Briefings in Bioinformatics, 00(00), 2021, 1–13

https://doi.org/10.1093/bib/bbab255
Problem Solving Protocol

Improved prediction of protein–protein interaction using a hybrid of functional-link Siamese neural network and gradient boosting machines

Satyajit Mahapatra and Sitanshu Sekhar Sahu

Corresponding author. Satyajit Mahapatra, Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi-835215, India. E-mail: satyajit6243@gmail.com

Abstract

In this paper, for accurate prediction of protein–protein interaction (PPI), a novel hybrid classifier is developed by combining the functional-link Siamese neural network (FSNN) with the light gradient boosting machine (LGBM) classifier. The hybrid classifier (FSNN-LGBM) uses the fusion of features derived using pseudo amino acid composition and conjoint triad descriptors. The FSNN extracts the high-level abstraction features from the raw features and LGBM performs the PPI prediction task using these abstraction features. On performing 5-fold cross-validation experiments, the proposed hybrid classifier provides average accuracies of 98.70 and 98.38%, respectively, on the intraspecies PPI data sets of *Saccharomyces cerevisiae* and *Helicobacter pylori*. Similarly, the average accuracies for the interspecies PPI data sets of the *Human-Bacillus* and *Human-Yersinia* data sets are 98.52 and 97.40%, respectively. Compared with the existing methods, the hybrid classifier achieves higher prediction accuracy on the independent test sets and network data sets. The improved prediction performance obtained by the FSNN-LGBM makes it a flexible and effective PPI prediction model.

Key words: Protein–protein interaction; Siamese architecture; functional-link Siamese neural network; light gradient boosting machine

Introduction

Identifying and characterizing protein–protein interactions (PPIs) is essential for understanding the biological processes in the cell. Knowledge from these studies facilitates the identification of therapeutic targets and the design of novel drugs [1, 2]. Analysis of the intraspecies protein interactions (interaction between proteins present within the organism) helps to understand different life processes such as hormone regulation, metabolism, etc. The interspecies interaction, otherwise called host-pathogen interaction, is the protein interaction between two organisms. In this process, the pathogen protein binds with the host protein, altering some biological activities inside the host cell (humans, animals and plants). Infectious diseases, such as COVID19, Ebola, Anthrax, Plague, HIV and Cholera, are caused by virus or bacterial pathogen protein interaction with the protein of human host, affecting the health or leading to the death of millions of people. In the case of plant, host interacting with fungal or bacterial pathogens such as Pseudomonas syringae or Magnaporthe grisea leads to colossal crop loss across the globe. Experimental approaches for PPI

Satyajit Mahapatra is a PhD scholar in the Department of Electronics and Communication, Birla Institute of Technology Mesra, Ranchi, India. His research interests include applied machine learning and genomic signal processing.
Sitanshu Sekhar Sahu is presently working as an Assistant Professor in the Department of Electronics and Communication Engineering at the Birla Institute of Technology Mesra, Ranchi, India. He has received his PhD Degree from NIT Rourkela, India in 2011. He has completed his Post-doctorate from Oklahoma State University, USA, in 2012–2014. He has been a recipient of the DFAIT GSEF Fellowship from Canada Govt. in 2008. He has published more than 60 research papers in reputed refereed international journals and conferences. His research interests include signal and image processing, bioinformatics, machine learning and computer vision.

Submitted: 5 October 2020; **Received (in revised form):** 26 November 2020

prediction remain expensive, laborious and time-consuming. In addition, they often have high levels of false-positive predictions [3, 4]. Therefore, computational methods have emerged as an alternative for high throughput identification and characterization of PPIs. To develop a computational model for PPI prediction, protein information related to structure, domain and sequence is required. Since information related to protein sequence is easily accessible, the computational model for PPI prediction using sequence information has attracted researchers' attention.

In recent years, numerous computational methods have been developed that use several baseline classifiers [4–15] and deep neural network (DNN)-based classifiers [3, 16–23] to predict PPIs as a binary classification task. Support vector machine (SVM) [4–7], *K* nearest neighbor [8], Forest classifier [9, 10], Extreme learning machines [11] and gradient boosting machines [12, 13] are some of the baseline classifiers used for PPI prediction. DeepPPI [16], DPPI [17], EsnDNN [18], DeepInteract [19] and RCNN [20] are some of the DNN-based classifiers used to predict protein interactions. These models are based on the features derived from protein sequences, using various descriptors such as autocovariance (AC) [5], conjoint triad (CT) [6], multi-scale continuous and discontinuous (MCD) [7], local descriptors (LD) [8], local phase quantization [9], pseudo amino acid composition (PseAAC) [12], position-specific scoring matrix (PSSM) [17], etc. In general, these features encapsulate specific characteristics of the protein sequences, which includes local pattern frequencies, physicochemical properties and the positional distribution of the amino acids. Many researchers have adopted the multi-feature fusion strategy for predicting PPI. The combination of multiple features results in high-dimensional features. Therefore, many feature selection techniques, such as mRMR [7], elastic net [12], L1-Regularized Logistic Regression [13], PCA [14] and Chi-square test [15], have been used as a preprocessing step to obtain the optimal feature subset for use with the baseline classifiers. Secondly, these high-dimensional features are also used with DNNs [16–23] because of their ability to extract high-level abstraction features. The use of these abstraction features yields high prediction accuracy. A hybrid classifier scheme for predicting protein interactions is developed by cascading the CNN with FSRF [24]. In this scheme, CNN is used to derive the abstraction features from the raw features that the FSRF uses for prediction. Compared with other existing methods, the hybrid of CNN and FSRF yielded improved performance.

Motivation and contribution

Although good prediction accuracies have been achieved by the existing machine learning methods listed in the literature, most of them are focused on intraspecies PPI prediction. Fewer efforts have been made for interspecies PPI prediction [25–29]. Siamese neural networks (NNs) are efficacious for tasks involving understanding the dynamic relationship between two entities [30]. In PPI prediction, for processing the input protein pairs, Siamese DNN architecture is preferred [16, 17, 20]. Due to the additional layers of abstraction, the amount of weight adjustment increases, making the training process of DNN complicated. The functional-link artificial NN (FLANN) reported in [31] is a flat single layer network that makes learning and weight adjustment simpler. In FLANN, the input feature dimensions are artificially expanded, and using the functional expansion mechanism (FEM), the enhanced feature space produces better discriminatory input patterns [32]. In this work, to mitigate the requirement of a large number of abstraction layers, an NN architecture suitable for the PPI prediction task is developed by integrating the FEM of FLANN with the Siamese NN architecture. The newly designed NN is known as the functional-link Siamese NN (FSNN). In several applications, such as character recognition [33], remote sensing [34], protein fold recognition [35], speech separation [36], etc., hybrid models built by cascading NNs with baseline classifiers such as SVM, RF and XGB (extreme gradient boosting machine) produce superior performance compared with a single classifier.

The recently proposed Light gradient boosting machines (LGBM) classifier has the benefits of high training speed and lower memory consumption while still achieving approximately the same accuracy compared with XGB and pGBRT [37]. The LGBM classifier has shown superior performance than the existing baseline classifier in bioinformatics and computational biology tasks [12, 13, 38]. Therefore, in this paper, a hybrid classifier is developed by combining the FSNN with LGBM for intraspecies and interspecies PPI prediction. The proposed classifier uses the fusion of features obtained using PseAAC and CT descriptors. The FSNN operates on the raw features to extract high-level abstraction features. LGBM uses these abstraction features to predict the interaction between proteins.

Materials and methods

Data sets

In this paper, 10 benchmark data sets are employed for the evaluation of the proposed hybrid classifier. The details are as follows:

- The intraspecies PPI database of *Saccharomyces cerevisiae* and *Helicobacter pylori* collected from [12] contains 11 888 pairs of protein (5594 positive interactions and 5594 negative interactions) and 2916 pairs of protein (1458 positive interactions and 1458 negative interactions), respectively.
- The interspecies interaction database of *Human-Bacillus* and *Human-Yersinia* is collected from [25]. There are 3094 interacting pairs and 9500 noninteracting pairs in the *Human-Bacillus* interaction data set. Interaction data set *Human-Yersinia* comprises 4097 interacting pairs and 12 500 noninteracting pairs. The number of samples in both positive and negative classes is not equal in the PPI data sets, resulting in an unbalanced data set. Therefore, a balanced data set is obtained by randomly selecting the negative samples, equal to the number of positive samples.
- Furthermore, the model is validated using four independent test data sets and two network data sets obtained from [12]. The independent test data set contains interacting pairs of *Caenorhabditis elegans* (4013 pairs), *Escherichia coli* (6954 pairs), *Homo sapiens* (1412 pairs) and *Mus musculus* (313 pairs) data sets. The network data set contains a one-core network that has 16 pairs of PPIs and a crossover network, which has 96 pairs of PPIs.

