International Journal of Scientific Research in Engineering & Technology
Volume4, Issue5 (September-October 2024), PP: 15-21.
www.ijsreat.com ISSN No: 2583-1240

Protein Structure Prediction Using FCNN

Aditya Das[1], Indrajeet Singh Yadav[2], Er. Harjasdeep Singh[3]
[1,2,3]Department of computer science, Malout Institute of Management and Information Technology, Punjab, India.

To Cite this Article
Aditya Das1, Indrajeet Singh Yadav2, Er. Harjasdeep Singh3, "Protein Structure Prediction Using FCNN" International Journal of Scientific Research in Engineering & Technology, Volume4, Issue5 (September-October 2024), PP: 15-21.

Abstract: In our work, we have done the cross-examination with mouse datasets and human datasets using a fully connected neural network (FCNN). This plays a vital role in protein sequencing and further examination of its helix structure. Secondly, our major agenda in our entire work was to focus on the result that we got when varying the length of the protein sequence while giving it as an input to our model. Recurrent Neural Networks can also be applied to prediction models as an alternative.

Key Word: Protein structure classifier, mouse dataset, Human dataset, protein data bank (PDB)

I.INTRODUCTION

Large molecules known as proteins play a variety of functions, including conveying and validating genetic information as well as signaling inside the cell. Long sequences of tiny molecules called amino acids are bound together to create the structure of proteins. There are 20 distinct types of amino acids, and the protein's fundamental structural component is its amino acid composition [3]. The length of the chain is a result of the flexible bonds between amino acids as well as the chemical and physical characteristics of various amino acids. to fold into a variety of forms. The secondary structure of a protein is made up of these forms. The FCNN was implemented using the deep learning framework Keras and Theano. Theano is the machine learning backend of Keras. This backend was used to code the Fourier layers. FCNN model is trained through two approaches first selecting the layers and then performing the Fourier pooling operation on them. This FCNN works much better on image datasets. Train and Test datasets both are used on human and mouse protein structure prediction with varying loss functions and varying window sizes. We have used different libraries: - pillow, torch, utils, FCNetwork, PyTorch, biopandas, TensorFlow, and RNNetwork. To forecast the protein secondary structure, we develop and train various machine learning models in this study. We contrast and explain the findings from sequential models like recurrent neural networks and deep learning models like fully connected neural networks (FCNN) (RNN). We also carry out three experiments utilizing our models to aid in the categorization of protein structures. Here is a list of what we contributed to the issue: A dataset of mouse and human proteins was cleaned and redundant data was eliminated. built-in code for PyTorch for importing PDB files; Constructed and trained two models: a recurrent neural network and a fully-connected neural network (FCNN) (RNN). The FCNN's root means squared error (RMSE) forpredictions is a respectable 0.22[4]; Two FCNNs were trained, one on a dataset of mouse proteins and the other on a dataset of human proteins, to conduct a cross-species comparison. Then, we put both models to the test using test sets with mice and people. The models fared quiet and also performed well on the cross-species test, but they performed best on the test sets for their species. This supports the idea that mice can serve as good model organisms in this situation.

II.METHODOLOGY

Feature selection: We have protein data bank files that contain information about amino acid sequences and information about secondary structures with their location for e.g. α-helix. From its helix structure information which is present in PDB files, we applied each kind of amino acid represented by one hot encoding and the same for our variables in a protein sequence.

Fully Connected Model: Using 200 input neurons, 20 hidden neurons, and 10 output neurons, we created a fully connected network. The term for this is the multilayer perceptron (MLP). A window of 10 successive amino acids in the protein is represented by the input vector. sequence. The amino acids which we are using as input we encode them in a standard basis vector and then flatten them into an input vector x of length 200, here we use the relu activation function [4]: -

$$(x) = \{0, x < 0 \\ x, x \geq 0\}$$

Hidden layer values as:-

$$h(I) = \sum_{j=1}^{200} w(I, J).\, x(J)$$

output vector as:-

$$o(I) = R\left(\sum_{j=1}^{20} w(I, J).\, h(J)\right)$$

With the input, we have given we get the solution in the form of the predicted probability of these 10 input values within the window at 10 different positions. In addition, a prediction for the entire protein by sliding the window in the middle position and extracting the middle value at each position.

