Computational and Structural Biotechnology Journal 23 (2024) 316–329

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

Deep-HPI-pred: An R-Shiny applet for network-based classification and prediction of Host-Pathogen protein-protein interactions

Muhammad Tahir ul Qamar [a,*,1], Fatima Noor [b,1], Yi-Xiong Guo [c], Xi-Tong Zhu [a], Ling-Ling Chen [a,*]

[a] State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Life Science and Technology, Guangxi University, Nanning 530004, China
[b] Integrative Omics and Molecular Modeling Laboratory, Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad 38000, Pakistan
[c] National Key Laboratory of Crop Genetic Improvement, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

ARTICLE INFO

Keywords:
Host–pathogen interactions
Deep learning
Multilayer perceptron
Neural networks
Topological features

ABSTRACT

Host-pathogen interactions (HPIs) are vital in numerous biological activities and are intrinsically linked to the onset and progression of infectious diseases. HPIs are pivotal in the entire lifecycle of diseases: from the onset of pathogen introduction, navigating through the mechanisms that bypass host cellular defenses, to its subsequent proliferation inside the host. At the heart of these stages lies the synergy of proteins from both the host and the pathogen. By understanding these interlinking protein dynamics, we can gain crucial insights into how diseases progress and pave the way for stronger plant defenses and the swift formulation of countermeasures. In the framework of current study, we developed a web-based R/Shiny app, Deep-HPI-pred, that uses network-driven feature learning method to predict the yet unmapped interactions between pathogen and host proteins. Leveraging citrus and *C*Las bacteria training datasets as case study, we spotlight the effectiveness of Deep-HPI-pred in discerning Protein-protein interaction (PPIs) between them. Deep-HPI-pred use Multilayer Perceptron (MLP) models for HPI prediction, which is based on a comprehensive evaluation of topological features and neural network architectures. When subjected to independent validation datasets, the predicted models consistently surpassed a Matthews correlation coefficient (MCC) of 0.80 in host-pathogen interactions. Remarkably, the use of Eigenvector Centrality as the leading topological feature further enhanced this performance. Further, Deep-HPI-pred also offers relevant gene ontology (GO) term information for each pathogen and host protein within the system. This protein annotation data contributes an additional layer to our understanding of the intricate dynamics within host-pathogen interactions. In the additional benchmarking studies, the Deep-HPI-pred model has proven its robustness by consistently delivering reliable results across different host-pathogen systems, including plant-pathogens (accuracy of 98.4% and 97.9%), human-virus (accuracy of 94.3%), and animal-bacteria (accuracy of 96.6%) interactomes. These results not only demonstrate the model's versatility but also pave the way for gaining comprehensive insights into the molecular underpinnings of complex host-pathogen interactions. Taken together, the Deep-HPI-pred applet offers a unified web service for both identifying and illustrating interaction networks. Deep-HPI-pred applet is freely accessible at its homepage: https://cbi.gxu.edu.cn/shiny-apps/Deep-HPI-pred/ and at github: https://github.com/tahirulqamar/Deep-HPI-pred.

1. Introduction

Host-pathogen interactions are crucial determinants in the dynamics of infectious diseases [1]. These interactions are primarily facilitated by protein-protein interactions (PPIs), which orchestrate every phase of disease progression, from the initial pathogen invasion to its eventual establishment within the host [2]. Citrus Huanglongbing (HLB), also known as citrus greening, represents a paradigmatic case of devastating

* Corresponding authors.
E-mail addresses: m.tahirulqamar@hotmail.com (M. Tahir ul Qamar), llchen@gxu.edu.cn (L.-L. Chen).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.csbj.2023.12.010
Received 22 October 2023; Received in revised form 11 December 2023; Accepted 12 December 2023
Available online 15 December 2023

plant diseases on a global scale [3]. Predominantly caused by the bacterium *Candidatus* Liberibacter asiaticus (*C*Las), along with *Ca.* L. americanus and *Ca*. L. africanus, HLB has emerged as a primary threat to citrus production, particularly in regions like Florida, USA [4]. CLas, the most prevalent among these pathogens, is vectored by the Asian citrus *psyllid Diaphorina* citri, leading to its colonization in the phloem tissues of citrus plants [5]. This invasion results in severe phytopathological symptoms and extensive agricultural damage. The economic ramifications of HLB are staggering, with the disease inflicting multibillion-dollar losses globally, debilitating the citrus industry's financial stability [6]. The pervasiveness of *C*Las and its profound impact on citrus crops underscore the necessity of an in-depth exploration of the PPIs between the host plants and the pathogen. Recently Yuan et al. [7] mapped a network of *Arabidopsis thaliana* interactions with the bacterial pathogen *Pseudomonas syringae*, identifying new components involved in the plant immune response and paving the way for future plant disease control strategies. In a similar vein, Dyer et al. [8] conducted a large-scale study on PPIs between human cells and *Bacillus anthracis*, the bacterium responsible for anthrax, revealing numerous potential therapeutic intervention targets. These studies underscore the vital role of host-pathogen PPIs in understanding the complex interplay between the host's defenses and the pathogen's invasion strategies. Therefore, these interactions form the cornerstone of research aiming to unravel the intricacies of infectious disease dynamics and serve as targets for potential therapeutic interventions.

Recent advancements in experimental methodologies, particularly in high-throughput screening and molecular imaging techniques, have significantly deepened our comprehension of host-pathogen PPIs [9]. These innovations have enabled more precise identification and analysis of PPIs, shedding light on the intricate mechanisms of infection and host response, and paving the way for novel therapeutic interventions. Despite this progress, such techniques often come with high costs in terms of resources, time, and labor. Given the multitude of potential protein-interacting partners, the reliance on these methods becomes even more challenging, underlining the necessity for more cost-effective and efficient alternatives [10,11]. In contrast, computational methods for predicting host-pathogen interactions present themselves as a fitting solution to this pressing demand [12]. These methods not only facilitate the detection of interactions but also have the potential to mine incomplete interaction maps for new discoveries, exponentially increasing the knowledge base regarding host-pathogen dynamics [12]. Recently, Kaundal et al. [13] developed deepHPI, an all-encompassing deep learning platform for the accurate prediction and visualization of HPPIs. Leveraging convolutional neural networks (CNN), their platform achieved a prediction accuracy of 96% on a curated dataset, significantly surpassing the performance of other methods on the same dataset. Complementing this work, Loaiza et al. introduced Pred-HPI [14], an integrated web server platform that employs sequence-based methods for the detection and visualization of host-pathogen interactions. PredHPI offered a sequence similarity-based approach and reached an accuracy of approximately 90% on a large-scale human-virus PPI dataset.

Despite the progress, these approaches primarily employed feature extraction strategies that represent host-pathogen protein pairs as fixed-length feature vectors, extracted from protein sequences. These sequence-based methods, while valuable, have limitations in terms of achieving high prediction accuracy, as they do not fully exploit the wealth of available structural and functional information. Recognizing these limitations, our study introduces a novel approach that leverages network-based integration methods for predicting host-pathogen interactions. Preliminary results have shown that this approach outperforms traditional sequence-based methods in terms of prediction accuracy scores.

To address these gaps, we introduced Deep-HPI-pred, an innovative R-shiny application that leverages advanced deep learning models. This application, for the first time, provides researchers with the autonomy to manually or automatically upload their training datasets for host-pathogen protein interaction prediction. The distinguishing characteristic of Deep-HPI-pred lies in its pioneering use of topological features for PPI prediction, as opposed to the conventional sequence-based features. This novel approach significantly enhances the sophistication and accuracy of predictions, marking a considerable advance in the bioinformatics field. Further, in order to substantiate the robustness and reliability of the Deep-HPI-pred model, our framework encompasses a rigorous validation protocol encompassing diverse biological systems, including plant-pathogen, human-virus, and animal-bacteria interactomes. This empirical evaluation across distinct host-pathogen pairs fortifies the scientific validity of the model's predictive prowess. In summary, as the premier R shiny application dedicated to predicting host-pathogen interactions, Deep-HPI-pred not only presents a groundbreaking tool for researchers, but also catalyzes a paradigm shift in our approach to understanding and predicting the intricate dynamics of infectious diseases.

2. Materials and methods

2.1. Gathering of true protein-protein interactions (PPIs) and data preprocessing

In this study, our initial host-pathogen PPI data was sourced from established host-pathogen interaction databases, including Pred-HPI [14] and GreeningDB [15]. These databases provided a broad spectrum of interactions, serving as a foundational reference for our research. However, to ensure a more robust and comprehensive dataset, we did not solely rely on these pre-compiled interactions. Recognizing the importance of data authenticity, we supplemented our database-derived PPIs with experimentally validated interactions. These additional PPIs were meticulously extracted from a wide range of peer-reviewed scientific literature. This step was vital for expanding the dataset and reinforcing its validity with real-world, experimentally confirmed interactions, thus substantially increasing the count of true positive PPIs in our study. The compiled data, encompassing both database-sourced and manually extracted PPIs, then underwent rigorous pre-processing. This crucial phase was aimed at maintaining the highest level of data integrity. We meticulously cleaned the combined dataset, carefully removing any duplicate entries to ensure a unique and non-redundant set of PPIs. This thorough cleaning process was instrumental in ensuring the reliability and accuracy of our dataset, thereby enhancing the overall quality of our analysis. By integrating these two diverse sources of data – established databases and manually verified literature – we aimed to strike a balance between the breadth and depth of our PPI dataset. This approach not only provided a comprehensive view of the host-pathogen interactions but also added a layer of validation to the interactions derived from databases, thereby bolstering the credibility and applicability of our study's findings.

2.2. Generation of negative protein-protein interactions (PPIs)

To augment our existing true positive PPIs dataset, the generation of negative PPIs was conducted. Negative PPIs, representing protein pairs that do not interact, serve as an important counterpart to positive interactions in our analysis [16]. The generation of negative PPIs was a two-step process. The first step involved creating a random pairing between the protein sets of the selected pathogen and host. Given that there are m pathogen proteins and n host proteins, the total possible pairings, assuming no interaction between any two proteins, are $m \times n$. Let $P = \{p_1, p_2, p_3 \ldots, p_m\}$ be the set of pathogen proteins and $H = \{h_1, h_2, h_3 \ldots, h_n\}$ be the set of host proteins. The total possible random pairings R between pathogen proteins and host proteins is represented by the Cartesian product of the two sets:

$$R = P \times H = \{(p_i, h_j) : p_i \in P \text{and} h_j \in \mathrm{H}\} \tag{3.1}$$

In the second step, a subset was randomly selected from these random pairings to construct the negative dataset. This subset selection ensured that the number of negative PPIs matched the number of true positive PPIs in our dataset. This balancing act serves as a cornerstone for the impending deep learning model training, underpinning an unbiased and balanced approach to the development of a robust predictive model. In current study, the generation of negative PPIs was integral to creating a balanced dataset for effective host-pathogen PPI prediction. This approach aligns with methodologies adopted in similar studies, ensuring robustness and validity in our predictive model. Specifically, Chen et al. [17] employed a comparable strategy in their research, where negative PPIs were systematically generated through random pairing, thereby maintaining an equivalent size of negative and positive datasets. This method of balancing is crucial, as underscored by Scott et al. [18], who highlighted the potential risks associated with imbalanced datasets in PPI predictions, such as overfitting and poor model generalization. Our methodology addresses these concerns by matching the number of negative PPIs with positive PPIs, thereby fostering a more accurate and unbiased predictive model.

