AKADÉMIAI KIADÓ

UNIVERSITY of DEBRECEN

International Review of Applied Sciences and Engineering

13 (2022) 1, 98–105

DOI:
10.1556/1848.2021.00315

ORIGINAL RESEARCH PAPER

Protein active site prediction for early drug discovery and designing

Aqsa Yousaf[1], Tahira Shehzadi[1*], Aqeel Farooq[2] and Komal Ilyas[3]

[1] Department of Computer Science, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan

[2] School of Science, Computing and Engineering Technologies, Swinburne University of Technology, VIC 3122, Australia

[3] Research Center for Modelling and Simulation, National University of Science and Technology, Islamabad, Pakistan

Received: May 23, 2021 • Accepted: July 23, 2021
Published online: September 7, 2021

ABSTRACT

Adenosine triphosphate (ATP) is an energy compound present in living organisms and is required by living cells for performing operations such as replication, molecules transportation, chemical synthesis, etc. ATP connects with living cells through specialized sites called ATP-sites. ATP-sites are present in various proteins of a living cell. The life span of a cell can be controlled by controlling ATP compounds and without the provision of energy to ATP compounds, cells cannot survive. Countless diseases treatment (such as cancer, diabetes) can be possible once protein active sites are predicted. Considering the need for an algorithm that predicts ATP-sites with higher accuracy and effectiveness, this research work predicts protein ATP sites in a very novel way. Till now Position-specific scoring matrix (PSSM) along with many physicochemical properties have been used as features with deep neural networks in order to create a model that predicts the ATP-sites. To overcome this problem of complex computation, this exertion proposes k-mer feature vectors with simple machine learning (ML) models to attain the same or even better performance with less computation required. Using 2-mer as feature vectors, this research work trained and tested five different models including KNN, Conv1D, XGBoost, SVM and Random Forest. SVM gave the best performance on k-mer features. The accuracy of the created model is 96%, MCC 90% and ROC-AUC is 99%, which are the same or even better in some aspects than the state-of-the-art results. The state-of-the-art results have an accuracy of 97%, MCC 78% and ROC-AUC is 92%. One of the benefits of the created model is that it is much simpler and more accurate.

KEYWORDS

sequence-based features, ATP-sites, XGBoost, Conv1D, MCC, AUC, ROC, ATP, PSSM

1. INTRODUCTION

Proteins perform a variety of functions within organisms, including catalytic metabolic reactions, DNA replication, response to stimuli, cell and organism structure, and transport of molecules from one place to another. A particular amino acid pattern specifies a specific binding site for specific ligands. Protein-ligand interactions are essential for a number of biological processes such as transportation through membrane [1] contraction of muscle [2], replication of DNA [3] and transcription of DNA [4]. Therefore, it is a guide for the precise identification of protein-ligand binding residues [5], leading to marking of protein function [6], and the development of new drugs for human diseases like cancer [7], diabetes [8], as well as Alzheimer's [9]. Among these binding sites, Adenosine-triphosphate (ATP) acting an important role as it is the straight energy source for most of the biological activities for mankind from bacteria to humans [10]. ATP intermingles with proteins through protein-

*Corresponding author.
E-mail: tahirashehzadi9@gmail.com

AKJournals

ATP-binding remains and delivers chemical energy to the protein by ATP hydrolysis. This energy helps proteins in performing their specific functions. Adenosine-5-triphosphate (ATP) also provides an important role in cell motility, DNA signaling and various metabolic processes. For chemical energy-boosting, a protein can provide a variety of biological functions. It is necessary to examine ATP residues in proteins as ATP-related debris can be used as interesting targets for chemotherapeutic agents. Therefore, the precise localization of Adenosine-5-triphosphate (ATP) sites residues is crucial for the analysis of protein function and drug modeling.

To determine the structure of a protein is relatively difficult as compared to find the sequences of the protein. Usha and Selvaraj [11] first identify adenine as well as guanine in proteins using structural information. This was the innovator work on molecular discrimination and recognition. In Ref. [12], ATPint was designed. This was the first ATP site predictor which was developed using sequences-based information. ATPint used PSSM [13] and 7 other sequential descriptors as features. ATPint used SVMs as a classifier. The accuracy of ATPint was 0.66 and Matthews correlation coefficient (MCC) was 0.33. In 2012, Yu et al. [14] developed the Adaboost model to address the imbalanced dataset problem. He achieved better performance than ATPint and ATP site [15]. He called this method TargetATP. Gao et al. [16] in 2019 attempted to extract input features most effective from PSSM and trained on support vector machine classifier in order to classify ATP and non-ATP sites. The area under the curve (AUC) of the classifying technique has a value of 0.899 which shows most of the information regarding binding sites is concealed in sequences. Arif et al. [17] in 2020 used the gradient boosting technique. It also includes two layers, the first was the KNN layer and the second was the threshold layer. It was thought that the deeper model would perform better. Hu et al. [18] in 2018 proposed a method called ATPbind. This method used a hybrid approach to predict ATP-sites in proteins. It means it used sequentially as well as structural information.

