International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 11 Issue 10, October, 2022, Impact Factor: 7.751

An Artificial Intelligence Deep Learning Model of Antiviral-HPV Protein Interaction Prediction

Dr. A. Jayanthila Devi[1], Dr. P.S. Aithal[2], Dr. Radha Mohan[3], Dr. Sudhanshu Maurya[4]

[1]Professor & Post-Doctoral Research Fellow, Institute of Computer Science and Information, Srinivas University, Mangalore, India
[2]Vice Chancellor, Srinivas University, Mangalore – 575001, India
[3]Post-Doctoral Research Fellow, Institute of Computer Science and Information, Srinivas University, Mangalore, India
[4]Postdoctoral Fellow, University Malaysia Perlis, Ministry of Education Malaysia.
Assistant Professor, School of Computing, Graphic Era Hill University, Bhimtal Campus, Uttarakhand, India

ABSTRACT

Many computer programmes can predict protein-protein interaction grounded with anamino acid sequence, although they tend to focus on species-specific interactions rather than cross-species ones. Homogeneous protein interaction prediction algorithms fail to find interactions between proteins from different species. In this research, we constructed an artificial intelligence deep learning model to encode the frequency of consecutive amino acids in a protein sequence. The deep learning model predicts human-viral protein interactions. The study used inartificial intelligence deep learning model and protein annotations to predict human-virus protein interactions. A simple but effective representation technique for predicting inter-species protein-protein interactions. The representation approach has several advantages, such as improving model performance, generating feature vectors, and applying the same representation to diverse protein types. The results of simulation shows that the proposed method achieves an accuracy of 98% than other methods.

Keywords: Deep Learning, Protein Interaction, Prediction Antiviral-HPV Protein

INTRODUCTION

Proteins physically interact with one another [2,3], allowing them to play vital roles in a wide range of aspects of life [1,2]. The molecular basis of various functions like trafficking, signal transduction, gene expression, metabolic regulation, proliferation and cell growth may be traced back to protein–protein interactions (PPIs) [4, 5]. It is not uncommon for particular interface residues to perform a more substantial role in protein binding than other residues in the same interface. These remnants are referred to as hotspots in some circles [6]-[10].

Generally, these hotspots are considered to be pre-arranged in terms of a protein state that are unbound, whereas bound protein is not. The notion is that a major percentage of the protein surface is inaccessible to binding as a result, and these sites on potential binding for a specific protein is imprinted already in unbound state.

The sites of PPI are required for the purpose of selective molecular identification as well as the formation of complexes [11, 12]. In order to understand and explain signal transduction networks, protein function, and develop new therapeutics, it is necessary to discover proteins that interact with one another. To characterise the PPI sites, researchers have used NMR and X-ray crystallography [13,14]. Despite the fact that these procedures are time-and money-consuming [15,16], they are effective. The use of various computational and machine learning methodologies including molecular dynamics [17,18] has made it possible to predict PPI sites to a greater extent. The methods on machine-learning us considered proven to be the most successful, as illustrated in Figure 1.

Figure 1: Machine learning methods in protein interaction

AI is one of the most recent advancements in neural networks, and it has been used to predict the location of PPIs in the past. The convolutional neural network (CNN) is a deep learning approach that can be used to train representations and extract the optimal features from input data. It is a good example of how deep learning may be applied to representation learning and feature extraction on AI. Various prediction methods are identified on PPI and this can be categorized into three categories based on the facts upon which they are based.

Based on a series of events that have occurred. Methods based on sequence information are used to extract properties from protein sequences to predict the protein-protein location. In order to forecast PPI sites, the amino acid composition as well as the position-specific scoring matrix (PSSM) are taken into consideration by PPiPP. With the use of long- and short-term memory, DLPred can learn qualities (LSTM).

approaches that are structural in nature. By examining the 3D structure of the complex proteins, it is feasible to gain a great deal of information regarding the protein complex interaction sites. Some predictors use 3D structural information from proteins to produce predictions regarding PPI sites, whereas others use only 2D structural information. ProMate incorporates all of the interface properties that are most important to users, with a 0.70 success rate. The method in [21] used protein structural data to get a success rate of 0.76 out of a possible 100 attempts.

