Deep-LPIN: An Innovative AI-Driven Model for Protein–Ligand Interaction Prediction

Ashok kumar R 1[a],Manikandan N 2[a],Jeeva B 3[b],Lohith V U 4[c]

[1,]*Assistant Professor, Department of Artificial Intelligence and Data Science, Paavai College of Engineering, Namakkal, Tamil Nadu.* salim26ptuniv.edu.in

[2,3,4]*Department of Artificial Intelligence and Data Science, Paavai College of Engineering, Namakkal, Tamil Nadu.*

manikandansalthi@gmail.com, jeevab893@gmail.com, lohivu999@gmail.com

Abstract—New drug research and development comes at a high cost, has high failure rate, and burden pharmaceutical companies and patients. The fight back to this is, as the name says, drug repurposed and it is to use approved drugs for new intended use. In this effort, important computational methods are also used to forecast how drug molecules bind with target proteins, lowering the cost and chances of drug development failure. However, in this work we propose the novel Deep Protein Ligands Interaction (Deep-LPIN) model using deep learning protocol to predict protein–ligand interaction. The system for Deep-LPIN is a combination of one-dimensional Graph Neural Network (GNN) and bi-directional long short term memory network (biLSTM). First, raw drug molecular sequences and target protein sequences are converted into dense vector representations and the model works on converting them. These representations are then fed to ResNet based 1D GNN modules to extract useful features. After creating the feature vectors, we combine the feature vectors and put them as input to the biLSTM network, then we use a multilayer perceptron (MLP) module to predict the protein–ligand interactions. We applied Deep-LPIN in training and testing using Binding DB and Davis datasets. Then, we compared its performance with the two baseline methods, namely Random Forest and SVM. The results indicated that Deep-LPIN also achieved higher accuracy than the two in predicting protein–ligand interactions. Its excellent results on different datasets suggest that the model can generalize and could be used for discovering new drug – target interactions and alternative applications of known drugs.

Index Terms— AI-Driven Prediction, biLSTM, Binding, MLP, ResNet 1D-GNN, Drug Discovery

I. INTRODUCTION

Proteins have an essential molecule in molecular biology; its structure determines the function [1]. The three-dimensional configuration of atoms in a protein molecule dictates its interactions with other molecules, its catalytic function, and its overall significance in biological processes [2]. The structure of a protein is necessary to understand the protein's role in cellular systems and its relationships.

Due to the importance for drug development and engineering, the precise prediction of protein structures is of importance [3]. The ability to accurately predict protein structure permits the building of medications that will selectively bind to and regulate its function. Precise structure prediction is one of the main factors that make protein engineering such an important field — the modification of protein sequence that changes its structure and function (protein engineering) very strongly relies on it [4].

Proteins amino acid sequence and their three dimensional conformation are correlated in a most intricate way. While the sequence specifies possible interactions between its parts, folding speed and environment, other molecules, even play a role in determining the final structure. Complexity of development of precise predictive algorithms stems from the interacting elements. A sequence is used to forecast protein function using computational techniques [5]. Often, these methodologies synthesize analyses of sequence and a structure, with energy computations.

Homology Modelling and Threading made up conventional techniques for the protein structure prediction. Furthermore, these approaches are based on the idea that similar protein sequences lead to similar three dimensional structures [6]. Homologous proteins with determined structure can be localized, thus allowing to build a model of a target proteins structure. It has already reached the state-of-the-art machine learning methodologies [7]. Even deep learning models have shown great success predicting protein structures by machine learning algorithms. These algorithms are trained extensively on established protein structures using large amounts of data to learn complicated pattern and correlation between sequence and structure. The coming of such AI driven techniques has completely changed the field of protein structure prediction.

Contemporary difficulties are handled more effectively by artificial intelligence [8]. Homology modelling and threading are conventional methods for predicting protein structure, but the proteins for which they can predict the structure can only guide them through limited templates. By utilizing AI driven methodologies, patterns and correlations between two sequence and structure, even without evidential homology, can be discerned. Not only is this specialization useful for predicting novel protein structures and probing proteins' exon universe, it is also an advantage for forecasting the structures of novel proteins [9].

