Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 9
September 2023

PPIBA-D: Protein Protein Interactions Binding Affinity Through Deep Learning Techniques

Pallavi M. Shanthappa
Department of Computer Science School of Computing Mysuru CampusAmrita Vishwa Vidyapeetham,India
Nikitha A
Department of Computer Science School of Computing Mysuru CampusAmrita Vishwa Vidyapeetham,India

Abstract—Protein-Protein interactions(PPI) are vital effective working for the machinery in organisms and are important across biological activities. In this work, deep learning and machine learning models were used for predicting binding affinities of PPI. Both machine learning and deep learning algorithms were implemented in python. Six classes namely antigen-antibodies, enzyme-inhibitors, G-proteins, receptors, other-enzymes and mis-cellaneous were used. Autocovariance and Autocorrelation forall six classes are calculated followed by predicting the binding affinity and dissociation constant. 1048 protein complexes con- taining various protein sequences made up the dataset utilizedfor training, validating and testing the Logistic Regression , Gaussian Naive Bayes , GridSearchCV, ADABoost, Random Forest, Decision Tree, SVM, KNN, Deep Belief Network, CNN, RNN, LSTM. The overall accuracy of the system was about 95.59 percentage. The LSTM technique has been used to successfully identify unique binding affinity on a web server known as PPIBA-D.

Keywords—Autocovariance, Autocorrelation, Binding free energy, Dissociation constant, Long-short term memory.

I. Introduction

Protein complexes are formed through various biological processes [1] [2]. It is challenging to comprehend the iden- tification mechanisms and binding affinity of these protein complexes using both molecular and biological computation. Numerous biological activities, Including signaling cascades, gene expression, enzyme activity, and the structure of cellsand tissues, rely significantly on PPIs [3].

The Isothermal titration calorimeter, surface plasmon res- onance, yeast two-hybrid system, and Forster/fluorescence resonance energy transfer are primarily used to study protein-protein interactions. These methods allow for the evaluation of the binding free energy and dissociation constant of inter- acting proteins [4]. Determining protein-protein binding affin- ity requires labor-intensive, expensive, and time-consuming experimental methods [5]. To assist researchers in selecting protein complexes of interest based on their binding affinity, Computational methods have been developed. Protein binding affinity refers to the degree of contact between two protein molecules and measures how strongly they bind to each other.It is often reported as the dissociation constant (Kd) of the protein-protein complex [6]. Proteins with high binding affin- ity have strong interactions with their binding partners and aremore likely to bind, whereas those with low binding affinity
have weak interactions and are more likely to dissociate from their binding partners [7].

Various computational techniques have been developed for predicting binding affinity based on force-field potentials [8]
[9] , empirical evaluation [10] [11] , and free energy perturba- tion [12] [13]. However, these scoring function-based methods often fail to accurately predict binding affinities for diverse datasets, As they are typically tested on small datasets. With the vast and diverse amount of biological data available, it is crucial to improve current computer-based procedures in order to extract maximal knowledge, which is essential for effective storage and maintenance of the data [14]. The author
L. Zhao et. al., 2020, developed GAN-based approach forlearning valuable features from both labelled and unlabeled data and for estimating drug-target binding affinity through convolutional regression [15]. T.Siebenmorgen et. al., 2020, isolated protein complex estimation of binding affinity to a hypothesised complex structure of

binding and unbinding of proteins using a spatial coordinate shows promising accuracy needed to make reliable predictions [16]. The pRank estimatesan AUC-ROC of 96.8% and surpasses the other machine learning techniques. Additionally, miSVM and pRANK two multiple instance classifiers perform higher than a traditional SVM. This increase in efficiency demonstrates the value of using multiple instance for learning in protein binding affinity [17].

Another author T.Galochkina et. al., 2021, created a novel technique, to classify each protein sequence's flexibility posi- tion using evolutionary data acquired from protein homologous sequences MEDUSA algorithm developed by author employs physico-chemical properties of amino acids as input for a convolutional neural network. It predicts flexibility in two, three and five classes after being trained on an X-ray structural dataset that is not redundant [18]. B.Oliva et. al., 2012, devel- oped BIPS prediction server, bases its inference solely on pair-to-pair similarities. When using pair of sequence similarity the cut-off between 30% and 70% identity yields results that are comparably produced predictions based on groupings of orthologous genes [19]. The P.Varadwaj et. al., 2017, devel- oped DeepInteract an integrative deep learning method for PPI prediction. The database of Interacting proteins was used to retrieve the PPIs that interacted with the classifiers of protein
complexes from Escherichia coli, Drosophila melanogaster, Homo sapiens, and Caenorhabditis elegans. The prediction accuracy for each complex was 97.01%, 90.85%, 94.47%, and
88.91%, respectively [20].