Since determining the 3D proteins structure is considered expensive and difficult, the amount of information available in the databases of protein structure that includes the Protein Data Bank (PDB) is far less than the amount of

information available in the databases of protein sequence like Protein Sequence Database (UniProt). As a result, the majority of techniques for predicting PPI sites make use of both structural and sequence information..
SPPIDER predicts PPI sites using RSA, sequence, and structural information, and it is seen that the prediction of RSA using the protein interaction fingerprintsare considered to improve the discrimination between noninteracting and interacting sites in a variety of protein interactions. – SPPIDER An overall number of 11 sequence and structure-specific features are used by IntPred. In order to train a artificial intelligencedeep learning classifier, paired kernels based on the sequence and structure of residue pairs collected by PAIRpred are used in conjunction with PAIRpred (Figure 1).

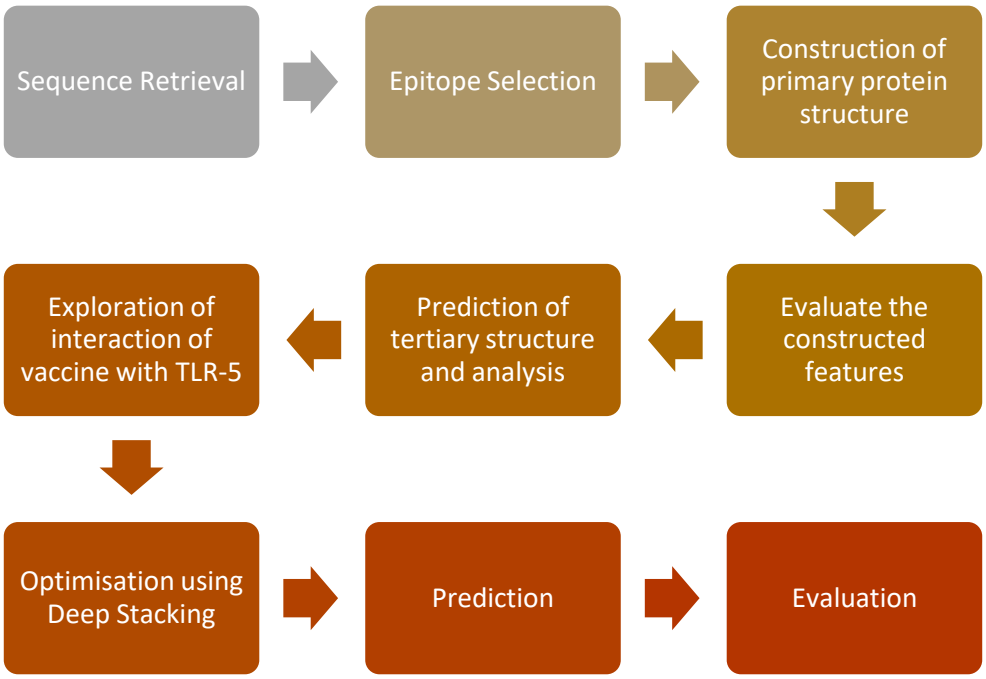

Figure 2: Design a protein-HPV peptide for the purpose of vaccination

The main contribution of the work involves the construction of a deep learning model to encode the frequency of consecutive amino acids in a protein sequence. The deep learning model predicts human-viral protein interactions.

BACKGROUND

In recent years, a variety of computer-based solutions have been used to overcome this challenge. Some of these projects have also focused on the creation of new machine-learning algorithms, which has been a focus of other projects. Using protein information, the frequency of any three consecutive are estimated for amino acid unit in protein sequences. PPIs have been demonstrated to be predicted solely by sequences [6]. The auto covariance (AC) [7] methods on the index distribution of the amino acid [8] are two different ways of describing a protein sequence that have been created to extract information on the physical and chemical properties of amino acids, as well as the frequencies and placements of amino acids. Various techniques have been used to reduce the dimensions of the features. Support vector machines (SVMs) and their variants [9, 10], and neural networks [12] and random forests [11] are all machine learning algorithms that have been employed in various applications. In a few articles, cross-validation results have been provided, but they have not been tested with other datasets [13, 14].

Deep-learning algorithm helps in recreation of neural connections in denser way and hence the processes of learning the human brain received a greater interest in implementation successfully various applications like image and PPIrecognition [15, 16], decision making [18] and natural language understanding [17]. Deep-learning algorithms have a lot easier time dealing with large amounts of complex data than traditional machine learning approaches [19], which is a significant advantage. High-throughput approaches, such as those used in bioinformatics, have necessitated the use of these algorithms in recent years [20–24].