Feature extraction

From the literature of the PPI study, it is observed that the fusion of multiple types of sequence-based features has shown better results compared with a single feature type [7, 12–15]. In this paper, pseudo amino acid composition (PseAAC) and conjoint triad (CT) descriptors are employed to transform the variable-length protein sequences, represented in alphabet form, into a numerical form of fixed-length. The dipoles and volume of side chains influence the electrostatic and hydrophobic properties of amino acids, which controls the interaction between proteins [6]. The CT descriptor extracts the sequence information by grouping the amino acids based on the dipoles and volume of side chains. The PseAAC descriptor extracts the correlation between residues of a certain distance which are useful for interaction study [12]. Therefore, the fusion of features obtained using PseAAC and CT descriptors is used to extract the pattern present in the protein sequences.

Pseudo amino acid composition

In the PseAAC descriptor [12], the amino acid correlation factor (λ) is integrated with the amino acid composition information. As a result, a $(20 + \lambda)$-dimensional feature vector (Z) is obtained, for each protein sequence, as defined in Equation 1

$$Z = \{z_1, z_2, z_3, \ldots\ldots\ldots\ldots, z_{20}, z_{20+1}, \ldots\ldots\ldots, z_{20+\lambda}\}^T \ (\lambda < L)\,. \quad (1)$$

The $(20 + \lambda)$ components are computed as

$$z_\eta = \left\{ \begin{array}{ll} \frac{f_\eta}{\sum_{\eta=1}^{20} f_\eta + \omega \sum_{\kappa=1}^{\lambda} \tau_\kappa}, & 1 \le \eta \le 20 \\ \frac{\omega \tau_{\eta-20}}{\sum_{\eta=1}^{20} f_\eta + \omega \sum_{\kappa=1}^{\lambda} \tau_\kappa}, & 21 \le \eta \le 20 + \lambda \end{array} \right\}, \quad (2)$$

where L represents the length of the protein sequence, ω represents the weight factor,f_η represents the occurrence frequency of η amino acids present in a protein sequence Z and τ_k represents the k-tier amino acid correlation information computed as

$$\tau_\kappa = \frac{1}{L - \kappa} \sum_{l}^{L-\kappa} F_{i,i+\kappa} \ (\kappa < L) \quad (3)$$

$$F_{i,i+\kappa} = \frac{1}{3} \left\{ [M(Re_i) - M(Re_{i+\kappa})]^2 + [H_A(Re_i) - H_A(Re_{i+\kappa})]^2 + [H_B(Re_i) - H_\tau(Re_{i+\kappa})]^2 \right\}, \quad (4)$$

where$M(Re_i)$, $H_A(Re_i)$ and $H_B(Re_i)$ are the side chain mass, hydrophobicity value and hydrophilicity value of the amino acid residue Re_i, respectively. The maximum value of λ should be less than the length of the shortest sequence present in the data set.

Conjoint triad

The CT descriptor [6] is based on the relationship between one amino acid with its adjacent amino acids. At first, the 20 amino acids are separated into seven clusters ({A, G, V}; {I, L, F, P}; {Y, M, T, S}; {H, N, Q, W}; {R, K}; {D, E}; {C}), based on the volume of the side chains and dipoles. The number of its respective group substitutes each amino acid present in the protein sequence. The three amino acids in succession are considered as a unit. Total, a set of combination which includes {1, 1, 1}; {1, 2, 1}; · · · {1, 7, 1}; · · · {1, 7, 7}; · · · {7, 7, 7} is formed. As a result, a 343-dimensional feature vector is obtained for each protein sequence.

Functional-link Siamese NN

The FSNN takes the features of a pair of protein sequences P1 (protein1) and P2 (protein2) as inputs, as shown in Figure 1, and gives a binary value O(P1, P2) as output, which indicates whether the two proteins are interacting or noninteracting. It consists of three modules: an input module, a feature extraction module and a prediction module. The input module, feature extraction module consists of two channels for processing the pair of inputs. The output of the two channels are combined and passed through the prediction module to obtain the probability score.

Input module

In this module, the features of protein sequences are artificially expanded using an FEM. Consider a data set D_{mn}, where m and n represent the number of samples and features, and k is the order of the functional expansion, respectively. For each input sample, $(2n + 1)$numbers of functionally expanded values are generated.

$$O(D_{mn}) = \{D_{mn}, \sin(\Pi D_{mn}), \cos(\Pi D_{mn}), \sin(2\Pi D_{mn}), \cos(2\Pi D_{mn}), \ldots\ldots\ldots\ldots, \sin(k\Pi D_{mn}), \cos(k\Pi D_{mn})\}\,.$$

Feature extraction module

In this module, each channel consists of two layers, as shown in Figure 1. The first layer contains three hidden units (HUs) connected in parallel to process the expanded inputs. The element-wise summation of the output of the first layer's HUs is passed through the HU of the second layer. In the HUs, the weighted sum of inputs is computed and passed through an activation function. In this paper, a rectified linear unit function (ReLU) activation function is used. For positive values, the ReLU function outputs the input directly, and for negative values, it outputs zero. Some neurons are dropped out (assigned zero) along with their connections to prevent over-fitting.

The HU begins with a dense layer and ends with a dropout layer. The output of an HU is defined as HU $=$ Dropout(Relu($Dense_N$())).

The computations of channel 1 are as follows:

$$C1_{l1} = HU(P1) + HU(\sin(\Pi P1)) + HU(\cos(\Pi P1)) + \ldots\ldots, + HU(\sin(k\Pi P1)) + HU(\cos(k\Pi P1))\,; C1_{l2} = HU(C1_{l1})\,,$$

where $C1_{l1}$ and $C1_{l2}$ represents the output of layer 1 and 2 of channel 1, respectively.

The computations of channel 2 are as follows:

$$C2_{l1} = HU(P2) + HU(\sin(\Pi P2)) + HU(\cos(\Pi P2)) + \ldots\ldots + HU(\sin(k\Pi P2)) + HU(\cos(k\Pi P2)); C2_{l2} = HU(C2_{l1})\,,$$

where $C2_{l1}$ and $C2_{l2}$ represents the output of layer 1 and 2 of channel 2, respectively.

The number of neurons (N) present in each HU of layer 1 is 256, and layer 2 is 128. The output of two channels undergoes an element-wise multiplication, i.e. ($C = C1_{l2} \odot C2_{l2}$) to infer the relationship between the pair of proteins [17, 20]. As a result, for each protein pairs, a 128-dimensional feature vector is obtained. Before passing through the prediction module, the abstraction features are normalized using a min-max normalization operator.

