Course No – CSE 4241

Course Title – Biomedical Engineering

Submitted By,

Md. Mahbubur Rahman
Roll: 1907022
4th Year 2nd term
Dept. of CSE
KUET, Khulna

Submitted To,

Safin Ahmmed
Lecturer
Dept. of CSE
KUET, Khulna

1. Introduction

Protein structure prediction is a crucial aspect of bioinformatics and computational biology. Proteins are essential biomolecules whose functions are largely determined by their three-dimensional structures, which are encoded by their amino acid sequences. Predicting the secondary structure (**2D structure**) of proteins— such as α-helices, β-sheets, and coils— provides insights into protein function, stability, and interaction. Efficient 2D structure prediction methods are vital for drug design, understanding disease mechanisms, and guiding experimental studies.

Recent advances in machine learning and deep learning have significantly improved prediction accuracy, allowing models to learn complex sequence-to-structure relationships from large datasets.

2. Methodology

2.1 Data Collection

- Protein sequences and corresponding 2D structures were collected from public databases such as **NCBI Protein** and **PDB (Protein Data Bank)**.
- **FASTA** format sequences were used, and the secondary structure labels (H: helix, E: strand, C: coil) were extracted from corresponding structure files and saved into a .csv format file for better usability. The .csv file contains the following information –
    i. sequences
    ii. 2d structure (3 amino acids)
    iii. 2d structure (8 amino acids)

2.2 Data Preprocessing

- **Sequence encoding:** Each amino acid was represented using **one-hot encoding.**
- **Sequence padding or truncation:** Sequences were standardized to a fixed length to allow batch processing.
- **Train-test split:** 80% of the dataset was used for training, 10% for validation, and 10% for testing. The training, validation and testing dataset was in different csv files for better control.

2.3 Model Architecture

We designed a **CNN-based 2D structure prediction model** that learns spatial patterns in protein sequences.

Input Representation

- Each amino acid in the protein sequence is converted to a **one-hot encoded vector** of size 20 (for the 20 standard amino acids).
- Sequences are padded to a fixed length for batch processing.
- Input shape: (batch size, sequence length, 20).

Convolutional Layers

- **Conv1:** 1D convolution with kernel size 3, input channels = 20, output channels = 128, padding = 1. Captures short-range sequential patterns (tripeptides).
- **Conv2:** 1D convolution with kernel size 5, input/output channels = 128, padding = 2. Captures medium-range motifs.
- **Conv3:** 1D convolution with kernel size 7, input/output channels = 128, padding = 3. Captures longer-range dependencies across residues.

Activation

ReLU applied after each convolution.

Dropout

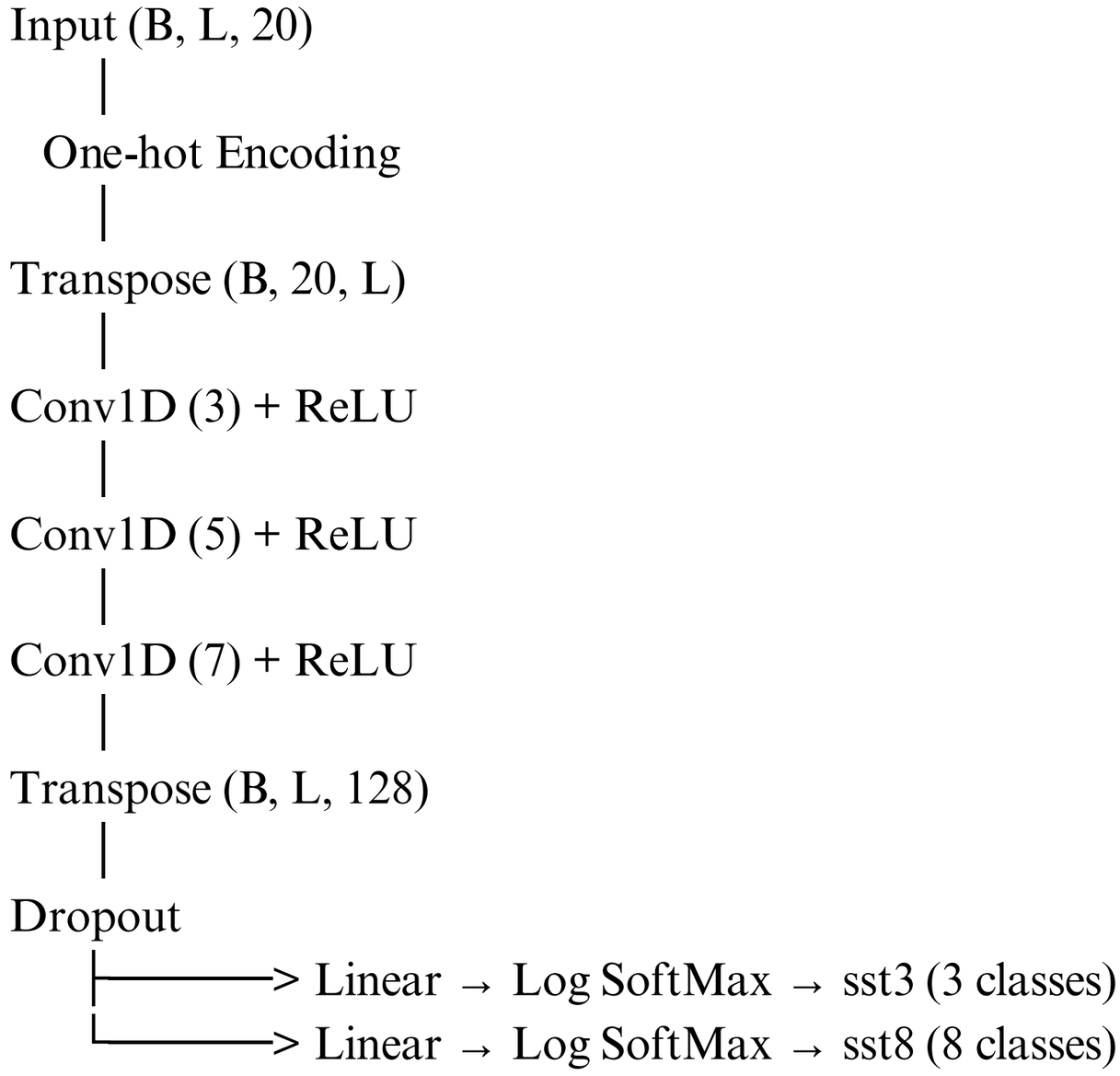
0.3 applied after convolutional layers to prevent overfitting.