The use of deep neural network models to forecast DNA polymorphisms that induce aberrant splicing in genome regulation function prediction, for example, has been proven effective. [25] Their method outperformed earlier models in terms of accuracy. The Deep Bind model, which is based on convolutional neural networks, can be used to predict the sequence specificities and binding motifs of DNA and RNA-binding proteins in a variety of situations. When it comes to determining the functional effects of noncoding mutations, human geneticists confront a significant uphill battle. As a result of Deep SEA development, it is now possible to predict reliably the effects of chromatin on alterations of protein sequence with sensitivity (single-nucleotide) from large-scale data, allowing for more precise gene targeting.

After that, when it came to estimating the function of non-coding DNA, the DnaQ model outperformed other models by more than 50%, according to the researchers. ABNs were used to predict protein secondary structures, and they were found to be accurate in predicting protein function with an accuracy of 80.7%. A DNN method can be used to forecast secondary structures, backbone angles, and solvent-accessible surface areas, amongst other things. According to a recent review that goes into detail on how they are being used, deep learning algorithms are being employed in computational biology.

PROPOSED METHOD

In this section, we learn and classify complex functions, simple modules of functions or classifiers are "stacked" on top of each other in a deep stacking network, which is a deep neural network (DSN). This is the fundamental concept underpinning the design of DSNs. Prior to the advent of supervised information, stacked operations were performed using a variety of different approaches, with the simplest modules frequently relying on supervised information. In many cases, the classifier output from lower modules and the properties of raw input data are combined for building features at higher levels for a stacked classifier.

As the foundation of the stacking module, a conditional random field (CRF) was utilised. The CRF architecture is refined by including the number of hidden states in order to achieve success in the prediction of PPI or protein synthesis where the information on segmentation may not be available in the dataset.

Deep Convex Network (DCN) is a name given to the DSN architecture that emphasizes the convex nature is useful in learning the network. Using supervisory information, each of the basic modules of DSN is placed on top of the others. Nonlinear sigmoidal nonlinear output is utilised instead of linear units. Because of the linearity of the output units in this regard, it is possible to develop an efficient and parallelizable approach to estimate output network weights based on hidden unit activity in the output network.

The convex term emphasises the convex optimization relevance in case of learning the output network weights and to distinguish it from other types of optimizations. Closed-form constraints, which arise with convexity between the input and output weights, play a crucial role in this situation. In addition to making learning the remaining network features (such as input network weights) substantially simpler, implementing these limits makes it possible to distribute batch-mode DSN learning across CPU clusters. DSN has also been used in more recent publications to emphasise the importance of stacking as a fundamental operation.

Architecture of DSN

The number of layers in a DSN can be any size, and each layer is a neural network with two sets of weights and a hidden layer that can be trained separately. In Figure , only four of these modules are represented, and each one is depicted in a distinct colour. It has been possible to successfully train and deploy up to a few hundred modules for protein synthesis classification investigations.

In the DSN lowest module, you'll find both linear and nonlinear input and output layers, as well as a nonlinear, hidden layer and a second, linear output layer. In the buried layer, sigmoidal nonlinearity is frequently used to achieve a desired result. Other nonlinearities, on the other hand, are also feasible. As a result, input units can be assigned values that are at least in part determined by the features extracted in protein, and output units are assigned with its respective intensity values, or something similar. When the DSN is utilised in conjunction with PPI recognition, PPI samples or features extracted from PPI can be employed as input units for the DSN. The classification targets are represented by the linear output units of this layer.

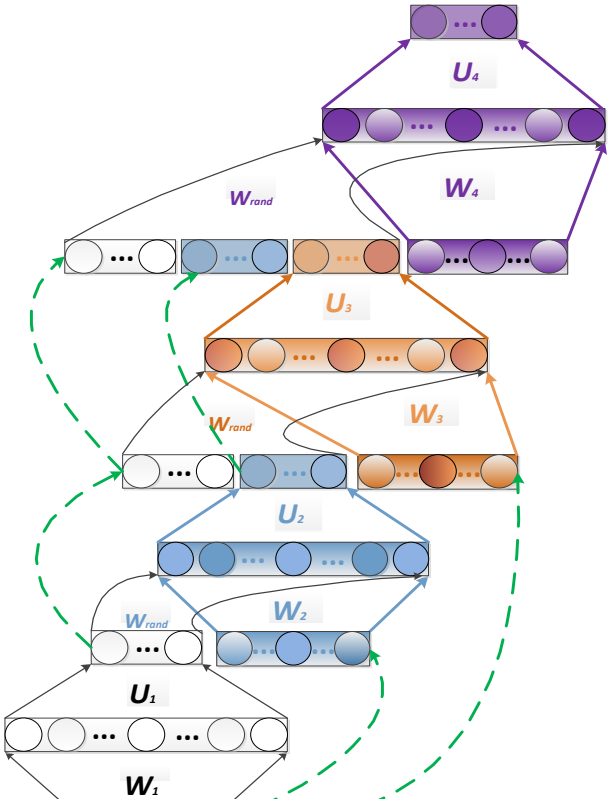

Figure 3: A DSN architecture

RESULTS AND DISCUSSIONS

Interactions between proteins are essential for the performance of many important biological processes. When compared to the other approaches that have been developed, proposed artificial intelligence deep learning is considered promising for predicting PPIs in the real world. Deep learning has risen in popularity as an area of machine learning in recent years, and it is now being employed in a broad variety of different industries.