DeepATP was built by Nguyen et al. [19] to predicted ATP-binding residues. It used two-dimension convolutional neural networks architecture. They addressed the issue of the imbalance dataset and acquire a very high AUC of 0.991 on the independent dataset. Deeper Bind is developed by Hassanzadeh and Wang [20] for DNA protein binding prediction. They used CNN and RNN. Song et al. [21] developed a technique for protein ATP-binding sites prediction by using two convolutional neural networks architectures which comprised a residual-inception as well as a multi-inception-based predictor. These two methods contain sequence-based information. The final prediction results from an ensemble learning method that takes input from the two predictors. Dataset227 and dataset429 are used in this study. In all methods either in Machine learning or deep learning, different sets of the same features have been used. One thing that is common in them is that all the methods used a Position-specific scoring matrix as their main features. It is a common perception that evolutionary features can predict better. If I summarize the literature work, ATP prediction has been done with SVM, KNN, RF and CNN with PSSM as features (Although other features have been used but they have very little effect on performance) and imbalanced problems handled by undersampling, oversampling and weighted cross entropy metric. Often the experiment is performed to collect data and data is obtained using sensors [22–24]. Once the data is obtained, then machine learning (ML) algorithms are applied to achieve prediction results [25].

This method provides a prediction method Instead of PSSM and uses 2-mer vector as a feature. It is found that the 2-mer feature [26] vector for our model is as follows: 1. A Cartesian product of two vectors having 20 universal amino acids in each, results in a vector of dimension 400. 2. After finding all the unique sequences of length two in the query, the algorithm searches them in a 2-mer feature vector, found in step 1. Replace the sub-sequences in the 2-mer feature vector which is present in the query vector with numeric value 1 and replace the other subsequences of the 2-mer feature vector which is not present in the query with numeric value 0. This results in a feature vector of our query sample and at this point, the accomplished vector is normalization. This feature vector is simply computed but has very high performance. Overall, five different models included KNN, Conv1D [27], XGBoost, SVM and Random Forest are trained and tested. SVM gave the best performance on these features. Our model gave an accuracy of 0.96, MCC 0.90 and ROC-AUC 0.99, which are the same or even better in some aspects than the state of the art results.

2. METHOD

2.1. Window size

A single amino acid is an example but the order and presence of other amino acids around it play an important role. The number of neighboring amino acids considered as a sample is called window size. For instance, if one takes seven amino acids on each side of the sample amino acid then the size of the query would be fifteen which is its window size. In the literature review, it can be seen that the window size affects results to some extent. Windows of sizes 15, 17, 19 and 21 are used. Among them, 21 works best.

2.2. Feature vector

A 2-mer vector is used as a feature in the proposed method. To compute a 2-mer vector, the following steps have to be followed. Find all the possible pairs of 20 universal amino acids. In this way, there are 400 distinct pairs of amino acids. A 2-mer vector is a 400 X 1-dimension vector with each place representing a pair of amino acids. The existence of a particular pair in a query is represented by 1 and absent by 0. If the pair exists more than once, then it is also counted by replacing 1 with the number of existences. This results in a feature vector of our query sample.

PSSM, which has been used as a feature vector, claims to find the evolutionary information in a query sequence. Evolutionary information can give useful features to predict ligand sites [28]. As no evolutionary information is obtained, therefore, computations are performed to achieve different pairs of amino acids in the window. From the results, it is hypothesized that to predict an amino acid as ATP or non-ATP, the information regarding the presence and order of different amino acid pairs around is useful.