This work is motivated from our previous work on ProAll-D server, which predicts allergens and non allergens of proteins using a deep learning approach [21] , and In silico method to predict multi-epitope design against norovirus were the study focuses on identifying novel epitopes against norovirus by us- ing various databases [22], and this work has developed the 3D structure of BRAF(V600E) via homology modelling, validated it via Ramachandran plot and verify3D, and implemented the docking method via virtual screening of protein with drug and simulation of molecular dynamics to further suggest powerful potent novel inhibitors for malignant melanoma management [23]. This paper presented CNN-MDR using a back propagation technique based on a convolutional neural network [24] , and melanoma cancer proposes a method for preprocessing, docking, binding, and protein computational simulation. The protein computational simulation identifies the optimum technique for treating [25]. This study details the strategies used to create a group of unique binding affinity prediction models, which make use of data obtained from the publically accessible server ISLAND [26]. Various machine learning techiniques and deep learning techniques are used to calcuate the binding free energy and dissociation constant for increased performance. F.soleymani et. al., 2022, focuses on contemporary deep learning methods that have been applied to challenges such as protein function prediction, protein- protein interaction and their locations, protein-ligand binding, and protein design [27].

II. Dataset

A total of 1048 protein datasets were acquired from several sources, including the National Centre for Biotechnology Information(NCBI) and RCSB (https://www.rcsb.org). Users can access the PPIBA-D predictor server and data by using the link https://doi:10.17632/vsbd5m9f93.1 .The dataset used in this study consists of several protein complexes from variousclasses.

In the below figure 1 depicts the protein sequence formatted as a fasta file. The symbol '<' signifies that the data is in fasta format and denotes the start of a new sequence. The header information following the '>' symbol contains the protein chains and scientific name of the protein. Only organisms classified as homosapiens are included in the analysis.

In protein classification, binding free energy and dissoci- ation constant of protein sequences were characterised using six protein classes [28]. There are 356 protein complexes of antigen-antibodies, enzyme-inhibitor of 145 complexes, other enzymes have 92 complexes, G-protein contains 137 protein complexes, receptor containing 97, and miscellaneous has 221 protein complexes. These classes are used to calculate auto-covariance and autocorrelation based on the

protein sequences using different machine leraning techiniques.

III. Methodology

Many deep learning and machine learning techniques have been implemented in Python. The logistic regression, ADA Boost algorithm, kNeighbors algorithm, Grid search CV, Gaus- sian naive Bayes algorithm, Support Vector Machine, Deci- sion Tree algorithm, Random Forest algorithm, Deep Belief Network(DBN), Recurrent Neural Network(RNN), Convolu- tional Neural Network(CNN) and Long Short-Term Mem- ory(LSTM).

A. Calculation of Binding Free Energy and Dissociation con-stant

The binding free energy (ΔG) and Dissociation con- stant(Kd) are used to express protein binding affinities. The following mathematical equation can be used to express re-
lationship between binding affinity and dissociation constant [26].

$$kd = [L][P]/[LP] \quad (1)$$

In the equation (1), The dissociation constant is Kd, the freeligand concentration is [L], the binding partner concentrationis [P], and the ligand-bound complex concentration is [LP]. Because a lower Kd shows that less ligand is need to occupy a significant number of binding sites, a lower Kd indicates a higher binding affinity. In contrast, a greater Kd denotes a lesser binding affinity.

The thermodynamic parameters known as dissociation con- stant (Kd) and free energy of dissociation (ΔG) reflect how strongly a ligand binds to a binding partner (such as a protein or enzyme). The energy change that takes place when the complex forms or dissociates is represented by the symbol (ΔG) With respect to Kd [26], the following equation holds:

$$\Delta G = -RTln(Kd) \quad (2)$$

In the equation (2) above (ΔG) represents binding free energy, Kd represents dissociation constant, R represents gas constant (1.987x10-3 kcal mol-1 K -1) and T represents
temperature (considered to be 30 degrees Celsius or room temperature) and ln is a natural logarithm. The equation
(2) negative sign represents the fact that the free energy of dissociation gets more negative as Kd drops, indicating a stronger binding (meaning the binding is more favorable). On the other hand, as Kd rises (signifying a weaker binding), the free energy of dissociation becomes less negative.

B. Autocovariance and Autocorrelation method

The relationship between the values in a time series can be evaluated statistically using autocovariance and autocor- relation. RMSE(Root Mean Squared Error) is a measure of the difference between the actual and predicted values of a regression model.

$$Cov(X_t, X_{t-h}) = (1/n) * \Sigma[(X_t - \mu) * (X_{t-h} - \mu)] \quad (3)$$

```
>2K8B_1|Chain A|Ubiquitin|Homo sapiens (9606)
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
>2K8B_2|Chain B|Phospholipase A-2-activating protein|Homo sapiens (9606)
ANQQTSGKVLYEGKEFDYVFSIDVNEGGPSYKLPYNTSDDPWLTAYNFLQKNDLNPMFLDQVAKFIIDNTKGQMLGLGNP

>4DRA_1|Chains A, B, C, D|Centromere protein S|Homo sapiens (9606)
HHHHHHMEEEAETEEQQRFSYQQRLKAAVHYTVGCLCEEVALDKEMQFSKQTIAAISELTFRQCENFAKDLEMFARHAKRT
TINTEDVKLLARRSNSLLKYITDKSEEIAQIN
>4DRA_2|Chains E, F, G, H|Centromere protein X|Homo sapiens (9606)
GSHMEGAGAGSGFRKELVSRLLHLHFKDDKTKVSGDALQLMVELLKVFVVEAAVRGVRQAQAEDALRVDVDQLEKVLPQLLLDF
```

Fig. 1. Sample protein sequence dataset in Fasta format.

