APPLICATION OF BILSTM FOR PROTEIN-PROTEIN INTERACTION PREDICTION

Project Submitted to the
SRM University AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
in
Computer Science & Engineering
School of Engineering & Sciences

submitted by

Sandeep Reddy Yaram (AP21110011538)

Chandra Sekhar Naidu Vattem(AP21110011576)

Sai Kumar Mundru (AP21110011624)

Under the Guidance of

Dr. Anirban Bhar

Department of Computer Science & Engineering
SRM University-AP
Neerukonda, Mangalgiri, Guntur
Andhra Pradesh - 522 240
May 2025

DECLARATION

I undersigned hereby declare that the project report **Application of BiLSTM for Protein-Protein Interaction Prediction** submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology in the Computer Science & Engineering, SRM University-AP, is a bonafide work done by me under supervision of Dr. Anirban Bhar . This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree of any other University.

Place	:	Date	: April 27, 2025
Name of student	: Sandeep Reddy Yaram	Signature	: Y. Sandeep Reddy
Name of student	: Chandra Sekhar Naidu Vattem	Signature	: V. Chandra
Name of student	: Sai Kumar Mundru	Signature	: M. Sai kumar.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SRM University-AP
Neerukonda, Mangalgiri, Guntur
Andhra Pradesh - 522 240

CERTIFICATE

This is to certify that the report entitled **Application of BiLSTM for Protein-Protein Interaction Prediction** submitted by **Sandeep Reddy Yaram , Chandra Sekhar Naidu Vattem, Sai Kumar Mundru** to the SRM University-AP in partial fulfillment of the requirements for the award of the Degree of Master of Technology in in is a bonafide record of the project work carried out under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Project Guide
Name : Dr. Anirban Bhar
Signature: Anirban Bhar ....

Head of Department
Name : Dr. Murali Krishna Enduri
Signature: ....

ACKNOWLEDGMENT

I wish to record my indebtedness and thankfulness to all who helped me prepare this Project Report titled **Application of BiLSTM for Protein-Protein Interaction Prediction** and present it satisfactorily.

I am especially thankful for my guide and supervisor Dr. Anirban Bhar in the Department of Computer Science & Engineering for giving me valuable suggestions and critical inputs in the preparation of this report. I am also thankful to Dr. Murali Krishna Enduri, Head of Department of Computer Science & Engineering for encouragement.

We extend our sincere thanks to all the faculty members in the CSE Department. We are also grateful to our classmates and their cheerful nature. We are also thankful to everyone who has helped us in this project. Special Thanks to our parents for their unwavering support and unwavering belief in our abilities and trust in us.

Lastly, we are extremely grateful to SRM University, Andhra Pradesh, for providing us with a platform to grow both academically and personally. The support and kindness from all the faculty members in the CSE Department have played a pivotal role in the success of our project and research. And our special thanks to Vice-Chancellor sir and Dean sir.

Sandeep Reddy Yaram , Chandra Sekhar Naidu Vattem, Sai Kumar Mundru ,
(Reg. No. AP21110011538, AP21110011576, AP21110011624)
B. Tech.
Department of Computer Science & Engineering
SRM University-AP

ABSTRACT

Protein–protein interactions (PPIs) between viral and human proteins play a vital role in the development and progression of infectious diseases. Viruses such as the dengue virus often exploit host cell processes by interacting with crucial human proteins, aiding in viral replication, immune system evasion, and the onset of disease symptoms. Gaining insight into these interactions is essential to understand host-pathogen relationships, disruptions in cellular signaling pathways, and identifying potential targets for antiviral therapies.

In this work, we introduce a deep learning-based approach designed to predict interactions between human and dengue virus proteins using only their amino acid sequences. This approach is inspired by recent advancements in sequence-based deep learning methods for protein interaction prediction [1]. The method involves representing protein sequences through k-mer encoding, which captures localized sequence motifs that are typically significant for binding specificity. Separate bidirectional long-short-term memory (BiLSTM) models are independently trained to generate 128-dimensional vector embeddings for both human and viral proteins, effectively learning meaningful sequence patterns.

These embeddings are then combined with a dataset containing experimentally verified interacting and non-interacting protein pairs. A fully connected neural network classifier is trained on the concatenated embeddings to predict the likelihood of interaction between a given human–dengue protein pair.

The proposed framework demonstrates strong predictive capabilities and offers a scalable, sequence-based, and alignment-free solution for host–virus PPI prediction. By facilitating the rapid identification of potential protein interactions, this method can help prioritize candidates for laboratory validation, deepen our understanding of dengue virus pathogenesis, and support the discovery of new antiviral therapeutic targets.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION TO THE PROJECT

Protein–protein interactions (PPIs) are fundamental to virtually all biological processes, including immune responses, cellular signaling, and pathogen-host interactions [3]. Viruses frequently employ host proteins to aid in their reproduction and elude immune responses in the context of infectious illnesses. Determining how viral and host proteins interact is therefore crucial to comprehending the molecular underpinnings of infection and creating efficient treatment plans. The dengue virus, a common arbovirus worldwide, causes millions of cases each year and presents serious public health issues. Comprehensive information of dengue–human protein interactions is still lacking, despite advancements in clinical care and vaccine research [7].

Co-immunoprecipitation and yeast two-hybrid screening are two expensive, time-consuming, and labor-intensive experimental techniques for PPI detection [4]. As a result, computer methods have drawn interest as effective, scalable substitutes for forecasting possible interactions between proteins. Conventional computational techniques frequently depend on manually curated interaction databases, structural similarity, or sequence alignment, which may not translate well to new protein sequences or emerging diseases.