Protein structure prediction has crossed the threshold of the novel epoch thanks to DL technologies [10]. The inspiration from the architecture and functionality of the human brain has shown incredible efficacy to the deep

979-8-3503-1502-8/23/$31.00 ©2023 IEEE

learning models in discovery of very intricate patterns from vast amount of data. Deep learning models can not only predict protein structures, but can also predict complicated relationships between amino acid sequences and their three dimensional shapes in order to go beyond how one could reliably predict in the past. It is shown that AlphaFold2 can predict the three dimensional structures of many unknown proteins [11]. This is a much more important advance for protein structure prediction with AlphaFold2, a protein structure prediction technology developed by Deep Mind. By precisely forecast protein structures, it has opened the gateways towards understanding protein function, speeding up drug discovery, and bring an understanding of biology and medicine.

II. RELATED STUDY

As a main tool for trying to figure out structures where there aren't a lot of experimental data, the Rosetta software suite is used widely by structural biologists to design or create protein structures. In [12] the computer methods in the Rosetta framework dedicated to modelling protein structures with NMR data are brought together. There we describe the fundamental computational strategies used to integrate NMR data formats with Rosetta. Recent advancements are reviewed, and more specifically it focuses on special tools developed to combine paramagnetic NMR and hydrogen-deuterium exchange data. Additionally, we note that chemical changes were incorporated into CS-Rosetta. In the paper, the methods for improving and supplementing the structure prediction of advanced AlphaFold2 algorithm are studied with NMR-guided Rosetta modelling tools.

An extensive analysis of techniques for predicting protein structures, their historical evolution and their present status are given in the paper [13]. The study begins with explaining older methods used in this field: homology modeling and threading. Then the research takes these proven techniques, and then builds on them focusing on the current emergence of machine learning algorithms, and specifically alpha fold and rosetta fold. With these methodologies, computing capacity is employed to reach the levels of prediction precision not previously possible. A crucial piece of the paper deals with ESMFold, a new tool that enhances computers' ability to predict protein structures greatly. Particularly in terms of unknown proteins or complex structures, the second part of the study carefully considers the pros and cons of these models to see how well they work. The techniques are compared in terms of how they work, how accurate they are and when they are suitable to use. This comparison framework facilitates a detailed understanding of the unique contributions and constraints of each method. A conclusion is drawn that provides potential future study directions within the subject and areas where improvement in prediction accuracy and expansion of application scope are needed in order to advance in the field.

Protein is an advanced computation method for predicting protein structures from amino acid sequences directly using deep learning, as shown in this study [14]. As structural bioinformatics and artificial intelligence advance, it is emphasized by the authors that accurate prediction of protein structures is indispensable in understanding biology, and there are many important implications for understanding biology and drug development as well as protein design which depend on accurate predictions. This research addresses the shortcomings of current methodologies and strives for the application of transformer topologies to boost the representation learning. Protein is an end to end transformer based architecture on integer encoded amino acid sequences. By following this approach the model is able to predict both the secondary and tertiary structures [15] from a single input sequence thus resulting in optimal prediction. The research identifies the thorough review of existing studies aimed at developing Protein, highlighting recent successes and challenges of deep learning application to predicting protein structures. However, according to the results, Protein is capable of predicting protein structures, though more improvements are still required to make the model more accurate especially for structurally more complex features. They introduce a benchmark system as well as visualization tools for analysis. The accessibility of the implementation and source code facilitates additional research and development in the domain.

III. METHODOLOGY

To enable rapid characterization of the novel pharmacidal agents, against particular targets, to speed up drug discovery. To determine the efficiency of performing a ranking on the protein ligand pairing in terms of number of simulations per second, we propose this metric. To check the performance of the proposed Deep Protein-Ligand Interaction (Deep-PLIN) model, it counts how many protein–ligand pairs it identifies as binding or not binding each second, and is called protein–ligand pairs per second (PLP per second).

The representation of a protein as it includes the amino acid sequences, the structural components and chemical properties is shown in Figure 1. Thus, in protein sequences, there are 15 different types of amino acids plus one unknown type, arriving at 19 unique types and represented by a 19 element one-hot vector. We frequently use a system called SMILES for taking a look a ligands' structure in a very simple line format, by atoms and bonds. Ligands have SMILES strings made up of 1 to 64 characters where each character is represented by a unique integer. There are three feature extraction modules and each of the three have a linear embedding layer. The protein pocket layer converts simple one-hot vector into many more compact dense vectors at the same time and ensure input format of the protein pocket and ligand match the size in the cross attention process. For the input data of proteins, pockets, and ligands, the embedding layer converts them to a 128 dimensional dense vector. And so in the case of embedding matrices that are 150 *128, we produce one of 1000 *256, and another of 150 *128.