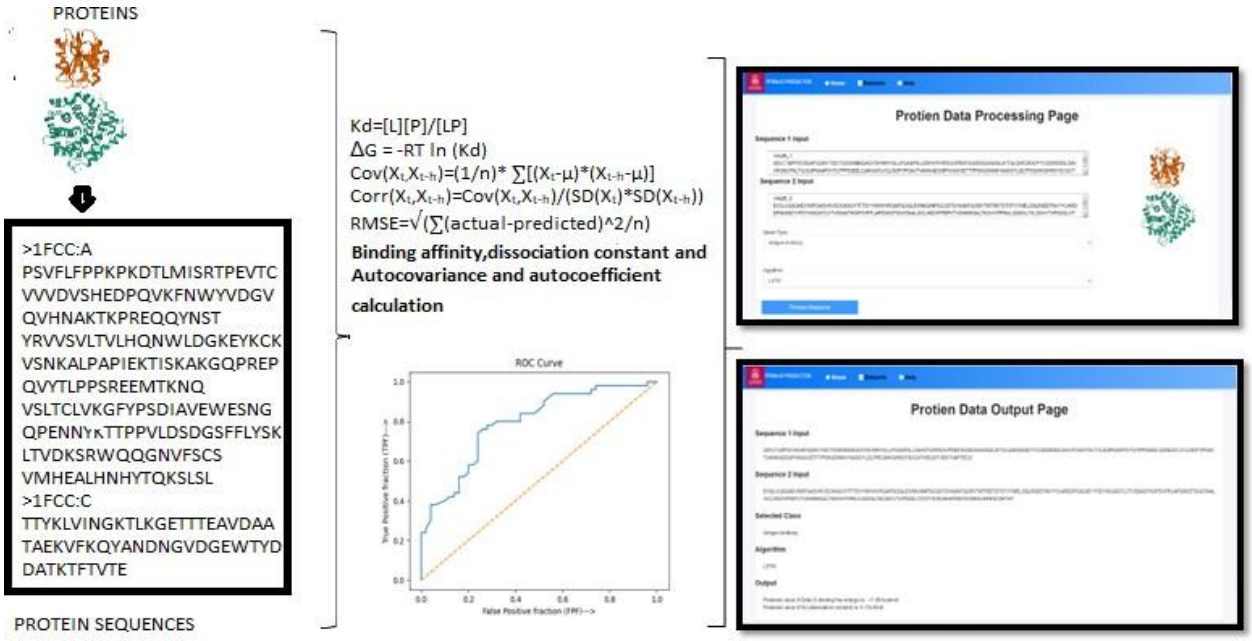


Fig. 2. Flow chart of proposed methodology.

Autocovariance is calculated using the above formula. In equation (3) n is the total number of time series observations. Xt and Xt-h are time series values at times t and t-h, and μ is the time series mean [29].

$$Corr(X_t, X_{t-h}) = Cov(X_t, X_{t-h}) / (SD(X_t) SD(X_{t-h}))] \quad (4)$$

Autocorrelation is calculated using the above formula. In equation (4) SD is time series standard deviation [29].

$$RMSE = (\Sigma(actual - predicted)^2 / n) \quad (5)$$

RMSE is calculated using the above formula. In equation (5), actual and predicted are the actual and predicted valuesof the response variable and the total number of observations is known as n [26].

C. Evaluations with other methods

The current work uses sequence based technique to cate- gorise protein complexes according to there binding affinities and this approach differs from structure-based approaches that have been suggested in the literature and are mostly used for predicting absolute binding affinity of protein complexes. These techniques have a number of drawbacks are: (i) Lot of descriptors are required [30]. (ii) Only rigid complexes exhibit good correlation [4]. (iii) It needs structural information [6].

Nonetheless, the current approach provides a number of benefits. (i) The characteristics are produced from sparsely utilised amino acid sequences. (ii) Different algorithms are used for each of the protein classes. (iii) They classify protein complexes into autocovariance and autocorrelation techniques.

(iv) Evaluates the binding affinity and dissociation constant value based on the algorithms and classes.

Table Ithe Results Of Autocovariance And Autocorrelation

Antigen-Antibody			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	0.43	0.12	0.6
ADA Boost	2.11	0.42	0.5
kNeighbors	0.36	0.08	0.3
GridSearchCV	-0.7	-0.14	0.2
Gaussian naive bayes	-0.71	-0.15	0.3
SVM	0.4	0.10	0.6
Decision tree	0.79	0.17	0.6
RandomForest	-1.44	-0.43	0.5
Deep Belief Network	-0.61	-0.13	0.6
RNN	-0.64	-0.10	0.3
CNN	-2.04	-0.41	0.6
LSTM	-0.43	-0.16	0.4
Enzyme-Inhibitor			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	0.31	0.07	0.3
ADA Boost	-0.41	-0.07	0.2
kNeighbors	2.65	0.35	0.4
GridSearchCV	-0.33	-0.06	0.5
Gaussian na¨ıve	-0.2	-0.2	0.4