$$F = \frac{C - \min(C)}{\max(C) - \min(C)}\,. \quad (5)$$

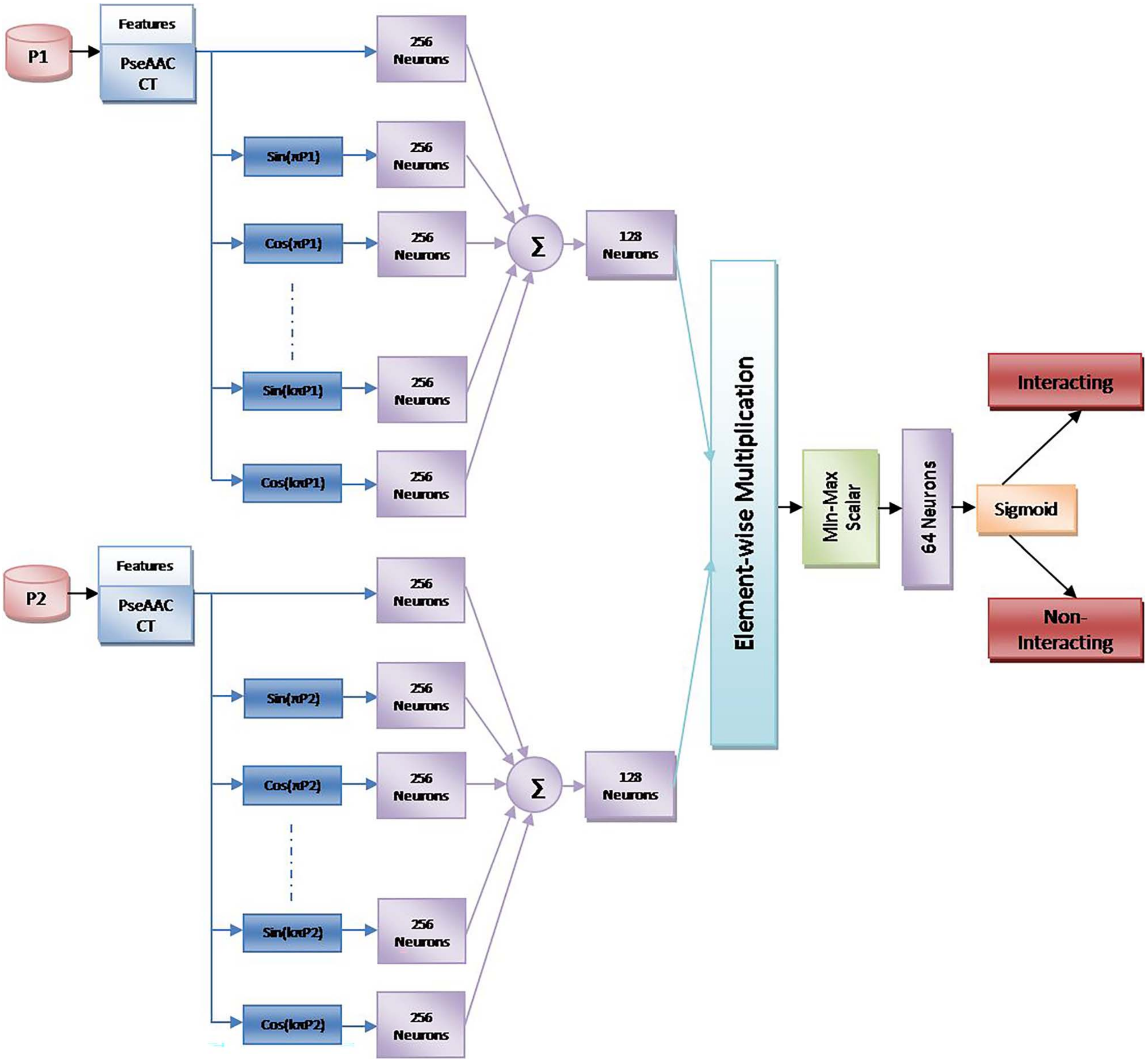


Figure 1. Proposed FSNNs for prediction of PPI (PseAAC, Pseudo amino acid composition features; CT, Conjoint triad features).

Prediction layer

This module consists of an HU with 64 neurons, followed by a single neuron with a sigmoid activation function that transforms the input vector Q ($Q = \mathrm{Dropout}(\mathrm{Relu}(\mathrm{Dense}_{64}(F)))$) of dimension d from the previous layer into an output score.

$$O(P_1, P_2) = \sum_{k=1}^{d} Q_k W_k + \beta_0, \tag{6}$$

where W_k are the weights corresponding to the input Q_k and β_0 is the bias vector.

Finally, the interaction probability between P_1 and P_2 is computed by $\frac{1}{1+e^{-O(P_1,P_2)}}$ and passed through the binary cross-entropy loss function given by

$$l\left(O(P_1,P_2), Y_{P_1P_2}\right) = -\left(Y_{P_1P_2} \log\left(O(P_1,P_2)\right) + \left(1 - Y_{P_1P_2}\right) \log\left(1 - O(P_1,P_2)\right)\right), \tag{7}$$

where $Y_{P_1P_2} = 1$ and $Y_{P_1P_2} = 0$ are the respective class labels for interacting and noninteracting.

Light gradient boosting machine

LGBM is a fast, distributed, high-performance, decision tree algorithm-based gradient boosting system developed by Microsoft in 2016 [37]. LGBM uses the histogram-based algorithm, which transforms continuous feature values into discrete values that fasten the training processes and minimize memory consumption. In contrast to other boosting algorithms, instead of a level-wise splitting approach, the LGBM uses a leaf-wise splitting approach. The leaf-wise algorithm reduces more losses compared with the level-wise algorithm, resulting in better accuracy. The leaf-wise growth leads to the development of an unbalanced decision tree which may lead to over-fitting. To prevent over-fitting, LGBM restricts the maximum depth during tree growth.

Consider a data set $D = \{(d_1, y_1); (d_2, y_2); \ldots..(d_K, y_K)\}$, where d and y represent features and class labels, respectively.

Step 1: Initialization of model $f_0(d)$

$$f_0(d) = argmin_h \sum_{k=1}^{K} L(y_k, h), \tag{8}$$

where k $(k = 1, 2, \ldots..K)$ represents the number of samples.

$L(y_k, h)$ denotes the loss function, computed as $L(y_k, h) = L(y, f(d)) = (y - f(d))^2$.

Step 2: Compute for N iterations to generate N weak learning models

(a) The gradient or the pseudo-residuals (r_{nk}) are computed as

$$r_{nk} = -\left[\frac{\partial L(y_k, f(d_k))}{\partial f(f(d_k))}\right]_{f(d)=f_{n-1}(d)}, \tag{9}$$

where n $(n = 1, 2, \ldots..N)$ represents the number of weak learning models.

(b) The residue $r_{nk}$is considered as the new feature of the sample. Fit a decision tree for $\{(d_1, r_{n1}), \ldots\ldots\ldots, (d_k, r_{nk})\}$ and create a new decision tree for $f_n(d)$. The area of the leaf nodes corresponding to $f_n(d)$ is R_{ln} (for $l = 1, 2, \ldots\ldots\ldots\ldots L$), where L is the number of leaf nodes in the classification tree $f_n(d)$.

(c) Compute the best-fit value of the leaf area

$$C_{lm} = argmin_c \sum_{d_k \epsilon R_{ln}} L(y_k, f_{n-1}(d_k + h)), \tag{10}$$

where $f_n(d)$ and $f_{n-1}(d)$ are the new model and current model, respectively.

I denotes the indicator function, where $I = \left\{\begin{smallmatrix} 1 \text{ if } d\epsilon R_{ln} \\ 0 \text{ if } d\notin R_{ln} \end{smallmatrix}\right\}$.

Step 3: The final additive model $F(d)$

$$F(d) = \sum_{n=1}^{N}\sum_{l=1}^{L} C_{ln} I(d\epsilon R_{ln}). \tag{11}$$

Hybrid classifier

The proposed hybrid classifier is constructed by replacing the final layer of FSNN with the LGBM classifier. The activation function (sigmoid) in the final layer of the FSNN provides an estimated probability of the input. The linear combination of outputs from the penultimate layer with trainable weights is the input to this activation function. For other classifiers, the output values of the penultimate layer can be used as input features. Figure 2 presents the architecture of the proposed hybrid classifier. First, in the input layer, the protein sequence features are expanded and used to train the FSNN. After the training is complete, the activation maps, otherwise called the abstraction features present in the penultimate layer, are predicted and used as input to the LGBM classifier.

Performance evaluation

For evaluating the performance of the FSNN-LGBM, the 5-fold cross-validation approach is employed. In this process, the whole data set is divided randomly into five separate nonoverlapping subsets of the same size. For testing, one subset is used, while the remaining four subsets are used for training. To ensure that each subset is used as a test set only once, this experiment is conducted five times, and the mean and standard deviation of these studies are taken as final results. For the evaluation of the proposed method, the output metrics used are as follows:

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

$$\text{Sensitivity (Sens)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (Spec)} = \frac{TN}{TN + FP}$$

$$\text{Precision (Prec)} = \frac{TP}{TP + FP}$$

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

where TP and TN are the correctly predicted number of interacting and noninteracting pairs, FP is the number of noninteracting pairs predicted as interacting and FN is the number of interacting pairs predicted as noninteracting.

Results and discussion

For each protein sequence, the PseAAC descriptor (with $\lambda = 3$ as optimal value) gives a 23-dimensional feature vector, and the CT descriptor gives a 343-dimensional feature vector. The features obtained using the two descriptors are combined, yielding a 366-dimensional feature vector for each protein sequence. Thus, for a pair of protein sequences, a 732-dimensional feature vector is obtained. This feature vector is used with the proposed hybrid FSNN-LGBM classifier to predict the interaction. The grid search is used to get the FSNN and LGBM classifiers' optimum parameters, outlined in Tables 1 and 2.