2.3. Assembly of dataset

The assembly of the final dataset was carried out by integrating both positive *P_set* and negative (*N_set*) PPIs. The negative PPI dataset, representing non-interactions, was carefully merged with the positive PPI dataset. This integrated approach not only maintains the balance between interaction and non-interaction data but also encapsulates the full range of possible protein interactions. Mathematically, the final dataset *F_set* is represented as the union of the positive PPI dataset (P_{set}) and the negative PPI dataset (*N_set*):

$$F\_set = P\_set \cup N\_set \tag{3.2}$$

This final dataset, *F_set*, encapsulates a broad spectrum of potential PPIs. It is primed for rigorous training within the deep learning model, offering a well-rounded, realistic, and balanced resource for the analysis. This strategic assembly of positive and negative interactions underpins the analytical framework, fostering a robust and comprehensive model tailored for effective host-pathogen interaction prediction.

2.4. Feature extraction: quantification of network topology parameters

The creation of the final PPI dataset marked the transition to a critical phase - feature extraction, which is the quantification of the topological parameters of the constructed network [19]. The intent behind this step is to decipher and numerically articulate the intricate structural patterns within the protein network. Our primary objective was to leverage network topology for the accurate prediction of interactions between host and viral proteins. To achieve this, current study strategically employed a range of centrality measures, namely Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, PageRank Centrality, Hub Score, and Eccentricity. This was achieved with the assistance of the topological algorithms incorporated within the Igraph package of R [20]. These features were not selected on the basis of their individual merits alone but were chosen for their collective ability to provide a comprehensive view of the network dynamics, crucial for understanding the intricacies of host-pathogen PPI networks. This methodological approach is supported by the work of Przulj et al. [21] and Ashtiani et al. [22], who have demonstrated the importance of network topology in understanding PPI networks and the utility of centrality measures in identifying key proteins within these networks. These topological features assist in identifying the potential interaction points and key nodes within the host-pathogen network.

These features, ranging from quantifying the number of direct connections a protein node has (Degree Centrality) to calculating the maximum distance from a node to all other nodes (Eccentricity), offer a comprehensive view of each protein's role and importance within the overall network structure. Degree Centrality and Hub Score, for instance, shed light on the interaction richness and the core component status of a protein respectively [22,23]. For a graph G with N vertices, the Degree Centrality (DC) of a vertex v is given by $DC(v) = \frac{Degree(v)}{N-1}$, where $Degree(v)$ represents the number of edges incident on vertex v. Similarly, Betweenness Centrality, Closeness Centrality, and PageRank Centrality provide insights into a protein's role as a network connector, its centrality, and its importance based on neighbouring proteins. Betweenness Centrality is defined as $BC(v) = \sum \frac{\sigma(s,t|v)}{\sigma(s,t))}$ for all $s \neq v \neq t$, where $\sigma(s,t)$ is the total number of shortest paths from node s to node t and $\sigma(s,t)$ is the number of those paths that pass through v. For a connected graph G with N vertices, the Closeness Centrality (CC) of a vertex v is given by $CC(v) = \frac{1}{\sum d(v,t)}$ for all t, where $d(v,t)$ represents the shortest-path distance between v and t.

On the other hand, Eigenvector Centrality allots relative scores to all nodes in the network, acknowledging that connections to high-scoring nodes contribute more to the overall node score [24]. The Eigenvector Centrality (EC) of a node i is defined as $EC(i) = \frac{1}{\lambda}\sum A_{ij} EC(j)$ for all j, where A_{ij} are the elements of the adjacency matrix and λ is a constant. PageRank Centrality is computed using the formula $PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(T_n)/C(T_n))$, where $PR(A)$is the PageRank of page A, $PR(T_i)$ is the PageRank of pages T_i which link to page A, $C(T_i)$is the number of outbound links on page T_i and d is a damping factor which set between 0 and 1. Finally, the Eccentricity (ECC) of a vertex v in a graph is the maximum distance from v to all other vertices, calculated by $ECC(v) = \max d(v,t) for\ all\ t \neq v$ where $d(v,t)$ is the shortest-path distance between the vertices v and t..

2.5. Deep learning models

Current study leveraged the potential of deep learning through the implementation of three distinguished models, namely Identity Convolutional Neural Networks (ID-CNN), Recurrent Neural Networks (RNN), and Multi-Layer Perceptrons (MLP). In R, the 'keras' package was employed, providing a high-level neural networks API on the TensorFlow platform. Additionally, 'caret' in R, facilitating access to numerous machine learning algorithms, was used in conjunction with 'tensorflow' and 'reticulate'. 'Tensorflow' offers an R interface for TensorFlow, while 'reticulate' enables Python integration within the R environment. Each model's unique architecture processes input data differently, offering diversified data interpretations, thereby improving the accuracy and reliability of the predicted host-pathogen PPIs.

2.5.1. ID-Convolutional Neural Networks (CNN)

The ID-Convolutional Neural Networks (CNN) model forms an integral part of the deep learning approach. The ID-CNN are primarily known for their application in image processing, but their utility extends to PPI prediction as well [25]. The core architecture of the ID-CNN model implemented is designed with multiple layers, each layer equipped with a specific role and function to perform [26]. The architecture commenced with a 2D convolution layer equipped with 16 filters of kernel size 1×2, employing the Rectified Linear Unit (ReLU) activation function $f(x) = \max(1,x)$. This layer was instrumental in the detection of initial patterns or features within the data, with its output, $O1$ given by $O1 = f(1 * F1)$. Following this, a second 2D convolution layer was added, this one featuring 32 filters and a kernel size of 1×1. The same ReLU activation function was used, aiding in the learning of more complex patterns based on the initial features. The output of this layer, $O2$ was calculated by $O2 = f(O1 * F2)$.

Post the convolutional operations, a global max pooling layer was implemented, denoted by $M = \max O2$.This step reduced the dimensionality of the model, mitigating computational complexity and overfitting. Subsequently, two dense layers were added to the architecture. The first of these layers consisted of 64 neurons, its output, $O3$, defined

by $O3 = f(M * W1 + b1)$. The final layer was a single neuron employing a sigmoid activation functiong$(x) = 1 + \exp(-x)1$, producing the model output:

$$P = g(O3 * W2 + b2) \tag{3.3}$$

The entire model was trained using the Adam optimizer and the binary cross-entropy loss function, a suitable choice for the binary nature of the task. This function is given by:

$$Loss = -(ylog(p) + (1-y)\log(1-P) \tag{3.4}$$

It's noteworthy that the input data to the model included degree centrality features for both pathogen and host proteins, which was normalized before being fed into the model.

2.5.2. Recurrent neural network

In the pursuit of host-pathogen protein interaction analysis, a Recurrent Neural Network (RNN) was incorporated, ideal due to the inherent sequential and interconnected nature of the protein data [27]. An RNN model was devised via the Keras library in R [28], comprising of an input layer, a hidden layer, and an output layer. The input layer is designed to accept feature vectors signifying both the pathogen and host protein's network centrality measures, reshaped suitably for an RNN. The hidden layer utilizes simple RNN units governed by the ReLU activation function. On a mathematical note, suppose the input sequence is $X = (x_1, x_2, \ldots, x_n)$. The new hidden state at time t or H_t is calculated by the equation

$$H_t = f(H_{t-1} \times W_h + x_t \quad \times W_x + b) \tag{3.5}$$

where f is the ReLU function, x_t represents the input at time t, and W_h, W_x, and b stand for the weights and bias parameters. The output layer is a dense layer featuring a single unit using a sigmoid activation function, fitting for the binary classification task as it computes probabilities. Mathematically, the output y_t at time t is given by the equation

$$y_t = sigmoid(H_t \times W_o + b_o) \tag{3.6}$$

with *sigmoid* as the sigmoid functions and W_o and $b_o$as the weight and bias parameters for the output layer respectively. The training of this model was accomplished using the Adam optimization algorithm, an efficient extension to stochastic gradient descent. The loss function applied was binary cross-entropy, appropriate for binary classification problems. On a mathematical level, assuming the target (true value) as T_t and the predicted output as y_t, the binary cross-entropy loss for the output at time t is calculated as

$$L_t = -T_t\log(y_t) - (1-T_t)\log(1-y_t) \tag{3.7}$$

This framework of equations directs the forward propagation, learning, and prediction process of the RNN model within the study.

2.5.3. Multilayer Perceptron (MLP)

The Multi-layer Perceptron (MLP), an artificial neural network type, was selected for this analysis owing to the networked architecture of the protein data. The MLP's ability to decipher complex, non-linear relationships through multiple layers of nodes made it a perfect fit for this study. Each node, symbolizing a neuron, is equipped with an activation function that aids in transforming the input data [29]. The MLP model employed for this study comprises two hidden layers, containing 10 and 20 neurons respectively. These hidden layers form the heart of the MLP, endowing the network with its capacity to abstract and transform the input data. As the input data flows through these hidden layers, the model unveils intricate patterns and structures, allowing for more accurate prediction outcomes.

A critical aspect of this model is the logistic activation function. Chosen for its ability to map any real-valued number into a range between 0 and 1, the logistic function is a prime choice for binary classification tasks. It proves ideal for differentiating between interaction and non-interaction within the network-based protein data. Mathematically, in the hidden layers, each neuron computes a weighted sum of the inputs, adds a bias term and applies the logistic activation function. If we denote the input to a neuron as $x = (x_1, x_2, \ldots, x_m)$, weights as $w = (w_1, w_2, \ldots, w_m)$, and bias as b, the output of the neuron y is computed as follows:

$$y = logistic\left(\sum_{i=1}^{m} w_i \times x_i + b\right) = \frac{1}{1 + e^{-(\sum_{i=1}^{m} w_i \times x_i + b)}} \tag{3.8}$$

The MLP model training was regulated by a learning rate of 0.01 and a convergence threshold of 0.1. These values were fine-tuned to optimize the learning process and to control the pace at which the model learns from the protein data. Balancing the learning rate was essential to avoid erratic learning behaviors and to prevent the model from learning too slowly. A convergence threshold of 0.1 was set to ensure the learning process perseveres until the error on the training data reaches an acceptably low point.

Additionally, the model utilized the backpropagation algorithm to adjust its weights and bias values. This adjustment minimizes the discrepancy between the actual and predicted outputs over numerous iterations, or epochs, thereby progressively enhancing the model's accuracy. During each epoch, the backpropagation algorithm implements two primary steps for each training sample: Forward Propagation: The predicted output (y_{pred}) is calculated using the current weights and bias values. On the other hand, the weights and bias terms are updated based on the gradient of the loss function with respect to the weights and bias in back propagation. If we denote the loss function as $L(y, y_{pred})$, where y is the actual output and (y_{pred}) is the predicted output, the weights and bias are updated as follows:

$$w_i = w_i - \eta \times \frac{\partial L(y, y_{pred})}{\partial w_i} \tag{3.9}$$

$$b = b - \eta \times \frac{\partial L(y, y_{pred})}{\partial b} \tag{3.10}$$

Here η is the learning rate and $\frac{\partial L(y,y_{pred})}{\partial w_i}$ and $\frac{\partial L(y,y_{pred})}{\partial b}$ are the gradients of the loss function with respect to the weights and bias, computed using the chain rule of differentiation.

2.6. Evaluation of models

To thoroughly assess the predictive competence of the models, six widely recognized evaluation metrics were employed: Precision, Accuracy, Sensitivity, Specificity, F1-score, and Matthew's Correlation Coefficient (MCC). These metrics provide a comprehensive understanding of the model's performance, considering both the positive and negative classes of the dataset, as well as the balance between them.