The PPI data for Pan was acquired by us. As a result of the lack of repetition in a dataset, where the true positive samples of the PPIs are considered specifically from the Human Protein Reference Database (HPRD). Several proteins found in diverse subcellular locations are coupled together to produce negative samples. The information on protein subcellular localization was obtained from the Swiss-Prot database, which was updated to version 57.3.

- There were no non-human proteins found in any of the samples examined.
- It was determined that sequences labelled with unclear or ambiguous subcellular location terms are not included in the analysis.
- Because of this, sequences that were labelled in more than one position were eliminated from consideration.
- A decision was made to exclude sequences with the annotation sequences and fragment with lesser residues of amino acid (<50) due to presence of the possibility of fragments in these sequences.

A total of 2,184 proteins were discovered in six different subcellular locations. It was possible to create a total of 36,480 negative pairings with random matching of the proteins on additional proteins discovered in other subcellular locations as well as by adding negative pairs to the mix. The benchmark dataset consists of positive (36,545) and negative (36,323) samples that have been purged of protein pairs containing rare amino acidsto create a more accurate representation of the data.

We randomly selected 7,000 pairs of positive and negative samples from the benchmark dataset for model validation, while the remaining samples from the benchmark dataset were used for pre-training. When it came to predicting the holdout test set, training and testing 10-CV models on a large pre-training dataset was necessary.

When creating non-redundant test sets, pairs with a pairwise identity of less than 25% of those in the training set were removed from the holdout test set in order to examine the model robustness. A pairwise identity of less than 25% was defined as follows: In order to achieve the best possible results, we trained our final PPI prediction model on the complete benchmark dataset before applying it to the external test sets.