Output Layers

- After convolutions, the feature map is transposed back to (batch size, sequence length, hidden dim) for residue-wise prediction.
- Two linear layers act as **dual heads**:
    - head3: predicts sst3 classes (3 outputs per residue).
    - head8: predicts sst8 classes (8 outputs per residue).

Log SoftMax is applied along the last dimension to generate log-probabilities.

Architecture Diagram



3. Results

The model was evaluated using confusion metrics, accuracy and loss.

3.1 Graphical Representation

- **Loss Curve:** Training vs. Validation loss over epochs.
- **Accuracy Curve:** Training vs. Validation accuracy over epochs.
- **Confusion Matrix:** Class-wise prediction performance.

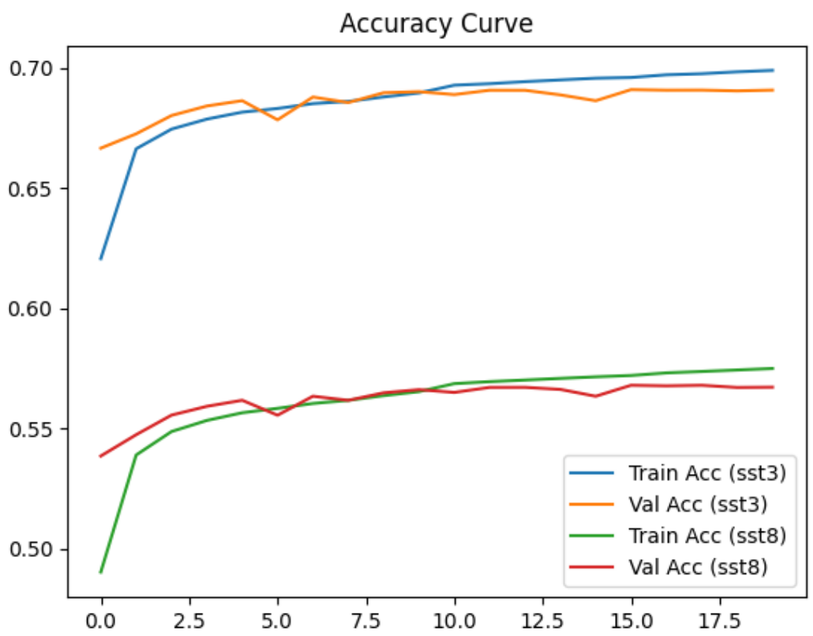


Figure 3.1 : Accuracy graph

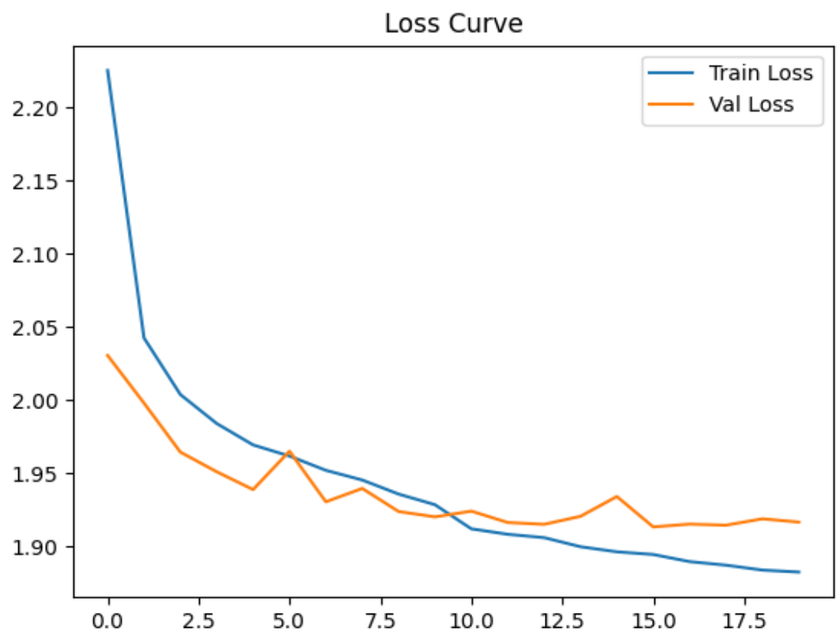


Figure 3.2: Loss graph

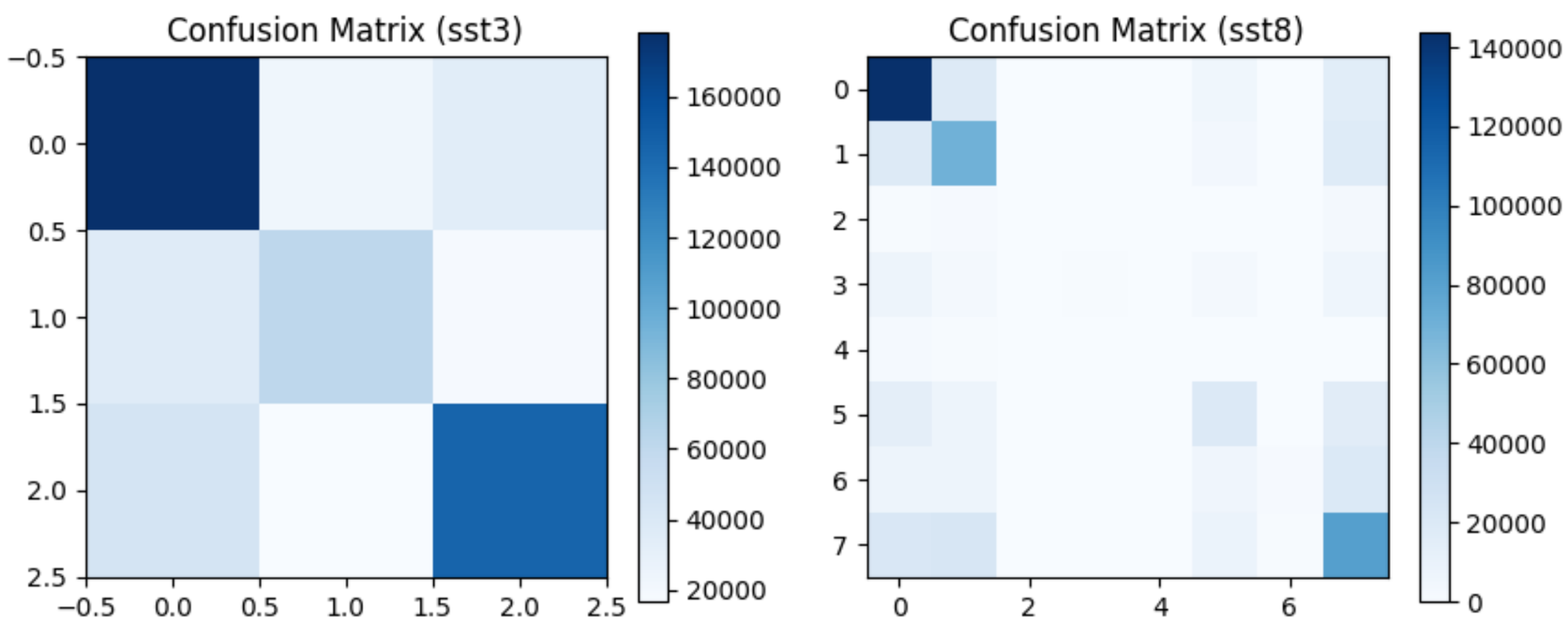


Figure 3.3: Confusion matrix

3.2 Evaluation Metrics

Metric	Test1 (cb513)	Test2 (ts115)
Accuracy (sst3)	70.53%	68.24%
Accuracy (sat)8)	58.26%	53.86%
Loss	1.8634	2.0132