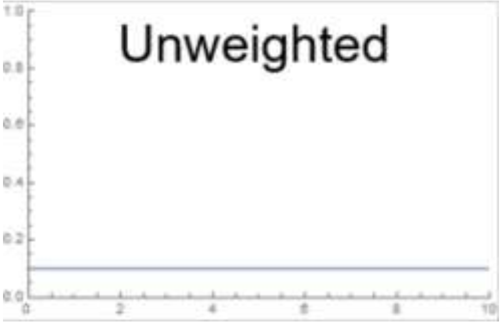
Loss Function for Train Model: - We have used three error functions as follows: -

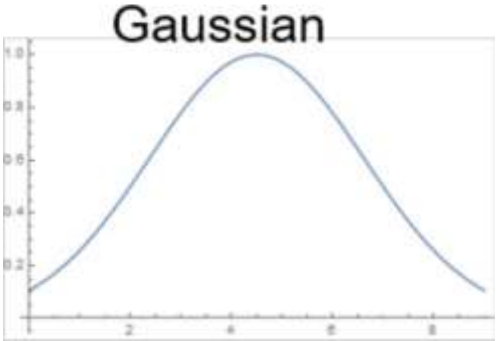
1. Unweighted loss: Each of a protein′s acids has an equal weight.

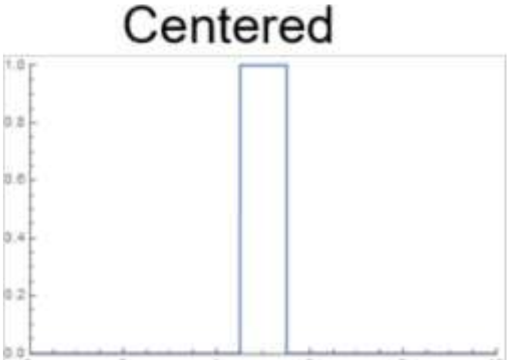
2. Gaussian loss: The weight at the center of the window is the greatest among all the weights and it gradually decreases when we move out of the central point.

3. Centered loss: the actual weight is concentrated at the centered position and other acids have negligible weight.

These loss functions are used for testing our hypothesis that the weight which is in the center plays a significant role in protein sequencing and helps in accurately anticipating where the helix structure will be found in the protein sequence.







Test Requirements: The loss functions mentioned above are used to build the models. On a narrow period of the pattern, they are given the discrepancy between the real and anticipated helices. To analyze the outcomes during testing, a standard error function is required. Our test criterion is the standardized RMSE throughout the full protein chain. By rotating the frame throughout the complete protein sequence in this instance and collecting helicity estimates for every point using the frame in which it is nearest to the center, the anticipated quantities of helicity are recreated. Every time the phrase "overall average loss" is used, it refers to this notion.

In the overall implementation, we selected FCNN i.e., fully connected neural network having 150+ input neurons,30+ occult neurons, and 5+ output neurons This is a multi-layer perceptron. Different sequence patterns are probably used by proteins with significantly different lengths and proteins with evolutionary histories spanning thousands of years to build their shapes. Initial tests where proteins were selected randomly from the archive fit extremely badly. To improve homology and symmetry in the collection, we produced life forms datasets for just two similar animals, mice, and humankind. These selections were still fairly big since the baseline dataset was so big.