In this work, we introduce a deep learning system that uses solely amino acid sequence information to predict human–dengue protein interactions. In our method, protein sequences are converted into k-mer representations, local sequential patterns are captured, and two bidirectional long short-term memory (BiLSTM) networks—one for dengue proteins and one for human proteins—are used to independently generate embeddings [6].These embeddings are then concatenated and passed through a fully connected neural network to predict the likelihood of interaction between the protein pairs. This sequence-based, alignment-free method offers a rapid and flexible tool for exploring host–virus interactomes and could contribute to identifying novel therapeutic targets in dengue virus infection [5].

Chapter 2

MOTIVATION

The primary use case for the project's output is the established computational framework itself. It can be used by researchers studying the pathophysiology of the Dengue virus to: This project, **Application of BiLSTM for Protein-Protein Interaction**, is motivated by the urgent need to expedite the identification of protein–protein interactions (PPIs) between viral and human proteins, particularly for infectious diseases like dengue. Conventional experimental techniques for PPI detection, like co-immunoprecipitation and yeast two-hybrid screening, are expensive and time-consuming, which restricts their scalability. Our team set out to address this real-world problem by fusing state-of-the-art machine learning techniques, specifically word2vec embeddings and BiLSTM networks, with biological sequence analysis. We had the opportunity to apply theoretical knowledge to a real-world, multidisciplinary challenge at the nexus of molecular biology and computer science thanks to this capstone project. We want to support the larger scientific endeavor to understand viral pathophysiology and discover new therapeutic targets by focusing on sequence-based, alignment-free prediction of PPIs. The project improved our skills in technical documentation, project planning, and group research while also fostering creativity and self-expression.

Chapter 3

LITERATURE SURVEY

3.1 INTRODUCTION: THE ROLE OF HUMAN-VIRUS PPIS

Protein-protein interactions (PPIs) are essential for almost all biological processes, including cellular communication and immunological responses. In the context of infectious diseases, interactions between viral and host proteins—often referred to as human-virus PPIs or HV-PPIs—are particularly significant. Viruses like Dengue Virus (DENV), which are required intracellular parasites, use the host's cellular apparatus to complete their life cycle and proliferate. They achieve this by interacting with specific host proteins to control processes like viral component movement, cell cycle regulation, and apoptosis. Understanding these molecular interactions is essential for determining host immunological responses, viral pathogenesis processes, and potential targets for antiviral therapy. Despite the importance, especially for diseases like DENV that are prevalent worldwide, little is known about the specific HV-PPIs at work.[2].

3.2 CHALLENGES IN PPI DETECTION

Traditional experimental techniques have been widely employed to detect PPIs, including co-immunoprecipitation followed by mass spectrometry, molecular dynamic simulations (restricted to proteins whose 3D structures are known), and yeast two-hybrid (Y2H) screens. Nevertheless, these methods are frequently costly, time-consuming, and labor-intensive, and they may have sensitivity or specificity issues. It is typically impractical to use them extensively to map whole host-virus interactomes. [4].

3.3 COMPUTATIONAL APPROACHES FOR PPI PREDICTION

Numerous computational techniques have been developed as effective and scalable substitutes for experimental methods in order to overcome their limitations in the prediction of possible PPIs.

3.3.1 Traditional Computational Methods

Earlier computational techniques frequently depended on data other than the core sequence, including structural similarities or databases of interactions already in existence. These methods, however, might not be able to anticipate interactions involving novel proteins or emerging diseases, such as new virus strains, or when structural information is absent.[1].

3.3.2 Sequence-Based Machine Learning Methods

Due to the quantity of sequence data and developments in machine learning (ML), prediction techniques that just use protein amino acid sequences have become increasingly popular in recent years. These techniques seek to extract patterns from the sequences that suggest the possibility of interaction.

- Feature Encoding:Converting variable-length protein sequences into fixed-size numerical representations appropriate for machine learning models is a major difficulty. Numerous encoding systems have been investigated, including:

  - Composition-based: K-mer frequencies (as employed in our experiment), dipeptide composition, or amino acid composition can be used.
  - Physicochemical Properties: Incorporating properties of amino acids.
  - Evolutionary Information: Using Position-Specific Scoring Matrices (PSSMs) derived from multiple sequence alignments.
  - Sequence Embedding: By treating amino acid sequences or k-mers as ”words” and ”sentences,” natural language processing techniques like word2vec and doc2vec have been modified to produce dense vector representations (embeddings) that capture sequence context. [5].

- Machine Learning Models: Various ML algorithms have been applied to the encoded features:
  - Support Vector Machines (SVM): Used with various encoding methods, like amino acid triplets and composition.
  - Random Forests (RF): Used in combination with doc2vec embeddings.
  - Deep Learning:Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM), have demonstrated potential in deep learning. While LSTMs are skilled at identifying long-range dependencies in sequential data, CNNs are better at identifying local patterns.

3.4 POSITIONING OUR PROJECT

Our project aligns with the state-of-the-art focus on sequence-based deep learning methods for HV-PPI prediction, specifically targeting Dengue virus[1]. Key aspects include:

1. Sequence-based: Relies only on amino acid sequences, making it broadly applicable.

2. Deep Learning: Employs BiLSTMs, known for capturing sequential patterns.

3. Embedding Strategy: Uses BiLSTM *autoencoders* trained independently on human and dengue k-mer sequences to generate embeddings, differing from LSTM-PHV's use of word2vec.

4. Classifier: Uses a fully connected neural network on concatenated embeddings, similar in principle to the final classification network in LSTM-PHV.

This approach provides an alternate deep learning methodology for the quick and scalable prediction of DENV-human PPIs by utilizing BiLSTM autoencoders for feature extraction from k-mers and a specialized classifier.

Chapter 4

DESIGN AND METHODOLOGY

This work intends to build and assess a computational model for protein-protein interaction (PPIs) between host proteins and Dengue virus (DENV) proteins. Whether deep learning models—more especially, Bidirectional Long Short-Term Memory (BiLSTM) networks and Fully Connected Neural Networks (FCNNs)—can efficiently predict these interactions using just amino acid sequence data is the main focus of research question.[1].