Precision (Pre) is a metric that reflects the exactness of the positive predictions made by the model [30]. It is calculated as the ratio of true positives to the sum of true positives and false positives. Precision is essential as it focuses on the proportion of true positive predictions among all positive predictions, a crucial measure when false positives can significantly skew results, as discussed in the work of Powers et al. [31]. Mathematically, it's represented as:

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)} \tag{3.11}$$

Accuracy (Acc) measures the proportion of total predictions that are correct, irrespective of whether they are positive or negative [32]. However, when used alongside the other metrics, it contributes to a holistic understanding of the model's effectiveness, as suggested by Johnson and Khoshgoftaar [33]. It is computed as the sum of true positives and true negatives divided by the total number of predictions, as shown:

$$Accuracy = \frac{TruePositives(TP) + TrueNegatives(TN)}{TotalPredictions} \quad (3.12)$$

Sensitivity (Sn), also known as recall or true positive rate, quantifies the ability of the model to correctly identify positive instances [34]. It is calculated as the ratio of true positives to the sum of true positives and false negatives, which are expressed as:

$$Sensitivity = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)} \quad (3.13)$$

Specificity (Sp) is the metric that reflects the model's ability to correctly identify negative instances [35]. It is the ratio of true negatives to the sum of true negatives and false positives, represented as:

$$Specificity = \frac{TrueNegatives(TN)}{TrueNegatives(TN) + FalsePositives(FP)} \quad (3.14)$$

F1-score is the harmonic mean of Precision and Sensitivity. It is particularly useful when dealing with imbalanced datasets as it takes both false positives and false negatives into account [36]. The F1-score provides a balanced view of the model's performance on both classes, ensuring that neither false positives nor false negatives are disproportionately affecting the model's evaluation, as indicated in studies by Kakkar et al. [37]. The formula for the F1-score is:

$$F1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (3.15)$$

Matthew's Correlation Coefficient (MCC) is a more robust metric as it considers all values in the confusion matrix (true positives, true negatives, false positives, and false negatives) [36]. The range of MCC is between -1 and $+1$, where $+1$ represents a perfect prediction, 0 indicates a random prediction, and -1 denotes an inverse prediction. The MCC is calculated as follows:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (3.16)$$

These performance evaluation metrics were critical in systematically comparing and assessing the predictive capabilities of the 1D-CNN, RNN, and MLP models in the context of host-pathogen interaction prediction. By utilizing these six metrics, it became possible to analyze each model's proficiency from various dimensions such as precision, accuracy, sensitivity, specificity, and overall correlation. It is important to underscore that this comprehensive evaluation framework facilitated the discernment of the best-suited deep learning model for predicting potential interactions. This strategic methodology considerably enhanced the selection process, ensuring the identification of the most adept model to navigate the complex landscape of host-pathogen interactions.

3. Results

3.1. Dataset assembly and preliminary analysis

In order to predict host-pathogen interactions, the Citrus-HLB (Huanglongbing) interactome was selected as the primary case study. The primary sources of data for this study included the Pred-HPI and GreeningDB databases. The Pred-HPI database, known for its predictive modeling features, was combined with GreeningDB, a specialized database for citrus greening disease interactions, to provide a comprehensive range of interaction data. The true PPI data between Citrus and the Candidatus Liberibacter asiaticus (CLas) bacteria was sourced from the Greening and Pred-HPI database. This resulted in the retrieval of 1011 host-pathogen PPIs (45 unique pathogen proteins and 359 unique host proteins) from these databases. Additionally, 5 experimentally validated interactions were incorporated into the dataset, leading to a total of 1016 true PPIs. The dataset comprised of 45 unique pathogen proteins and 359 unique host proteins, accumulating a network of 404 nodes with a potential for 1016 possible interactions.

3.2. Construction of negative interaction sets with varied ratios

In this study, a common strategy was adopted to generate negative Protein-Protein Interactions (PPIs) based on traditional approaches found in the literature. This strategy involved creating random pairings between distinct protein sets - specifically, the selected pathogen proteins and host proteins. Subsequently, from these random pairings, subsets were strategically chosen to represent negative interactions. The importance of this lies in creating a comparative base against which positive interactions can be evaluated, thereby enhancing the predictive power of the developed models.

To further enrich the analysis and examine the robustness of the models, negative interactions were selected with varying ratios: 1:1 and 1:10. This led to the formation of two distinct datasets. The first, balanced, set comprised 1016 positive and 1016 negative PPIs, while the second, imbalanced, set encompassed 1016 positive and 10160 negative PPIs. The implementation of varying negative to positive ratios was essential in evaluating the robustness and resilience of the predictive models under different levels of data imbalance, a common scenario in biological studies. In the development of predictive models for host-pathogen PPIs, the selection of an appropriate class ratio is pivotal to balance between model performance and realistic representation of the dataset. For this study, a 1:10 class ratio was strategically chosen, aiming to effectively address the challenges presented by highly imbalanced datasets and to ensure adequate representation of the minority class. This decision is rooted in evidence from the literature, where extremely imbalanced ratios like 1:25 or 1:100, commonly used in PPI studies, have been shown to introduce a significant bias towards the majority class, leading to a high incidence of false negatives, as elucidated by Chen et al. [17]. Conversely, ratios that closely approximate equal distribution, such as 1:1, often do not accurately reflect the actual distribution encountered in biological datasets, where non-interacting pairs are generally more prevalent. The adoption of a 1:10 ratio represents this balance, reducing the risk of bias towards the majority class while still mirroring the typical distribution in biological datasets. Supporting this selection, preliminary experiments within the study demonstrated that a 1:10 ratio optimally balances sensitivity and specificity. This finding is in alignment with the results presented by Lei et al. [38], where similar class ratios have been shown to enhance model performance in accurately identifying true positives without disproportionately increasing false negatives. Hence, this ratio has been found to effectively represent the real-world distribution of PPI datasets, ensuring sensitivity towards the minority class and enhancing the overall reliability of the predictive models.

3.3. Performance of the features in modeling host-pathogen interactions

The feature extraction process was integral to the deployment of the 1D-CNN, RNN, and MLP models. Seven key features, namely, Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, PageRank Centrality, Hub Score, and Eccentricity were selected to comprehensively describe each host-pathogen protein pair. For each host-pathogen protein pair, these attributes were extracted and consolidated into a unified vector. This process resulted in comprehensive Protein-Protein Interaction (PPI) feature vectors, embodying the synergistic characteristics of both host and pathogen proteins. Each feature was individually assessed across the three models to deduce their unique contribution to the predictive accuracy of the models, fostering a comprehensive understanding of the interplay between individual features and model performance.

An effective evaluation of the models necessitated the careful construction of the independent datasets. Contrary to common practice where cross-validation is predominantly used due to its superior performance with unseen datasets, a different strategy was implemented for

this study. Specifically, a fifth of the PPIs, encompassing both positive and negative interactions, were randomly allocated to form the independent dataset. The remainder of the PPIs, involving both types of interactions, were amalgamated into the training set. This approach, though unconventional, ensured a robust and comprehensive evaluation of the models' performance.

3.3.1. Assessment of 1D-CNN model at different class ratios

Analyzing the performance of the 1D-CNN model at different class ratios resulted in significant findings. The model's efficacy was assessed using multiple protein features, and statistical parameters such as Sensitivity, Specificity, Precision, Accuracy, F1-score, and MCC were evaluated (Fig. 1). Table 1 provides an exhaustive assessment of the 1D-CNN model's performance at a class ratio of 1:1. During independent testing, the Degree Centrality feature exhibited strong performance with a Sensitivity, Specificity, and Precision of 0.8374, 0.8423, and 0.8415 respectively. The model's Accuracy and F1-score stood at 0.8399 and 0.8395, showcasing balanced categorization capabilities. Additionally, the MCC value of 0.6798 denotes reliable prediction of both classes. However, the standout feature at the 1:1 class ratio was Eigenvector Centrality. During independent testing, this feature achieved the highest Accuracy, F1-score, and MCC values, which were 0.8522, 0.8648, and 0.7171 respectively, indicating the model's proficiency in discerning true positives and negatives. Similar trends were observed during 5-fold cross-validation. Degree Centrality showed an Accuracy of 0.8529, an F1-score of 0.8568, and an MCC of 0.7074. Meanwhile, Eigenvector Centrality continued to excel with an Accuracy of 0.8388, an F1-score of 0.8493, and an MCC of 0.6844.

On the other hand, the model's performance with a more imbalanced class ratio of 1:10 is elucidated in Table 1. With this ratio, the Eigenvector Centrality and Hub Score features emerged as significant, recording high accuracy levels in both independent testing and 5-fold cross-validation. These results suggest that the model can uphold its performance even with a heavily skewed class distribution.

3.3.2. Comprehensive analysis of RNN model performance across features and class ratios

The RNN model's performance was thoroughly assessed across two disparate class ratios of 1:1 and 1:10, with each protein feature being evaluated individually (Fig. 2). This extensive appraisal facilitated a clear distinction of the model's strengths and areas requiring improvement. At the 1:1 class ratio, several noteworthy findings were observed. The feature that stood out the most was Eigenvector Centrality, which achieved the highest metrics in the majority of categories. These included an Accuracy of 0.8694 during independent testing, a similarly impressive F1-score of 0.8716, and an MCC of 0.7393, the highest among all features. The Degree Centrality, although not leading, displayed considerable efficacy, particularly with its Sensitivity of 0.8088 and an MCC of 0.6887. However, Closeness Centrality demonstrated difficulties in this scenario, especially indicated by a negative MCC during independent testing, signifying a possible shortcoming of the RNN model when handling this feature.

The analysis at the class ratio of 1:10 provided further intriguing insights. Eigenvector Centrality maintained its robust performance, securing an Accuracy of 0.9482 and an F1-score of 0.9720, marking its consistent influence on the RNN model's effectiveness. Degree Centrality not only exhibited exceptional Sensitivity and Precision but also contributed to an overall Accuracy of 0.9268. The Hub Score also performed consistently, exhibiting an Accuracy of 0.9477, reinforcing the model's overall effectiveness at the 1:10 ratio. On the other hand, the

