Assalamu alaikum everyone. Myself Md Liad Hossain , Roll: 1809008 .

Today I’m going to present my thesis on protein secondary structure prediction using graph neural network with ensemble learning under the supervision of prof. Dr. A.B.M Aowlad Hossain sir.Without further delay let’s get into it.

These are the outlines that will be discussed in the presentation.

Since the thesis is all about the prediction of protein structure, so first of all we need to know about proteins and why we need it. Proteins are building blocks of life. The basic structure of protein is a chain of amino acids and it’s functionality depends totally on it’s structure. For that, we need to understand the protein structure very well and understanding the protein’s structure will help in drug design and development, disease understanding and treatment, biological innovation and many more.

These are some literature review of the previous works on protein secondary structure prediction. First of all in 2012, Zhang and some other authors of the team applied convoultional neural network and bidirectional recurrent neural network on protein’s primary structure and devised the 8 states of secondary structure. But they obtained only 71.4% accuracy. Then in 2018, Kathuria and his other team members successfully predicted alpha and non-alpha structures of the protein by applying Random forest method and ROC curve. But here the prediction is limited to binary class classification. After that in 2019, Toussie and his team increased the accuracy of prediction by applying support vector machine and artificial neural network. But they converted the 8-state of secondary structure to 3-state. At last in 2021, Nahid and some other authors implemented graph neural network with support vector machine to predict the 8-states of secondary structure. And they achieved a great accuracy of 89%. But the limitation was that the model couldn’t perform well for the larger dataset. They achieved this accuracy only for the smaller datasets.

After literature review the major problems discovered and needed to be addressed are -

Previously most of the models were implemented for alpha and non-alpha structure prediction or 3-state of secondary structure prediction. But the obtained accuracies for the 8-state prediction were not very high and performed very poor for larger datasets.

So, the objectives of our proposed system are-

1. To improve the accuracy for 8-state prediction.
2. To research and implement the graph neural network model.
3. To study and implement the ensembling of SVM and Random Forest methods and
4. To perform well on both smaller and larger datasets.

This is the overview of our woking procedure. Firstly, we have created a graph from the primary sequences, then we have extracted the features. We have passed the feature to the neural network in forms of message. Then the model has predicted the secondary structure.

Firstly we have preprocessed the data by removing the non-standard amino acids and eliminating the dataset of mismatched length of primary and secondary sequences. Then we have applied orthogonal encoding to both primary and secondary structures.

These are the encoding vectors of all the 20 types of amino acids of the primary sequence.

Then we have represented the primary sequences in form of graph. In the graph each amino acid acts as a node and contains encoded information as it’s feature and adjacent amino acids are connected with an edge.

Then we have aggregated the information of the neighborhood. We have iterated the graph. For the first amino acid, the feature of the next amino acid is added to it’s own feature, for all the internal amino acids, the feature of the previous and next amino acids are added to it’s own feature and for the last amino acid the feature of the previous amino acid is added.

After that a sliding window technique is applied to train the dataset and window padding is applied to avoid missing sequences at the terminal nodes.

Then conformation parameter has been added as another feature vector. Conformation parameters represent the proportions that amino acids tend to form secondary structures.

Then we have passed the feature vectors of the smaller datasets to the support vector machine and larger datasets to the random forest model to predict the secondary structure.

We have used jupyter notebook for the implementation of all of these steps and wps office to observe the datasets properly.

In the result, we have observed the accuracy for different window size and achieved the best accuracy for window size 17 which is 86.02%.

These are the various performance parameters we have observed in the result analysis part.

We have tested the model for different hyperparameters and achieved the best accuracy in support vector macine for C=1.5 and gamma=0.1. Here C represents the penalty for misclassifying a data and gamma represents the decision region. For the random forest we have achieved the best accuracy for 100 n\_estimators and 42 random state. Here n\_estimators represents the number of decision trees in the forest and random state controls both the randomness of the bootstrapping of the samples and best split of the features.

This is the comparison of our method with some other traditional methods. We can see that the model has achieved higher accuracy than the others but except one. It was 89% and our model has acieved 86% but the difference is that model is only for smaller datasets and our model is for both smaller and larger datasets.

These are the impacts of the project outcome on the environment and it’s sustainability.

There are 6 complex engineering problems and 4 complex engineering activites have been addressed in this project.

There are some limitations in our method. It takes a little bit longer time to predict. Graph neural network may not capture the dynamic nature of the protein effectively and it may not be able to compensate because of our limitations of understanding the relationship between primary and secondary structure.

So, if we conclude the research, we have explored the large complex molecules of the human body. For that we have applied graph neural network which is very efficient for feature extraction. At last we have ensembled support vector machine and random forest which lead to better accuracy for both smaller and larger datasets.

There are many future scopes of this research such as-

We can achieve better accuracy by combinig neural networks with other models and other optimization techniques can also be used. 3D graphing of the protein structure can improve the accuracy and also we can do protein tertiary structure prediction. We can also select the feature differently to improve the accuracy.

Thank you.