The overall design follows a three-stage deep learning approach:

1. **K-mer Embeddings:**Pre-trained vector representations help to embed overlapping 4-mers from protein sequences.

2. **Embedding Generation:**To generate sequence-level representations, bi-LSTM autoencoders are taught independently on DENV and host protein embeddings.

3. **Interaction Classification:**Concatenated protein embeddings are used to train an FCNN classifier, which predicts binary interaction labels.

This sequence-based, alignment-free approach offers scalability and does not require known structural or homology information.

4.1 MODEL COMPONENTS AND DATASET

4.1.1 Bidirectional Long Short-Term Memory (BiLSTM) Networks

BiLSTM networks are useful for modeling protein sequences when the context from both directions is biologically significant because they can process sequences both forward and backward.[4].

Each LSTM unit updates its states using the following equations:

$$\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget gate)} \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input gate)} \\
\tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad \text{(Candidate cell state)} \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \text{(Cell state update)} \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output gate)} \\
h_t &= o_t \odot \tanh(c_t) \quad \text{(Hidden state update)}
\end{aligned}$$

In a BiLSTM, the final output is the concatenation of forward and backward hidden states:

$$h_t^{\text{bi}} = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

4.1.2 Fully Connected Neural Networks (FCNN)

A FCNN layer transforms an input vector x using:

$$y = f(Wx + b)$$

Activation Functions:

- **ReLU:** $\text{ReLU}(x) = \max(0, x)$ — Used in hidden layers.

- **Sigmoid:** $\sigma(x) = \frac{1}{1+e^{-x}}$ — Used in the output layer.

Regularization with Dropout:

$$\text{Dropout}(x) = \begin{cases} 0 & \text{with probability } p \\ \frac{x}{1-p} & \text{otherwise} \end{cases}$$

Loss Function: Binary Cross-Entropy

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

4.1.3 Dataset Details

Data Sources and Preprocessing:

- **Positive Samples:** Known human-Dengue virus (DENV) protein-protein interactions (PPIs) were collected from curated databases such as HPIDB and UniProt [7, 8]. The protein sequences with more than 90% sequence similarity were eliminated using the CD-HIT tool to cut down on redundancy and possible bias from highly similar sequences.

- **Negative Samples:** Subcellular localization data for the human and DENV proteins were used to create high-confidence negative samples. Proteins in different cellular compartments are less likely to interact, according to the reasoning. Examples of negative interactions were chosen from pairs of human and DENV proteins that are known to reside in different subcellular locations. In order to achieve an initially balanced dataset of interacting and probably non-interacting pairs, the number of such negative pairs was calculated to equal the number of positive samples.

- **K-mer Embeddings Preparation:** First, overlapping subsequences of length k=4 (4-mers) were created from protein sequences. A pre-trained Word2Vec model was then used to map these 4-mers to dense vector representations. In particular, 128-dimensional embedding vectors were created for every distinct 4-mer using the Continuous Bag-of-Words (CBOW) architecture, which was trained over 100 iterations.

- **Fixed Input Representation:** For every protein, a fixed-length representation was made to be fed into the BiLSTM models. The 100 most common 4-mers in a protein sequence were selected from the list of all 4-mers. A matrix was created by retrieving their matching 128-dimensional Word2Vec embeddings. To ensure that each protein had a consistent input dimension of 100×128, this matrix was either truncated or padded with zero vectors if fewer than 100 distinct top k-mers were discovered.

- **Protein-Level Embedding Extraction:** Each protein's 100×128 k-mer embedding matrix was fed into the corresponding trained BiLSTM autoencoder (one model for DENV, one for humans). The encoder portion's output produced a final, fixed-length, 128-dimensional embedding vector that represented the entire protein after averaging over the 100 time steps.

- **Classification Dataset Construction:** Human and DENV proteins were paired to create the final dataset that was used to train the interaction classifier. A 256-dimensional feature vector was produced by concatenating the corresponding 128-dimensional BiLSTM embeddings for each pair. The generated negative pairs were given a binary label of 0; known positive interactions were given a label of 1.

4.1.4 Evaluation Measures

To evaluate model performance, the following metrics were used: sensitivity (recall), specificity, accuracy, AUC, and AUPRC [1].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(t)\right) dt$$

$$\text{AUPRC} = \int_0^1 \text{Precision}(r)\, d(\text{Recall}(r))$$

Where:

$$\text{TP} = \text{True Positives}, \quad \text{TN} = \text{True Negatives}$$

$$\text{FP} = \text{False Positives}, \quad \text{FN} = \text{False Negatives}$$

$$\text{TPR} = \text{True Positive Rate (Sensitivity)}$$

$$\text{FPR} = \text{False Positive Rate (1 - Specificity)}$$

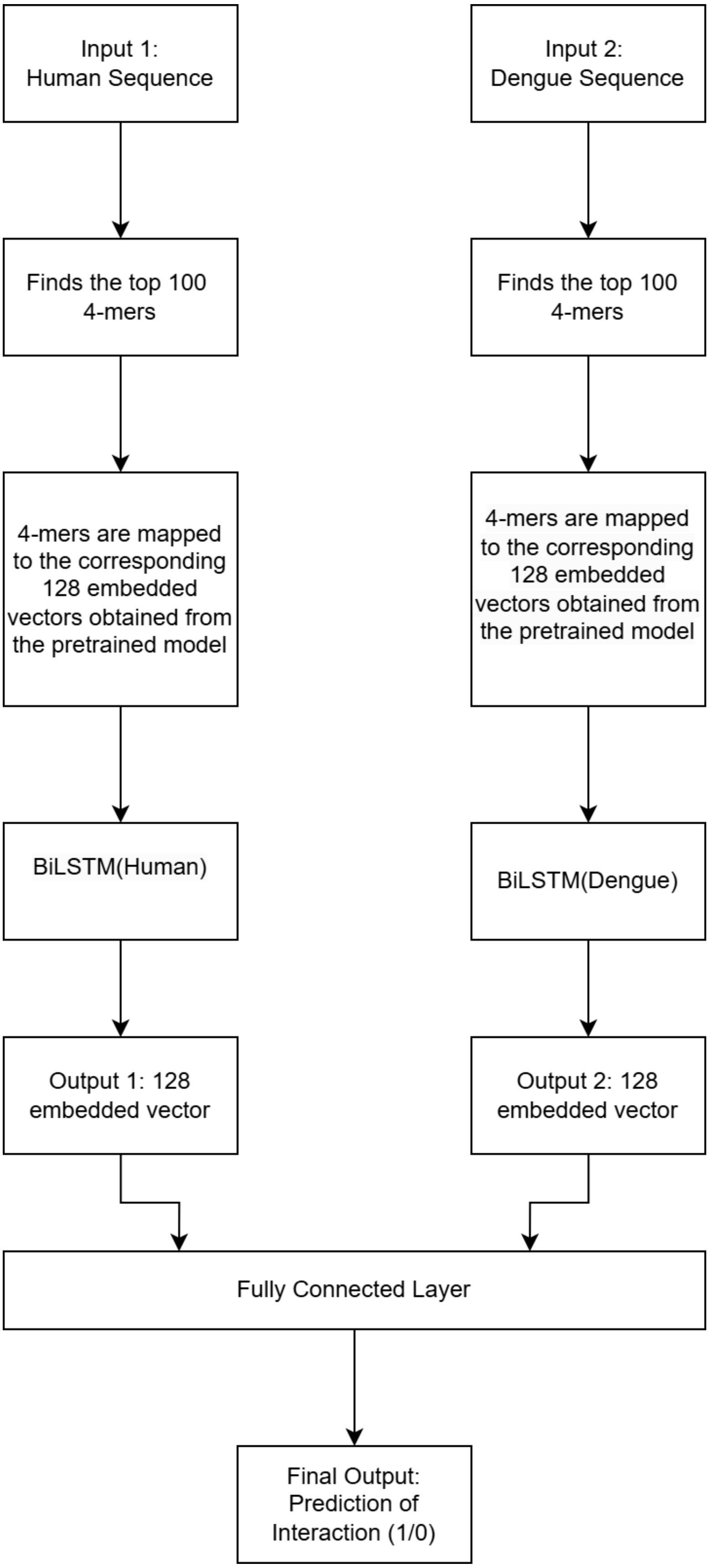
Input 1:
Human Sequence
Input 2:
Dengue Sequence
Finds the top 100 4-mers
Finds the top 100 4-mers
4-mers are mapped to the corresponding 128 embedded vectors obtained from the pretrained model
4-mers are mapped to the corresponding 128 embedded vectors obtained from the pretrained model
BiLSTM(Human)
BiLSTM(Dengue)
Output 1: 128 embedded vector
Output 2: 128 embedded vector
Fully Connected Layer
Final Output: Prediction of Interaction (1/0)

Figure 4.1: Workflow of PPI Prediction using BiLSTM and FCNN

Chapter 5

IMPLEMENTATION

5.1 DATA ACQUISITION AND PREPARATION

A dataset of known interacting and non-interacting human-DENV protein pairs with their sequences and binary interaction labels (from `Interactions.csv`) as well as amino acid sequences for human proteins (from `Human.csv`) and dengue virus proteins (from `Dengue.csv`) comprised the main input data. Additionally, `kmers_embedding_human_dengue.txt` was imported to provide 128-dimensional pre-computed embeddings for 4-mers formed from these sequences. The `load_kmer_embeddings` function was used to load the embeddings into a dictionary that mapped each 4-mer string to its corresponding numerical vector.

5.2 FEATURE ENGINEERING: K-MER SEQUENCE REPRESENTATION

The `process_sequence` function was used to process each protein sequence in order to provide a matrix representation that could be fed into BiLSTM models. All overlapping subsequences of length 4 (4-mers) were to be extracted, the top 100 most frequent 4-mers had to be chosen, and each had to be mapped to its appropriate 128-dimensional embedding. A zero vector was employed in the absence of a 4-mer. To guarantee that (100, 128) had a consistent shape, the output was padded or truncated. An all-zero matrix was returned by short sequences.

5.3 HYPERPARAMETER TUNING OF LSTM AUTOENCODER MODELS

Key hyperparameters were tuned using Keras Tuner. These included:

- **Input sequence length:** Defined using `hp.Choice('max_kmers', ...)`. For human proteins: [100, 200, 300], for dengue: [100, 150, 200].
- **LSTM Units:** Ranged from 32–128 for the first BiLSTM layer and 16–64 for the second.
- **Dropout Rate:** Tuned between 0.0–0.5 to mitigate overfitting.
- **Learning Rate:** Chosen logarithmically between 1e-4 and 1e-2.

A matrix with the shape `(max_kmers, 128)` was accepted by the model input layer. To avoid overfitting, dropout layers were interleaved between three stacked BiLSTM layers. For reconstruction, a `TimeDistributed(Dense(128, activation='relu'))` layer was employed.

Protein k-mer embeddings were used as the input and target for each model configuration during training, along with the Adam optimizer and mean squared error loss. The best weights were restored by using early stopping. A vector of 128 dimensions per protein was obtained by averaging the final embeddings across the time dimension.

5.4 FINAL EMBEDDING GENERATION: BILSTM AUTOENCODERS

Two separate BiLSTM autoencoders were built for human and DENV proteins using the `build_bilstm_autoencoder` function. Input was the (100, 128) k-mer matrix. The encoder had stacked BiLSTM layers (e.g., 64 and 32 units) with dropout. The decoder mirrored the encoder, followed by a `TimeDistributed(Dense(128, activation='relu'))` output layer.