bayes			
SVM	1.06	0.16	0.2
Decision tree	-1.71	-0.30	0.4
RandomForest	-0.82	-0.19	0.4
Deep Belief Network	0.39	0.08	0.2
RNN	-0.70	-0.09	0.6
CNN	0.21	0.07	0.2
LSTM	-0.5	-0.08	0.3
Other Enzymes			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	-0.36	-0.05	0.6
ADA Boost	-0.45	-0.08	0.4
kNeighbors	-0.07	-0.01	0.5
GridSearchCV	0.61	0.10	0.3
Gaussian na¨ıve bayes	0.07	0.01	0.4
SVM	-0.23	-0.05	0.3
Decision tree	-0.1	-0.02	0.6
RandomForest	-3.2	-0.51	0.2
Deep Belief Network	-0.87	-0.15	0.2
RNN	-0.08	-0.01	0.2
CNN	-1.02	-0.33	0.4
LSTM	0.5	0.1	0.2

Results And Discussion

Autocovariance and Autocorrelation for all six classes arecalculated followed by predicting the binding affinity anddissociation constant. The specific values of 0.43 for auto-covariance and 0.12 for autocorrelation indicate a moderatelevel of correlation between the sequence of antigen-antibody. The output of Autocovariance and autocorrelation are de- picted from TABLE I protein complex classes of Antigen-antibody, Enzyme-Inhibhitor and other enzymes are calculatedfollowed by predicting the binding affinity and dissociation constant.

A. Web server for predicting binding affinitiy

A web server called PPIBA-D predictor has been created to use model for machine learning and deep learning techniques to forecast binding affinity and dissociation constant.

The quick and user-friendly Python Flask framework is usedin its development.

Fig. 3. Commands for executing the webserver

Fig. 4. Interface of PPIBA-D predictor.

Firstly the path of the web app folder should be copied and pasted in cmd. Then within the Binding affinity folder, there is a python file namely "app.py" to execute this the second com-mand has to be followed. Once the above commands are exe- cuted, we get the link of the local host ("http://127.0.0.1:3121/)

B. Performance evaluation

The dataset was divided into 80% for training and 20% for testing purposes. TABLE II shows the results of the models based on the training and testing data, with True Positives (TP) and True Negatives (TN) for the expected binding affinity anddissociation constant. False negatives (FN) and False positives (FP) were also designated for incorrectly predicted binding affinity and dissociation constant respectively. Precision, which is the ratio of correctly predicted samples of total number instances returned, and recall, defined as the proportion of true positives that were accurately detected to all true positives, were calculated. The F1 score, which is a measure of a model'saccuracy, was calculated using the formula [(2 *(Accuracy * Recall)/(Precision + Recall)]. Additionally, a Python function was used to compute confusion metrics based on the true labels, predicted labels, and a threshold for prediction.

Table Ii The Results Of Autocovariance And Autocorrelation

Receptor Containing			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	3.36	0.55	0.6
ADA Boost	0.65	0.14	0.4
kNeighbors	-1.67	-0.24	0.2
GridSearchCV	-0.07	-0.02	0.4
Gaussian naïve bayes	-1.41	-0.23	0.4
SVM	-0.02	-0.00	0.2
Decision tree	-1.5	-0.23	0.5
RandomForest	1.40	0.30	0.5
Deep Belief Network	0.48	0.13	0.2
RNN	-0.19	-0.05	0.2
CNN	-1.61	-0.26	0.2
LSTM	0.39	-0.33	0.5
G-Protein			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	1.03	0.25	0.3
ADA Boost	0.92	0.19	0.5
kNeighbors	0.91	0.19	0.6
GridSearchCV	0.75	0.15	0.6
Gaussian naïve bayes	-0.23	-0.03	0.2
SVM	1.13	0.22	0.6
Decision tree	0.59	0.10	0.4
RandomForest	-1.11	-0.21	0.3
Deep Belief Network	-0.3	-0.07	0.2
RNN	0.63	0.10	0.6
CNN	0.04	0.01	0.3
LSTM	-0.23	-0.06	0.5
Miscellanous			
Methods	**Autocovariance**	**Autocorrelation**	**RMSE**
Logistic Regression	-1.39	-0.17	0.4
ADA Boost	0.65	0.29	0.4
kNeighbors	-1.36	-0.21	0.5
GridSearchCV	-2.56	-0.41	0.5
Gaussian naïve bayes	-0.56	-0.16	0.4
SVM	-0.87	-0.22	0.4
Decision tree	-2.85	-0.45	0.4
RandomForest	0.57	0.11	0.4

Deep Belief Network	-0.07	-0.03	0.6
RNN	0.55	0.10	0.2
CNN	-0.23	-0.06	0.2
LSTM	1.51	0.24	0.6

The output of Autocovariance and autocorrelation are de- picted from TABLE II protein complex classes of Recep-tor containing, G-protein and Miscellanous are calculated followed by predicting the binding affinity and dissociation constant.

