IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 21, 2020, accepted September 20, 2020, date of publication September 30, 2020, date of current version October 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027887

Identification of Pathway-Specific Protein Domain by Incorporating Hyperparameter Optimization Based on 2D Convolutional Neural Network

ALI GHUALM, XIUJUAN LEI, (Member, IEEE), YUCHEN ZHANG, (Graduate Student Member, IEEE), SHI CHENG, (Member, IEEE), AND MIN GUO

School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

Corresponding author: Xiujuan Lei (xjlei@snnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972451, Grant 61672334, and Grant 61902230; and in part by the Fundamental Research Funds for the Central Universities under Grant GK201901010.

ABSTRACT Pathway-specific protein domain (PSPD) are associated with specific pathways. Many protein domains are pervasive in various biological processes, whereas other domains are linked to specific pathways. Many human disease pathways, such as cancer pathways and signaling pathway-related diseases, have caused the loss of functional PSPD. Therefore, the creation of an accurate method to predict its roles is a critical step toward human disease and pathways. In this study, we proposed a deep learning model based on a two-dimensional neural network (2D-CNN-PSPD) with a pathway-specific protein domain association prediction. In terms of the purposes of a sub-pathway, its parent pathway and its super pathway are linked to the Uni-Pathway. We also proposed a dipeptide composition (DPC) model and a dipeptide deviation (DDE) model of feature extraction profiles as PSSM. Then, we predicted the proteins associated with the same sub-pathway or with the same organism. The DDE model and DPC model of the PSSM feature profile input was associated with our proposed 2D-CNN method. We deployed several parameters to optimize the model's output performance and used the hyperparameter optimization approach to find the best model for our dataset based on the 10-fold cross-validation results. Ultimately, we assessed the predictive performance of the current model by using independent datasets and cross-validation datasets. Therefore, we enhanced the efficiency of deep learning methods. PSPD is involved in any known pathway and then follow the association in different stages of the pathway hierarchy with other proteins. Our proposed method could identify 2D-CNN-PSPD with 0.83% sensitivity, 0.92% specificity, 87.27% accuracy, and 0.75% accuracy. We provided an important method for the analysis of PSPD proteins in the proposed research, and our achievements might promote computational biological research. We concluded our proposed model architecture in the future, the use of the latest features, and the multi-one structure to predict different types of molecules, such as DNA, RNA, and disease-pathway specific proteins associations.

INDEX TERMS Molecular structure prediction, deep learning, convolutional neural network, deep learning, evolutionary knowledge, multiple features, features extraction.

I. INTRODUCTION

The awareness on the spatial approach of different residue pairs in two-dimensional protein data strictly restricts the features for possible topologies of expected protein structures, which makes it useful in the predictive settings of De novo [1] and the similar recognition of fold, irrespective of eventual application, depends on its importance. The prediction of high-principled contact remains challenging, especially for small groups of proteins. In particular, similarities between amino acid substitution patterns on a couple of locations suggest the interaction of residues in a structure [2]. The structural, evolutionary, and functional units of proteins constitute pathway-specific protein domains (PSPDs). PSPDs are crucial elements in complex human disease. Therefore, the domain-based annotation of pathways requires a quantitative method that can incorporate not only sequence similarities but also domain pathway association specificity [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhipeng Cai.

PSPD are preserved that can develop, function, and exist independently of the remaining parts of a protein chain in each protein sequence and tertiary structure. Each specific pathway domain is a compact, three-dimensional structure that can be stable and folded independently. Various structural domains consist of several proteins. In a variety of proteins, one domain can appear. PSPD proteins can be recombined to produce proteins that have different roles in various arrangements. In general, the length of domains varies from approximately 50 amino acids to 250 amino acids [4]. Consequently, protein folding must be guided along a certain folding path. The forces that drive this examination are probably a grouping of local and global factors, and their consequences are encountered in different phases [5]. The combined sequence, structure, and feature analysis also help us recommend a TIM barrel phylogeny. Based on these results, we can explore different pathway theories and enzyme production by mapping known TIM barrels in major metabolic pathways [6].