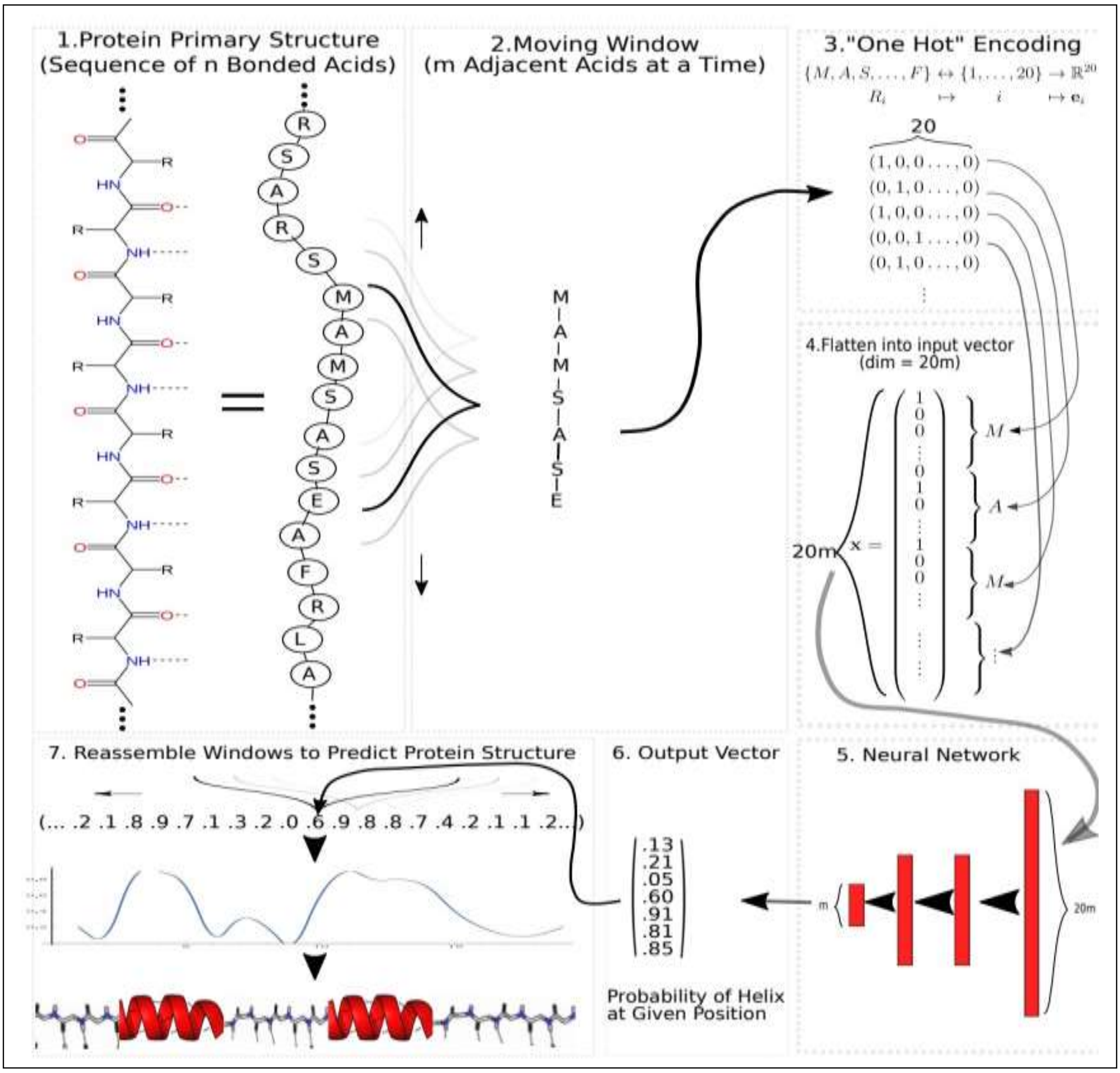


Fig: working of FCNN

In the working FCNN model, first, we pick up a protein sequence of n-bonded acids and the moving window of defined size takes the m- adjacent acids that are sub-sequence of primary protein at a time.

One hot encoding will be performed on the window and after this output values of encoding are flattened into an input vector of specified dimensions which can be said to be the hidden layer. After processing all protein sequences many output values are produced after encoding those will be flattened in a defined number of hidden layers.

After performing encoding, we got all the hidden layers these layers will be connected fully one by one. An output vector or layer is connected to a hidden layer which represents the probabilities of the helix at a given position.

According to the output vector, FCNN resembles windows to predict protein structure. It represents all the 20 possible acids that can be sequenced in protein structure.

The acid which has high probabilities among all will be predicted in the protein sequence of secondary structure.

Protein Data Bank (PDB): It contains different varieties of proteins of different species, and it is usually a large data set. From PDB we only select the data set of closely related animals, mice and humans for the purpose to increase the similarity and pattering and reduce the complexity. We acquire structural data of 30000 human proteins and 6000 mouse proteins and then we divided the data into training and test sets. By assigning randomly 80% of the data to the training set and the rest of the data to the test set.