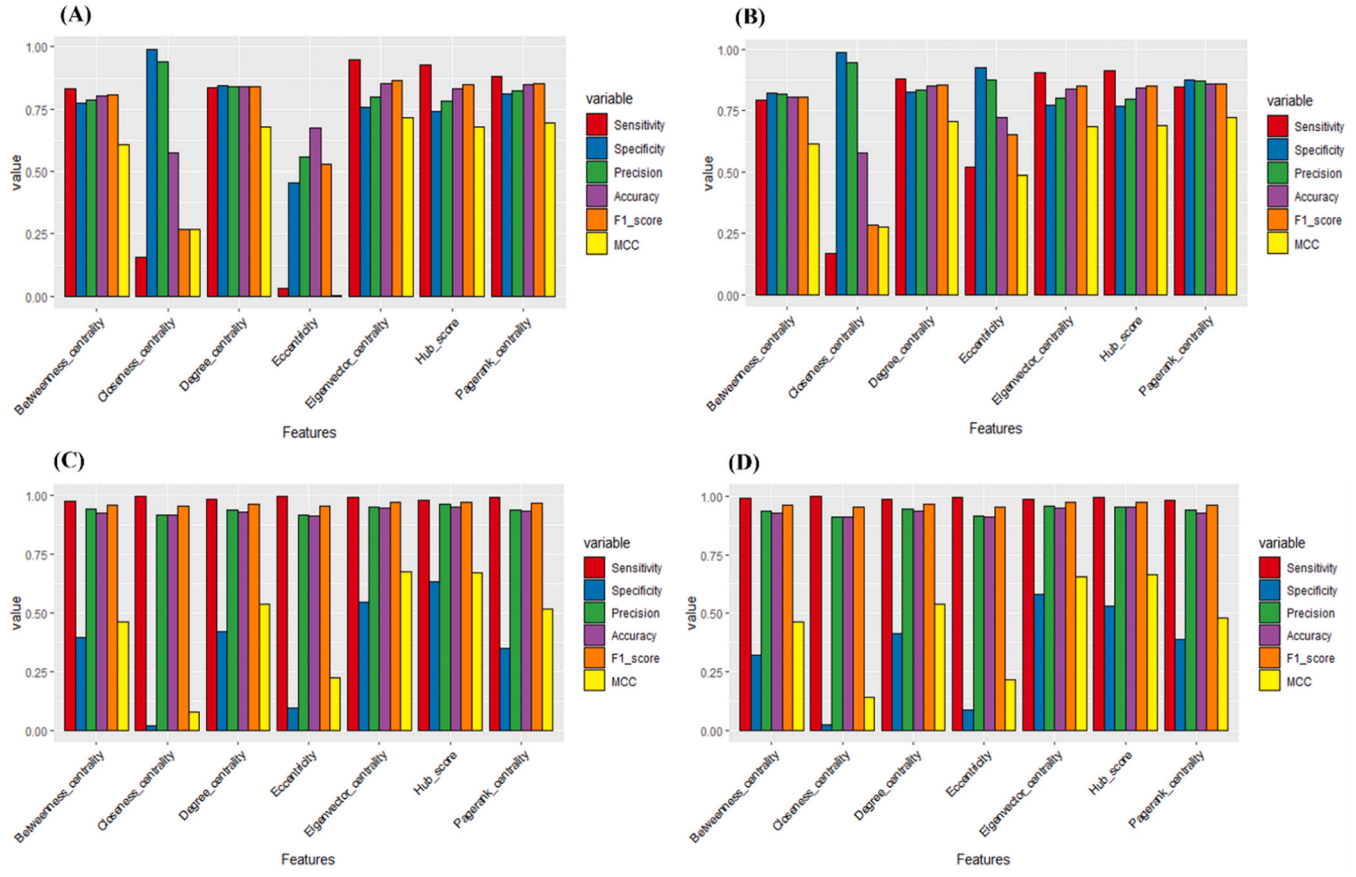

Fig. 1. Performance evaluation results of CNN model at class ratios 1:1: (**A**) Independent testing (**B**) 5-Fold cross validation. Performance evaluation results at different class ratios 1:10: (**C**) Independent testing (**D**) 5-Fold cross validation.

Table 1
1D-CNN model performance for each protein feature at class ratio of 1:1 and 1:10. Performance values of independent testing and 5-fold cross validation are shown within a same cell as follow: independent testing / 5-fold cross validation.

Network features Class ratio 1:1	Sensitivity	Specificity	Precision	Accuracy	F1-score	MCC
Degree_centrality	0.8374 / 0.8787	0.8423 / 0.8272	0.8415 / 0.8366	0.8399 / 0.8529	0.8395 / 0.8568	0.6798 / 0.7074
Betweenness_ centrality	0.8325 / 0.7928	0.7733 / 0.8222	0.7860 / 0.8181	0.8029 / 0.8075	0.8086 / 0.8046	0.6069 / 0.6163
Closeness_centrality	0.1576 / 0.1678	0.9901 / 0.9889	0.9411 / 0.9479	0.5738 / 0.5784	0.2700 / 0.2841	0.2667 / 0.2767
Eigenvector_centrality	0.9458 / 0.9056	0.7586 / 0.7721	0.7966 / 0.8001	0.8522 / 0.8388	0.8648 / 0.8493	0.7171 / 0.6844
Eccentricity	0.0314 / 0.5184	0.4546 / 0.9265	0.5571 / 0.8768	0.6723 / 0.7224	0.53 / 0.6510	0.005 /0.4878
Hub_score	0.9261 / 0.9130	0.7389 / 0.7696	0.7800 / 0.7994	0.8325 / 0.8413	0.8468 / 0.8520	0.6769 / 0.6905
Pagerank_centrality	0.8817 / 0.8468	0.8128 / 0.8750	0.8248 / 0.8716	0.8472 / 0.8609	0.8523 / 0.8581	0.6962 / 0.7237
Class ratio 1:10						
Degree_centrality	0.9856 / 0.9885	0.4227 / 0.4112	0.9400 / 0.9445	0.9303 / 0.9367	0.9623 / 0.9660	0.5355 / 0.5398
Betweenness_ centrality	0.9774 / 0.9899	0.3939 / 0.3192	0.9433 / 0.9359	0.9259 / 0.9291	0.9600 / 0.9621	0.4610 / 0.4644
Closeness_centrality	0.9977 / 1	0.0215 / 0.0234	0.9184 / 0.9107	0.9161 / 0.9109	0.9561 / 0.9532	0.0769 / 0.1423
Eigenvector_centrality	0.9925 / 0.9856	0.5475 / 0.5794	0.9524 / 0.9588	0.9486 / 0.9484	0.9721 / 0.9720	0.6744 / 0.6561
Eccentricity	0.9950 / 0.9960	0.0945 / 0.0854	0.9177 / 0.9155	0.914 / 0.9129	0.9548 / 0.9540	0.2265 / 0.2178
Hub_score	0.9808 /0.9931	0.6323 / 0.53178	0.9638 / 0.9548	0.9491 / 0.9510	0.9722 / 0.9735	0.6698 / 0.6652
Pagerank_centrality	0.9931 / 0.9823	0.35 / 0.3871	0.9397 / 0.9410	0.9357 / 0.9279	0.9656 / 0.9612	0.5150 /0.4814

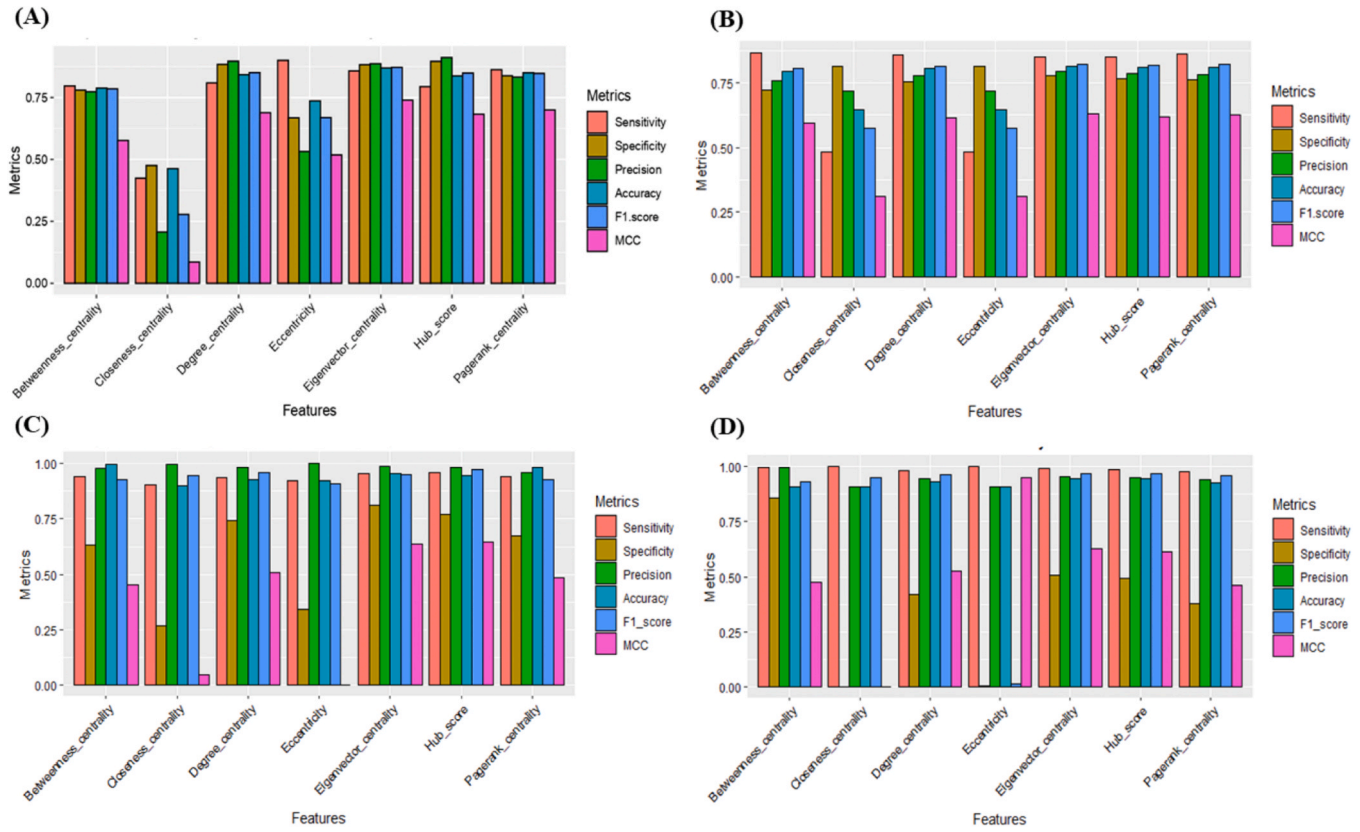

Fig. 2. Performance evaluation results of RNN model at class ratios 1:1: (**A**) Independent testing (**B**) 5-Fold cross validation. Performance evaluation results at different class ratios 1:10: (**C**) Independent testing (**D**) 5-Fold cross validation.

Closeness Centrality again presented unique challenges. It demonstrated a high Sensitivity of 0.9029 during independent testing but a strikingly low MCC of 0 in the 5-fold cross-validation. This anomaly exposes potential weaknesses in the model when operating under this class ratio and with this feature.Table 2.

3.3.3. Performance evaluation of MLP model for protein classification using different network features and class ratios

For the class ratio of 1:1, the model demonstrated varying degrees of effectiveness depending on the protein network feature used (Fig. 3). The Eigenvector Centrality feature outshone the others, achieving an impressive accuracy rate of 0.9334 in independent testing and 0.9487 in 5-fold cross-validation. Its sensitivity and specificity scores were also quite high (0.9605 / 0.9578 and 0.9064 / 0.9126 respectively), indicating its ability to correctly identify positive classes and correctly reject the negative ones. Moreover, its high F1-Score (0.9352 / 0.9209) suggests a balanced precision and recall, and the robust MCC (0.8682 / 0.8894) demonstrates the model's quality in terms of binary classification. The Hub Score feature also yielded substantial results, with an accuracy of 0.9261 / 0.8745 and a sensitivity of 0.9310 / 0.9234, reinforcing the reliability of these two features. Conversely, the Closeness Centrality feature performed less favorably, displaying an accuracy of 0.5689 / 0.5246 and sensitivity of 0.9605 / 0.9213, hinting at its weaker ability to accurately predict protein classes.

Table 2
RNN model performance for each protein feature at class ratio of 1:1 and 1:10. Performance values of independent testing and 5-fold cross validation are shown within a same cell as follow: independent testing / 5-fold cross validation.