Upon passing through the embedding matrices, the later cross attention processes will receive the complex

relationship between protein and ligand properties as input. In cross-attention technique, the model can assess the importance of different protein and ligand properties toward predicting binding affinity and hence what the most important interactions are for complex formation. Deep-PLIN is a method that is different from traditional methods that rely on fixed properties, and allows Deep-PLIN to adapt to the uniqueness of different protein–ligand systems. In addition, dense vector representations enable the model to tackle huge amounts of data while increasing the computing speed, and thus promotes the use of the model in virtual screening and drug development.

Dilated convolutions effectively acquire the multi scale contextual information by enlarging the receptive field of the kernels with various dilation rates. Using this feature, diluted convolutions were applied to investigate such multi-scale, long range intra-molecular interactions in protein and ligand sequences. After embedding layer, the protein feature extraction is performed with a dilated convolution block falling of four dilated convolution layers: 32, 64, 64, and 128 kernel respectively. The first dilated convolution layer has a kernel with size 3, while each of its subsequent dilation rates is 1, 2, 4, 8 and 16. Meanwhile, the dilated convolution block, in addition to the cross-attention layer, is introduced after a ligand feature extraction. This block contains 3 dilated convolutional layers with kernel size (32, 64, 128), respectively. Following each layer, the convolutions with kernel size of 3 are applied and use four dilation rates of 1, 2, 4, and 8 respectively. Next we apply a max pooling layer after either applying the 1D dilated convolution block or 1D conventional convolution block.

As the implementation of I-NN relies on data, i.e., I-NNs learn how to make connections from a set of data, I-NNs are extremely sensitive to how good and how much training data a I-NN has. We think that these models use hierarchies of features which they have acquired to generate their predictions. In contrast to a number of earlier neural networks, the proposed I-NN can learn to recognize those features in input data and is more flexible with regards to different types of changes in the input data.

The message passing iteration uses unique weights for each core ($Core_i$) implying that different cores do not share weights. This architecture is based on the idea that we have $Core_i$ that will transform an input latent representation L) to an output latent representation L_j. The weights need to be spread out which may ‘over burden’ the independent I-NNs within each core and thus impair them from completing this change successfully.

In training, we used a triangle waveform cyclical learning rate (C-LR) that cycles 5 epochs every. The base learning rate used by the C-LR was determined by a decay function, and an integration of the C-LR method with a schedule for learning rate decay was made. To speed up training, we used a learning rate of 0.1 and lowered it according to the number of workers (N). We validated the convergence 8 epochs of training and evaluated the validation accuracy. Convergence is assessed using a 1% subset of the 80% test set as a proxy. By using this method, there was a reliable link between how the model performed on the entire 80% test set and how well the model did on that 80% test set for checks after each training cycle without significantly increasing the overall training time.