The 5-fold cross-validation results obtained on benchmark intraspecies (*S. cerevisiae*, *H. pylori*) and interspecies (*Human-Bacillus*, *Human- Yersinia*) data sets are presented in Table 3. From the sensitivity and specificity values, it is observed that the proposed FSNN-LGBM can distinguish the positive and negative samples effectively.

The 5-fold cross-validation accuracy of the FSNN-LGBM is compared with the standard FSNN, LGBM and other baseline classifiers, such as SVM, random forest (RF) and Adaboost (AB) across the four data sets. The proposed model is also compared with Siamese DNN-architecture called DeepPPI reported in [16] alongside the baseline classifiers. From the results presented in Figure 3, it is observed that with less number of abstraction layers, the proposed FSNN produces equivalent or superior performance compared with DeepPPI when used upon a fusion of PseAAC and CT features. Secondly, compared with the baseline classifiers such as SVM, RF and AB, the LGBM yields superior performance. Therefore, using the LGBM classifier on the high-level abstract features obtained using FSNN leads to an improvement in prediction accuracy. On the *S. cerevisiae* and *H. pylori* data sets, the FSNN-LGBM achieves an accuracy of 98.70 and 98.38%, which is about 5 and 3.5% higher than the individual accuracy of FSNN and LGBM. An accuracy of 98.52 and 97.40% is obtained

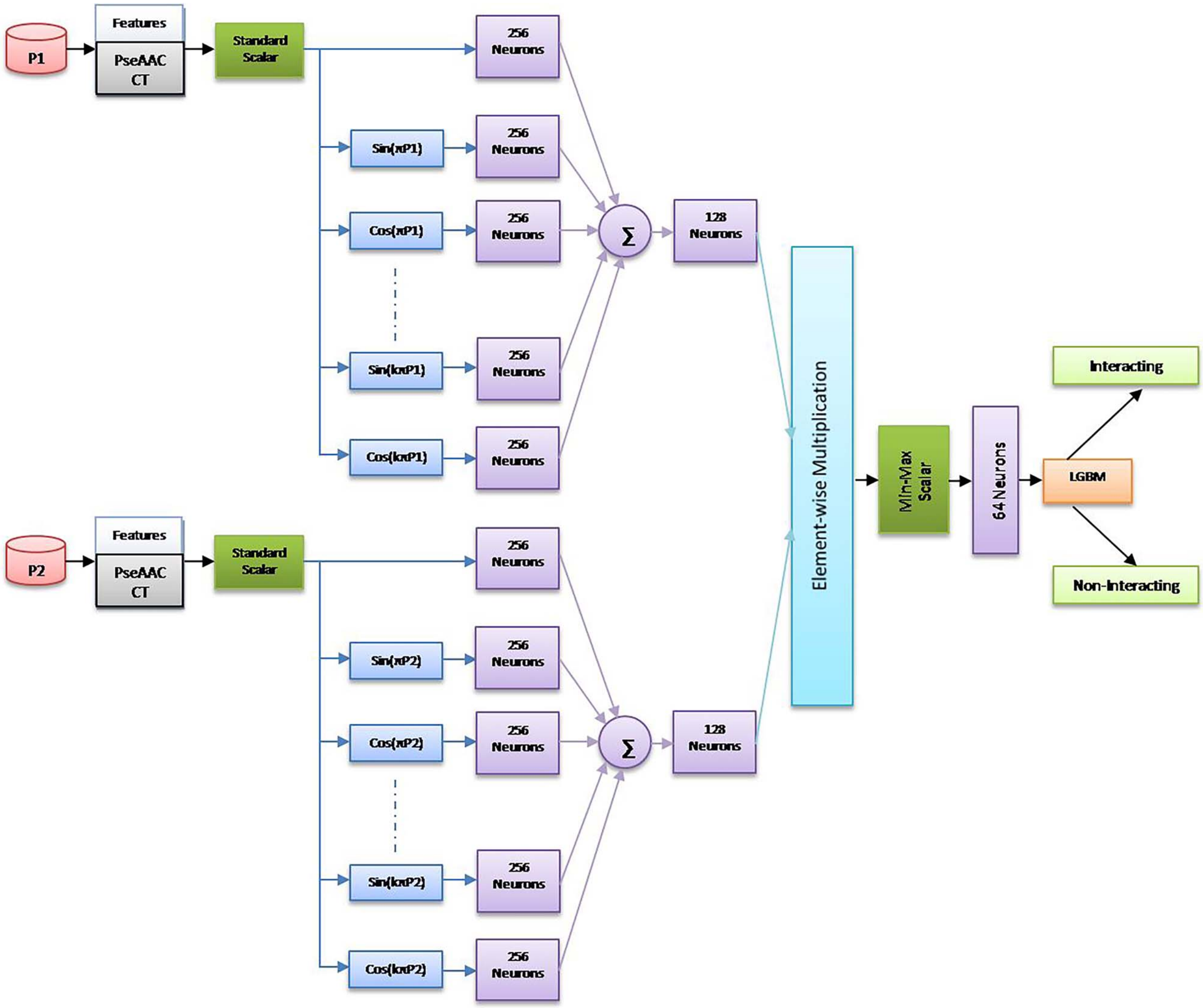


Figure 2. Schematic diagram of hybrid classifier (FSNN-LGBM).

Table 1. Parameters used for simulation of FSNN

Hyperparameter name	Range	Optimal value
Order of functional expansion (K)	1, 2, 3	1
Learning rate	1, 0.1, 0.01, 0.001	0.01
Batch size	16, 32, 64, 128	64
Momentum rate	0.8, 0.9	0.9
Weight initialization	uniform, normal, glorot_normal, glorot_uniform	glorot_normal
Weight regularization	L1, L2	L2
Adaptive learning rate method	SGD, RMSprop, Adam	SGD
Activation		ReLU, Sigmoid
Dropout rate	0.1, 0.2, 0.5	0.2
Loss function		binary_crossentropy
Epochs	10, 20, 30, 40, 50, 100	50

on the *Human-Bacillus* and *Human-Yersinia* data set, which is about 7 and 5% higher than the individual accuracy of FSNN and LGBM.

Furthermore, to assess the performance of the classifiers, receiver operating characteristics (ROC) is analyzed. ROC is the plot between the true positive rate (TPR) and the false positive rate (FPR). The comparison of ROC curves between the proposed hybrid classifier and other baseline classifiers using the *S. cerevisiae* and *Human-Yesrinia* data set is presented in Figures 4 and 5. From the ROC curve, it is observed that the proposed hybrid classifier has a high TPR and low FPR, which indicates its high discrimination capability. It provides that the area under the curve (AUC) for the *S. cerevisiae* data set is 0.997 and the *Human-Yesrinia* data set is 0.993, respectively.

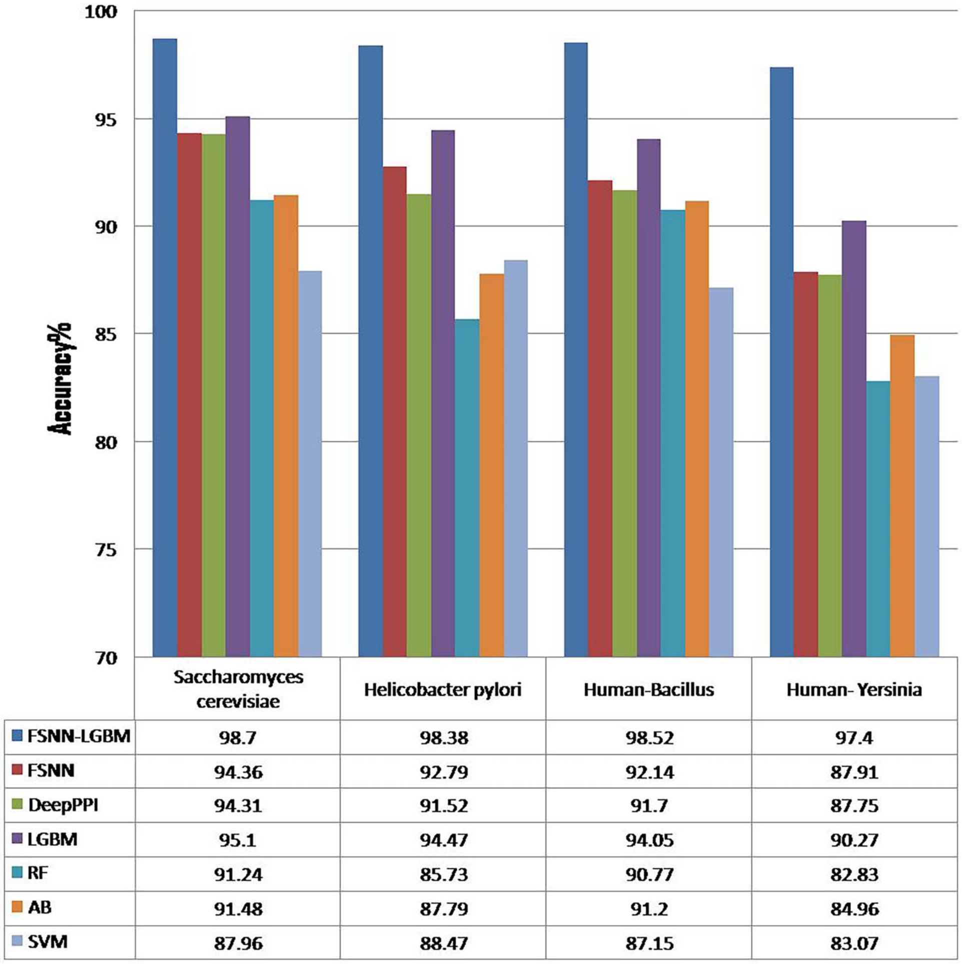
Table 2. Parameters used for simulation of LGBM