Network features Class ratio 1:1	Sensitivity	Specificity	Precision	Accuracy	F1-score	MCC
Degree_centrality	0.8088/ 0.8572	0.8839 / 0.7519	0.8965 / 0.7761	0.8423/ 0.8045	0.8504/ 0.8143	0.6887/ 0.6130
Betweenness_ centrality	0.7969/ 0.8640	0.7799/ 0.7220	0.7733 / 0.7568	0.7881/ 0.7930	0.785 / 0.8067	0.5766/ 0.5924
Closeness_centrality	0.4242 / 0.4808	0.4755 / 0.8125	0.2068 / 0.7189	0.4630 / 0.6467	0.2781 / 0.5758	-0.0860 / 0.3109
Eigenvector_centrality	0.8571 / 0.8503	0.8826 / 0.7789	0.8866 / 0.7938	0.8694 / 0.8146	0.8716 / 0.8208	0.7393 / 0.6313
Eccentricity	0.9 / 0.4808	0.6678 / 0.8125	0.5320 / 0.7189	0.7364 / 0.6467	0.6687 / 0.5758	0.5182 / 0.3109
Hub_score	0.7939 / 0.8505	0.8959 / 0.7656	0.9113 / 0.7841	0.8374 / 0.8081	0.8486 / 0.8159	0.6823 / 0.6186
Pagerank_centrality	0.8622 / 0.8608	0.8380 / 0.7610	0.8325 / 0.7832	0.8497 / 0.8109	0.8471 / 0.8200	0.6999 / 0.6254
Class ratio 1:10						
Degree_centrality	0.9369 / 0.9850	0.7413 / 0.4175	0.9851 / 0.9442	0.9268 / 0.9334	0.9604 / 0.9641	0.5050 / 0.5268
Betweenness_ centrality	0.9425 / 0.9950	0.6302 / 0.8571	0.9784 / 0.9950	0.9259 / 0.9308	0.9601 / 0.9631	0.4535 / 0.4769
Closeness_centrality	0.9029 / 1	0.2666 / 0	0.9945 / 0.9090	0.8988 / 0.9090	0.9465 / 0.9523	0.0464 / 0
Eigenvector_centrality	0.9568 / 0.9900	0.8106 / 0.5095	0.9877 / 0.9528	0.9482 / 0.9464	0.9720 / 0.9710	0.6366 / 0.6288
Eccentricity	0.9205 / 0.9996	0.34 / 0.0032	1 / 0.9093	0.9205 / 0.9090	0.0013 / 0.0151	0.9205 / 0.9523
Hub_score	0.9607 / 0.9895	0.7697 / 0.4945	0.9828 / 0.9514	0.9477 / 0.9445	0.9716 / 0.9701	0.6457 / 0.6136
Pagerank_centrality	0.9429 / 0.9801	0.675 / 0.3782	0.9615 / 0.9403	0.9808 / 0.9253	0.9286 / 0.9597	0.4857 / 0.4614

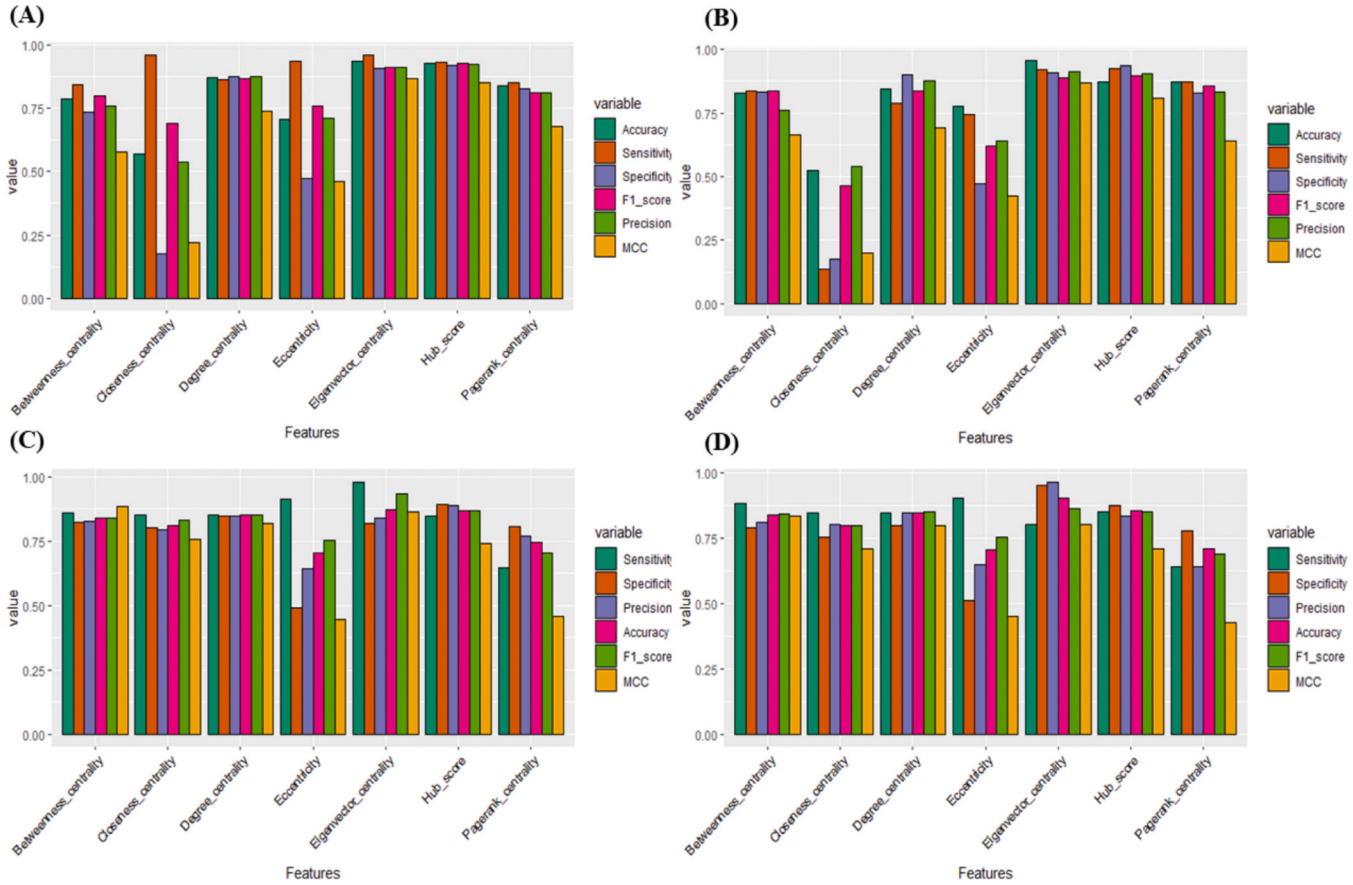

Fig. 3. Performance evaluation results of MLP model at class ratios 1:1: (**A**) Independent testing (**B**) 5-Fold cross validation. Performance evaluation results at different class ratios 1:10: (**C**) Independent testing (**D**) 5-Fold cross validation.

When the class ratio was adjusted to 1:10, the Eigenvector Centrality continued to lead in performance, particularly in sensitivity (0.9794 / 0.9651) and precision (0.9946 / 0.950739). This high sensitivity means the model could correctly identify a high proportion of actual positives, while the precision indicates that out of the classes predicted positive, most were accurate. The Degree Centrality also showed steady performance across multiple metrics like sensitivity (0.8522/0.8468), specificity (0.8472/0.8508), and accuracy (0.8497/0.8488). However, for this class ratio, Pagerank Centrality emerged as the underperforming feature with lower sensitivity (0.6453 / 0.6427) and accuracy scores

(0.7266 / 0.7105), implying a lower correct prediction rate and a higher false-negative rate. In essence, these results elucidate the considerable impact of different protein network features on the MLP model's performance, as well as the role of class ratios in the efficacy of these predictions. The superior performance of Eigenvector Centrality feature indicates its potential usefulness in practical applications, although the differing results among other features underscore the importance of careful feature selection in building effective protein classification models.Table 3.

3.4. Evaluation of models

The evaluation aimed to assess the performance of the MLP model across several network-based features. The features tested included Degree_centrality, Betweenness_centrality, Closeness_centrality, Eigenvector_centrality, Pagerank_centrality, Hub_score, and Eccentricity. These features are commonly used in network analysis to measure the importance of individual nodes within a network. In the context of protein-protein interaction networks, these features assists in the identification of key proteins that play critical roles in biological processes.

On independent testing, the Eigenvector_centrality feature stood out for its performance, achieving the highest accuracy (0.8736) and Matthews Correlation Coefficient (MCC, 0.8647) among all the features tested. The Eigenvector Centrality of a protein in a network measures the extent to which it is connected to other highly connected proteins. This feature is often used in biological network analysis to identify proteins that, although they may not have many connections themselves, are connected to key proteins in the network [39]. In this context, the application of Eigenvector Centrality features has garnered attention, particularly in the study of host-pathogen protein interactions. Studies such as Khorsand et al. [40] and Cui et al. [41] have demonstrated the efficacy of this approach in unraveling complex interaction dynamics, suggesting new pathways for therapeutic intervention. The model's performance remained high during cross-validation, achieving an accuracy of 0.905 and an MCC of 0.8020. The high cross-validation performance indicates that the model generalizes well to unseen data and is not overfitted to the training data, which adds confidence to the robustness of our model.

In order to apply these findings in a predictive manner, each protein pair was given a score based on their Eigenvector Centrality features. The pair's features were then normalized and input into the trained MLP model. The overall architecture of MLP model was also presented in Fig. 4. This produced a probability of interaction for each protein pair, which can be interpreted as a likelihood score for each pair's interaction. A threshold of 0.9 was set for these probability values to focus on the most significant interactions (**Supplementary File 1:** Table S1). Protein pairs with a score above this threshold were predicted to interact with high probability. This approach is common in binary classification tasks where it is more important to accurately predict one class (interacting pairs, in this case). It allows for better control over the trade-off between precision and recall, ensuring that predictions made are reliable and biologically significant.

Of significant importance, our study goes beyond the realm of predicting the likelihood of positive interactions. Instead, our model's implementation serves as an expansive exploration, unveiling hidden interactions that might not have been initially evident. This dynamic capability underscores the inherent value of our approach, which extends beyond the validation of established interactions to the prospect of uncovering novel, previously unexplored interactions. These newfound interactions, accompanied by calculated likelihoods, not only enrich our comprehension but also stimulate a more profound exploration of potential molecular dialogues within the intricate domain of host-pathogen interactions. In conclusion, through rigorous evaluation and thoughtful application, the MLP model trained on Eigenvector Centrality features proved to be a powerful tool for predicting host-pathogen interactions. This methodological approach could be useful for future research in this field, particularly for studies seeking to understand complex host-pathogen dynamics.

3.5. Developing an interactive R shiny application

Building upon the foundations of our approach, we have successfully translated our methodology into an interactive and user-friendly tool: R Shiny application named Deep-HPI-pred. This innovative platform empowers users to seamlessly engage with our predictive models and gain insights into host-pathogen interactions. By providing an intuitive interface and leveraging advanced deep learning models, Deep-HPI-pred offers an accessible means for researchers to explore, predict, and understand the intricate dynamics of PPIs between hosts and pathogens. This application acts as a bridge, converting intricate computational methodologies into a user-friendly tool that catalyzes the advancement of host-pathogen interaction studies. This application is distinctive in the sense that it allows users to either manually or automatically upload the training dataset, facilitating a more streamlined operation (Fig. 5).

Table 3
MLP model performance for each protein feature at class ratio of 1:1 and 1:10. Performance values of independent testing and 5-fold cross validation are shown within a same cell as follow: independent testing / 5-fold cross validation.