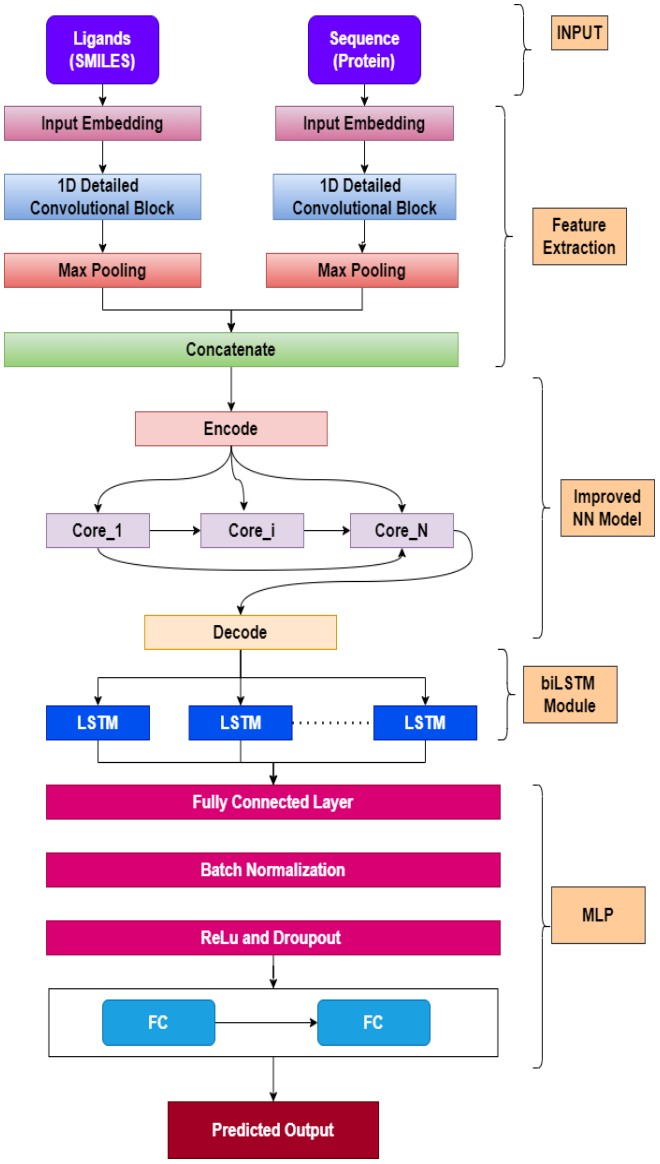

Figure.1 Proposed Deep-PLIN Model Architecture

We provide the concept of 'R-NN states' in order to utilize the correlation between distantly located interacting residues for prediction. Particular states of the I-NN (improved neural network) are these states which correspond to sequence places known to interact as pairs. We characterize these states in the I-NN as any interactive match state as shown in Fig 1. So ideally we want to predict the annotation for a certain place in the sequence and without knowing the place's within the set, to which position it belongs to. It would allow us to modify the likelihood that that position is classified as a state by using pre-calculated log-odds scores. Evenly, the issue lies in two parts: first that the information we want, that is, whether one location is part

of an interacting pair, is precisely the information we are trying to predict at a time that is otherwise unknowable.

Algorithm: Improved I-NN Procedure

1. Input: Extracted features for encoding $f_0 \ldots . f_k$
2. Output: Decoded Trace-back vector
3. Initialization: $f_0(0) = 1, f_k(0) = 0 \; for \; m > 0$
4. $f_j(i) = e_j(x_i) \max x_k \left(f_k(i-1) a_{kj}\right)$
   IF j is not an R-NN State
5. $e_j(x_i) \max x_k \left(f_k(i-1) a_{kj}\right) + E * Q(x_i'; x_i)$
   IF j is an R- NN
6. $P(x, \pi^*) = \; max_k(f_k(L) a_{k0});$
   $\pi_L^* = argamx_k(f_k(L) a_{k0})$
7. Trace-back: $\pi_{i-1}^* = ypt_i(\pi_i^*)$

An iterative methodology, which enhances the prediction of interacting residues, is presented. Then, we first obtain a tentative forecast of interacting residues from the I-NN output. Although not precise in this prediction, it should be enough as a basis upon which to recognize probable R-NN states. In addition, the log odds associated with these projected R-NN states can be used to further annotate these specific spots. Each round uses the new annotations to further improve the prediction of interacting pairs and the R-NN states, and this process can be repeated multiple times. We repeat this, so that the model will slowly increase its understanding about how remote residues interact, and thus more accurately fold and proteins function.

The biLSTM module works in a similar way to the GNN basics with the only difference that the consecutive feature outputs produced by two ResNet based GNN modules are outputted in this case but before reached the final output, they are averaged. It can learn long term dependencies on the inputs using BiLSTM (bidirectional long short memory). It is a network that processes molecules, proteins, as a forward and reverse chain, equivalent to molecular and protein datasets. After that starts concatenation of the output from both directions giving it a singular output vector. The biLSTM output vector is flattened into MLP module and fed into three sequential FC layers to obtain final output. In the end, the output is sent through a sigmoid function that gives a binary classification result (1 or 0).

This design makes it easier for the model to learn complicated relationships between molecular and protein features by achieving better prediction accuracy. The advantage in using ResNet-based GNN modules is that it is easier to extract the detailed features out of the raw input data. The advanced characteristic is extracted from unprocessed data by employ using of ResNet based GNN modules. These properties are processed by the biLSTM which acquires long range relationship that are missed out by simpler models. The final MLP module translates the acquired representations into a binary classification output that is easy and interpretable.