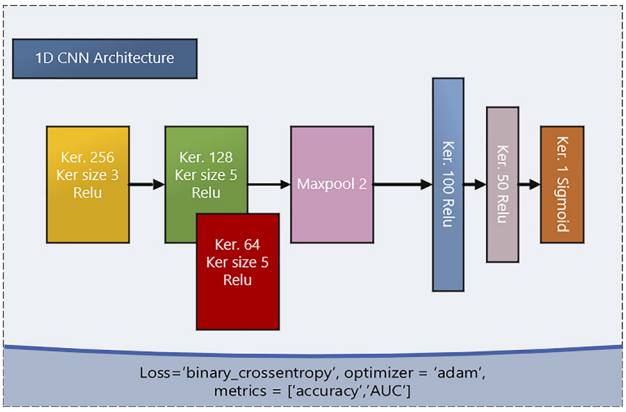

Fig. 1. 1-D Convolutional neural network

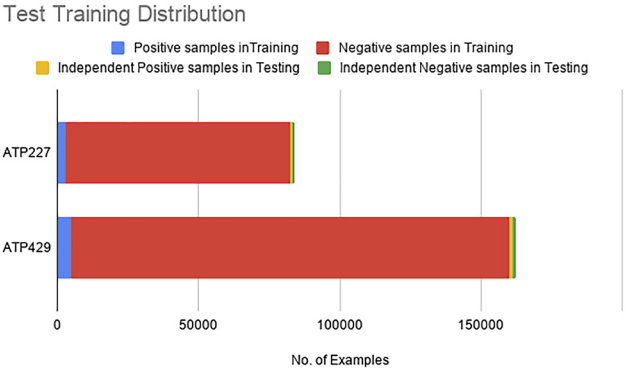

Fig. 2. Test and training samples distribution

2.3. Models

Five different machine learning algorithms such as Random Forest, K-nearest neighbors (KNN) algorithm ($K = 3$), XGBoost, Support Vector Machines (SVM), 1-D Convolutional neural networks are used to test K-mer features in ATP-site prediction.

The 1-D Convolutional neural network used has the architecture illustrated in Figure 1.

3. EXPERIMENTS

3.1. Datasets

There are two latest benchmark datasets called ATP227 and ATP429. The datasets contain many sequences. Each sequence has a variable number of amino acids. Each amino acid is one example and has a label corresponding to it. ATP site prediction is a severe data imbalance problem. The ratio between positive and negative examples in ATP-227 is 1:24 and 1:25 in ATP429. It means it is highly imbalanced. For testing, 20% of the positive class and stochastically obtained 20% samples of the positive class are considered. For the negative class, an approach based on the same number of negative samples as positive samples were taken. Finally, a balanced independent dataset for testing is formed (Fig. 2).

3.2. Position-specific scoring matrix (PSSM)

Following are the steps to compute the Position-specific scoring matrix (PSSM) (Fig. 3).

1. Find multiple sequence alignment of input query with a huge database. In NCBI Blast non-redundant databases or Swissport databases are available.
2. Find the top N highly scored sequences from the MSA.
3. Calculate Position Specific Scoring Matrix from these sequences, resulted from MSA.

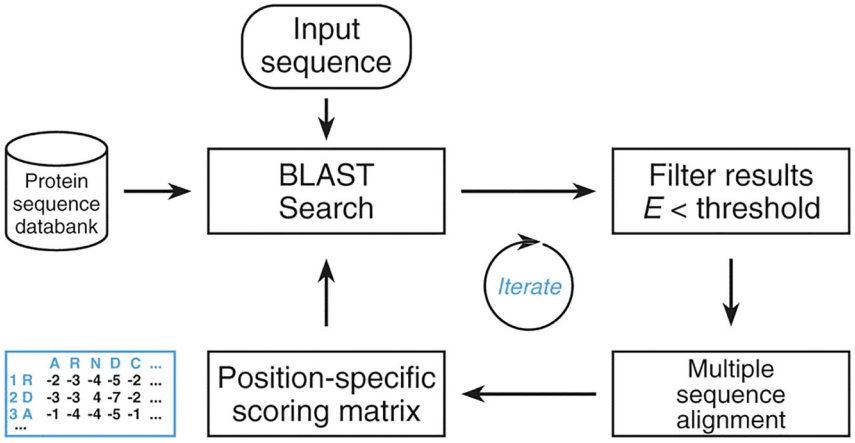

Fig. 3. PSI Blast used to find PSSM

4. Use this PSSM profile as a query this time and again search in the database.
5. Repeat steps 2–4 for three times.

Evolutionary information can give useful features to predict ligand sites [24]. This work does not find evolutionary information through PSSM, rather it is preferred to have computed information regarding different pairs of amino acids in the window. Change in amino acids occurs after a long time in history so it can be stated that to predict an amino acid as ATP or non-ATP, the information regarding the order and presence of different amino acid pairs is useful. From the results, it is hypothesized that some specific amino acid pairs surround an ATP site. 20 amino acids make all proteins. So, there can be 20 X 20 = 400 pairs of amino acids. The vectors of 400 amino acid pairs are termed a 2-mer vector. To find the 2-mer of a query, all pairs of amino acids of length two present in the window of a query are found. Now place value 1 in the 2-mer vector if the corresponding 2-mer pair exist in query otherwise 0. After Normalization, this 2-mer vector is a feature vector for one example (Fig. 4).