The Adam optimizer and MSE loss were used to compile each model. With batch sizes of 64 and 22, respectively, training was conducted for 180 epochs (for humans) and 300 epochs (for dengue). EarlyStopping tracked the loss of training. `model.predict(X)` and `np.mean(embeddings, axis=1)` were used to obtain the final embeddings.

Outputs were saved to `Human_BiLSTM_Final_Embeddings.csv` and `Dengue_BiLSTM_Final_Embeddings.csv`, and models were saved in `.keras` format.

5.5 INTERACTION PREDICTION: FULLY CONNECTED NEURAL NETWORK (FCNN)

The `load_and_preprocess`function loaded the dataset called `Interactions.csv`. Based on similar sequences, human and dengue embeddings were combined to create a dataset with 256 features (128 each) and a binary label. The dataset was divided into 80% training and 20% testing using `train_test_split`.

The FCNN architecture with input size 256, dense layers (512, 256, 128 units), Batch Normalization, Dropout (0.2), and a final sigmoid output was defined by the `build_model` function.

The accuracy metric, binary crossentropy, and Adam (lr=0.0001) were used to create the model. 300 epochs and a batch size of 32 were used for training, and the best model was stored to `best_model.keras` by `ModelCheckpoint`.

5.6 MODEL EVALUATION

From `best_model.keras`, the best model was loaded. On both training and test sets, the `evaluate_model` function computed accuracy, sensitivity, specificity, AUC, and AUPRC. The saved file `fc_final_model.keras` included the final trained model.

Chapter 6

HARDWARE/ SOFTWARE TOOLS USED

6.1 SOFTWARE ENVIRONMENT

The Python computer language (version 3.x suggested) was used to develop the Human-Dengue Protein-Protein Interaction prediction framework. The following open-source libraries were crucial to the project's success in data manipulation, machine learning, and deep learning tasks:

- TensorFlow/Keras: (Version 2.x) Served as the core deep learning framework. It was used for:
  - Defining the architecture of the Bidirectional LSTM (BiLSTM) autoencoders (`Input`, `Bidirectional`, `LSTM`, `Dense`, `TimeDistributed`, `Dropout`, `Model`).
  - Defining the architecture of the Fully Connected Neural Network (FCNN) classifier (`Sequential`, `Input`, `Dense`, `BatchNormalization`, `Dropout`).
  - Compiling the models (`compile`) with specified optimizers (Adam) and loss functions (Mean Squared Error for autoencoders, Binary Crossentropy for classifier).
  - Managing training processes with callbacks (`EarlyStopping` , `ModelCheckpoint`).
  - We have used keras-tuner (For hyper tuning kmers, Dropout, Learning rate etc.)

- Scikit-learn (sklearn): Provided crucial machine learning utilities, including:
  - Splitting datasets into training and testing sets (model_selection. train_test_split)
  - Calculating performance evaluation metrics such as confusion matrix, ROC AUC score, and Precision-Recall AUC (`metrics`).
- Pandas: Used extensively for data handling and preprocessing:
  - Loading datasets from CSV files (`read_csv`).
  - Manipulating data structures (`DataFrame`).
  - Merging different datasets based on protein sequences (`merge`).
  - Handling duplicate entries (`drop_duplicates`).
- NumPy: Provided fundamental support for numerical operations:
  - Creating and manipulating multi-dimensional arrays required for model inputs and outputs.
  - Performing mathematical operations, such as averaging embeddings (`mean`).
- Matplotlib: Used for visualizing the training progress, specifically plotting accuracy and loss curves over epochs (`pyplot`).
- Pickle: Utilized for saving and loading the Python dictionary containing the k-mer embeddings (`dump`, `load`).

6.2 HARDWARE CONFIGURATION

The hardware setup required for our project includes:

- Processor (CPU): Data preprocessing, workflow management, and possibly model inference or smaller-scale training are the typical uses for a multi-core CPU (such as the AMD Ryzen series or Intel Core i5/i7).

- Memory (RAM): Large dataset loading and the storage of intermediate data structures and model parameters during training necessitate a substantial amount of RAM (e.g., 16GB, 32GB, or more). The quantity of the protein sequence data, embeddings, and neural network model complexity all affect the precise requirement.

- Graphics Processing Unit (GPU) (Recommended) : To speed up the training process, high-performance GPUs (like the NVIDIA GeForce RTX series, Quadro, or Tesla) with a sizable amount of VRAM (such 8GB, 12GB, 24GB, or more) are essential.

- Storage: To store the datasets, code, created embeddings, and stored model files, there must be enough disk space (ideally on an SSD for faster data access).

Chapter 7

RESULTS & DISCUSSION

The results of applying the methods outlined in Chapter 4 are presented in this chapter. The findings include the performance of the Fully Connected Neural Network (FCNN) classifier trained for predicting Human-Dengue protein-protein interactions (PPIs) and the training results of the Bidirectional LSTM (BiLSTM) autoencoders utilized for protein embedding creation.

7.1 BILSTM TRAINING AND EMBEDDING GENERATION

Two distinct BiLSTM autoencoder models were built and trained, one for Dengue virus protein sequences and one for human protein sequences, utilizing the `kmers_embedding_human_dengue.txt` file′s 4-mer representations.

- Training Process: The ′adam′ optimizer and the′mean squared error′ (MSE) loss function were used to train both models. Early halting was used in response to the training loss to avoid overfitting.
  - The training loss (MSE) was plotted against the training epochs to track the Human BiLSTM autoencoder′s training progress. The reduction in reconstruction error as the model learned to represent the k-mer sequences of human proteins is graphically depicted in this plot.