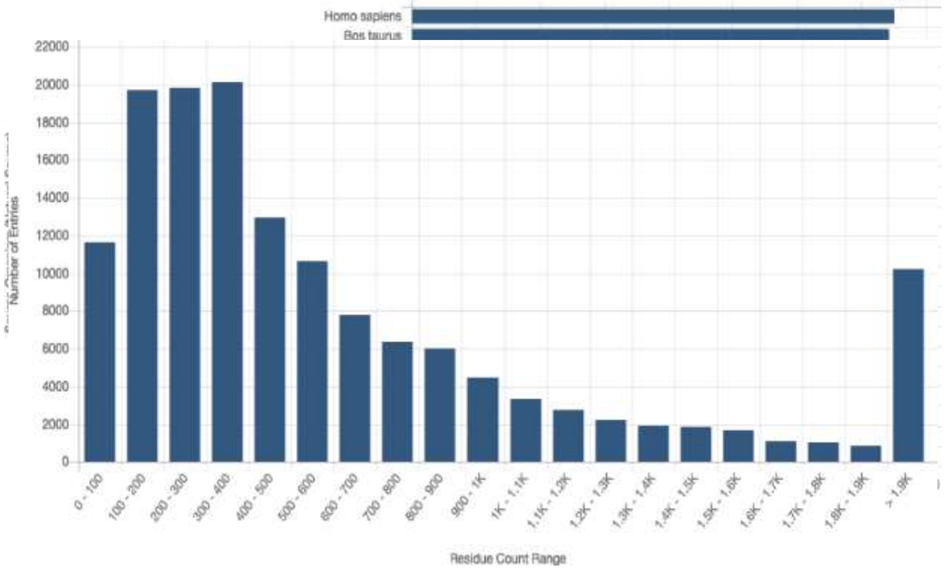


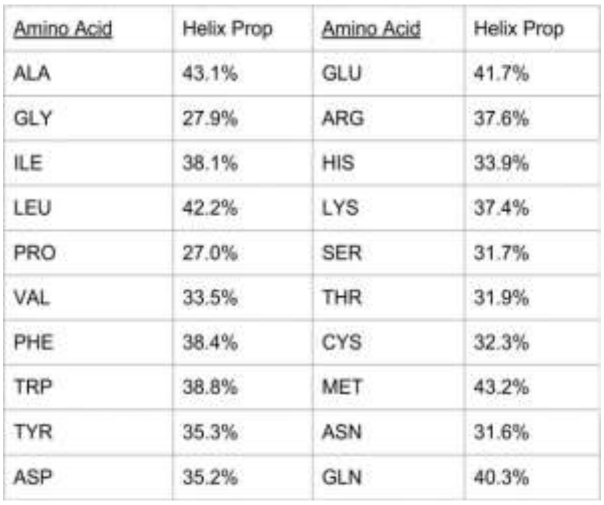
Fig: protein's set of species

Helix probabilities: we have implemented a model, and we got the probability for each helix in given amino acid.

$$P(Helix/R) = H_R/N_R$$

Where NR is the position in the type R dataset, where HR is the total number of positions in a type R helix.

Now we found out the probability of helix present in a particular position and it is considered as a baseline using this baseline, we can estimate the average loss in the mouse dataset of 0.48.



Amino Acid	Helix Prop	Amino Acid	Helix Prop
ALA	43.1%	GLU	41.7%
GLY	27.9%	ARG	37.6%
ILE	38.1%	HIS	33.9%
LEU	42.2%	LYS	37.4%
PRO	27.0%	SER	31.7%
VAL	33.5%	THR	31.9%
PHE	38.4%	CYS	32.3%
TRP	38.8%	MET	43.2%
TYR	35.3%	ASN	31.6%
ASP	35.2%	GLN	40.3%

Fig: helix probabilities for each acid

Comparison Between Two Training Models: By comparing training models we compare the error or loss function. We implement our first model on Keras in which loss is significantly high because the prediction is very close to the noise and remains centered.

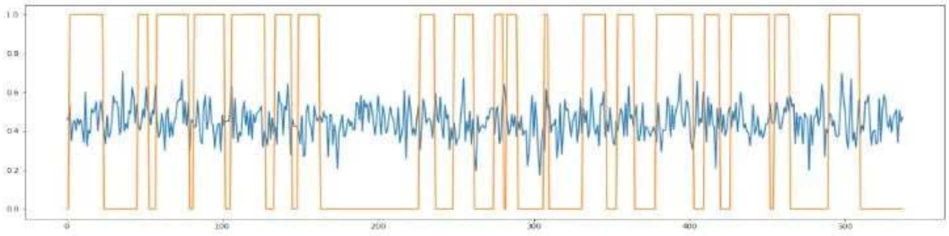


Fig: keras model

And our second model is implemented in PyTorch in which prediction majorly aligns with the actual value so it does not contain noise, in addition, it gives the loss which is diminished.