In this work, K-mer as input features for ATP-site prediction and found wonderful results are used. K-mer has less computation and time complexity than PSSM but gives better results even with a simple machine learning algorithm.

Even with $k = 2$, it is included that another less complex feature such as 1-hot encoding. The result for sure would be better. From the results, two-three benchmark datasets ATP168, ATP227 and ATP429 are obtained. ATP227 and ATP429 are preferable options. As it can be stated, the number of negative examples is 23–25 times less than the positive examples. Although methods do exist to solve this problem, the model trained on a balanced dataset always performs better than the model trained on an imbalanced dataset. Even though it cannot make an exact balance dataset but this research can produce a dataset that has at least more positive examples than the described data set has. Till now the problem is severely imbalanced. On the other hand, there are a lot of options, opportunities, ways, customs and manners to achieve a solution to this problem. An important benefit is that a good and accurate prediction can lead to the flourishing of the pharmaceutical industry as a compound can be created globally. Moreover, lives can be saved from the created drug compound accordingly.

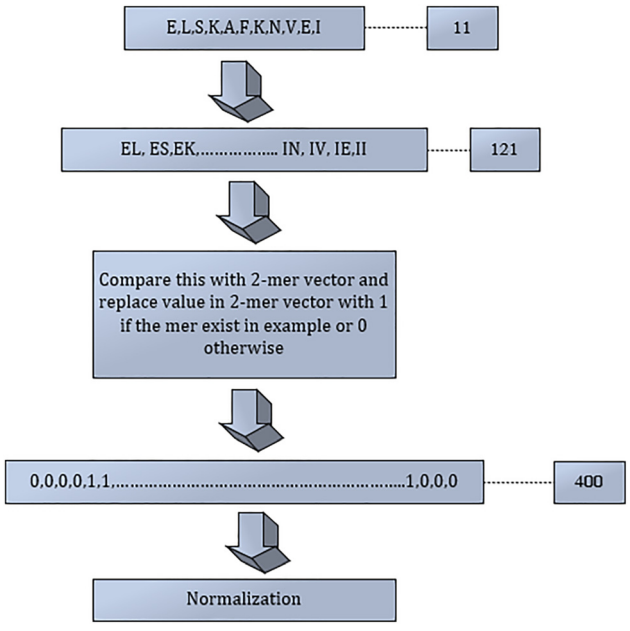

Fig. 4. Computing 2-mer of window size 11

3.3. SVM

Five different ML algorithms are used to test K-mer features in ATP site prediction as Random Forest, K-nearest neighbors (KNN) algorithm ($K = 3$), XGBoost, Support Vector Machines (SVM), 1-D Convolutional neural networks. Finally, SVM gave the best performance on these features. Furthermore, PSSM is computed using PSI-blast, Swissprot database, blosum90 metric, iteration = 3, e-value = 0.001, word-size = 2 and threshold = 2. The obtained accuracy of the created model is 96%, MCC 90% and ROC-AUC is 99% which are the same or even better in some aspects than the state-of-the-art results. The state-of-the-art results have an accuracy of 0.97, MCC 0.78 and ROC-AUC is 0.92. Our proposed method is simpler and more accurate.

3.4. Performance evaluation

Accuracy, Mathew's Correlation Coefficient (MCC), Specificity (S_p) Sensitivity (S_n) and F1 score matrix are used as a tool to measure performance and predict the behavior of the proposed model, Here, TP denotes true positive values, FP denotes false positive values, TN denotes true negative values and FN denotes false negatives values.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

$$S_n = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$S_P = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

PR and ROC curves are used for the evaluation of different models' performance. It is a graph between the false positive (FP) and true positive (TP) rates at numerous thresholds. With a large number of negative samples, precision is probably better so the PR curve and F1 score are considered. With a large number of Positive samples – the ROC curve is probably better so an AUC score from the curve is indicated.

Table 1. Window size performance

Window size	F1-score	ACC	MCC	ROC-AUC
W-S = 13	0.97	0.97	0.94	1.00
W-S = 15	0.98	0.98	0.96	1.00
W-S = 17	0.99	0.99	0.97	1.00
W-S = 19	0.99	0.99	0.98	1.00
W-S = 21	0.99	0.99	0.99	1.00

3.5. Results and discussion

The evaluation of this experiment is executed in Python language. To find the best window size an experiment is conducted keeping other parameters constant. The first column shows window size value and the next columns show F1-score, Accuracy, MCC value and ROC-AUC value according to window size. Only one classifier is employed which is Random Forest in order to see the window size effect on the data. The window sizes of 13, 15, 17, 19 and 21 are used in the test and found that a 2-mer window size of 21 gives the best results on the dataset ATP227 using Random Forest classifier (Table 1).

