Protein Secondary Structure Prediction Based on Triple Fusion Explainable Model

Supervised By Dr. Grace Ugochi Nneji
Cecilia Chu

Chengdu University of Technology, CDUT Sino-British Collaborative Education

Abstract

Proteins are basic biomolecules that perform multiple functions in living organisms [1]. This project has developed a new deep learning Model, Triple Fusion Explainable Model, which is capable of deeply extracting the local features and long-range dependencies of protein sequences to improve the accuracy and efficiency of protein secondary structure classification. The prediction accuracy rate of this model is 95.64%, the AUC is 98.82%, and the f1 score is 95.55%. The experimental results show that this model is robust and universal in predicting the secondary structure of proteins and is a valuable AI-based tool in the medical environment.

Figure 1: Protein Sequence From the Dataset

Dataset & Data Process

- **About dataset**: In this project, two independent datasets from Kaggle were utilized, namely Protein Secondary Structure and Protein Data Bank (PDB). The Protein Secondary Structure dataset was employed for model training, while the PDB dataset was used for evaluating the generalization capability of the model.

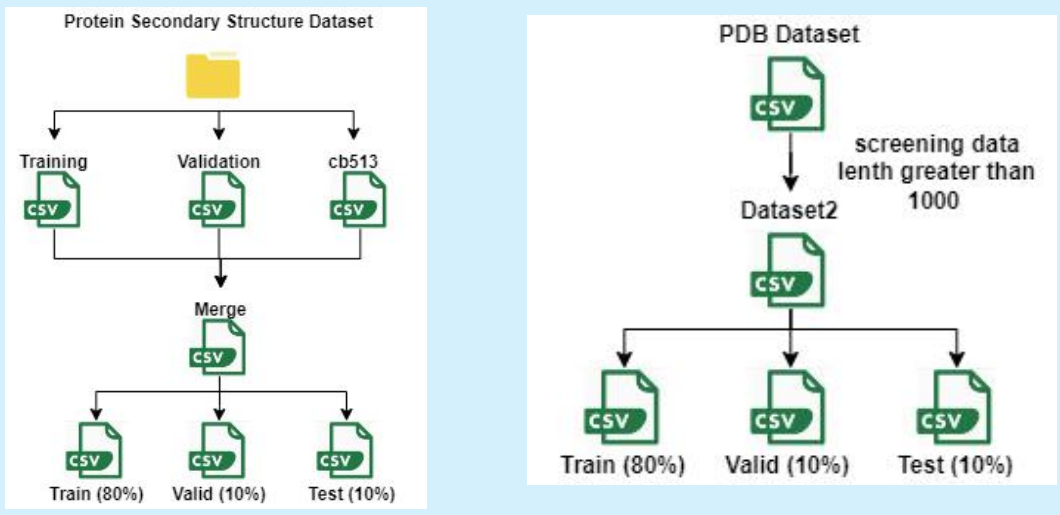


Figure 2:Data preprocessing to balance the protein sequences

Deployment

- Home page shows introduction of the project, how to use the GUI to make predictions, the model used in the project and the type of protein secondary structure.
- Click the **<<Model Prediction>>** button one home page, and leads user to prediction page.
- Upload a file or input one protein sequence and click **<<Predict Secondary Structure>>**
- Click the **<<Result Visualization>>** and **<<SHAP Visualization>>** buttons to check the details about the model.

Home Page **Prediction Page**

Figure 5:Deployment on Web

Implementation & Results

- The model integrates Convolutional Structural Predictor (CSP) model, Memory-based Structure Network model, and Self-Attention Convolutional Structural Predictor (Attention-CSP) model. Furthermore, SHAP was incorporated to explain the model's results.
- The training is evaluated with Accuracy, Loss, Precision, Recall, Specificity, F1-Score, ROC, Confusion Matrix.

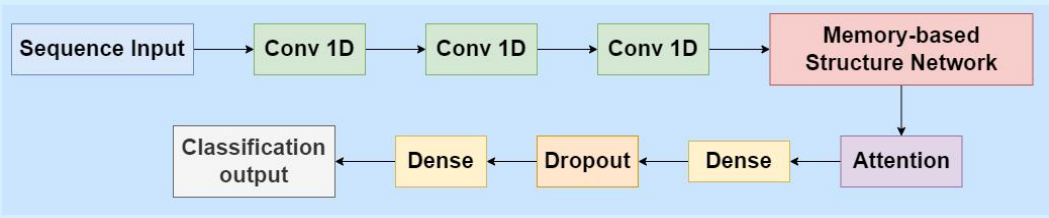
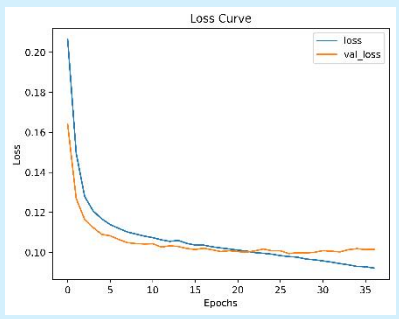
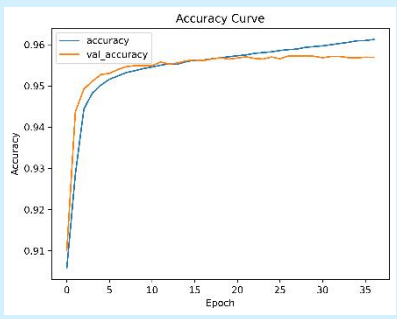


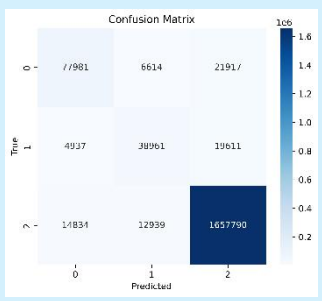
Figure 3: Model Overview



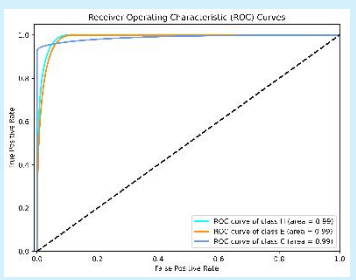
Loss



Accuracy



Confusion Matrix



ROC

Figure 4: Training Results Summary

Conclusion

- Developed Triple Fusion Explainable Model for protein secondary structure prediction.
- Achieved high accuracy and performance metrics.
- Improved efficiency.
- Requires better multi-class classification.
- Future work: enhance multi-class capabilities.
- Validate on more public datasets.

Reference

[1] M. A. Sofi and M. A. Wani, 'Improving Prediction of Protein Secondary Structures using Attention-enhanced Deep Neural Networks', in 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), Mar. 2022, pp. 664–668. doi: 10.23919/INDIACom54597.2022.9763114.