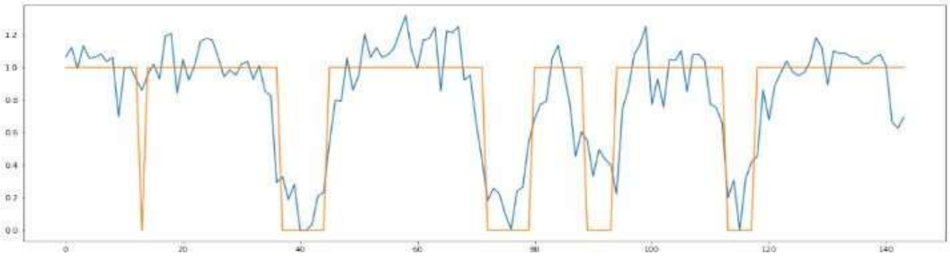
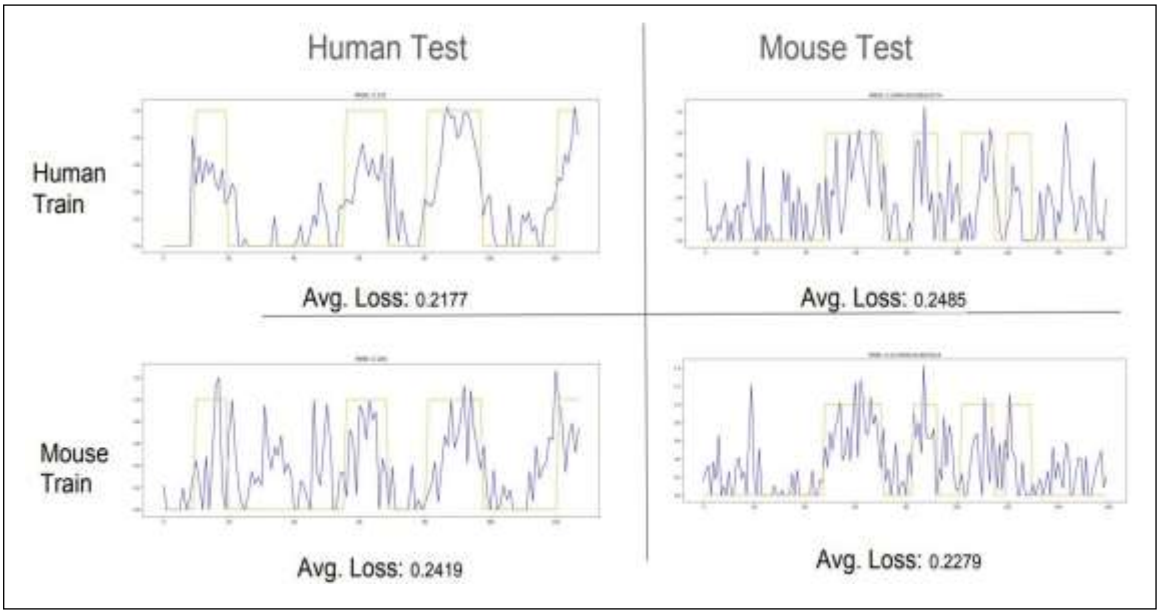


Fig: pytorch model

Where the yellow line represents the actual value and the blue line represents the prediction made by the model.

Comparison between Human and Mouse dataset: We took the trained FCNN model for train set of both humans as well as mouse and test them on its test data set as well as on another one's test dataset which cross validate by calculating RMSE (root mean square error).

As the result we found a good cross-species fit which the large similarity between human protein and mouse protein sequence.