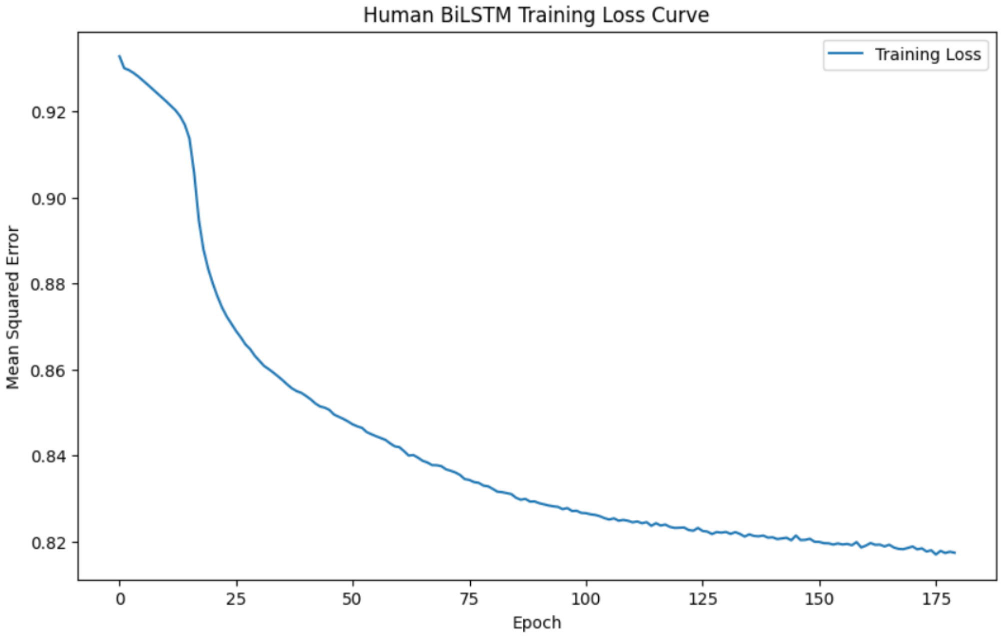

Figure 7.1: BiLSTM Loss Curve for Human Sequences

- Plotting the Dengue BiLSTM autoencoder's training loss (MSE) against training epochs allowed for the tracking of the training process.

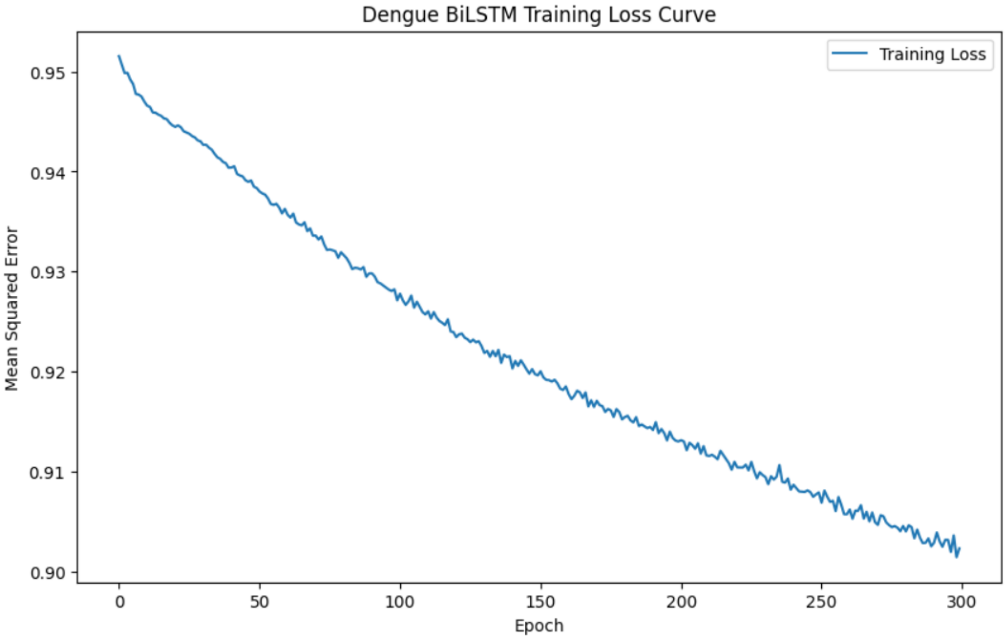

Figure 7.2: BiLSTM Loss Curve for Dengue Sequences

- Embedding Generation: The encoder sections of the top-performing autoencoder models (based on early stopping) were employed when training was finished. The corresponding encoders received the processed protein sequences, which had 100 k-mers each. To create a single 128-dimensional vector embedding for every protein, the output embeddings (shape: num_proteins, 100, 128) were averaged across the sequence length dimension (axis=1).

- Output Files: Both the dengue and human proteins final 128-dimensional embeddings were preserved. Additionally, the autoencoder models that were trained were stored.

7.2 FCNN CLASSIFIER TRAINING AND PERFORMANCE EVALUATION

In order to train an FCNN classifier to predict interaction labels (1 for interacting, 0 for non-interacting), the produced 128-dimensional BiLSTM embeddings for the human and dengue proteins were concatenated to create 256-dimensional feature vectors.

- Training Process: The Adam optimizer (learning rate 0.0001) and binary cross-entropy loss were used to train the FCNN classifier across 300 epochs. During training, the model's performance on the validation (test) set was tracked, and `ModelCheckpoint` was used to save the weights that achieved the lowest validation loss to `best_model.keras`.
  - The training and validation accuracy curves were plotted against epochs. The below figure shows the model's accuracy on the training and validation sets as training progressed.