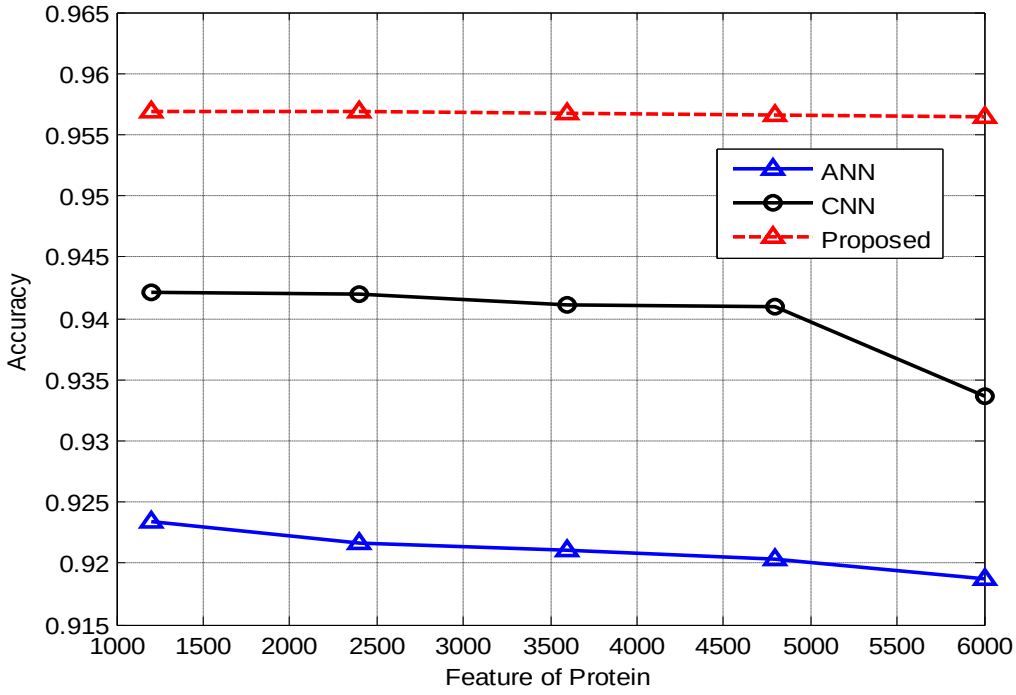

Figure 4: Accuracy

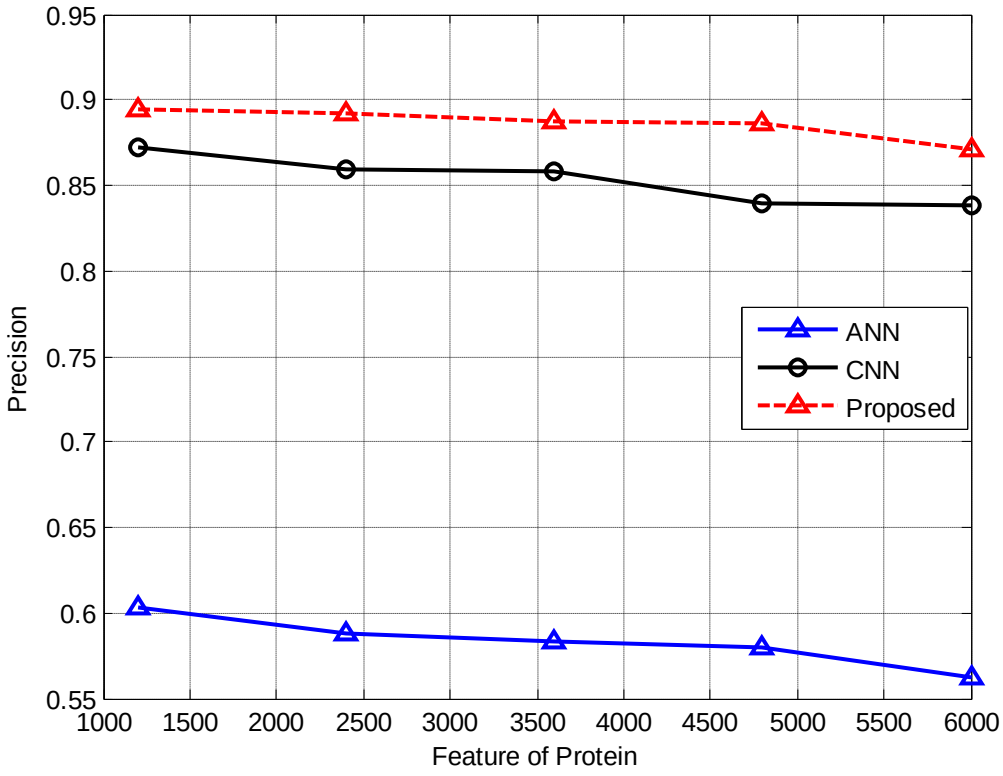

Figure 5: Precision

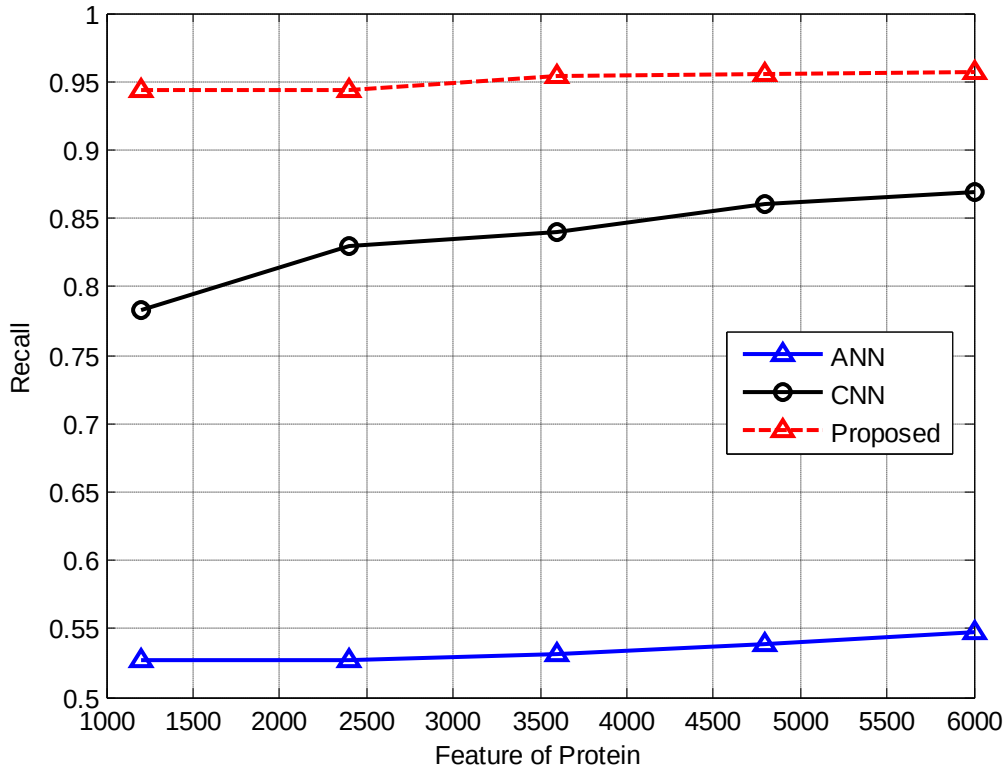

Figure 6: Recall

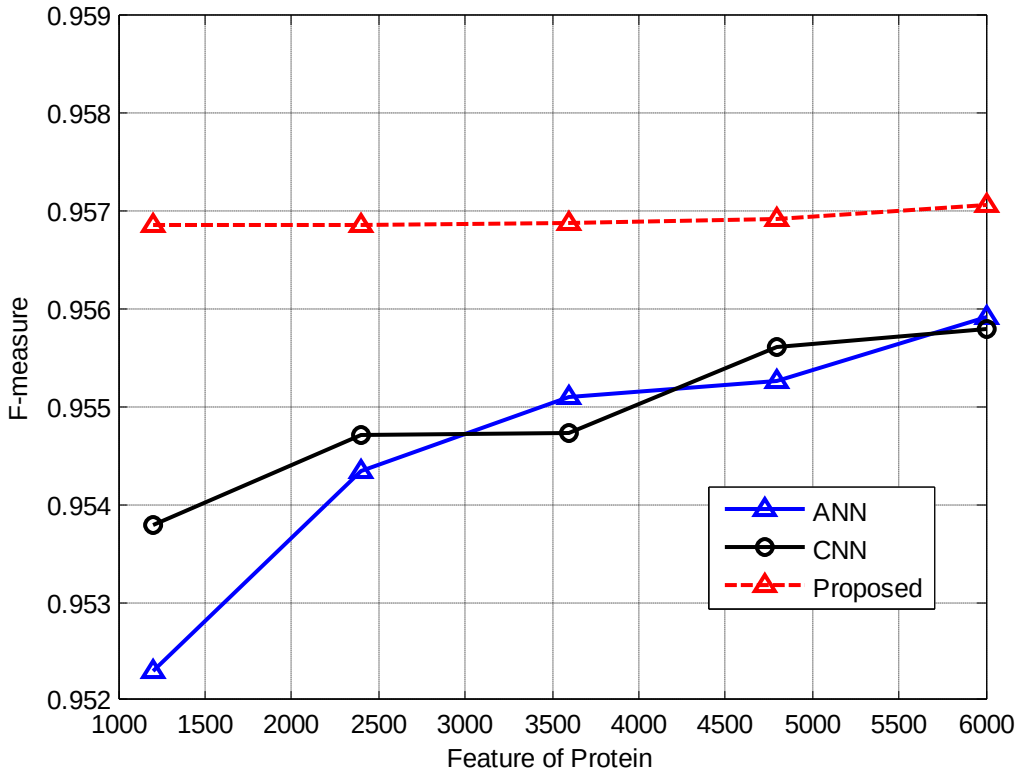

Figure 7: F-measure

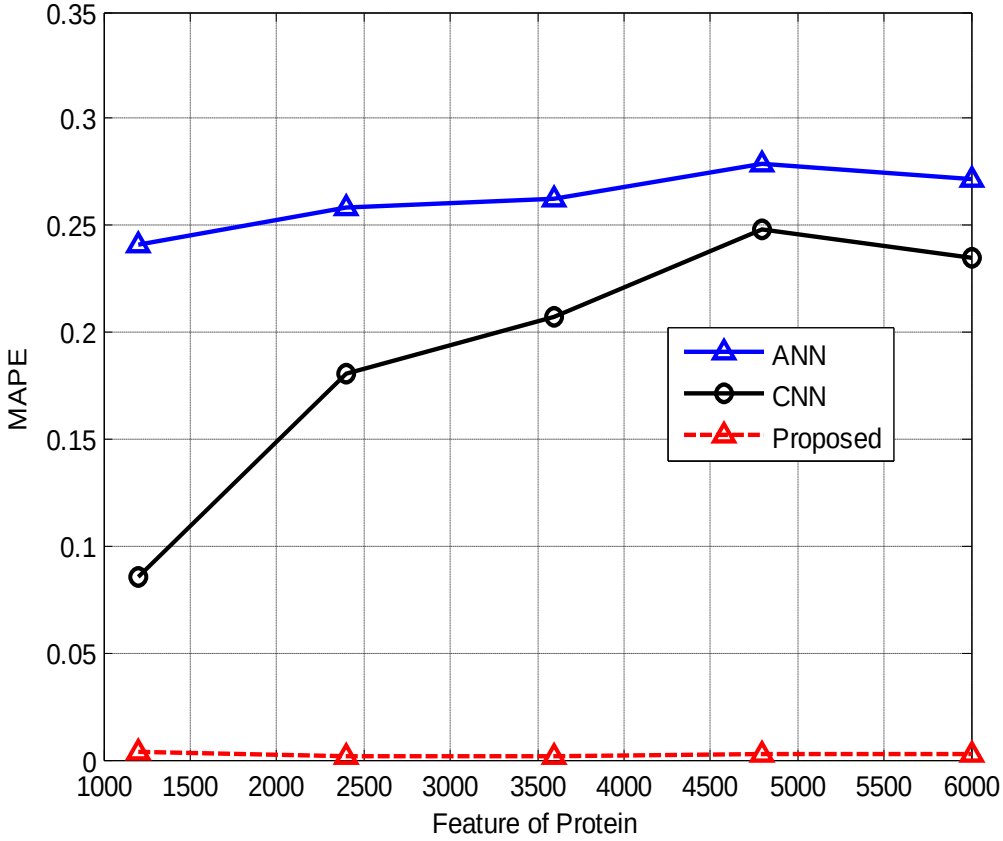

Figure 8: MAPE

The results (Figure 4- Figure 8) have found that different kinds of protein complexes have different residue interface propensities, which have been used to increase the accuracy of PPI site prediction in a variety of investigations. It is commonly used as a variable in predicting models because of its high predictability. In this experiment, the propensity of residue binding is used to screen the true positive samples, where the accuracy of prediction was significantly improved as a result. Despite the fact that our technique is unconventional, our findings demonstrate that it is reasonable to incorporate binding propensity for the reduction of false positives.

The presence of polar residues near interfaces, with the exception of arginine, is statistically unfavourable. The high binding propensity of arginine and another polar residue, histidine, was demonstrated to establish the relationship in an optimal way between polarity and hydrophobicity, as well as the association between polarity and binding propensity. According to the findings of this study, residues that are more likely to be located inside proteins have a higher affinity for binding. Interestingly, this