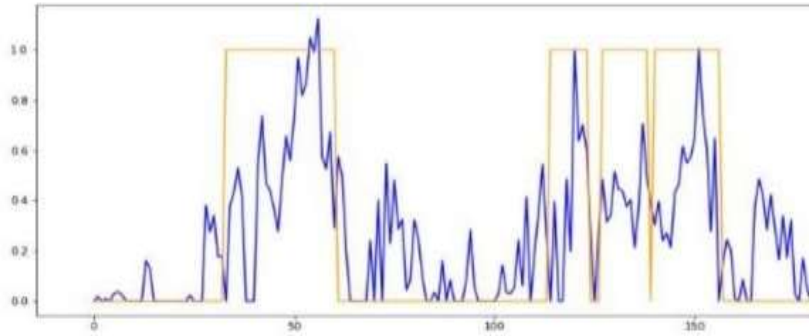


Window size Comparison: On the mouse dataset, we ran an experiment with our FCNN model employing a hidden layer with 41 neurons and window widths of 6, 11, and 13. These windows were chosen because the -helix requires 3.5 amino acids for each turn, making 3,4, and 5 rotations represented by these sizes, accordingly. According to our hypothesis, a window size of 11 or 15 should perform better than a shorter window in forecasting. The mean loss was 0.1938 for the window size of 7, while losses were correspondingly 0.2280 and 0.2590 for the window sizes of 10 and 13.

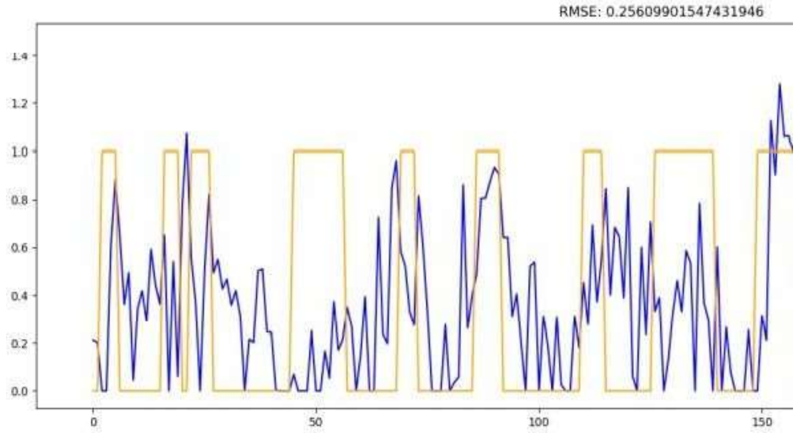
The precise secondary structures because longer lengths make more information available. The average loss was 0.1932 for the window size of 7, while losses were correspondingly 0.2280 and 0.2590 for the window sizes of 10 and 13 [4]. These findings defied our prediction that a window size of 11 or 15 would result in a loss reduction, as a higher frame size allow the model to make more accurate structural predictions. Instead, the window size with the best forecast was the smallest. 7 is probably not far from the smallest size that would produce decent results. For instance, a window size of 2 assumes no context and is comparable to utilizing the baseline, which had an average loss of 0.46.

III. RESULT & ANALYSIS

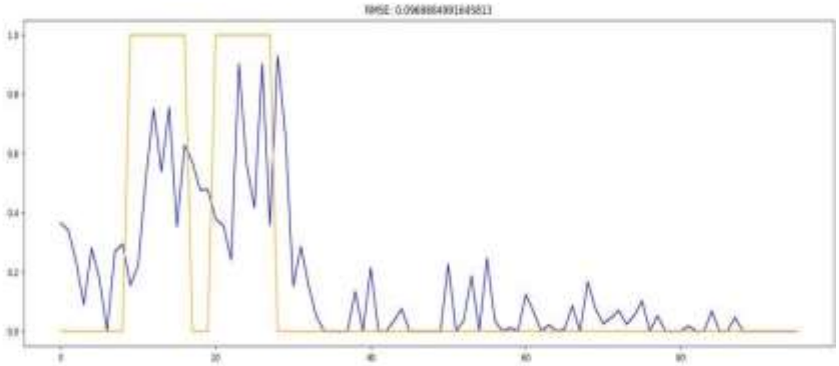
An example through mouse prediction: We demonstrate a few examples of the FCNN's predictions from the mouse test set. In the graph, the x-axis represents the index of the protein sequence and the helix's probability on the y-axis.
The line which is in orange color depicts the helices' actual presence. The likelihood of a helix is shown by the blue line in the model's forecast.



Pretty excellent outcome Between the second and third helix, notice the extremely noticeable dip in the blue line.



A Loss = 0.256. a product of average quality. Spikes are used to denote several helices; however, it is not evident where they start and stop. Small troughs between helices are unclear, and some helices are completely missing.