4. Model Architecture Diagram

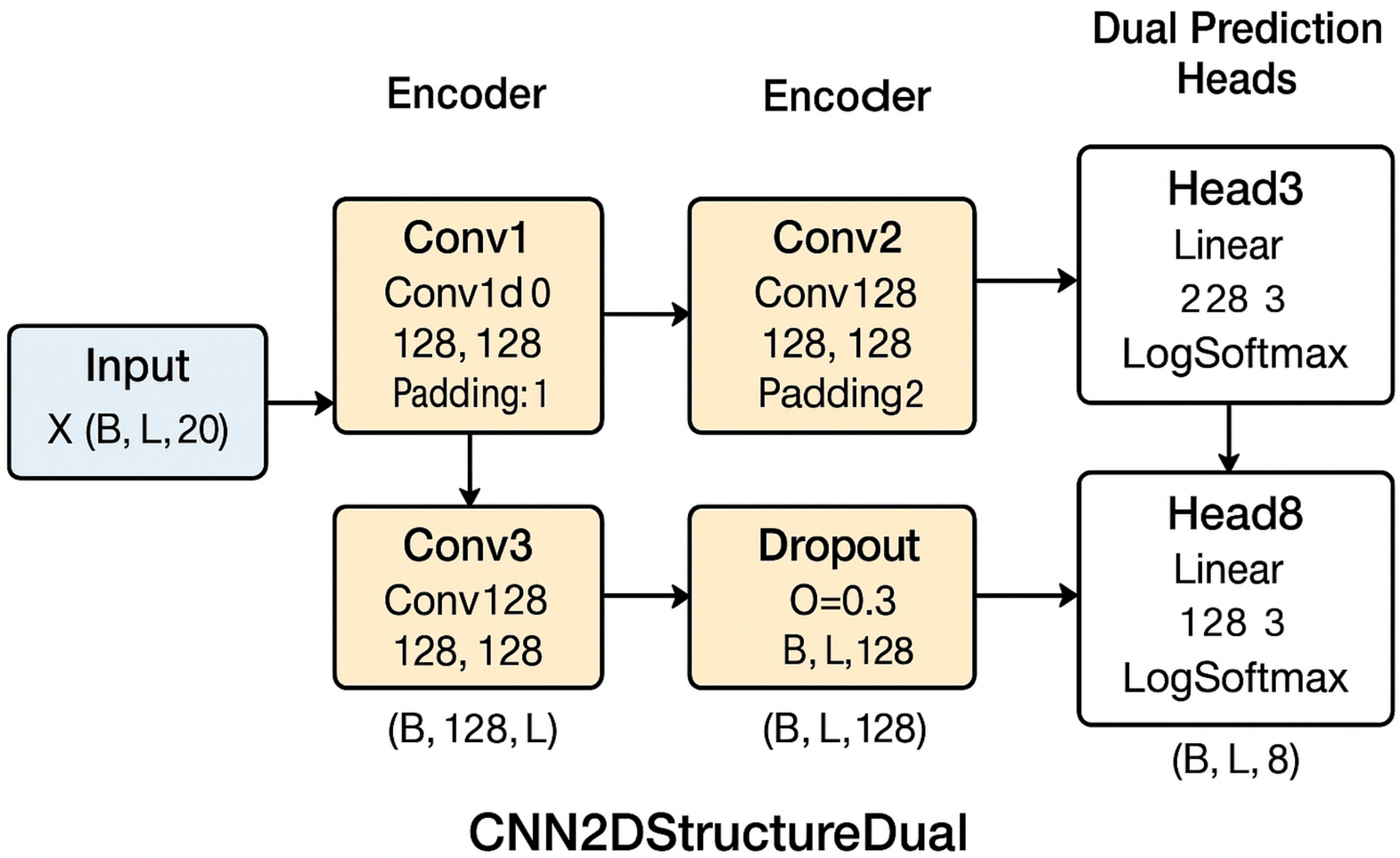


Figure 4.1: CNN-based Protein 2D Structure Prediction Model

5. Discussion

The model achieves **good prediction accuracy**, demonstrating that CNNs can capture sequential patterns and local motifs in protein sequences.

Limitations:

- Limited dataset size may restrict generalization.
- Padding/truncation may lose important sequence context.
- Complex structures might require attention-based models (e.g., Transformers) for better performance.

Potential improvements:

- May use **hybrid models** combining CNNs and LSTMs or Transformers.
- May incorporate evolutionary information using **PSSM (Position-Specific Scoring Matrix)**.
- May expand the dataset to include more diverse protein families.