This is because the important components for predicting precisely are the efficacy of the architecture in capturing complex interactions between molecular and protein characteristics. With the aid of ResNet based GNN modules, the model will pull important features from the raw input data, allowing us to know it better the biological processes involved. Such simple models can be enhanced by the biLSTM which is able to detect long range dependencies that are often neglected by the other simple models. In complicated biological systems, the complications involved in interactions between chemicals proteins can include multiple temporal or spatial dimensions.

IV. RESULTS AND DISCUSSION

.The proposed Deep-PLIN model was trained using Binding DB and Davis dataset. These datasets are open to public. As of August 27, 2020, the most recent version of the Binding DB database contains 3, 367, 337 experimentally determined binding affinities of 8, 005 target proteins and 875, 232 small pharmacological compounds. For training and testing our Deep-PLIN model, we randomly chose 80% of the prepared Binding DB data as training and reserve the remaining 20% as independent test. During the training phase, we isolated 10% of the training dataset as a validation and the remaining as a training subset for optimization of hyper parameters. We established several testing sets varying with respect to the amount of exposure to the training data to assess the generalizability of the model. The 'Drug unseen', 'Protein unseen' and 'None seen' set respectively comprise of pharmaceuticals not present in the training set, proteins absent in the training set, and drugs and proteins not visible in the training set together.

For predicting the protein–ligand binding affinity, we trained the Deep-PLIN model using the mean squared error loss function. Accordingly, we articulate the Mean Squared Error (MSE) loss function as follows.

$$MSE_{PLIN} = \frac{1}{N}\left(y_{k(actual)} - \hat{y}_{k(predicted)}\right)^2 \qquad (1)$$

In the above equation, N denotes total number of samples in the training dataset, $y_{k(actual)}$ denotes the empirically obtained binding affinity of sample k and $\hat{y}_{k(predicted)}$ denotes predicted binding affinity of sample k. Then, the Adam optimization algorithm with a constant learning rate equal to 0.001 is used to optimize this loss function. The final (optimal) model was the model that had the minimal error on the validation set after being trained, then researched different regularization strategies for the model developed. To combat over fitting and improve generalization, we had implemented L2 regularization or weight decay on the network weights. The weight decay parameter was picked by using a grid search method, which has been optimized to pick one that will minimize the performance of the validation set. We also used early stopping, checking the validation loss and stopping training once the loss stops improving for a number of epochs. This

made the model less likely to overfit the training data and ensures that the final model remains able to generalize on new data.