Hyperparameter name	Range	Optimal value
Booster		gbtree
Learning rate	0.05, 0.1, 0.15, 0.2	0.2
Gamma		0
Max_depth	6, 8, 10, 15	10
Tree method		Auto
NumParallael tree	500, 1000	1000

Table 3. Five-fold cross-validation results of the FSNN-LGBM on the intraspecies and interspecies data sets

Data set	Acc (%)	Sens(%)	Spec (%)	Prec (%)	MCC (%)	AUC
S. cerevisiae	98.70 ± 0.22	98.28 ± 0.39	99.12 ± 0.30	99.11 ± 0.26	97.41 ± 0.45	0.997
H. pylori	98.38 ± 0.48	98.26 ± 0.65	98.50 ± 0.47	98.30 ± 0.64	96.78 ± 0.96	0.992
Human-Bacillus	98.52 ± 0.37	98.51 ± 0.31	98.54 ± 0.44	98.55 ± 0.35	97.06 ± 0.73	0.995
Human-Yersinia	97.40 ± 0.36	97.73 ± 0.40	97.07 ± 0.51	97.09 ± 0.49	94.80 ± 0.92	0.993

Note: Results in the table are in the form of Mean ± Standard Deviation from the 5-folds.



	Saccharomyces cerevisiae	Helicobacter pylori	Human-Bacillus	Human- Yersinia
🟦 FSNN-LGBM	98.7	98.38	98.52	97.4
🟥 FSNN	94.36	92.79	92.14	87.91
🟩 DeepPPI	94.31	91.52	91.7	87.75
🟪 LGBM	95.1	94.47	94.05	90.27
🟦 RF	91.24	85.73	90.77	82.83
🟧 AB	91.48	87.79	91.2	84.96
🟦 SVM	87.96	88.47	87.15	83.07

Figure 3. Comparison of accuracy (%) of the hybrid classifier with baseline and DNN-based classifiers.

Comparison with existing prediction methods

For evaluation of the performance of the hybrid model, it is compared with the existing prediction methods on the four data sets. Using the *S. cerevisiae* data set, the proposed method is compared with several classical machine learning methods (MLD + RF [4], MCD + SVM [7], LightGBM [12], PSSM+RoF [39], GE + WSRC [40]) and deep learning-based methods (DeepPPI [16], DPPI [17], EsnDNN [18], DeepInteract [19], RCNN [20], DNN-LCTD [22], CNN-FSRF [24]). The comparison result is presented in Table 4. Among these methods, PSSM+ROF [39] provides the

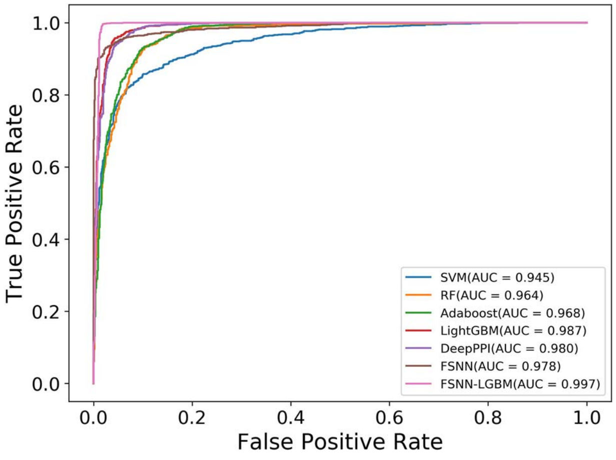


Figure 4. Comparison of ROC curves of different classifiers using the *S. cerevisiae* data set.

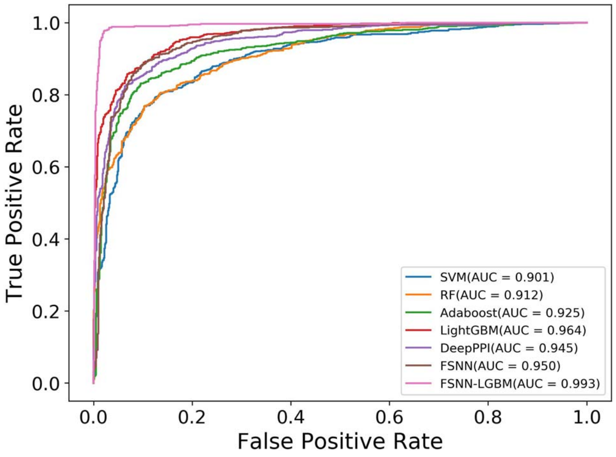


Figure 5. Comparison of ROC curves of different classifiers using the *Human-Yesrinia* data set.

best accuracy among the classical methods, and CNN-FSRF [24] provides the best accuracy for the deep learning-based methods. The proposed method achieved an enhancement of 1.64% compared with PSSM+ROF and 0.95% to the CNN-FSRF. The performance comparison of the proposed method with existing methods using the *H. pylori* data set is presented in Table 5. The best among the classical method is GE + WSRC [40], and the best accuracy is achieved by CNN-FSRF [24] in deep learning-based approach. Again, the proposed hybrid classifier has shown superior results compared with these methods. It achieves an improvement in an accuracy of 5.55% compared with GE + WSRC and 9.42% compared with CNN-FSRF. Performance comparison of the proposed method with the existing method using *Human-Bacillus* and *Human-Yersinia* is presented in Tables 6 and 7, respectively. The proposed method achieves an improvement in accuracy of 13.12 and 6.82% on the *Human-Bacillus* data set and 12.8 and 10.1% on the *Human-Yersinia* data set, compared with the LBE + RF [25] and LD + DNN [28].

Comparison of prediction performance on independent test sets

The independent test is based on the presumption that if a large number of interacting proteins have evolved in a correlated manner in one organism, then it is probable that their respective orthologs will interact. The *C. elegans*, *E. coli*, *H. sapiens* and *M. musculus* have shown orthologs to the *S. cerevisiae* [4]. Therefore, the entire *S. cerevisiae* data set is used as a training set, and the *C. elegans*, *E. coli*, *H. sapiens* and *M. musculus* are employed as independent test sets. The independent test sets comprise of only interacting pairs. Thus, only accuracy (ACC percent) is calculated and, compared with existing methods, the comparison is reported in Table 8. The proposed hybrid classifier is compared with five deep learning (CNN-FSRF [24], DPPI [17], DeepPPI [16], EsnDNN [18], DNN-LCTD [22]) and four classical approaches (GcForest-PPI [10], GTB-PPI [13], LightGBM [12], MLD-RF [4]) for prediction. It is observed that, relative to existing methods, the proposed method produces better performance, demonstrating that the proposed method has a better generalization capability.

Comparison of prediction performance on PPI network data sets

A computational model for PPI prediction is considered suitable for practical applications if it is capable of predicting PPI networks [6]. Therefore, to evaluate the efficacy of the proposed method, two PPI network data sets, i.e. the one-core network and the crossover network data sets, collected from [12] are employed. In the one-core network data set, the human CD9 is the core protein that interacts with 16 other satellite proteins. The crossover network consists of several multi-core and/or one-core networks with dynamic interactions between these networks. This data set contains 96 interacting protein pairs that are associated with the formation and growth of tumors in humans. The proteins of *S. cerevisiae* have orthologs with *human* proteins [4]. Therefore, the *S. cerevisiae* data set is used for training the model, while the one-core and crossover network data sets are used as test sets. A graphical representation of the prediction results achieved by the FSNN-LGBM on the one-core and the crossover network data sets is presented in Figure 6 and 7, respectively. Each node in the graph represents a protein. The protein pairs connected with a solid line represent that they are predicted as interacting by the model and connections with a dotted line represent that the model predicts them as noninteracting.