1AB0 protein. Loss = 0.097. A favorable outcome. clearly recognizes helices and their lengths, but it makes an error in anticipating when in the sequence those spots will be. This outcome accurately predicts that there won't be any helices for the remainder of the sequence.

IV. RELATED WORK

Through neural net and support vector machine-based algorithms, Yang [8] analyzed algorithmic productivity and computation times for static protein prediction. Yang experimented with various frames and hidden layer sizes in the ranges of 1 to 21, 0 to 125, and so on. The network of neural technique reached its best with a performance of 67.42% accuracy at 15 amino acid frame size and 75-unit hidden neurons. The actual quality of neural networks beats that of the SVMs, even though SVMs performed better at convergence and tended not to overfit. Malekpour [1] used three networks of neurons that made advantage of various sequence similarity profiles to enhance the segmental semi-Markov models approach that was already in use. To forecast the secondary structures of protein molecules SSMMs were applied to the neural net results. The suggested system predicted correctly the protein molecules 75.35% of the time. In jones’s [4] effort, a PSIPRED technique, the accuracy rate changed from 76.5% to 78.3%. Place scoring matrices are used by PSPIRED. Creating a sequence profile, predicting the starting secondary structure, then refining the projected secondary structure make up the product’s three primary steps. King [7] discusses PROMIS, a deep learning software that used generalized principles that describe the link between the primary and secondary structure of protein molecules to forecast secondary structures in proteins with an efficiency of 60%. To forecast the protein secondary structures in three and eight different categories using an ensemble of bidirectional Recurrent neural networks, Pollastri [2] developed two novel forecasters. The accuracy of the forecasters, which were evaluated on three distinct test sets, is 78%.

V. CONCLUSION

In our whole implementation, we concluded one important thing the similarities we have in the protein structure of mice and humans. In the time of cross-validation, we got the least amount of deviation when we tested the protein of the mouse in the model of human training model vice versa in the case of mouse and human. When the model is trained with varying the length of protein, we get an unimaginable result. We considered the entire implementation as giving unexpected results as we saw in the case of the mouse and human datasets. Keras model and PyTorch are the two models we examined respectively. The Keras model's entire model’s prediction is concentrated in the center so that was not a good sign but in the case of PyTorch implementation getting good results entire the prediction is aligned with the actual data so we observed that the prediction is quite acceptable. And deviation is very less. Meanwhile, the loss we got while cross- examining was also pretty low. When the model was trained Human dataset and tested on a mouse dataset the deviation of the predictions was quite low loss was very significant. Same as in the case of mice and humans.

VI. FUTURE WORK

The ultimate objective, which is now unachievable, and will be true in the future is to be capable to anticipate its entire secondary, tertiary, and quaternary structure from the protein sequence alone.

To forecast a protein's 3D structure and to better comprehend the intricate functions of proteins, protein structures are thought to be a crucial and required milestone.

References

[1]. *MALEKPOUR, S. A., NAGHIZADEH, S., PEZESHK, H., SADEGHI, M., AND ESLAHCHI, C. Protein secondary structure prediction using three neural networks and a segmental semi-Markov model.*
Mathematical Biosciences 217, 2 (2009), 145–150

[2]. *POLLASTRI, G., PRZYBYLSKI, D., ROST, B., AND BALDI, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Structure, Function, and Bioinformatics 47, 2 (2002), 228–235*

[3]. *SIDHARTH MALHOTRA AND ROBIN WALTERS Secondary Protein Structure Prediction Using Neural Networks.*

[4]. *JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices 1 1edited by g. von Heine. Journal of Molecular Biology 292, 2 (1999), 195–202*

[5]. *YOO, P. D., ZHOU, B. B., AND ZOMAYA, A. Y. Machine learning techniques for protein secondary structure prediction: an overview and evaluation. Current Bioinformatics 3, 2 (2008).*

[6]. *SERVICES, P. S. D. D. Protein structure. Technical Brief 8 (2009).*

[7]. *KING, R. D., AND STERNBERG, M. J. Machine learning approach for the prediction of protein secondary structure. Journal of molecular biology 216, 2 (1990), 441–457.*

[8]. *YANG, J. Protein secondary structure prediction based on neural network models and support vector machines. Departments of Electrical Engineering, Stanford University (2008)*