C. *The calculated confusion metrics*

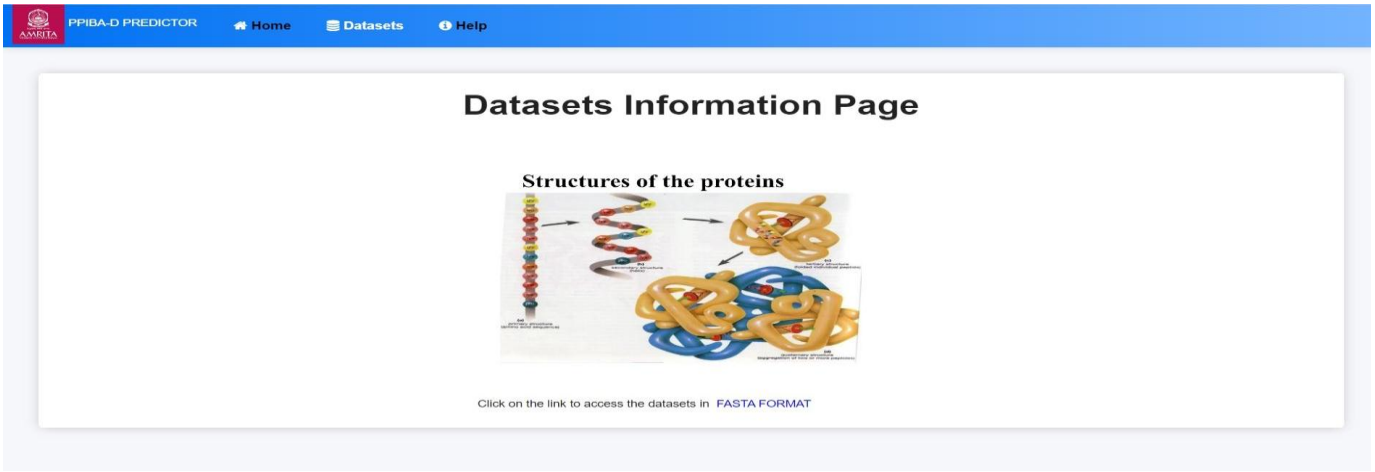
True Positive Rate (TPR): Sensitivity or recall refers to the proportion of true positives that the classifier properly identifies as such. False Positive Rate (FPR): The percentage of true negatives that the classifier incorrectly perceives as positives. The function then calculates the AUC using the NumPy library's trapezoidal rule with the step size serving as the x-coordinate spacing and the TPR values serving as the y-coordinates.

The Gaussian Naive Based of the naive Bayes algorithm, Bayes is a statistical classification predicting model. This algo-rithm produced a 94.78 percent accuracy for our dataset (Table 1). An ensemble method is called AdaBoost, or Adaptive Boosting. In supervised learning, boosting is used to reduce bias and variation. This model had a 93.45% accuracy rate. Using sigmoid function, which transfers a linear combination of the features to a value between 0 and 1, The logistic regression is used to create a model that predicts the likelihood that a protein sequence would have the desired quality,which has accuracy of 92.36 percent. In order to translate the data into a higher-dimensional feature space where separating hyperplanes could be simpler to discover, SVM can also be utilised The SVM method produced a 95.36 percent accurate result for our dataset. KNN is a non-parametric technique that does not call for training a model. Where as KNN creates predictions based on the labels of the Kneighbours of a novel protein sequence using a distance measure to identify those neighbours, Which produce results with a 93.38 percent accuracy. In order for GridSearchCV to function, a grid of all possible hyperparameter combinations must be created. Cross-validation is then used to assess how well the model performs with each set of hyperparameters. GridSearchCV algorithm produced results with a 93.43 percent accuracy. Decision Tree Classifier and Random Forest Classifier were similar to SVM, logistic regression and KNN, producing an accuracy of 95.67 and 95.14 percent. CNNs can be effective for protein sequence prediction because they can automatically learn local patterns or motifs in the protein sequences, which can be informative for predicting the protein function. The accuracy of this method was 95.17 percent. RNN may be trained to recognise patterns, such protein motifs or domains, that appear over numerous amino acids. The accuracy of RNN method was 95.37 percent. The accuracy of the LSTM model and DBN was 95.98 and 95.23 percent, respectively.

TABLE III. Reveals the conclusions from an analysis of the accuracy and confusion metrics performance based on the used algorithms. The LSTM approach clearly performs better and more consistently across all measures.

TABLE IV.Shows that in general, the models perform better with the 80-20 split than with the 40-60 and 70-30 splits. The LSTM model also performs very well, with an accuracy of 95.98 with the 80-20 split and 93.89% with the 40-60 split. The LSTM method was applied to a well known data set for the determination of proteins binding affinities and dissociation constants where a protein must be categorised into protein classes. When compared to other algorithms, LSTM provides much faster and more highly specified performance. Using performance evaluation metrics, all the strategies that were put into use were evaluated and contrasted. LSTM was the best-performing model with a 95.98% accuracy rate. A more advanced form of the RNN is the LSTM (recurrent neural

Fig. 5. Execution of PPIBA-D Predictor .