Network Features Class ratio 1:1	Accuracy	Sensitivity	Specificity	F1_Score	Precision	MCC
Degree_centrality	0.8694 / 0.8448	0.8620 / 0.7881	0.8768 / 0.9014	0.8684 / 0.8355	0.875 / 0.8889	0.7389 / 0.6941
Betweenness_centrality	0.7881 / 0.8325	0.8423 / 0.8275	0.7339 / 0.8374	0.7990 / 0.8316	0.76 / 0.8358	0.5797 / 0.6650
Eigenvector_centrality	0.9334/ 0.9487	0.9605 / 0.9578	0.9064 / 0.9126	0.9352 / 0.9209	0.9112 / 0.9334	0.8682 / 0.8894
Hub_score	0.9261 / 0.8745	0.9310 / 0.9234	0.9211 / 0.9382	0.9264 / 0.8978	0.9219 / 0.9051	0.8522 / 0.8089
Eccentricity	0.7044 / 0.6427	0.9359 / 0.7782	0.4729 / 0.7441	0.76 / 0.7105	0.6397 / 0.6894	0.4613 / 0.4253
Pagerank-centrality	0.8399 / 0.8570	0.8522 / 0.8721	0.8275 / 0.8091	0.84184 / 0.8101	0.8317 / 0.8566	0.6800 / 0.6421
Closeness-centrality	0.5689 / 0.5246	0.9605 / 0.9213	0.1773 / 0.1345	0.6902 / 0.6189	0.5386 / 0.4621	0.2218 / 0.2
Class ratio 1:10						
Degree_centrality	0.8522/ 0.8468	0.8472/ 0.8508	0.8480 / 0.8502	0.8497/ 0.8488	0.8501/ 0.8483	0.8195/ 0.7980
Betweenness_ centrality	0.8592 / 0.8832	0.8226 / 0.7909	0.8277 / 0.8093	0.8374 / 0.8370	0.8398 / 0.8443	0.8851 / 0.8345
Closeness_centrality	0.8522 / 0.8660	0.8029 / 0.7944	0.8122 / 0.8325	0.8275 / 0.8015	0.8317 / 0.80	0.7559 / 0.711
Eigenvector_centrality	0.9794 / 0.9651	0.8195 / 0.8411	0.9946 / 0.950739	0.8736 / 0.905	0.9352 / 0.8642	0.8647 / 0.8020
Eccentricity	0.9113 / 0.9018	0.4926 / 0.5122	0.6423 /0.6494	0.7019 / 0.7070	0.7535 / 0.7549	0.4448 / 0.4498
Hub_score	0.8472 / 0.8351	0.8916 / 0.8734	0.8865 / 0.8698	0.8694 / 0.8542	0.8664 / 0.8516	0.7396 / 0.7099
Pagerank_centrality	0.6453 / 0.6427	0.8078 / 0.7782	0.7705 / 0.7441	0.7266 / 0.7105	0.7024 / 0.6894	0.4593 / 0.4253

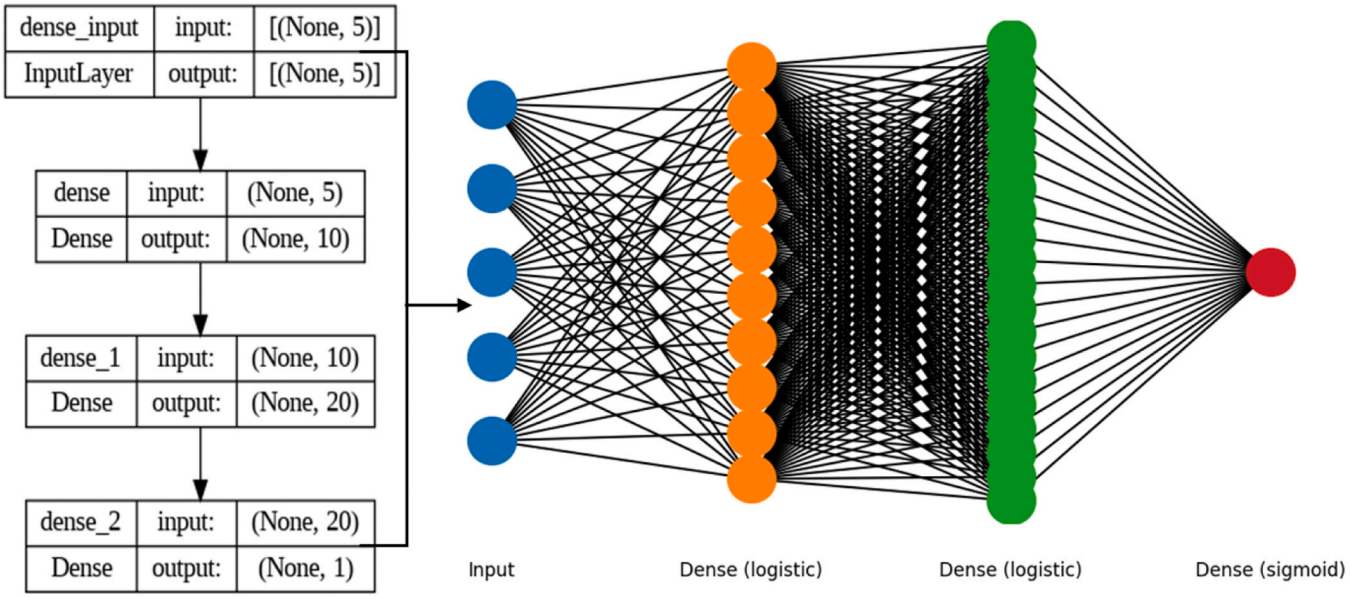

Fig. 4. MLP architecture used to train the host-pathogen PPIs models.

Furthermore, it is engineered to predict PPIs leveraging the topological features of the proteins, which is a significant step forward given that traditional prediction methods often lack this level of sophistication.

What makes Deep-HPI-pred stand out is its ability to present intricate host-pathogen protein interactions and the corresponding probability scores in a simplified, user-friendly interface. This aspect bridges the gap between complex data analysis and its interpretability, thereby promoting a broader comprehension and accessibility of the information. Notably, Deep-HPI-pred is the first-ever application of its kind dedicated to the network-based classification and prediction of host-pathogen protein interactions. This development marks a significant milestone in bioinformatics and opens new opportunities for researchers to delve deeper into the dynamics of biological interactions. The application of deep learning models combined with topological analysis presents a novel approach to study and predict PPIs. This not only broadens the scope of current research methodologies but also paves the way for future advancements in this rapidly evolving field of study. By fostering a deeper understanding of pathogenesis and disease progression, Deep-HPI-pred contributes to the advancement of the scientific community's collective knowledge. It demonstrates the potential of integrating machine learning with biological data analysis, thereby setting a precedent for future research in this direction.

3.5.1. Interface and data input in the deep-HPI-pred application

Upon initiation of the Deep-HPI-pred application, the user is presented with a straightforward interface, deliberately designed for easy navigation and usability (Fig. 5 **(A)**). A feature of the application allows users to upload their specific data files in CSV format through a conveniently located sidebar. In contrast, there is also an option to use pre-existing demo data. This flexibility caters to the distinct requirements of individual research, providing a tailored experience for interaction predictions. Moreover, the incorporation of demo data serves as an effective tool for users to familiarize themselves with the application's functionality swiftly.

3.5.2. Implementing training and prediction with deep learning models

After completion of the data upload process, the application prompts the "Train and Predict" feature (Fig. 5**(B, C)**). This launches the deep learning model previously trained on an extensive dataset of host-pathogen protein interactions. The model performs a comprehensive analysis of the uploaded data to accurately predict potential interactions. This application feature transforms the data input into valuable insights, opening avenues for the elucidation of complex host-pathogen interaction dynamics.

3.5.3. Visualizing and interpreting predicted interactions

The next phase of interaction with the application is through the "Results" tab, which is designed for data interpretation and visualization (Fig. 5**(D)**). The tab displays predicted interaction data and a network visualization of these interactions. The visual depiction facilitates better comprehension of the interaction networks and is particularly valuable in understanding complex interaction dynamics.

3.5.4. Conducting GO enrichment analysis for enhanced biological comprehension

The final component of the Deep-HPI-pred application, the "GO Analysis" tab, provides an effective tool for conducting GO enrichment analysis (Fig. 5**(E, F)**). Unlike other tools, which often lack the provision of GO terms and related analyses, Deep-HPI-pred fills this critical gap. The resulting Network Graph visualizes the outcomes of the GO analysis, offering an intuitive representation of the enriched terms. By collating and categorizing gene products based on shared GO annotations, the application presents an understanding of the biological implications of the predicted interactions. Moreover, to facilitate further in-depth investigations, researchers can conveniently download both the GO table and the corresponding plot, enhancing the tool's utility for advanced research endeavors.

Thus, the development and implementation of the Deep-HPI-pred R Shiny application constitutes a significant milestone in host-pathogen interaction studies. Its ability to integrate deep learning models, employ diverse protein interaction data, and provide interactive and comprehensible output in the form of network visualizations and GO enrichment analysis, creates a uniquely accessible platform for researchers across multiple disciplines. The application, by transforming raw data into insightful knowledge, effectively supports the intricate process of deciphering the complexities inherent in host-pathogen interactions.

4. Benchmarking

In this study, the versatility and effectiveness of the Deep-HPI-pred applet were demonstrated in accurately predicting PPIs across a range