3.6. Effects of under-sampling

To cope with the problem of an imbalanced training dataset, the first solution is under-sampling. In this case, majority class samples are randomly thrown and are kept equal to the samples of the minority class. In this way, a balanced dataset of a lesser number of samples is prepared. The problem with this case is that data is being lost which is not a preferred thing in machine learning. In this table, the first column shows five different machine learning algorithms (Random Forest, K-nearest neighbors (KNN) algorithm (K = 3), XGBoost, Support Vector Machines (SVM), 1-D Convolutional neural networks) to test K-mer features and next columns show their performance values. These values show the effect of under-sampling. Here, Random Forest gives better performance (Table 2).

3.7. Effects of over-sampling

The second solution is oversampling. In this case, the minority class samples are replicated until the number of samples becomes equal to the number of majority class samples. The best thing about the strategy is that no data is lost here. The minority class samples are replicated to a number equal to majority class samples, so that both classes would have the same number of samples. The result shows over-sampling is performing better among the three strategies. In this table, the first column shows five different machine learning algorithms (Random Forest, K-nearest neighbors (KNN) algorithm (K = 3), XGBoost, Support Vector Machines (SVM), 1-D Convolutional neural networks) to test K-mer features and next columns show their performance values. These values show the effect of over-sampling. Among the five models Conv1D, KNN, Random Forest and SVM work almost the same. But SVM is slightly better than others (Table 3).

3.8. Effects of weighted loss function

The third solution is the weighted loss function. In this case, the loss function is responsible for handling the imbalanced dataset problem. During training, whenever an example is misclassified, the loss function penalizes the classifier more if the misclassified example is from a minority class and vice versa. The weightage of penalty is specified during training and it is based on the ratio of imbalance in the number of samples of the classes. In this case, the number of positive samples is almost two thirds lesser than the number of negative samples. The weightage of the classes is accordingly. In this table, the first column shows five different machine learning algorithms (Random Forest, K-nearest neighbors (KNN) algorithm (K = 3), XGBoost, Support Vector Machines (SVM), 1-D Convolutional neural networks) to test K-mer features and next

Table 2. Under-sampled performance

Model	ACC	F1	Prec.	MCC	Sen	Spec	PR-AUC	ROC-AUC
Conv1d	0.76	0.77	0.75	0.53	0.80	0.73	0.80	0.82
KNN	0.72	0.75	0.68	0.45	0.82	0.62	0.80	0.78
XGBoost	0.67	0.65	0.68	0.33	0.63	0.70	0.72	0.71
Random Forest	0.79	0.79	0.79	0.79	0.80	0.79	0.88	0.86
SVM	0.76	0.76	0.77	0.53	0.76	0.77	0.84	0.83

Table 3. Over-sampled performance

Model	ACC	F1	Prec.	MCC	Sen	Spec	PR-AUC	ROC-AUC
Conv1d	0.70	0.69	0.71	0.39	0.67	0.72	0.77	0.77
KNN	0.66	0.48	0.98	0.42	0.32	0.99	0.89	0.81
XGBoost	0.67	0.65	0.69	0.35	0.62	0.73	0.74	0.73
Random Forest	0.64	0.43	0.98	0.39	0.28	1.0	0.90	0.87
SVM	0.92	0.91	1.0	0.85	0.84	1.0	0.99	0.98

Table 4. Weighted loss function performance

Model	ACC	F1	Prec.	MCC	Sen	Spec	PR-AUC	ROC-AUC
Conv1d	0.70	0.69	0.71	0.39	0.67	0.72	0.77	0.77
KNN	0.66	0.48	0.98	0.42	0.32	0.99	0.89	0.81
XGBoost	0.67	0.65	0.69	0.35	0.62	0.73	0.74	0.73
Random Forest	0.64	0.43	0.98	0.39	0.28	1.0	0.90	0.87
SVM	0.92	0.91	1.0	0.85	0.84	1.0	0.99	0.98

columns show their performance values. These values show the effect of a weighted loss function. Here Support Vector machine works better in case of a weighted loss function (Table 4).