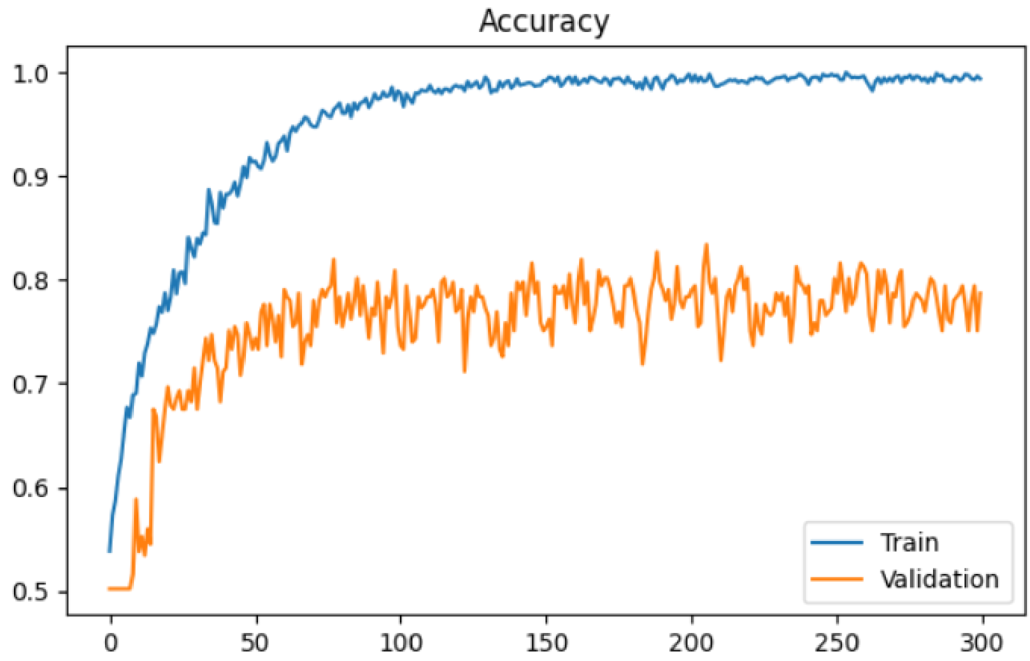

Figure 7.3: Accuracy for Training and Validation Data

- The training and validation loss curves were also plotted against epochs. The above figure illustrates the decrease in the binary

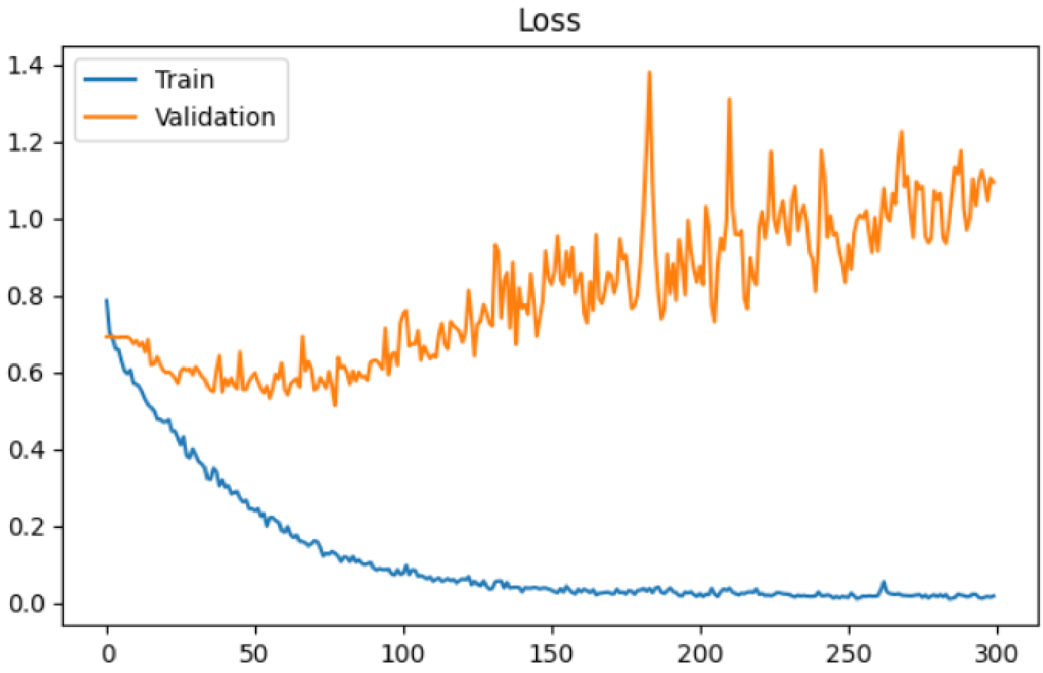

Figure 7.4: Loss for Training and Validation Data

cross-entropy loss for both training and validation sets over the epochs.

- Performance Evaluation: The best saved model (`best_model.keras`) was evaluated on both the balanced training set and the unseen balanced testing set. The following performance metrics were calculated:

Metric	Training Set	Testing Set
Accuracy	1.0000	0.8412
Sensitivity	1.0000	0.8551
Specificity	1.0000	0.8273
AUC	1.0000	0.9241
AUPRC	0.9991	0.9327

Table 7.1: Performance Metrics of the Model on Training and Testing Sets

Chapter 8

CONCLUSION

The goal of this study was to create and assess a deep learning framework that uses solely amino acid sequence information to predict protein-protein interactions (PPIs) between humans and the Dengue virus (DENV). Two steps were taken in the methodology: first, separate Bidirectional Long Short-Term Memory (BiLSTM) autoencoders trained on k-mer representations were used to generate 128-dimensional vector embeddings for both human and DENV proteins; second, a Fully Connected Neural Network (FCNN) classifier was trained on the concatenated embeddings of human-DENV protein pairs in order to predict the likelihood of interaction.