makes sense because hydrophobic residues on protein surfaces have a tendency to get interacted with the residues on other proteins surfaces, which makes sense. A notable exception to this norm is alanine with simply a methyl on its side chain, which has been employed in alanine scanning mutagenesis to great effect.

In the case of tryptophan, for example, it possesses a significant hydrophobic side chain that is critical in the folding structure of proteins as well as the formation of binding sites. Charged residues are more likely than non-charged residues to bind to residues that are divalently charged in the opposite direction. The amino acid arginine has been discovered as a prevalent residue in considering the known sites of protein interaction, which is a result of its broad spectrum of function. In our study, argonine was also discovered to be strongly bound, although it was found to be more likely to be present on the protein surface than on the protein inside.

Our convolutional deep learning model is able to reliably identify protein interaction sites when using this newly updated data set as a training set. It is possible that false positive interaction pairs were present in the original positive data, which could impede efforts to increase the accuracy of PPI site prediction. PPI site prediction research may find success if the number of false-positive interaction samples is reduced in comparison to previous studies.

CONCLUSIONS

The application of deep learning models for protein sequence encoding is discussed in this article. The proposed model utilizing the deep learning approach enables better and optimal prediction of interaction between the viral proteins and human. Using a artificial intelligence deep learning model and protein annotations, researchers were able to predict the interaction of protein-protein chain between the viral cells and human beings. This is hence considered as a successful and a simple method to predict the interactions protein-protein polymerase chain between different species Additionally, the representation approach can be used to represent a range of protein types, which can be useful for improving model performance and producing feature vectors, among other things. In future, the convolutional deep learning models and updated data sets, the study can be able to attain an AUC of 0.9 or more, which is higher than those existing methods.

REFERENCES

[1]. Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., ... & Zheng, M. (2020). Transformer CPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, *36*(16), 4406-4414.

[2]. Arivazhagan, N., Somasundaram, K., Vijendra Babu, D., Gomathy Nayagam, M., Bommi, R. M., Mohammad, G. B., ... & Prabhu Sundramurthy, V. (2022). Cloud-Internet of Health Things (IOHT) Task Scheduling Using Hybrid Moth Flame Optimization with Deep Neural Network Algorithm for E Healthcare Systems. Scientific Programming, 2022.