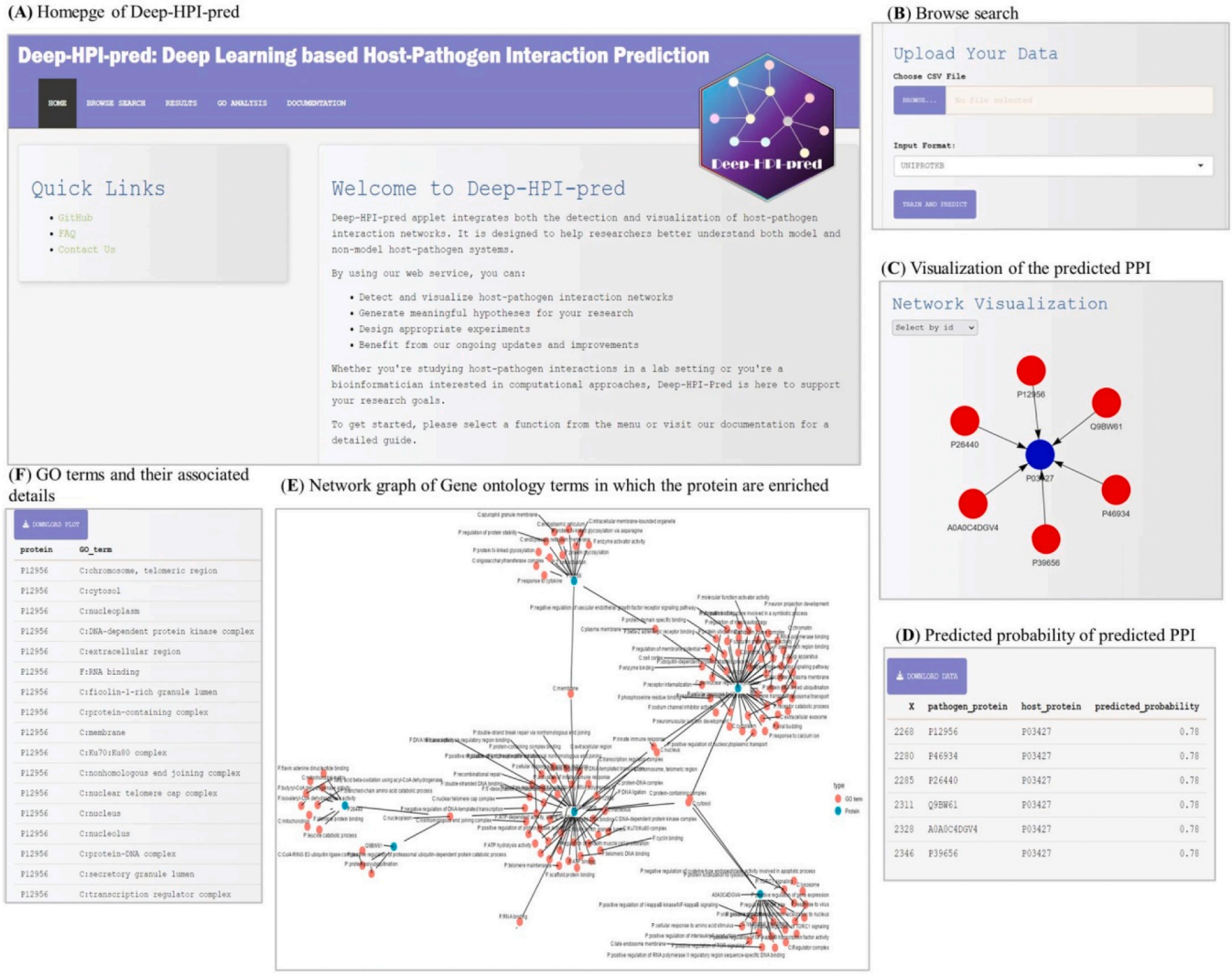

Fig. 5. (**A**) User interface of Deep-HPI-pred (**B**) Deep-HPI-pred Interface showcasing data upload functionality (**C**) Result's tab of Deep-HPI-pred. The blue color nodes within network indicate the Host-proteins while red color nodes represented the pathogen proteins (**D**) The probability of interactions among host and proteins are presented in form of table (**E**) The 'GO Analysis' Tab in Deep-HPI-pred demonstrating the visualization of a network graph and (**F**) the corresponding Gene Ontology (GO) Enrichment Analysis table.

of biological contexts. PPIs were analyzed across three distinct host-pathogen pairs, each representing a unique interplay between different types of organisms - bacteria, plants, and animals. The dataset of experimentally verified HPI interactions was curated from existing literature, and then classified into distinct categories: plant-bacteria, plant-fungi, animal-bacteria [14], and human-virus.

4.1. *Mus musculus and Burkholderia mallei*

Burkholderia mallei is known to cause glanders, a rare infectious disease that affects horses, mules, and donkeys. It can also infect humans and is considered a potential bioterrorism agent. In this study, our Deep-HPI-pred model achieved an accuracy of 96.6%, sensitivity of 94.9%, specificity of 87.3%, F1 score of 88.7%, precision of 81.4%, MCC of 90.9%, and AUC of 95.1% in predicting PPIs between *M. musculus* (mouse) and *B. mallei*. This highlights the model's potential in assisting researchers in understanding the intricate interplay between these two organisms and could provide valuable insights into the mechanism of glanders infection in various species.

4.2. *Arabidopsis thaliana and Golovinomyces orontii*

Golovinomyces orontii is known to cause powdery mildew in *A. thaliana*, a model organism in plant biology. Accurate prediction of PPIs is crucial for deciphering the molecular basis of plant-pathogen interactions and developing strategies for disease resistance. Our Deep-HPI-pred model demonstrated high accuracy in predicting these interactions, with an accuracy of 98.4%, sensitivity of 89.3%, specificity of 93.2%, F1 score of 95%, precision of 92.4%, MCC of 93.3%, and AUC of 92.8%.

4.3. *Arabidopsis thaliana and Pseudomonas syringae*

P. syringae is a well-known pathogenic bacterium that infects a wide range of plant species, causing bacterial speck disease in *A. thaliana*. Understanding PPIs between these two species is vital for developing new methods of disease control. Our Deep-HPI-pred model achieved an accuracy of 97.9%, sensitivity of 95%, specificity of 93.7%, F1 score of 91.8%, precision of 92.4%, MCC of 94.8%, and AUC of 98.2% in predicting the PPIs between *A. thaliana* and *P. syringae*.

4.4. Homo sapiens and SARS-Cov-2

Additionally, we also evaluated our model on the host-pathogen interactions between *H. sapiens* (human) and SARS-CoV-2 (virus), a critically important relationship in light of the recent global pandemic. Given the recent global impact of the COVID-19 pandemic, understanding the interactions between human proteins and SARS-CoV-2 viral proteins is of paramount importance. In our analysis of the interactions between *H. sapiens* and SARS-CoV-2, the Deep-HPI-pred model demonstrated its proficiency in capturing the complex relationships characteristic of host-pathogen interactions. Given the recent global impact of the COVID-19 pandemic, understanding the interactions between human proteins and SARS-CoV-2 viral proteins is of paramount importance. Our Deep-HPI-pred model showed excellent performance in predicting the interactions between human proteins and SARS-CoV-2 viral proteins, with an accuracy of 94.3%, sensitivity of 97.5%, specificity of 98.7%, F1 score of 96.5%, precision of 96.2%, MCC of 95.3%, and AUC of 97.8%. These results demonstrate the strong predictive power of our Deep-HPI-pred model in deciphering the intricate network of interactions that take place between human proteins and SARS-CoV-2 viral proteins.

The finding revealed potential interactions between SARS-CoV-2 proteins and human host factors, which may play a pivotal role in the viral life cycle, including entry, replication, and evasion of host immune defenses. Notably, the model predicts that the SARS-CoV-2 Envelope protein interact with human proteins involved in the endocytic pathway, a route well-documented for coronavirus entry into host cells [42]. Such interactions could potentially facilitate viral entry by altering endosomal trafficking. Additionally, interactions have been predicted between the SARS-CoV-2 Membrane glycoprotein and proteins located in the ER-to-Golgi intermediate compartment (ERGIC), which is instrumental in viral assembly and trafficking [43]. These predictions suggest a mechanism by which SARS-CoV-2 could hijack host cellular machinery to bolster its replication process. Further results also suggest potential interactions between the ORF1a polyprotein of SARS-CoV-2 and components of the human innate immune system, particularly proteins involved in interferon signaling. This aligns with findings such as those from McClainet al. [44], which indicate the virus's ability to modulate interferon-driven responses, a strategy that likely facilitates viral replication and contributes to pathogenesis. An extensive literature review was undertaken to provide context for each predicted interaction. While direct evidence of these specific interactions has not been previously documented, the literature corroborates the biological plausibility of the proposed mechanisms. For example, Hekman et al. [45] demonstrated that the SARS-CoV-2 Nucleoprotein's interaction with host RNA processing bodies suggests a potential for viral modulation of host mRNA processing, an avenue supported by recent findings. Thus, our Deep-HPI-pred model has demonstrated a high level of accuracy and reliability in predicting host-pathogen interactions across diverse biological contexts, underscoring its potential as a valuable tool for researchers studying infectious diseases and host-pathogen interactions.

Additionally, in the comparative analysis of Deep-HPI-pred with existing servers such as Pred-HPI and DeepHPI, several key differences emerge that highlight the enhanced capabilities of Deep-HPI-pred. Specifically, Deep-HPI-pred has predicted a total of 9673 interactions between SARS-CoV-2 virus and human proteins (**Supplementary File 3:** Table S1**)**. In contrast, PredHPI, focuses primarily on sequence-based methods for detecting host-pathogen interactions, has predicted 6654 interactions (**Supplementary File 3:** Table S2**)**, and notably, these predictions do not include probability information for each interaction. The higher number of interactions predicted by Deep-HPI-pred suggests its increased sensitivity in detecting potential host-pathogen interactions, which is crucial for comprehensive understanding and exploration of SARS-CoV-2 pathogenesis. Moreover, Deep-HPI-pred provides probability scores for each predicted interaction, offering a quantifiable measure of confidence in the predictions. This feature is particularly valuable for researchers, as it allows prioritization of interactions based on their likelihood, facilitating targeted experimental validation. Such probability information is absent in the predictions made by Pred-HPI and DeepHPI, which limits the ability to assess the confidence level in each predicted interaction. Additionally, Deep-HPI-pred enhances the utility of its predictions by providing GO information. This inclusion allows for immediate biological interpretation of the interactions, offering insights into potential biological processes and molecular functions involved, thereby enriching the understanding of the interaction landscape. In comparison, Deep-HPI, another existing tool in this domain, does not provide specific interaction results in a format that is directly comparable to Deep-HPI-pred and Pred-HPI. Therefore, the comparison focuses on the methodological approach and the theoretical framework of these models.

In the similar vein, GreeningDB, another host-pathogen interaction database, offers a more specialized scope, focusing specifically on citrus greening disease (Huanglongbing or HLB). It compiles data primarily relevant to HLB, including genomic, transcriptomic, and proteomic information. While invaluable for HLB research, GreeningDB's utility is confined to this particular disease. Deep-HPI-pred, however, extends its applicability beyond a singular disease context, enabling broader investigations of HPIs across multiple biological systems. This key distinction underscores Deep-HPI-pred's potential as a versatile platform for a more generalized understanding of host-pathogen dynamics, applicable to a diverse range of infectious diseases.

Another recent study by Yang et al. [46] employed a transfer learning approach using multi-scale convolutional neural layers to predict human-virus PPIs. While this approach effectively captures the complex features of protein sequences, Deep-HPI-pred extends beyond sequence-based predictions. It integrates topological features and GO information, offering a more holistic view of the interaction landscape. Additionally, Deep-HPI-pred's application is not limited to human-virus interactions but encompasses a broader range of host-pathogen systems, demonstrating its versatile and comprehensive predictive capabilities. Similarly, in comparison with DeepViral [47], a deep learning-based method for predicting novel virus-host interactions from protein sequences and infectious disease phenotypes, Deep-HPI-pred showcases distinct advantages. DeepViral focuses on novel virus-host interaction predictions using protein sequences and disease phenotypes, leveraging a deep learning approach. While DeepViral's integration of infectious disease phenotypes offers a unique perspective, Deep-HPI-pred's methodology is distinguished by its utilization of MLP models based on a comprehensive evaluation of topological features and neural network architectures. This not only enhances the prediction accuracy but also provides a more detailed understanding of the underlying protein interaction mechanisms. Furthermore, Deep-HPI-pred's consistent performance across diverse host-pathogen systems, including plant-pathogens, human-viruses, and animal-bacteria, as evidenced by its high accuracy rates, illustrates its robustness and adaptability in various biological contexts.

To sum up, Deep-HPI-pred stands out in its ability to not only predict a larger number of host-pathogen interactions but also to provide critical additional information like probability scores and GO annotations. These features significantly contribute to its utility as a research tool, offering a more nuanced and informed approach to exploring host-pathogen interactions compared to tools like Pred-HPI, Deep-HPI, and DeepViral. The advanced algorithms utilized by Deep-HPI-pred enable the analysis of extensive interaction datasets, highlighting proteins that are central and often critical in these processes. This analytical capability is instrumental in deepening our understanding of the molecular mechanisms of viral infections, as supported by studies like those conducted by Barman et al. [48], which employed state-of-art techniques to identify key viral interaction proteins. Furthermore, the fusion of biological network analysis with deep learning heralds a transformative era in clinical and personalized medicine, particularly in managing viral

diseases. This integration enables the unraveling of complex host-pathogen interactions at a molecular level, paving the way for targeted therapeutic strategies and more individualized treatment approaches in combating viral infections.