As shown in Figure 6, the proposed FSNN-LGBM classifier correctly predicts all the PPI in the single-core network; 95 out of 96 interactions are correctly predicted in the crossover network data set, as shown in Figure 7. A comparative analysis of the prediction results obtained by the proposed method with the existing methods on two network data sets is provided in Table 9. It is observed that the proposed method has achieved equivalent or better results compared with the existing methods.

It is inferred from the comparative analysis that the proposed method achieves better results on both intraspecies and interspecies data sets than the existing classical machine learning and deep learning methods. Furthermore, the assessment on the independent test sets and network data sets shows its generalization ability in interaction prediction.

Prediction of human-SARS-CoV-2 PPI

COVID-19 (Coronavirus Disease-19), a disease caused by the SARS-CoV-2 virus, was declared as a pandemic by the World Health Organization on 11 March 2020. The FSNN-LGBM is used to construct a Human-SARS-CoV-2 PPI prediction model that can be used to predict new interacting protein pairs between humans and SARS-CoV-2. The Human-SARS-CoV-2 Positive Data PPI data set is obtained from the Intact database

Table 4. Performance comparison of FSNN-LGBM with existing methods on *S. cerevisiae* data set

Method	Acc (%)	Sens (%)	Prec (%)	MCC (%)	AUC(%)
MLD + RF [4]	94.72	94.34	98.91	85.99	NA
MCD + SVM [7]	91.36	90.37	91.94	84.21	97.07
LightGBM [12]	95.07	92.21	97.82	90.30	98.75
DeepPPI [16]	94.43	92.06	96.65	88.97	97.00
DPPI [17]	94.55	92.24	96.68	NA	NA
EsnDNN [18]	95.29	95.12	95.45	90.59	97.00
DeepInteract [19]	92.67	86.85	98.31	85.96	NA
RCNN [20]	97.09	97.17	97.00	94.17	NA
DNN-LCTD [22]	93.11	92.40	93.75	86.24	97.95
CNN-FSRF [24]	97.75	99.61	95.89	96.04	97.54
PSSM+RoF [39]	97.06	95.23	98.85	94.18	97.11
GE + WSRC [40]	96.82	93.63	100	93.83	96.88
Proposed Method (FSNN-LGBM)	98.70	98.28	99.11	97.41	99.70

Note: NA means not available.

Table 5. Performance comparison of FSNN-LGBM with existing methods on *H. pylori* data set

Method	Acc (%)	Sens (%)	Prec (%)	MCC (%)	AUC(%)
MLD + RF [4]	88.30	92.47	NA	79.19	NA
LightGBM [12]	89.03	89.99	88.36	78.14	95.34
DeepPPI [16]	86.23	89.44	84.32	72.63	NA
CNN-FSRF [24]	88.96	91.86	86.86	78.09	89.08
PSSM+RoF [39]	89.69	88.53	90.66	79.42	90.07
GE + WSRC [40]	92.83	89.32	96.13	86.65	93.75
Proposed Method (FSNN-LGBM)	98.38	98.26	98.30	96.78	99.20

Note: NA means not available.

Table 6. Performance comparison of FSNN-LGBM with existing methods on *Human-Bacillus* data set

Method	Acc (%)	Sens (%)	Prec (%)	MCC (%)	AUC(%)
LBE + BN [25]	78.7	73.0	42.0	43.4	83.70
LBE + NB [25]	82.5	52.8	47.8	39.7	82.10
LBE + RF [25]	85.4	24.0	67.0	34.0	86.80
AAC + BN [25]	77.4	51.7	37.3	30.3	79.00
LBE + j48 [25]	80.06	31.2	39.6	23.9	54.10
LD + DNN [28]	91.7	89.5	93.9	83.5	96.37
Proposed Method (FSNN-LGBM)	98.52	98.51	98.55	97.06	99.50

Table 7. Performance comparison of FSNN-LGBM with existing methods on *Human-Yersinia* data set

Method	Acc (%)	Sens (%)	Prec (%)	MCC (%)	AUC(%)
LBE-BN [25]	76.10	73.5	38.6	40.1	81.30
LBE-NB [25]	80.9	45.5	43.2	32.8	78.60
LBE-RF [25]	84.6	16.0	66.3	27.3	83.50
AAC-BN [25]	80.0	52.4	42.1	34.9	75.6
LBE-j48 [25]	80.1	27.9	37.1	20.8	51.70
LD-DNN [28]	87.3	84.2	90.4	74.9	94.99
Proposed Method (FSNN-LGBM)	97.40	97.73	97.09	94.80	99.30

Table 8. Performance comparison of FSNN-LGBM with existing methods on the Independent test set

Data set→ Algorithm↓	*C. elegans* (Acc %)	*E. coli* (Acc %)	*H. sapiens* (Acc %)	*M. musculus* (Acc %)
Proposed Method (FSNN-LGBM)	97.06	96.98	98.44	98.40
MLD-RF [4]	87.71	89.30	94.19	91.96
GcForest-PPI [10]	96.01	96.30	98.58	99.04
LightGBM [12]	90.16	92.16	94.83	94.57
GTB-PPI [13]	92.42	94.06	97.38	98.08
DeepPPI [16]	93.77	91.37	94.84	92.19
DPPI [17]	95.51	96.66	96.24	95.84
EsnDNN [18]	93.22	95.10	95.00	94.06
DNN-LCTD [22]	93.17	94.62	94.18	92.65
CNN-FSRF [24]	96.41	95.47	98.65	93.27

Table 9. Performance comparison of FSNN-LGBM with existing methods on PPI network data sets

Data set→ Algorithm↓	One-core network	Crossover network
Proposed method (FSNN-LGBM)	16/16	95/96
SVM-CT [6]	13/16	73/96
GcForest-PPI [10]	16/16	94/96
LightGBM-PPI [12]	15/16	89/96
GTB-PPI [13]	15/16	92/96

Note: The numerator represents the number of correctly predicted interactions, and the denominator represents the total number of interactions.

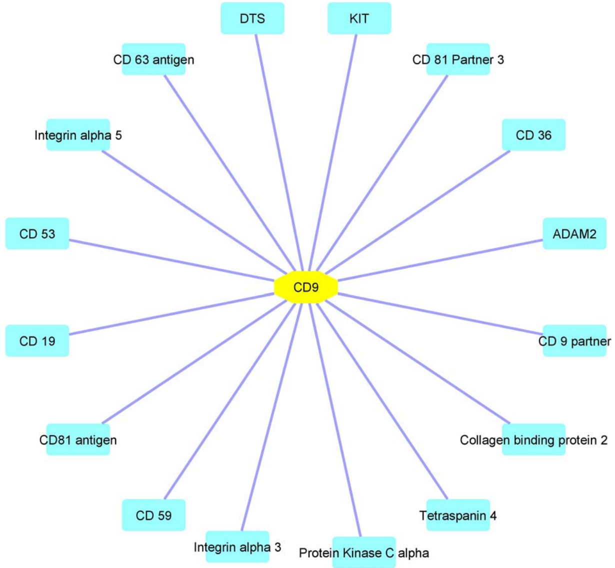


Figure 6. Graphical representation of the prediction result achieved by the FSNN-LGBM on the one-core network data set. Core and satellite proteins are colored yellow and green, respectively.

[41], which contains human protein interactions with SARS-CoV2 and SARS-CoV, as well as some interactions with other members of the Coronaviridae community. A total of 4658 interacting pairs (UniProt ID pairs) are extracted, of which about 85% (4000 samples) are used as a positive data set and the remaining 15% (658 samples) are used as an independent test set. Due to the unavailability of an experimentally confirmed noninteracting data set (i.e. negative samples), a negative data set is prepared according to the procedure specified in [25]. The protein sequences are randomly paired, one from the host (human) and the other from the pathogen organism (SARS-CoV2) for which there is no evidence of an interaction. The FSNN-LGBM classifier is trained using 4000 positive and 4000 negative samples, utilizing the 5-fold cross-validation approach, and then evaluated using the independent test set. Accuracy of 98% obtained in both training and testing as listed in Table 10 demonstrates the effectiveness of the proposed method in analyzing the Human-SARS-CoV-2 protein pairs which are suspected to be interacting. It is important to take note that the samples suspected to be interacting do not appear in the negative data set.