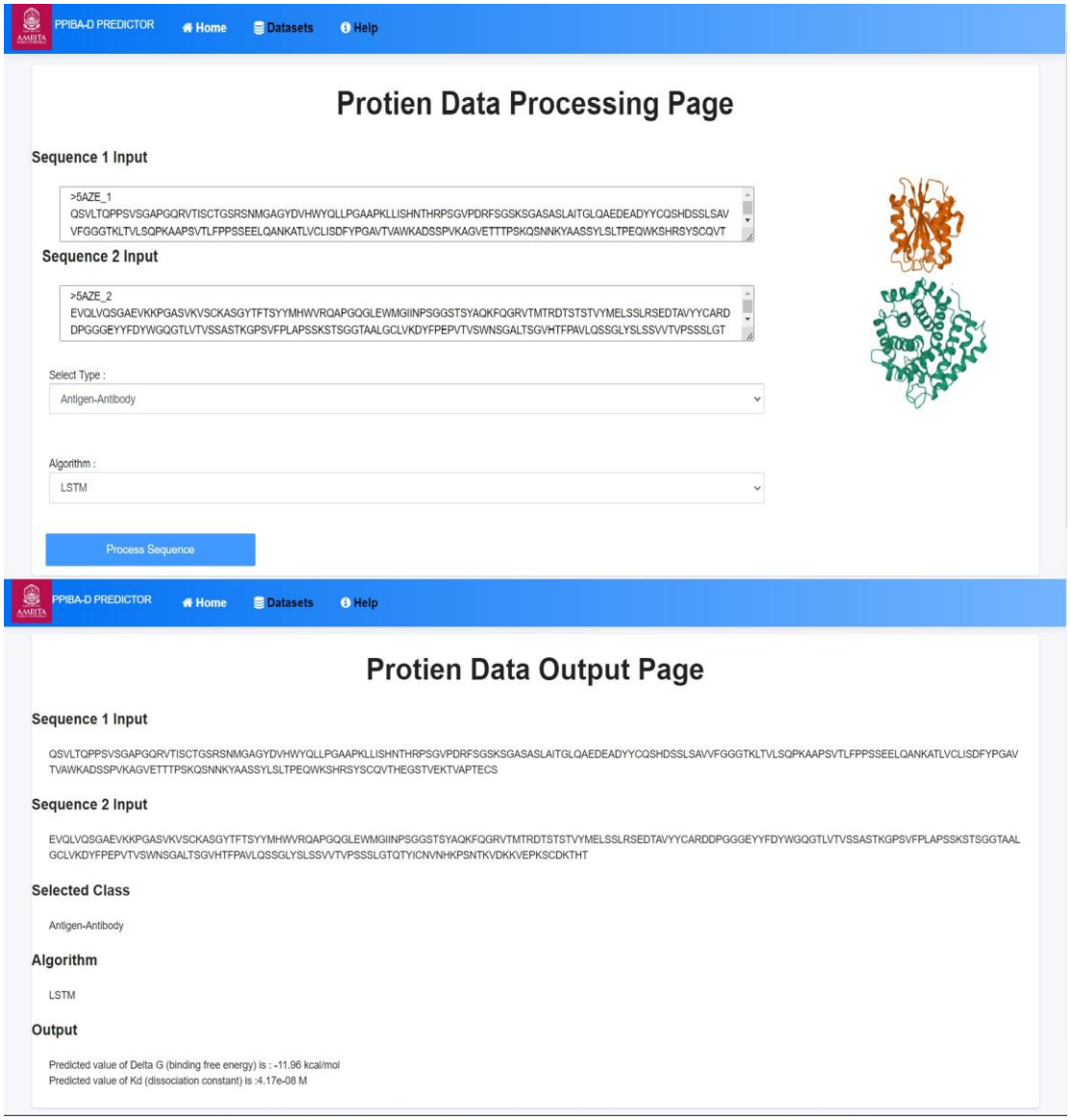


Fig. 6. The Data-set section.

Table Iii Performance Evaluation Of Various Methods

Methods	Accuracy	Precision	Recall	F1-score
Logistic Regression	92.36	1.00	0.86	0.92
SVM	95.36	0.91	1.00	0.95
kNeighbors	93.38	1.00	0.88	0.93
ADA Boost	93.45	1.00	0.87	0.93
GridSearchCV	93.43	1.00	0.88	0.94
Decision Tree	95.67	0.91	1.00	0.95
Gaussian na¨ıve bayes	94.78	1.00	0.90	0.95
RandomForest	95.14	0.91	1.00	0.95
RNN	95.37	1.00	0.91	0.95
Deep Belief Network	95.23	1.00	0.91	0.95
CNN	95.17	0..91	1.00	0.95
LSTM	95.98	0.91	1.00	0.95

network). Figure 7 shows that TPR rises, FPR rises as well, until at a certain point, which is around 0.95 then the graph becomes constant. Area under curve (AUC) is the region of ROC curve between (0,0) and (1,1). The overall threshold values for the model's performance are effectively aggregated by AUC.