3.9. Comparison of three strategies and their best results

In this section, an experimental comparison from the results is made which consists of the shortcomings in the form of three strategies (under-sampling, weighted loss function and over-sampling). The blue bar represents under-sampling, the red bar represents weighted loss function and the yellow bar represents oversampling. The evaluation parameters are accuracy, F1-score, precision, MCC, sensitivity, specificity, PR-curve and ROC curve. The first group of bars is for Accuracy, the second group of bars is for F1-score, the third group of bars is for precision, the fourth group of bars is for MCC, the fifth group of bars is for specificity, the sixth group of bars is for PR-curve and the seventh group of bars is for ROC curve. It can be seen that oversampling gives the best results. Among the models, SVM is the one whose performance is best (Fig. 5).

3.10. Comparison among the previous model and the proposed

The experimental results of the proposed method are compared with other high-tech methods. Sensitivity, specificity, accuracy and MCC are being used in order to provide a comparison of the results. The first column shows eight previous state-of-the-art methods are chosen in order to compare with the proposed technique and the next columns show their performance value. From the results reported in Table 5, it can be seen that the proposed predictor accomplishes higher values in sensitivity, specificity and MCC compared to the other eight methods. It outperforms all other methods in terms of sensitivity and MCC. In the case of specificity, the performance of our proposed method is improved significantly as compared to all other eight methods.

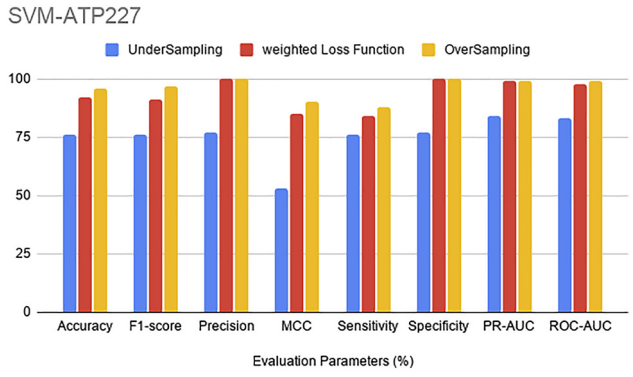

Fig. 5. Comparison of three different strategies

Table 5. Comparison among models

Model	ACC	Sen	Spec.	MCC	ROC-AUC
ATPint	0.648	0.539	0.651	0.078	0.627
ATPsite	0.961	0.361	0.988	0.433	0.854
NsitePred	0.967	0.444	0.982	0.461	0.861
TargetATP	0.798	0.539	0.651	0.078	0.627
TargetATPsite	0.965	0.433	0.988	0.50	0.872
TSC-ATP	0.962	0.419	0.985	0.46	0.860
ATPbind	0.974	0.630	0.990	0.677	0.915
DeepATP	0.96	0.256	0.993	0.384	
Ensemble predictor	0.976	0.552	0.991	0.606	0.922
Proposed predictor	0.941	0.88	1.0	0.89	0.98

4. CONCLUSION

The main goal of this research work is to create an algorithm that performs better in predicting the ATP-sites required for creating medicines for different diseases. PSSM along with some other physicochemical and structural properties of sequences have been used in ATP prediction. Computing PSSM is a computationally complex procedure especially if it is totaled from a large number of sample sequences of around 80,000 to 150,000. To find a PSSM of a sequence, it has to be aligned with a very large database having millions of sequences. From this, 2-mer feature vectors are found which are just counting the number of amino acid pairs in the sequence. This feature vector is simply computed but has a very high performance even on simple machine learning algorithms. Using the indicated approach, there is no longer a need to use deep learning (DL) except in the case in which the highest performance is required and there are enough PSSM computing resources available.

As the conferred approach used a 2-mer feature vector, therefore, this is the start of using some other feature vector than PSSM which gives better performance in predicting ATP site prediction. A model is created in this lab which predicts with accuracy of 90–99% and uses fewer resources

compared to other alternative approaches. Future work can include creating a better algorithm to have higher efficiency. Moreover, some other deep learning architecture can be used to observe if the performance is enhanced. Furthermore, the value of k can be increased. For instance, the value of k between 2 and 3 provides a feature vector of around 8,000. Although it increases complexity somehow it can be possible that results would be much better as compared to the results when $k = 2$. Even with $k = 2$, it can include another less complex feature such as 1-hot encoding to see if the results got better. Better ATP site prediction can help in creating effective compounds and drugs that can save lives.

REFERENCES

[1] J. Schneider, K. Korshunova, F. Musiani, M. Alfonso-Prieto, A. Giorgetti, and P. Carloni, "Predicting ligand binding poses for low-resolution membrane protein models: perspectives from multiscale simulations," *Biochem. Biophysical Res. Commun.*, vol. 498, no. 2, pp. 366–74, 2018. https://doi.org/10.1016/j.bbrc.2018.01.160.