Feature visualization

The main advantage of the NN is its ability to extract discriminative features (abstraction features) from the raw features. The *t*-SNE [42] (t-Distributed Stochastic Neighbor Embedding) is a powerful tool commonly used with the NN to visualize and compare abstraction features and the raw features. In the current study, the abstraction features derived by the FSNN is used as an input to the LGBM classifier to enhance the accuracy of PPI prediction. Therefore, visualizing the distribution of the raw and the abstraction features will provide an insight into the reason for the enhancement of accuracy. A comparison of the *t*-SNE plot of original features and abstraction features of the *S. cerevisiae* data set and the Human-Yesrinia data set is given in Figures 8 and 9. From the figures, it can be seen that the original features of positive and negative samples are overlapping, whereas the abstraction features are discriminative. Thus, it is inferred that the proposed FSNN architecture efficiently extracts meaningful information from the raw features pertinent to the interaction. The *t*-SNE plot is implemented in python using the *scikit-learn* library.

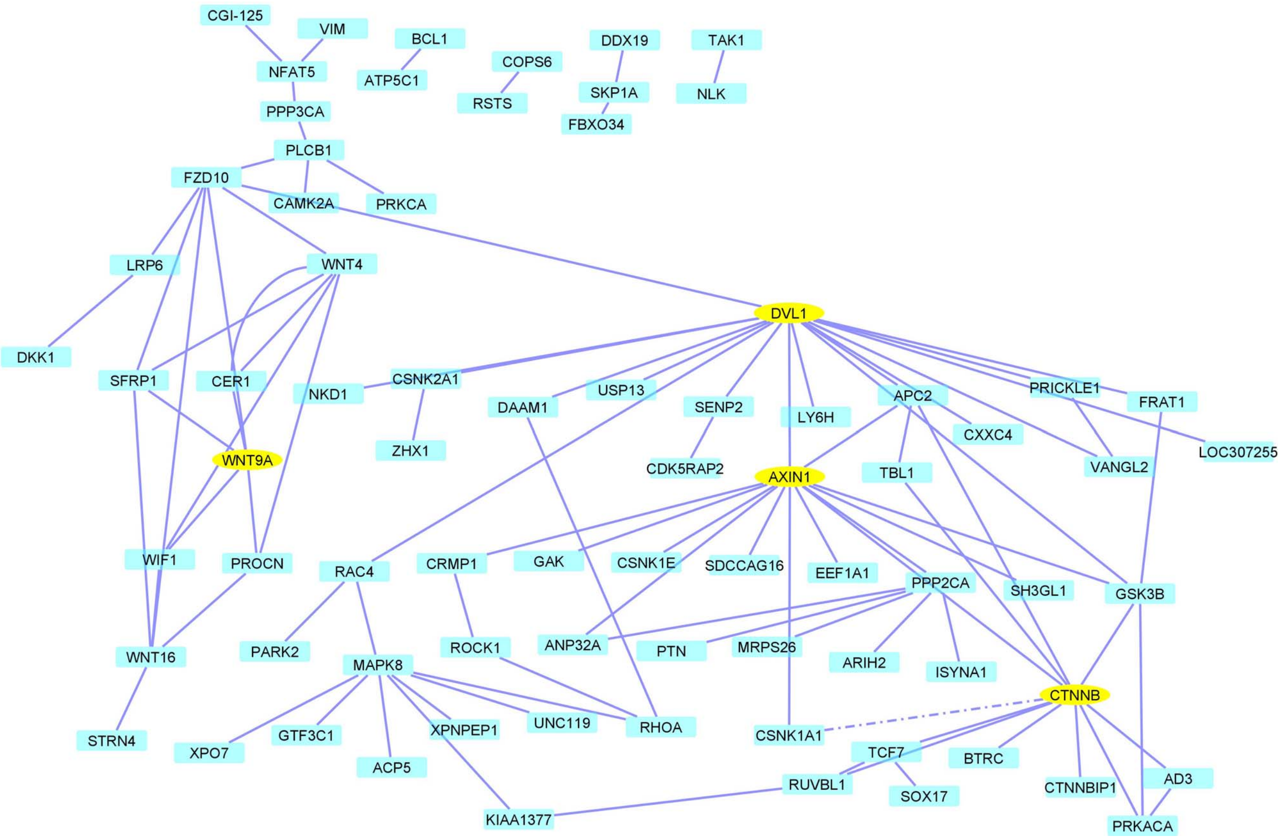


Figure 7. Graphical representation of the prediction result achieved by the FSNN-LGBM on the crossover network data set. The solid lines are the true prediction, and the dotted lines are the false prediction.

Table 10. Performance of the FSNN-LGBM on Human-SARS-CoV-2 PPI

Data set (Human-SARS-CoV-2)	Acc (%)	Sens(%)	Spec (%)	Prec (%)	MCC (%)	AUC
Five-fold CV	98.86 ± 0.22	98.60 ± 0.35	99.12 ± 0.25	99.13 ± 0.24	97.73 ± 0.45	0.998
Independent test	Accuracy: 98.50%					

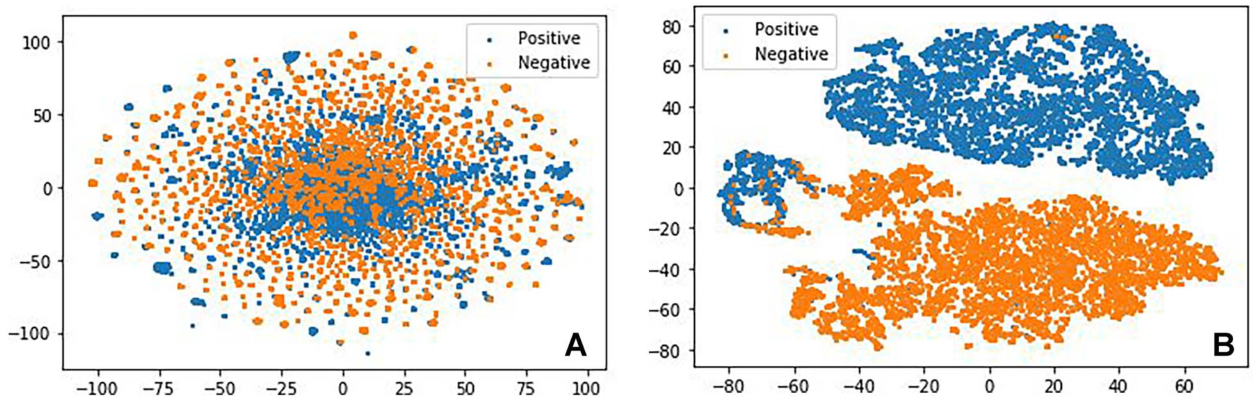


Figure 8. *t*-SNE plots of *S. cerevisiae* data set (**A**) raw input features, (**B**) abstracted features present in the penultimate layer. Blue color dots represent positive interactions, and orange dots represent negative interactions.

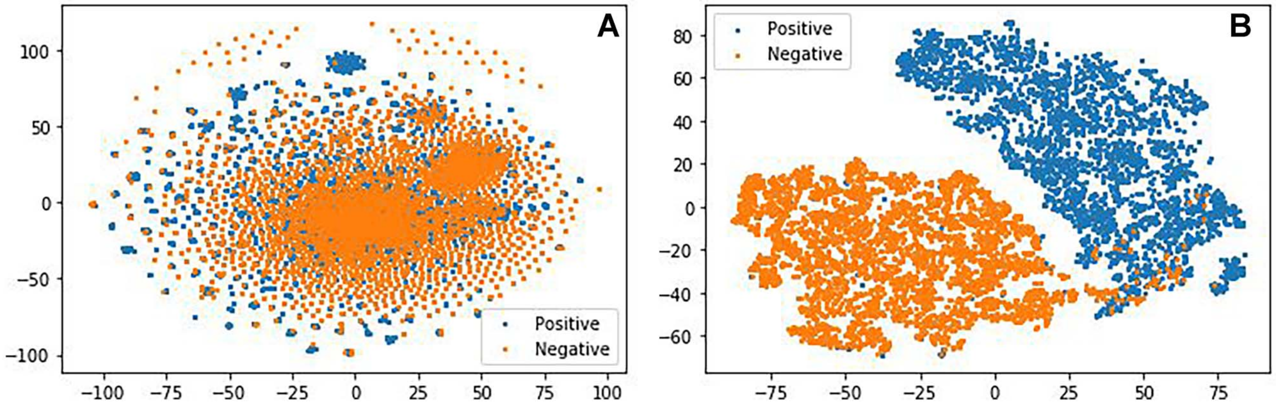


Figure 9. t-SNE plots of *Human-Yesrinia* data set (**A**) raw input features, (**B**) abstracted features present in the penultimate layer. Blue color dots represent positive interactions, and orange dots represent negative interactions.