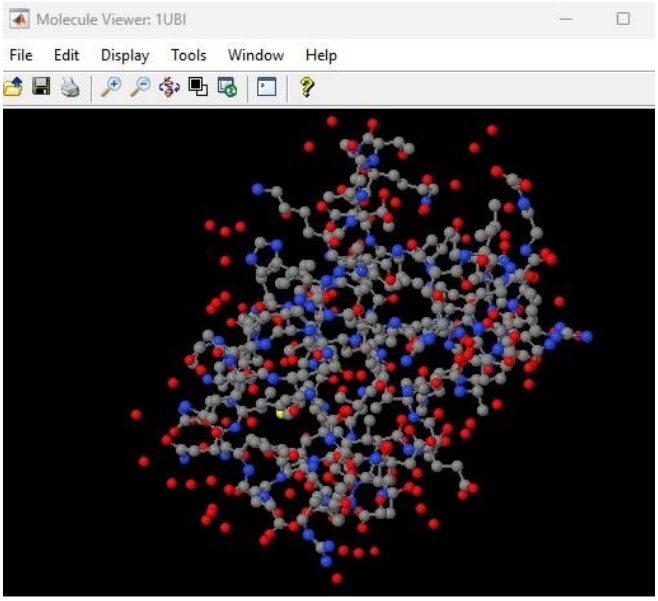

Figure.2 Prepared Dataset View using MATLAB Viewer

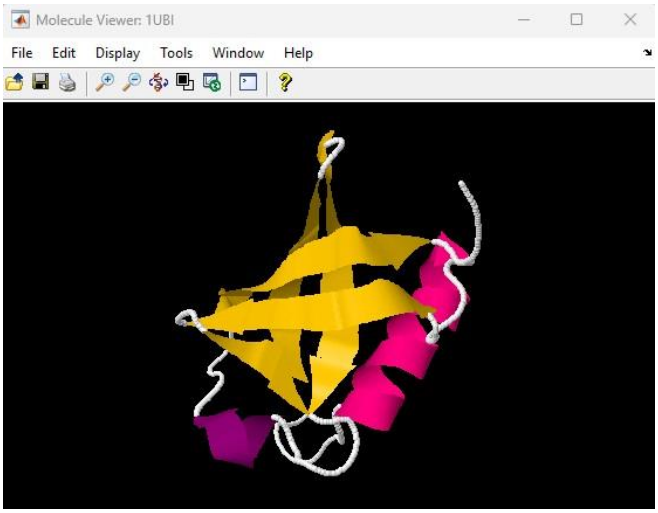

Figure.3 Concatenate Features of SMILES and Sequence View using MATLAB Viewer

The Figures 2 and 3 shows the visual details of the protein and the ligands features from the collected database. The figure2 shows the protein points of the cleaned and preprocessed data form database. The figure3 shows the proposed model based feature extracted and concatenate of both protein and ligands structure of the dataset.

The Figure.4 shows the integration of Binding DB and Davis dataset structural details taken for the proposed Deep-PLIN models training process. In this process, the total dataset are divided by 80% and 20% for training, testing and validation process. The C.E and H denotes the cumulative count of the relative feature structural arrangements.

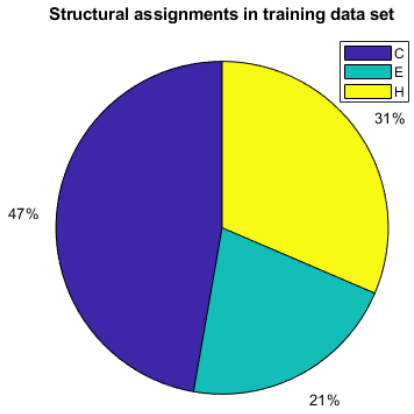

Figure. 4 Proposed Model Training dataset Structural Assignment

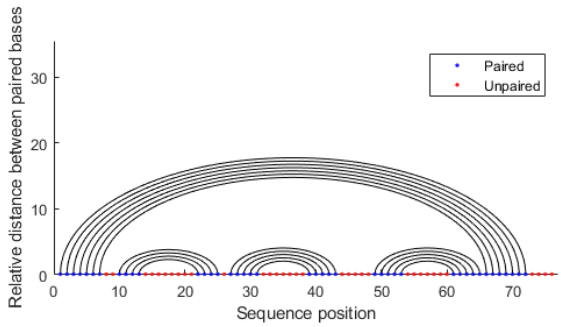

Figure.5 Proposed I-NN model based paring structure

Figure.5 shows the proposed I-NN Method based core structure from $core_i$ to $core_N$ using MSE_{PLIN} loss function. These pairing is formed based on the process of encode and decode manner. The figure 6 illustrations of the proposed model correlated prediction heat map relate to the relative delta index. Using this figure we proven our proposed model's predicted and classification error achieve lower value compared to the other conventional methods like Random forest, SVM.