The outcomes proved that this sequence-based, alignment-free method was feasible. The diminishing training loss curves show that the BiLSTM autoencoders successfully learned compressed representations from the protein k-mer sequences. Strong predictive performance was attained by the following FCNN classifier on the unseen test set after it was trained on balanced data with experimentally confirmed interactions and non-interactions (using the generated embeddings as features). Accuracy of 0.8412, Sensitivity of 0.8551, Specificity of 0.8273, Area Under the ROC Curve (AUC) of 0.9241, and Area Under the Precision-Recall Curve (AUPRC) of 0.9327 were among the test set's key performance indicators.

Based just on sequence data, these results indicate that the suggested deep learning framework may effectively differentiate between Human-Dengue protein combinations that interact and those that do not. The resilient performance on the test set indicates good generalization capabilities for predicting possible novel interactions, even if the model received perfect scores on the training data, demonstrating a great ability to learn the patterns within that particular set. This framework provides an effective and scalable computational tool to support experiments.

8.1 LIMITATIONS

While promising, the current approach has limitations:

- **Embedding Method:** The average of BiLSTM outputs after using k-mers (more precisely, 4-mers) may not adequately capture all intricate long-range dependencies or structural details present in protein sequences that may affect interactions.

- **Data Dependency:**The quality and representativeness of the input k-mer embeddings and the known interaction data (`Interactions.csv`) used to train the classifier are key factors that affect performance. Model bias may be impacted by the definition of non-interacting pairs (e.g., random pairing versus experimentally confirmed non-interactions).

- **Model Interpretability:**BiLSTMs and FCNNs are two examples of deep learning models that can operate as ″black boxes,″ making it difficult to identify the precise sequence patterns behind a given interaction prediction.

8.2 SCOPE OF FURTHER WORK

Based on the current work, several avenues for future research can be explored to potentially enhance prediction accuracy and biological insight:

8.2.1 Future Directions

Multi-modal learning: Beyond primary sequences, incorporate a variety of data kinds. This could entail integrating data from pre-existing human protein interaction networks (host interactome data) or merging the sequence-based embeddings produced in this study with structural information, such as potentially anticipated structures from models like AlphaFold2. By combining these many perspectives, proteins may be represented more comprehensively and interaction prediction accuracy may be increased.

Dynamic interaction modeling: Expand the model to take viral infection's dynamic nature into consideration. The model may be able to predict stage-specific PPIs by incorporating temporal data, such as host or viral protein expression levels assessed at various stages of DENV infection (for example, from transcriptomics or proteomics time-course studies), providing more detailed information on the infection's progression.

Experimental validation: Generate a prioritized list of new, highly confident Human-Dengue PPI predictions using the predictive model. To verify the biological interactions, these selected candidates should thereafter undergo experimental validation utilizing well-established wet-lab techniques (such as yeast two-hybrid, co-immuno precipitation, and surface plasmon resonance). In order to convert computer forecasts into useful biological knowledge, this step is essential.

Enhanced Embedding Strategies: Investigate different or additional sequence embedding methods, such utilizing huge pre-trained protein language models (e.g., ESM, ProtBERT), altering k-mer lengths, or attention mechanisms inside the LSTMs.

Alternative Architectures: Examine various model architectures for classification, such as graph neural networks, particularly if a multi-modal approach includes network or structural data.

8.3 SOCIAL RELEVANCE

With millions of cases each year and the potential for serious sequelae, dengue virus infections represent a substantial global public health burden. Developing effective countermeasures requires an understanding of the molecular mechanisms behind DENV infection, namely how the virus manipulates host cellular processes through protein interactions. By offering a computational method that helps speed up the identification of important Human-Dengue PPIs, this study makes a contribution. This findings can assist identify new targets for antiviral medication development or therapeutic treatments, ultimately aiding in the fight against Dengue fever, by possibly emphasizing novel interactions essential for viral replication or host immune evasion. Additionally, the time and expense involved in extensive experimental screening can be decreased by the effectiveness of computational prediction.

8.4 APPLICABILITY OF FINDINGS

The established computational framework itself is the main use case for the results of this project. Researchers looking at the pathophysiology of the Dengue virus can use it to:

- **Generate Hypotheses:** Predict potential interaction partners between human and DENV proteins based on sequence alone.
- **Prioritize Experiments:** Guide experimental validation efforts by ranking potential interactions, allowing researchers to focus resources on the most promising candidates.
- **Explore the Interactome:** Facilitate a broader exploration of the potential Human-Dengue interactome than might be feasible through experiments alone.

The underlying technology shows a versatile sequence-based strategy beyond Dengue virus, including k-mer representations, BiLSTM autoencoders for sequence embedding, and an FCNN for interaction categorization. This approach might be modified and used to forecast PPIs in different host-virus systems or even in other kinds of biomolecular interaction prediction challenges where the main information available is sequence data.

REFERENCES

[1] S. Tsukiyama, M. M. Hasan, S. Fujii, and H. Kurata, "LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021, Art. no. bbab228.

[2] S. Yang, C. Fu, X. Lian, et al., "Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework," *mSystems*, vol. 4, no. 2, pp. e00303–18, 2019.

[3] M. D. Dyer, T. M. Murali, and B. W. Sobral, "The landscape of human proteins interacting with viruses and other pathogens," *PLoS Pathogens*, vol. 4, no. 2, p. e32, 2008.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[6] X. Yang, S. Yang, Q. Li, et al., "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 153–161, 2020.

[7] M. G. Ammari, C. R. Gresham, F. M. McCarthy, et al., "HPIDB 2.0: a curated database for host-pathogen interactions," *Database*, vol. 2016, p. baw103, 2016.

[8] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, 2017.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

LIST OF PUBLICATIONS

[1] **Mundru, S. K., Vattem, C. S. N., Yaram, S. R. and Bhar, A.** (2025) Application of BiLSTM for Protein-Protein Interaction Prediction, *Proceedings of the 24th International Conference on Bioinformatics (InCoB 2025) (prepared for submission)*.