Conclusion

In this paper, a new hybrid approach that combines the FSNN with the LGBM is presented to predict PPI efficiently. The fusion of sequence information features conjoint triad (CT), and pseudo amino acid composition (PseAAC) descriptors are used as input to the hybrid classifier. The FSNN utilizes nonlinear transformation techniques to extract abstraction features from the raw features of the protein sequences. The extracted features used with the LGBM classifier enhances the prediction accuracy. When assessed with several standard intraspecies and interspecies PPIs, the FSNN-LGBM performs substantially well compared with existing methods. In addition, on independent test sets, the hybrid classifier achieved higher accuracy than existing methods. The prediction results on network data sets indicate that it can provide new insights into the signaling pathway analysis, the prediction of drug targets and the understanding of disease pathogenesis. Although the proposed method provided better results than existing methods, it comes at the expense of an increased computational burden.

Key Points

- A hybrid classifier termed FSNN-LGBM is developed by combining the functional link Siamese neural network (FSNN) with light gradient boosting machines (LGBM) for PPI prediction.
- FSNN extracts high-level abstraction features from raw protein sequence features.
- LGBM classifier uses abstraction features for predicting PPIs.
- FSNN-LGBM achieved improved accuracy on interspecies and intraspecies PPI data sets.

Availability of data and codes

The datasets and codes used for this study are available on request to the corresponding author.

Acknowledgements

This work has been carried out in the Signal Processing Lab, Department of Electronics and Communication Engineering of Birla Institute of Technology, Mesra, Ranchi

Funding

Science and Engineering Research Board, Department of Science and Technology Government of India (Grant No. ECR/2017/000345).

References

1. Petta I, Lievens S, Libert C, *et al*. Modulation of protein–protein interactions for the development of novel therapeutics. *Mol Ther* 2016;**24**(4):707–18.
2. Skrabanek L, Saini HK, Bader GD, *et al*. Computational prediction of protein–protein interactions. *Mol Biotechnol* 2008;**38**(1):1–17.
3. Sun T, Zhou B, Lai L, *et al*. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 2017;**18**(1):277.
4. You ZH, Chan KC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS One* 2015;**10**(5):e0125811.
5. Guo Y, Yu L, Wen Z, *et al*. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res* 2008;**36**(9):3025–30.
6. Shen J, Zhang J, Luo X, *et al*. Predicting protein–protein interactions based only on sequences information. *Proc Natl Acad Sci* 2007;**104**(11):4337–41.
7. You ZH, Zhu L, Zheng CH, *et al*. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics* 2014;**15**(S15):S9.
8. Yang L, Xia JF, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept Lett* 2010;**17**(9):1085–90.
9. Wong, L., You, Z. H., Li, S., *et al*. Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor. In *International Conference on Intelligent Computing* (pp. 713–20), 2015. Springer, Cham.
10. Yu B, Chen C, Wang X, *et al*. Prediction of protein-protein interactions based on elastic net and deep forest. *Expert Syst Appl* 2019;**176**:114876.doi: 10.1016/j.eswa.2021.114876.
11. You ZH, Lei YK, Zhu L, *et al*. Prediction of protein-protein interactions from amino acid sequences with ensemble

extreme learning machines and principal component analysis. *BMC Bioinformatics* 2013;**14**(S8):S10.

12. Chen C, Zhang Q, Ma Q, *et al.* LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom Intel Lab Syst* 2019;**191**:54–64.
13. Yu B, Chen C, Zhou H, *et al.* GTB-PPI: predict protein-protein interactions based on L1-regularized logistic regression and gradient tree boosting. *Genomics, Proteomics Bioinforma* 2021. doi: 10.1016/j.gpb.2021.01.001.
14. Göktepe YE, Kodaz H. Prediction of protein-protein interactions using an effective sequence based combined method. *Neurocomputing* 2018;**303**:68–74.
15. Wang L, You ZH, Xia SX, *et al.* An improved efficient rotation forest algorithm to predict the interactions among proteins. *Soft Computing* 2018;**22**(10):3373–81.
16. Du X, Sun S, Hu C, *et al.* DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J Chem Inf Model* 2017;**57**(6):1499–510.
17. Hashemifar S, Neyshabur B, Khan AA, *et al.* Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 2018;**34**(17):i802–10.
18. Zhang L, Yu G, Xia D, *et al.* Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 2019;**324**:10–9.
19. Patel S, Tripathi R, Kumari V, *et al.* DeepInteract: deep neural network based protein-protein interaction prediction tool. *Current Bioinformatics* 2017;**12**(6):551–7.
20. Chen M, Ju CJT, Zhou G, *et al.* Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* 2019;**35**(14):i305–14.
21. Wang X, Wang R, Wei Y, *et al.* A novel conjoint triad auto covariance (CTAC) coding method for predicting protein-protein interaction based on amino acid sequence. *Math Biosci* 2019;**313**:41–7.
22. Wang J, Zhang L, Jia L, *et al.* Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. *Int J Mol Sci* 2017;**18**(11):2373.
23. Yao Y, Du X, Diao Y, *et al.* An integration of deep learning with feature embedding for protein-protein interaction prediction. *PeerJ* 2019;**7**:e7126. http://doi.org/10.7717/peerj.7126.
24. Wang L, Wang HF, Liu SR, *et al.* Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Sci Rep* 2019;**9**(1):1–12.
25. Kösesoy İ, Gök M, Öz C. A new sequence based encoding for prediction of host–pathogen protein interactions. *Comput Biol Chem* 2019;**78**:170–7.
26. Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLoS One* 2014;**9**(11): e112034. https://doi.org/10.1371/journal.pone.0112034.
27. Zhou X, Park B, Choi D, *et al.* A generalized approach to predicting protein-protein interactions between virus and host. *BMC Genomics* 2018;**19**(6):568.
28. Mahapatra S, Sahu SS. Boosting predictions of Host-Pathogen protein interactions using Deep neural networks. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science* (SCEECS), 2020, pp. 1–4. doi: 10.1109/SCEECS48394.2020.150.
29. Chen H, Li F, Wang L, *et al.* Systematic evaluation of machine learning methods for identifying human-pathogen protein-protein interactions. *Brief Bioinform* 2021;**22**(3):bbaa068. https://doi.org/10.1093/bib/bbaa068.
30. Bromley J, Bentz JW, Bottou L, *et al.* Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 1993;**7**(4):669–88.
31. Pao YH. *Adaptive Pattern Recognition and Neural Networks*. 1989, Chapter 8, pp. 197–22. Addison-Wesley, Reading, MA.
32. Naik B, Obaidat MS, Nayak J, *et al.* Intelligent secure ecosystem based on metaheuristic and functional link neural network for edge of things. *IEEE Transactions on Industrial Informatics* 2019;**16**(3):1947–56.
33. Weldegebriel HT, Liu H, Haq AU, *et al.* A new hybrid convolutional neural network and eXtreme gradient boosting classifier for recognizing handwritten Ethiopian characters. *IEEE Access* 2019;**8**:17804–18.
34. Dong L, Du H, Mao F, *et al.* Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—subtropical area for example. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2019;**13**:113–28.
35. Liu B, Li CC, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform* 2020**21**(5):1733–41. https://doi.org/10.1093/bib/bbz098.
36. Wang Y, Wang D. Towards scaling up classification-based speech separation. *IEEE Trans Audio Speech Lang Process* 2013;**21**(7):1381–90.
37. Ke G, Meng Q, Finley T, *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. 2017, 3149–57. Curran Associates Inc., Red Hook, NY, USA.
38. Zhang Y, Xie R, Wang J, *et al.* Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform* 2019;**20**(6):2185–99.
39. Zhu HJ, You ZH, Shi WL, *et al.* Improved prediction of protein-protein interactions using descriptors derived from PSSM via gray level co-occurrence matrix. *IEEE Access* 2019;**7**: 49456–65.
40. Huang YA, You ZH, Chen X, *et al.* Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinformatics* 2016;**17**(1):184.
41. Orchard S, Ammari M, Aranda B, *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2014;**42**(D1):D358–63.
42. Maaten LVD, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008;**9**(Nov):2579–605.