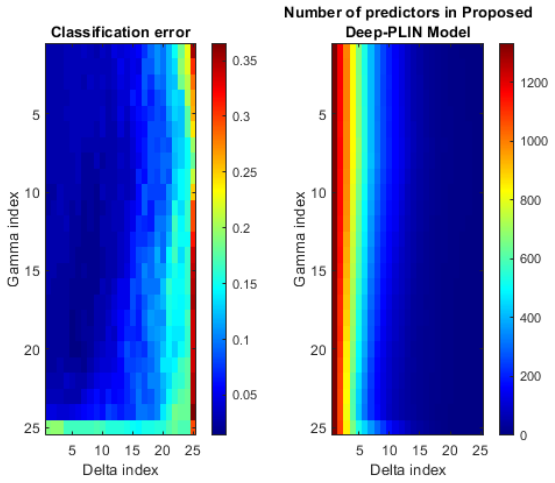

Figure.6 Proposed Deep-PLIN Model's Prediction Efficiency

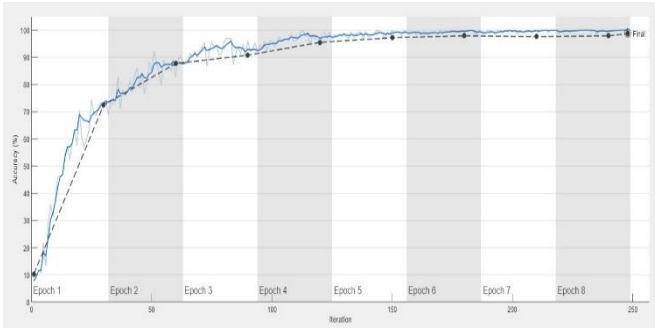

Figure.7 Proposed Deep-PLIN Model Accuracy Curve

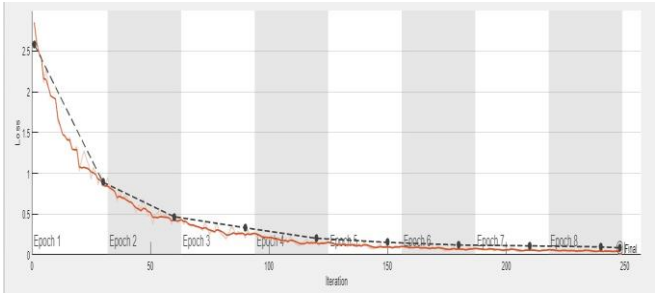

Figure.8 Proposed Deep-PLIN Model Loss Curve

The Figure 7 shows the proposed model's accuracy curve. In this curve we notice that the testing accuracy reached with the best optimal value within Epoch 5 with respect of 100th iteration. This value shows our proposed Deep-PLIN models classification accuracy efficacy. The figure.8 proves the model's reached the lowest error value of 0.1 within 4th epoch.

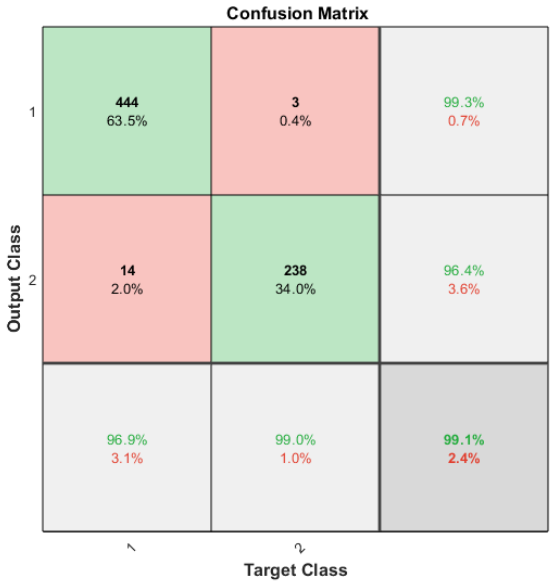

Figure.9 Proposed Model's Confusion Matrix

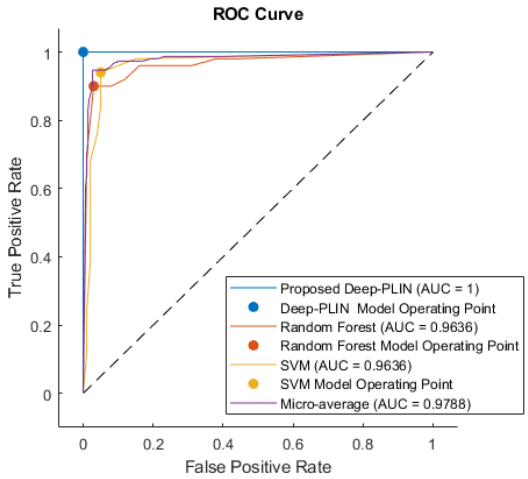

Figure.10 juxtaposition of Proposed ROC Curve with other conventional Models

We show the loss and metric progress as a function of training on the Binding DB training dataset in Figure 8. After 7 epochs, the validation loss stops declining, this means the training of the model stops. With further training, overfitting will become evident with an increase of validation loss. The values of Area under the Receiver Operating Characteristic curve (AU-ROC) metric obtained for the 0.98 for training and 0.97 for validation. In Figures 9 and 10, the results suggest that Deep-LPIN yielded 0.018 AUROC more than Random Forest and SVM. We can clearly see that AUROC values are nearly 0.9 for all models, which indicates that Deep-LPIN is most effective in predicting on the Binding DB set.

V. CONCLUSION

To classify one dimensional sequence data of proteins and medicinal compounds we successfully created the Deep-LPIN model. In the first place, we utilize predefined embedding methods to embed the raw drug molecular SMILES strings and target protein sequences into dense vector representations. Initially, we apply the pre-trained embedding methods on the raw drug molecular SMILES strings and target protein sequences and represent them as compact vector forms. Next, we use one dimensional GNN to gather features from the encoded dense vector representations by feed these encoded vector representations to head modules and ResNet based modules, separately. Once the feature vectors are combined they are fed into the biLSTM network followed by MLP module to predict active or inactive. The training data beyond protein structural knowledge allows us to speed up the drug discovery process and increase the success rate if the convergence conditions are met.