[2] Q. Yu, C. Gratzke, Y. Wang, A. Herlemann, F. Strittmatter, B. Rutz, C. G. Stief, and M. Hennenberg, "Inhibition of prostatic smooth muscle contraction by the inhibitor of G protein-coupled receptor kinase 2/3, CMPD101," *Eur. J. Pharmacol.*, vol. 831, pp. 9–19, 2018. Available: https://doi.org/10.1016/j.ejphar.2018.04.022.

[3] R. Kaempfer, "Ribosome cycle emerges from DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 8, pp. 470-470, 2017/08/01 2017 https://doi.org/10.1038/nrm.2017.59.

[4] N. A. Becker, T. L. Schwab, K. J. Clark, and L. J. Maher, III, "Bacterial gene control by DNA looping using engineered dimeric transcription activator like effector (TALE) proteins," *Nucleic Acids Res.*, vol. 46, no. 5, pp. 2690–6, 2018. https://doi.org/10.1093/nar/gky047.

[5] Q. Wu, Z. Peng, Y. Zhang, and J. Yang, "COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W438–42, 2018. https://doi.org/10.1093/nar/gky439.

[6] D. Toti, L. Viet Hung, V. Tortosa, V. Brandi, and F. Polticelli, "LIBRA-WA: a web application for ligand binding site detection and protein function recognition," *Bioinformatics*, vol. 34, no. 5, pp. 878–80, 2018. https://doi.org/10.1093/bioinformatics/btx715.

[7] L. Villanueva, L. Silva, D. Llopiz, M. Ruiz, T. Iglesias, T. Lozano, N. Casares, S. Hervas-Stubbs, M. J. Rodríguez, J. L. Carrascosa, J. J. Lasarte, and P. Sarobe, "The Toll like receptor 4 ligand cold-inducible RNA-binding protein as vaccination platform against cancer," *OncoImmunology*, vol. 7, no. 4, 2018, e1409321, https://doi.org/10.1080/2162402X.2017.1409321.

[8] J. P. Berger, R. SinhaRoy, A. Pocai, T. M. Kelly, G. Scapin, Y.-D. Gao, K. A. D. Pryor, J. K. Wu, G. J. Eiermann, S. S. Xu, X. Zhang, D. A. Tatosian, A. E. Weber, N. A. Thornberry, R. D. Carr. "A comparative study of the binding properties, dipeptidyl peptidase-4 (DPP-4) inhibitory activity and glucose-lowering efficacy of the DPP-4 inhibitors alogliptin, linagliptin, saxagliptin, sitagliptin and vildagliptin in mice," *Endocrinol. Diabetes Metab.*, vol. 1, no. 1, 2018, e00002, https://doi.org/10.1002/edm2.2.

[9] T. Kamimura, N. Isobe, and Y. Yoshimura, "Effects of inhibitors of transcription factors, nuclear factor-κB and activator protein 1, on the expression of proinflammatory cytokines and chemokines induced by stimulation with Toll-like receptor ligands in hen vaginal cells," *Poult. Sci.*, vol. 96, no. 3, pp. 723–30, 2017. https://doi.org/10.3382/ps/pew366.

[10] J. Sun, and K. Chen, "NSiteMatch: prediction of binding sites of nucleotides by identifying the structure similarity of local surface patches," *Comput. Math. Methods Med.*, vol. 2017, 2017, 5471607-5471607. https://doi.org/10.1155/2017/5471607.

[11] S. Usha, and S. Selvaraj, "Structure-wise discrimination of adenine and guanine by proteins on the basis of their nonbonded interactions," *J. Biomol. Struct. Dyn.*, vol. 33, no. 7, pp. 1474–92, 2015. https://doi.org/10.1080/07391102.2014.958759.

[12] J. Song, G. Liu, J. Jiang, P. Zhang, and Y. Liang, "Prediction of protein-ATP binding residues based on ensemble of deep convolutional neural networks and light GBM algorithm," *Int. J. Mol. Sci.*, vol. 22, no. 2, 2021, https://doi.org/10.3390/ijms22020939.

[13] S. Wang, M. Li, L. Guo, Z. Cao, and Y. Fei, "Efficient utilization on PSSM combining with recurrent neural network for membrane protein types prediction," *Comput. Biol. Chem.*, vol. 81, pp. 9–15, 2019. https://doi.org/10.1016/j.compbiolchem.2019.107094.

[14] D.-J. Yu, J. Hu, Z.-M. Tang, H.-B. Shen, J. Yang, and J.-Y. Yang, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing*, vol. 104, pp. 180–90, 2013. https://doi.org/10.1016/j.neucom.2012.10.012.