[3]. Peng, C., Han, S., Zhang, H., & Li, Y. (2019). RPITER: a hierarchical deep learning framework for ncRNA–protein interaction prediction. *International journal of molecular sciences*, *20*(5), 1070.

[4]. Yao, Y., Du, X., Diao, Y., & Zhu, H. (2019). An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ*, *7*, e7126.

[5]. Jamasb, A. R., Day, B., Cangea, C., Liò, P., & Blundell, T. L. (2021). Deep learning for protein–protein interaction site prediction. In *Proteomics Data Analysis* (pp. 263-288). Humana, New York, NY.

[6]. Zeng, M., Zhang, F., Wu, F. X., Li, Y., Wang, J., & Li, M. (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics*, *36*(4), 1114-1120.

[7]. Mariappan, L. T., & Yuvaraj, N. (2020). Analysis On Cardiovascular Disease Classification Using Machine Learning Framework. Solid State Technology, 63(6), 10374-10383.

[8]. Zhao, Q., Zhao, H., Zheng, K., & Wang, J. (2022). HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, *38*(3), 655-662.

[9]. Richoux, F., Servantie, C., Borès, C., & Téletchéa, S. (2019). Comparing two deep learning sequence-based models for protein-protein interaction prediction. *arXiv preprint arXiv:1901.06268*.

[10]. Wang, J., Zhao, Y., Gong, W., Liu, Y., Wang, M., Huang, X., & Tan, J. (2021). EDLMFC: an ensemble deep learning framework with multi-scale features combination for ncRNA–protein interaction prediction. *BMC bioinformatics*, *22*(1), 1-19.

[11]. Khadidos, A., Khadidos, A. O., Kannan, S., Natarajan, Y., Mohanty, S. N., & Tsaramirsis, G. (2020). Analysis of COVID-19 Infections on a CT Image Using DeepSense Model. Frontiers in Public Health, 8.

[12]. Natarajan, Y., Srihari, K., Dhiman, G., Chandragandhi, S., Gheisari, M., Liu, Y., ... & Alharbi, H. F. (2021). An IoT and machine learning-based routing protocol for reconfigurable engineering application. IET Communications.

[13]. Zhang, B., Li, J., Quan, L., Chen, Y., & Lü, Q. (2019). Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing*, *357*, 86-100.

[14]. Li, Y., Golding, G. B., & Ilie, L. (2021). DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics*, *37*(7), 896-904.

[15]. Tsubaki, M., Tomii, K., & Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, *35*(2), 309-318.

[16]. Syed, S. A., Sheela Sobana Rani, K., Mohammad, G. B., Chennam, K. K., Jaikumar, R., Natarajan, Y., ... & Sundramurthy, V. P. (2022). Design of Resources Allocation in 6G Cybertwin Technology Using the Fuzzy Neuro Model in Healthcare Systems. Journal of Healthcare Engineering, 2022.

[17]. Maheshwari, V., Mahmood, M. R., Sravanthi, S., Arivazhagan, N., ParimalaGandhi, A., Srihari, K., ... & Sundramurthy, V. P. (2021). Nanotechnology-Based Sensitive Biosensors for COVID-19 Prediction Using Fuzzy Logic Control. Journal of Nanomaterials, 2021.

[18]. Zheng, S., Li, Y., Chen, S., Xu, J., & Yang, Y. (2020). Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, *2*(2), 134-140.

[19]. Zhang, H., Guan, R., Zhou, F., Liang, Y., Zhan, Z. H., Huang, L., & Feng, X. (2019). Deep residual convolutional neural network for protein-protein interaction extraction. *IEEE Access*, *7*, 89354-89365.
[20]. Wekesa, J. S., Luan, Y., Chen, M., & Meng, J. (2019). A hybrid prediction method for plant lncRNA-protein interaction. *Cells*, *8*(6), 521.
[21]. Zhang, D., & Kabuka, M. (2019). Multimodal deep representation learning for protein interaction identification and protein family classification. *BMC bioinformatics*, *20*(16), 1-14.
[22]. Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184-192.
[23]. Liu, J., & Gong, X. (2019). Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC bioinformatics*, *20*(1), 1-11.
[24]. Lv, Z., Ao, C., & Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. *Proteomics*, *19*(14), 1900119.
[25]. Cong, Q., Anishchenko, I., Ovchinnikov, S., & Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. *Science*, *365*(6449), 185-189.