VI. REFERENCES

[1] Chen, Lingtao, Li, Qiaomu, Nasif, KaziFahim Ahmad, Xie, Ying, Deng, Bobin, Niu, Shuteng, Pouriyeh, Seyedamin, Dai, Zhiyu, Chen, Jiawei, and Xie, Yixin. 2024. "AI-Driven Deep Learning Techniques in Protein Structure Prediction". Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/ijms25158426

[2] Bertoline, Letcia M. F., Lima, A. N., Krieger, J., and Teixeira, Samantha K.. 2023. "Before and after AlphaFold2: An overview of protein structure prediction". Frontiers in Bioinformatics. https://doi.org/10.3389/fbinf.2023.1120370

[3] Jnes, Jrgen and Beltro, Pedro. 2024. "Deep learning for protein structure prediction and designprogress and applications". Springer Nature. https://doi.org/10.1038/s44320-024-00016-x

[4] Szelogowski, Daniel. 2023. "Deep Learning for Protein Structure Prediction: Advancements in Structural Bioinformatics". Cold Spring Harbor Laboratory. https://doi.org/10.1101/2023.04.26.538026

[5] Wuyun, Qiqige, Chen, Yihan, Shen, Yifeng, Cao, Yang, Hu, Gang, Cui, Wei, Gao, Jianzhao, and Zheng, Wei. 2024. "Recent Progress of Protein Tertiary Structure Prediction". Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/molecules29040832

[6] Qin, Yiming, Chen, Zihan, Ye, Peng, Xiao, Ying, Zhong, Tian, and Yu, Xi. 2024. "Deep learning methods for protein structure prediction". Wiley. https://doi.org/10.1002/mef2.96

[7] Wang, Y.. 2024. "Advancements in Protein Structure Prediction: Novel Bioinformatics Algorithms and Applications". None. https://doi.org/10.54254/2753-8818/2024.la18241

[8] Lau, Maggie. 2024. "Progress and Research Trends of Artificial Intelligence Incorporation in Protein Structure Prediction". International Conference on Multimodal Interaction. https://doi.org/10.1109/ICMI60790.2024.10586012

[9] Prasad, Dr.Smrity, Nandhini, Dr. N., Singh, Rajesh, Anuradha, D. A., Varshitha, Lakshmi, Averineni, and Debnath, Sandip. 2023. "Perspectives of machine learning on protein structure prediction and function". None. https://doi.org/10.1109/ICACITE57410.2023.10183157

[10] Altunkulah, Elif and Ensari, Yunus. 2023. "PROTEIN STRUCTURE PREDICTION: AN IN-DEPTH COMPARISON OF APPROACHES AND TOOLS". EskiehirTeknikniversitesiBilimveTeknolojiDergisi - C YaamBilimleriVeBiyoteknoloji. https://doi.org/10.18036/estubtdc.1378676

[11] Peng, Chunxiang, Liang, Fang, Xia, Yuhao, Zhao, Kailong, Hou, Minghua, and Zhang, Guijun. 2023. "Recent Advances and Challenges in Protein Structure Prediction". Journal of Chemical Information and Modeling. https://doi.org/10.1021/acs.jcim.3c01324

[12] Saharkhiz, Saber, et al.. 2024. "The State-of-the-Art Overview to Application of Deep Learning in Accurate Protein Design and Structure Prediction". Topics in current chemistry. https://doi.org/10.1007/s41061-024-00469-6

[13] Yang, Zhenyu, Zeng, Xiaoxi, Zhao, Yi, and Chen, Runsheng. 2023. "AlphaFold2 and its applications in the fields of biology and medicine". Springer Nature. https://doi.org/10.1038/s41392-023-01381-z

[14] Li, Jiaxuan, Wang, Lei, Zhu, Zefeng, and Song, Chen. 2023. "Exploring the Alternative Conformation of a Known Protein Structure Based on Contact Map Prediction". American Chemical Society. https://doi.org/10.1021/acs.jcim.3c01381

[15] Qiu, Xinru, Li, Han, Steeg, Greg Ver, and Godzik, Adam. 2024. "Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development". Biomolecules. https://doi.org/10.3390/biom14030339