5. Conclusion

Traditional experimental techniques for host-pathogen interaction prediction, though effective, have proven to be labor-intensive, expensive, and time-consuming. To address this challenge, we have introduced Deep-HPI-pred, an R/Shiny application that provides a computational approach for predicting hitherto unmapped interactions between host and pathogen proteins. By harnessing the power of network-driven feature learning, Deep-HPI-pred, as demonstrated through our case study using citrus and *C*Las bacteria training sets, offers a promising alternative for accelerating the discovery of PPIs. In our research, we employed a comprehensive evaluation of various neural network architectures and topological features, the results of which led us to adopt the MLP models for HPI prediction. Notably, the MLP model using the Eigenvector Centrality topological feature exhibited exemplary performance, achieving an overall MCC value exceeding 0.80 when tested on independent validation datasets. Beyond its capacity for interaction prediction, Deep-HPI-pred further enriches our understanding of the dynamics within host-pathogen interactions by providing GO term information for each protein. This added layer of information presents an insightful view of the system and enhances our comprehension of the overall biological processes at play. Furthermore, in the benchmarking studies conducted, the Deep-HPI-pred model demonstrated its robustness and reliability across various host-pathogen systems. The model exhibited a remarkable performance in predicting interactions between different host-pathogen pairs, including plant-virus, human-virus, plant-bacteria, and animal-bacteria. Specifically, the model achieved an accuracy of 98.4% and 97.9% for plant-pathogen interactions, 94.3% for human-virus interactions, and 96.6% for animal-bacteria interactions. These results not only validate the efficacy of our model but also highlight its potential as a versatile and comprehensive tool for understanding the complex dynamics of host-pathogen interactions across different biological systems. While MLPs have demonstrated robust performance in this study, their structure is not inherently optimized for contrastive learning, which is increasingly recognized for its efficacy in unsupervised and semi-supervised scenarios. This limitation suggests a potential area for future improvement of the model, where integrating contrastive learning techniques could expand its capabilities to handle and learn from the vast amounts of unlabeled data in biological research. Such enhancements would not only address a key limitation but also significantly enrich the model's utility in understanding and predicting complex host-pathogen interactions. In conclusion, the introduction of Deep-HPI-pred represents a significant stride in the field of bioinformatics. By integrating detection and visualization of interaction networks into a single user-friendly platform, it equips researchers with a powerful tool for understanding both model and non-model host-pathogen systems. This advancement is expected to aid in the generation of hypotheses, the design of appropriate experiments, and ultimately, in the development of disease control and prevention strategies.

Availability of data and materials

Project name: Deep-HPI-pred. Project home page: https://cbi.gxu.edu.cn/shiny-apps/Deep-HPI-pred/. Programming language: R and HTML. The resources and data used during the current study are available in the GitHub repository, https://github.com/tahirulqamar/Deep-HPI-pred.

Funding

This work was supported by the Starting Research Grant for High-level Talents from Guangxi University and Postdoctoral research platform grant of Guangxi University.

CRediT authorship contribution statement

Muhammad Tahir ul Qamar and Fatima Noor: Data Curation, Methodology, Software, Formal Analysis, Investigation, Visualization, Writing - Original Draft. **Yi-Xiong Guo and Xi-Tong Zhu**: Software, Validation, Writing - Review & Editing. **Ling-Ling Chen**: Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.010.

References

[1] Kuo Z-Y, Chuang Y-J, Chao C-C, Liu F-C, Lan C-Y, Chen B-S. Identification of infection-and defense-related genes via a dynamic host-pathogen interaction network using a Candida albicans-zebrafish infection model. J innate Immun 2013;5(2):137–52.
[2] Garbutt CC, Bangalore PV, Kannar P, Mukhtar M. Getting to the edge: protein dynamical networks as a new frontier in plant–microbe interactions. Front Plant Sci 2014;5:312.
[3] J.J. Da Graça, L.Korsten. Citrus huanglongbing: Review, present status and future strategies. In: Naqvi SAMH, editor. Diseases of Fruits and Vegetables Volume I: Diagnosis and Management. The Netherlands: Kluwer Academic; (2004). pp. 229–45.
[4] Andrade M, Li J, Wang N. Candidatus Liberibacter asiaticus: virulence traits and control strategies. Trop Plant Pathol 2020;45:285–97.
[5] Pandey SS, Hendrich C, Andrade MO, Wang N. Candidatus Liberibacter: From movement, host responses, to symptom development of citrus Huanglongbing. Phytopathology® 2022;112(1):55–68.
[6] Hoddle MS, Hoddle CD, Morgan DJ, Milosavljević I. Successful Biological Control of Asian Citrus Psyllid, Diaphorina citri, in California. In: Van Driesche RG, Winston RL, Perring TM, Lopez VM, editors. Contributions of Classical Biological Control to the US Food Security. USDA FHAAST: Forestry, and Biodiversity. Washington; 2022. p. 127–45.
[7] Yuan M, Xin XF. Bacterial Infection and Hypersensitive Response Assays in Arabidopsis-Pseudomonas syringae Pathosystem. Bio Protoc 2021;11(24):e4268.
[8] Dyer MD, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, et al. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PloS One 2010;5(8):e12089.
[9] Lian X, Yang X, Yang S, Zhang Z. Current status and future perspectives of computational studies on human–virus protein–protein interactions. Brief Bioinforma 2021;22(5):bbab029.
[10] Westermann AJ, Cross-species JVogel. RNA-seq for deciphering host–microbe interactions. Nat Rev Genet 2021;22(6):361–78.
[11] Balotf S, Wilson R, Tegg RS, Nichols DS, Wilson CR. Shotgun proteomics as a powerful tool for the study of the proteomes of plants, their pathogens, and plant–pathogen interactions. Proteomes 2022;10(1):5.
[12] Jain A, Mittal S, Tripathi LP, Nussinov R, Ahmad S. Host-pathogen protein-nucleic acid interactions: A comprehensive review. Comput Struct Biotechnol J 2022.
[13] Kaundal R, Loaiza CD, Duhan N, Flann N. deepHPI: a comprehensive deep learning platform for accurate prediction and visualization of host–pathogen protein–protein interactions. Brief Bioinforma 2022;23(3):bbac125.
[14] Loaiza CD, Kaundal R. PredHPI: an integrated web server platform for the detection and visualization of host–pathogen interactions using sequence-based methods. Bioinformatics 2021;37(5):622–4.
[15] Loaiza CD, Duhan N, Kaundal R. GreeningDB: A Database of Host–Pathogen Protein–Protein Interactions and Annotation Features of the Bacteria Causing Huanglongbing HLB Disease. Int J Mol Sci 2021;22(19):10897.
[16] Singhal M, Resat H. A domain-based approach to predict protein-protein interactions. BMC Bioinform 2007;8:1–19.

[17] Chen H, Li F, Wang L, Jin Y, Chi C-H, Kurgan L, et al. Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions. Brief Bioinforma 2021;22(3):bbaa068.
[18] Scott MS, Probabilistic GJ Barton. prediction and ranking of human protein-protein interactions. BMC Bioinform 2007;8(1):1–21.
[19] Wang S, Wu R, Lu J, Jiang Y, Huang T, Cai YD. Protein-protein interaction networks as miners of biological discovery. Proteomics 2022;22(15-16):2100190.
[20] Csardi G, Nepusz T. The igraph software package for complex network research. Inter, Complex Syst 2006;1695(5):1–9.
[21] Pržulj N, Wigle DA, Jurisica I. Functional topology in a network of protein interactions. Bioinformatics 2004;20(3):340–8.
[22] Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, Mirzaie M, Jafari M. A systematic survey of centrality measures for protein-protein interaction networks. BMC Sys. Biol 2018;12(1):1–17.
[23] Eryilmaz H, Pax M, O'Neill AG, Vangel M, Diez I, Holt DJ, et al. Network hub centrality and working memory performance in schizophrenia. Schizophrenia 2022;8(1):76.
[24] A. Ali, V.R. Hulipalled, S. Patil, editors. Centrality measure analysis on protein interaction networks. 2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET); 2020: IEEE.
[25] V. Chaubey, M.S. Nair, G.N. Pillai, editors. Gene expression prediction using a deep 1D convolution neural network. 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 2019: IEEE.
[26] Patiyal S, Dhall A, Raghava GP. A deep learning-based method for the prediction of DNA interacting residues in a protein. Brief Bioinforma 2022;23(5):bbac322.
[27] Fang C, Moriwaki Y, Li C, Shimizu K. Prediction of antifungal peptides by deep learning with character embedding. IPSJ Trans Bioinforma 2019;12:21–9.
[28] Arnold TB. kerasR: R Interface to the Keras Deep Learning Library. J Open Source Softw,2 2017;(14):296.
[29] Yang X, Yang S, Lian X, Wuchty S, Zhang Z. Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. Bioinformatics 2021;37(24):4771–8.
[30] Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. 200805756 arXiv Prepr arXiv 2020. 200805756.
[31] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 201016061 arXiv Prepr arXiv 2020. 201016061.
[32] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16(5):412–24.
[33] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6(1):1–54.
[34] Broadley RW, Klenk J, Thies SB, Kenney LP, Granat MH. Methods for the real-world evaluation of fall detection technology: A scoping review. Sensors 2018;18(7):2060.
[35] Alves MA, Castro GZ, Oliveira BAS, Ferreira LA, Ramírez JA, Silva R, Guimarães FG. Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Comput Biol Med 2021;132:104335.
[36] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 2020;21(1):1–13.
[37] Kakkar B, goyal M, Johri P, Kumar Y. Artificial Intelligence-Based Approaches for Detection and Classification of Different Classes of Malaria Parasites Using Microscopic Images: A Systematic Review. Arch Comput Methods Eng 2023:1–20.
[38] Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, et al. A deep-learning framework for multi-level peptide–protein interaction prediction. Nat Commun 2021;12(1):5465.
[39] Negre CF, Morzan UN, Hendrickson HP, Pal R, Lisi GP, Loria JP, et al. Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci USA 2018;115(52). E12201-E8.
[40] Khorsand B, Savadi A, Naghibzadeh M. Comprehensive host-pathogen protein-protein interaction network analysis. BMC Bioinform 2020;21:1–22.
[41] Cui Y, Cai M, Stanley HE. Discovering disease-associated genes in weighted protein–protein interaction networks. Phys A: Stat Mech its Appl 2018;496:53–61.
[42] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. cell 2020;181(2):271–80. e8.
[43] Cortese M, Laketa V. Advanced microscopy technologies enable rapid response to SARS-CoV-2 pandemic. Cell Microbiol 2021;23(7):e13319.
[44] McClain MT, Constantine FJ, Henao R, Liu Y, Tsalik EL, Burke TW, et al. Dysregulated transcriptional responses to SARS-CoV-2 in the periphery. Nat Commun 2021;12(1):1079.
[45] Hekman RM, Hume AJ, Goel RK, Abo KM, Huang J, Blum BC, et al. Actionable cytopathogenic host responses of human alveolar type 2 cells to SARS-CoV-2. Mol Cell 2020;80(6):1104–22. e9.
[46] Yang X, Yang S, Lian X, Wuchty S, Zhang Z. Transfer learning via multi-scale convolutional neural layers for human-virus protein-protein interaction prediction. Bioinformatics 2021;37(24):4771–8.
[47] Liu-Wei W, Kafkas Ş, Chen J, Dimonaco NJ, Tegnér J, Hoehndorf R. DeepViral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. Bioinformatics 2021;37(17):2722–9.
[48] 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.