Table Ivanalysis Of Training And Testing Based Prediction Results From Different Methods

	80:20	70:30	40:60
Methods	**Accuracy**	**Accuracy**	**Accuracy**
Logistic Regression	92.36	91.23	90.46
ADA Boost	93.45	92.12	90.43
kNeighbors	93.38	92.23	91.46
GridSearchCV	93.43	92.21	91.67
Gaussian na¨ıve bayes	94.78	93.48	92.64
SVM	95.36	94.11	90.16
Decision tree	95.67	93.48	91.69
RandomForest	95.14	92.62	91.26
Deep Belief Network	95.23	94.03	92.49
RNN	95.37	93.16	92.46
CNN	95.17	93.27	92.17
LSTM	95.98	94.79	93.89

The LSTM model performs best for our data observations since it has highest AUC which denotes it has a largest area under the curve. Along with the other methods, the LSTM method has been used in the software developmentfor the server PPIBA-D to forecast probable binding affinity. The quick and user-friendly Python flask framework is usedin its development. The LSTM is a quick recurrent neural network model for protein binding free energy and dissociation constant

prediction. LSTM works best with time series or sequential data and is generally used to tackle long-term relianceissues. As protein dataset is similarly ordered in a sequential style, the LSTM would be a better method for predicting binding affinity with increased performance. Additionally, a few deep learning and machine learning techniques have been evaluated for categorization requirements.

As indicated in Figure 4. There are three separate sections, which are Home, Datasets, and Help. The user enters a pairof protein sequences in the input area, then selects a class and algorithm. The results will be displayed as depicted in Figure

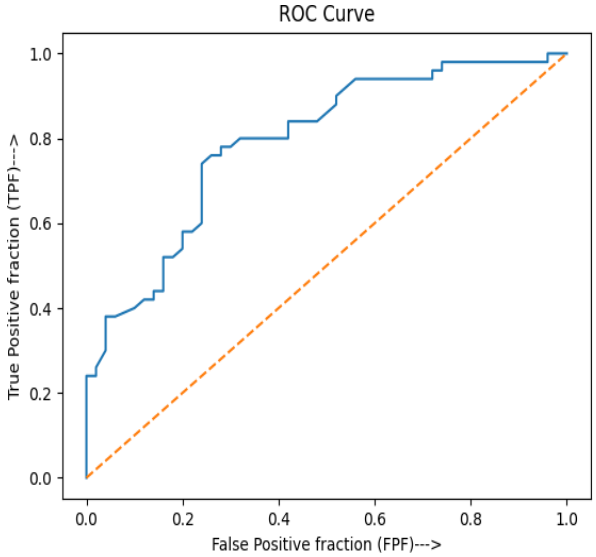


Fig. 7. ROC curve for the LSTM model

5. Predicting the binding affinity and dissociation constant of protein sequences. As shown in Figure 6. We have uploaded the fasta file formatted data sets used in our research in the dataset section.
The user has to enter the protein sequence such as sequence 1 and sequence 2 and select the class and algorithm of your choice in the home section. The model predicts the valueof ΔG (binding free energy) and value of kd(dissociation constant). In command promt it display the value of auto-covariance , autocorrelation and RMSE value depending on the classes and sequences we choose to run.

IV. Conclusion

A protein sequence based method is created that can beused to forecast the binding affinities and dissociation constant of protein complexes. The main objective of this study is to improve the accuracy of the method while keeping the earlier considerations of machine learning and deep learning models. The autocovariance and autocoefficient for all the six classes are predicted for the machine learning techiques such as The Logistic Regression, Gaussian Naive Bayes, GridSearchCV, ADA Boost, Random Forest, Decision Tree, Support Vector Machine, k-Nearest Neighbour's. Deep learning models like Deep Belief Network, CNN, RNN and LSTM model to pedict the binding affinity and dissociation constant of the protein sequences. Furthermore, the AUC value of LSTM (the best performance) was 0.95(95 precentage). Overall this study is the only one to use the mentioned techniques to assess protein binding affinity and dissociation constant which will serve asa model for future predictions of protein binding affinity.

References

[1] S. Jones and J. M. Thornton, "Principles of protein-protein interactions.," *Proceedings of the National Academy of Sciences*, vol. 93, no. 1, pp. 13–20, 1996.M. Nooren and J. M. Thornton, "Diversity of protein–protein interac- tions," *The EMBO journal*, vol. 22, no. 14, pp. 3486–3492, 2003.

[2] M. M. Gromiha, K. Yokota, and K. Fukui, "Energy based approach for understanding

the recognition mechanism in protein–protein complexes," *Molecular BioSystems*, vol. 5, no. 12, pp. 1779–1786, 2009.H. Moal, R. Agius, and P. A. Bates, "Protein–protein binding affinity prediction on a diverse set of structures," *Bioinformatics*, vol. 27, no. 21,pp. 3002–3009, 2011.

[3] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computa- tional prediction of protein–protein interactions," *Molecular biotechnol-ogy*, vol. 38, pp. 1–17, 2008.

[4] P. L. Kastritis and A. M. Bonvin, "On the binding affinity of macro- molecular interactions: daring to ask why proteins interact," *Journal of The Royal Society Interface*, vol. 10, no. 79, p. 20120835, 2013.

[5] L. Fielding, "Nmr methods for the determination of protein–ligand dissociation constants," *Progress in Nuclear Magnetic Resonance Spec- troscopy*, vol. 51, no. 4, pp. 219–242, 2007.