[15] Y. Chen, Y. Tang, and H. Wang, "Quantification of ATP in cell by fluorescence spectroscopy based on generalized ratio quantitative analysis model," *Spectrochim. Acta A: Mol. Biomol. Spectrosc.*, vol. 263, 2021, 120170-120170. https://doi.org/10.1016/j.saa.2021.120170.

[16] Y. Gao, P. Zhang, A. Cui, D.-Y. Ye, M. Xiang, and Y. Chu, "Discovery and anti-inflammatory evaluation of benzothiazepinones (BTZs) as novel non-ATP competitive inhibitors of glycogen synthase kinase-3β (GSK-3β)," *Bioorg. Med. Chem.*, vol. 26, no. 20, pp. 5479–93, 2018. https://doi.org/10.1016/j.bmc.2018.09.027.

[17] M. Arif, S. Ahmad, F. Ali, G. Fang, M. Li, and D.-J. Yu, "TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree," *J. Comput. Aided Mol. Des.*, vol. 34, no. 8, pp. 841–56, 2020. https://doi.org/10.1007/s10822-020-00307-z.

[18] J. Hu, Y. Li, Y. Zhang, and D.-J. Yu, "ATPbind: accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons," *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 501–10, 2018. https://doi.org/10.1021/acs.jcim.7b00397.

[19] T.-T.-D. Nguyen, N.-Q.-K. Le, R. M. I. Kusuma, and Y.-Y. Ou, "Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network," *J. Mol. Graph. Model.*, vol. 92, pp. 86–93, 2019. https://doi.org/10.1016/j.jmgm.2019.07.003.

[20] H. R. Hassanzadeh, and M. D. Wang, "DeeperBind: enhancing prediction of sequence specificities of DNA binding proteins," *bioRxiv*, 2017, 99754-99754. https://doi.org/10.1101/099754.

[21] C. Song, G. Liu, J. Song, and J. Jiang, "A novel prediction method of ATP binding residues from protein primary sequence BT – advances in neural networks – ISNN 2019," in *Chemistry of Polymeric Metal Chelates*, Cham, H. Lu, H. Tang, and Z. Wang, Eds., Springer International Publishing, 2019, pp. 548–55.

[22] W. Alhalabi, A. Farooq, A. Alhudali, and L. Khafaji, "Smart electrical design of medical center to vary field parameters: sensor network in improving health care," *J. Eng. Appl. Sci.*, vol. 14, no. 3, pp. 879–86, 2019. https://doi.org/10.36478/jeasci.2019.879.886.

[23] A. Farooq, W. Alhalabi, and S. M. Alahmadi, "Traffic systems in smart cities using LabVIEW," *J. Sci. Technol. Pol. Manage.*, vol. 9, no. 2, pp. 242–55, 0000. https://doi.org/10.1108/JSTPM-05-2017-0015.

[24] A. Farooq, M. Seyedmahmoudian, B. Horan, S. Mekhilef, and A. Stojcevski, "Overview and exploitation of haptic tele-weight device in virtual shopping stores," *Sustainability (Basel, Switzerland)*, vol. 13, no. 13, p. 7253, 2021. https://doi.org/10.3390/su13137253.

[25] A. Farooq, and R. Aftab, "Performance study and evaluation of a solar PV testbed system using LabVIEW," *Int. Rev. Appl. Sci. Eng.*, vol. 10, no. 1, pp. 113–23, 2019. https://doi.org/10.1556/1848.2018.0012.

[26] T. Shehzadi, A. Majid, M. Hameed, A. Farooq, and A. Yousaf, "Intelligent predictor using cancer-related biologically information extraction from cancer transcriptomes," in *International Symposium on Recent Advances in Electrical Engineering & Computer Sciences (RAEE & CS)*, vol. 5, pp. 1–5, 2020. https://doi.org/10.1109/RAEECS50817.2020.9265692.

[27] Q. Chao, J. Tao, X. Wei, Y. Wang, L. Meng, and C. Liu, "Cavitation intensity recognition for high-speed axial piston pumps using 1-D convolutional neural networks with multi-channel inputs of vibration signals," *Alexandria Eng. J.*, vol. 59, no. 6, pp. 4463–73, 2020. https://doi.org/10.1016/j.aej.2020.07.052.

[28] A. Farooq, M. Seyedmahmoudian, and A. Stojcevski, "A wearable wireless sensor system using machine learning classification to detect arrhythmia," *IEEE Sensors J.*, vol. 21, no. 9, pp. 11109–16, 2021. https://doi.org/10.1109/JSEN.2021.3062395.