomes. The integration of explainability techniques enhances the model's transparency, making it a valuable tool for viral surveillance and research.

Future Scope:

- **Incorporation of Additional Data:**
  Expanding the dataset to include more viral sequences can improve the model's robustness.

- **Advanced Encoding Techniques:**
  Exploring alternative encoding methods may capture more intricate patterns in genomic data.

- **Real-Time Surveillance:**
  Deploying the model in real-time systems can aid in the early detection of emerging viral variants.

References:

1. **Explainable Deep Neural Networks for Novel Viral Genome Prediction**
   [Link](LINK)