[6] N. Horton and M. Lewis, "Calculation of the free energy of association for protein complexes," *Protein Science*, vol. 1, no. 1, pp. 169–181, 1992. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of molecular biology*, vol. 280, no. 1, pp. 1–9, 1998.

[7] S. Qin, X. Pang, and H.-X. Zhou, "Automated prediction of protein association rate constants," *Structure*, vol. 19, no. 12, pp. 1744–1751, 2011.

[8] J. Audie and S. Scarlata, "A novel empirical free energy function that explains and predicts protein–protein binding affinities," *Biophysical chemistry*, vol. 129, no. 2-3, pp. 198–211, 2007.

[9] X. H. Ma, C. X. Wang, C. H. Li, and W. Z. Chen, "A fast empirical approach to binding free energy calculations based on protein interface information," *Protein engineering*, vol. 15, no. 8, pp. 677–681, 2002.

[10] Y. Su, A. Zhou, X. Xia, W. Li, and Z. Sun, "Quantitative prediction of protein–protein binding affinity with a potential of mean force consider- ing volume correction," *Protein Science*, vol. 18, no. 12, pp. 2550–2558,2009.

[11] S. M. Sekhar, G. Siddesh, M. Raj, and S. S. Manvi, "Protein class prediction based on count vectorizer and long short term memory," *International Journal of Information Technology*, vol. 13, no. 1, pp. 341–348, 2021.

[12] L. Zhao, J. Wang, L. Pang, Y. Liu, and J. Zhang, "Gansdta: Predicting drug-target binding affinity using gans," *Frontiers in genetics*, vol. 10,p. 1243, 2020.

[13] T. Siebenmorgen and M. Zacharias, "Computational prediction of protein–protein binding affinities," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 10, no. 3, p. e1448, 2020.

[14] F. u. A. Afsar Minhas, E. D. Ross, and A. Ben-Hur, "Amino acid com- position predicts prion activity," *PLoS computational biology*, vol. 13, no. 4, p. e1005465, 2017.

[15] Y. Vander Meersche, G. Cretin, A. G. de Brevern, J.-C. Gelly, andT. Galochkina, "Medusa: prediction of protein flexibility from sequence,"

[16] *Journal of molecular biology*, vol. 433, no. 11, p. 166882, 2021.

[17] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, and B. Oliva, "Bips: Biana interolog prediction server. a tool for protein–protein interaction inference," *Nucleic acids research*, vol. 40, no. W1, pp. W147–W151, 2012.

[18] S. Patel, R. Tripathi, V. Kumari, and P. Varadwaj, "Deepinteract: deep neural network based protein-protein interaction prediction tool," *Current Bioinformatics*, vol. 12, no. 6, pp. 551–557, 2017.

[19] P. M. Shanthappa and R. Kumar, "Proall-d: protein allergen detection using long short term memory-a deep learning approach," *ADMET and DMPK*, vol. 10, no. 3, pp. 231–240, 2022.

[20] P. M. Shanthappa, R. Suravajhala, P. Suravajhala, G. Kumar, and

[21] N. Melethadathil, "In silico based multi-epitope vaccine design against norovirus," *Journal of Biomolecular*

Structure and Dynamics, pp. 1–11, 2022.

[22] M. Pallavi and P. K. HS, "In-silico analysis to determine the efficient drug for malignant melanoma using molecular dynamics," *Biomedical and Pharmacology Journal*, vol. 13, no. 3, pp. 1463–1470, 2020.

[23] P. Fahad and M. Pallavi, "Prediction of human health using machine learning and big data," in *2018 International Conference on Communi- cation and Signal Processing (ICCSP)*, pp. 0195–0199, IEEE, 2018.

[24] Sidharth, K. Priya, and M. Pallavi, "Adaptive approach to melanoma cancer using mapk pathway," in *2017 International Conference on*

[25] *Communication and Signal Processing (ICCSP)*, pp. 1463–1467, IEEE, 2017.

[26] W. A. Abbasi, A. Yaseen, F. U. Hassan, S. Andleeb, and F. U. A. A. Minhas, "Island: in-silico proteins binding affinity prediction using sequence information," *BioData Mining*, vol. 13, no. 1, pp. 1–13, 2020.

[27] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski, and D. Spinello, "Protein–protein interaction prediction with deep learning: A compre- hensive review," *Computational and Structural Biotechnology Journal*, 2022.

[28] K. Yugandhar and M. M. Gromiha, "Protein–protein binding affinity prediction from amino acid sequence," *Bioinformatics*, vol. 30, no. 24, pp. 3583–3589, 2014.

[29] X. Wang, R. Wang, Y. Wei, and Y. Gui, "A novel conjoint triad auto covariance (ctac) coding method for predicting protein-protein interaction based on amino acid sequence," *Mathematical biosciences*, vol. 313, pp. 41–47, 2019.

[30] F. Tian, Y. Lv, and L. Yang, "Structure-based prediction of protein– protein binding affinity with consideration of allosteric effect," *Amino Acids*, vol. 43, pp. 531–543, 2012.