PSPDs are used as a common absorption by a variety of hybrid feature extraction molecules, such as protections, brain receptors, and ion channels. The cellular large uptake of DNA–chitosan nanoparticles is also one of the principal scaffolded proteins. It contains many cholesterol-rich pathways, e.g., the caveolae pathway [7]. Several studies have found that a lack of function of PSPDs likely affects a wide variety of human diseases, such as cancer, cancer pathways, Alzheimer's disease, and others [8]. PSPD proteins have attracted many researchers because of their essential role in human diseases. We chose (HPV) L1 as the carrier of a new peptide subunit vaccine and inserted it into sites of natural variation in the L1 proteins of several HPV strains by inserting coding sequences of the desired epitopes. The original compatibility of these epitopes is maintained with disulfide, which combines their endpoints with molecular models and may contribute to the preservation of the direction of the folding of L1 [9]. In this section, we add the opening bar of GeoFold and use experimental data to verify the effects of pathway simulation. If seam motions are used without applying the preceding procedure, the results of GeoFold are consistent with experimental data. To improve the kinetic and thermodynamic stability of proteins, we understand the new protein development model in terms of how disulfide linkages can be used for engineering [10]. In view of high-order 3D genome conformation, the DNA loci of these mutations. Yi Shi *et al.* presented the details on the 3D genome may be much more prosperous compared to the current neoantigen prediction processes for the amino acid sequence. This study, therefore, explores in retrospect the neoantigens' DNA origin in the sense of the 3D conformation, both immune-positive and negative, and reveals some results that are worthy of consideration. Yi Shi *et al.*, have integrated 3D genome data into a collection of peptide coding schemes, and have developed a group of deep sparse, neural network selection (DNN-GFS) model which is tailored and customized for the prediction task of neoantigen. The proposed DNN-GFS method, along with other machine-learning methods, and generates priority antigens, as well as useful intermediate functions such as vcf annotation, neoantigen-enumeration of candidates [11]. The advancement of DNA sequencing technology and a wide range of sequencing data have been provided over the last few years, providing unparalleled possibilities for advanced association studies among somatic point mutations and types and subtypes of cancer that can lead to a more accurate SMCC classification. However, the current SMCC processes present major obstacles to improving classification efficiency, such as high data sparsity, limited volumes of sample size and the implementation of simple linear classifications. The benefits and capabilities of the DeepGene model for gene processing based on somatic point mutation and suggest that the model can be applied to other complex genotype-phenotype interaction studies that believe support several related areas. For future research, DeepGene model deploy for other broad and complex data, and expand our training data collection, in order to further develop the classification result [12]. Accurate association with high order spatial chromatin folding, somatic co-mutations were important in protein coding genes. As per SCH regions are also enriched the preserved mutational signatures and sequences of DNA flanking these co-mutations as well as CTCF binding sites. The genetic variations in the same SCH appear to disrupt genes that drive cancer that participate in the signaling pathways. The present paper shows that high-quality spatial chromatin organisation, during tumor growth, can lead to the somatic mutations of certain cancer genes. These SCHs share some common characteristics such as identical transformational signatures, preserved neighboring sequences flanking points of mutation and capable of perturbing genes involved in various molecular pathways. We also characterized SCHs from various cancer forms, including point-mutation signatures, conservation of flank sequences of point mutations and interruptions in driver mutations signaling pathways [13]. Protein development is modeled as a series of pathways through which one possible degree of conformational freedom is applied to each step in each direction. The cuts represent a network of simultaneous equilibrium and are a directed acyclic diagram. Finite simulations of differences in this map simulate native unfolding pathways [14].

Machine-based diagnostic systems may be useful to help clinicians recognize patients with PD. In this work, the performance of PD-based machine-learning techniques are assessed based on the symptoms of dysphony [15]. CNN models are trained using these images as inputs and training groups as outputs. In addition, were different classifiers trained with the pathologist-estimated fibrosis score (PEFS) as inputs and training classes as outputs [16]. Empirical studies on current methods have been conducted, but none of them have found a solution to avoid the loss of information about amino acid sequences in PSSM profiles. Here, to address this issue, we present a revolutionary approach by utilizing a recurrent neural network (RNN) architecture [17]. Most PSPD proteins have been published, but PSPD proteins have yet to be identified using machine training technology.

Doing so is difficult; as such, we are motivated to develop a precise model. Other researchers used low neural networks in earlier years to resolve computer biology problems. For example, our [18] developed QuickRBF to create radial basis (RBF) networks and applied this package to a range of biological problems, including the classification of membrane proteins. Some researchers used a deep learner in molecular research, such as pathological prediction [19], pathway cancer prediction, cancer disease prediction [20], or secondary protein sequence based structures [21], because deep learning has been successfully applied in a variety of fields. Although these studies have very good findings, we assure that some biological applications by using 2D CNN.

Based on the advantages of deep learning, we suggested that a 2D convolutional neural network (CNN) could be used to identify PSPD proteins based on feature extraction models, such as DDE, DPC. The basic theory has been successfully applied to detect proteins in electron transport [22] and examine the relationship of diseases, pathways, and human variants based on the ModSNP database material [23] and HumanCyc (Romero *et al.*, 2004), which is a computerized metabolic network database for humans [24]. Therefore, the present study extends this approach to the molecular functioning of PSPD proteins. The contributions of this paper are as follows.

(i) We establish a deep learning system for recognizing the PSPD functions in protein sequences, which have significantly improved beyond conventional machine learning algorithms in our model.
(ii) We conduct the first computer-based research to classify PSPD proteins and provide biologists with useful knowledge.
(iii) We also perform cross-validation and independent tests for high-precision PSPD proteins that form the foundation for future research on PSPD proteins.
(iv) We propose PSPD sources and methods for additional research on the application of the 2D CNN framework design in protein prediction.

II. MATERIALS AND METHODS

A. DATASETS COLLECTION

In this study, an approach was employed to investigate the datasets obtained from the NCBI respiratory, which is one of the biotechnology information's extensive tools. First, with the keyword "pathway-specific proteins," and second, the query "non pathway-specific cancer proteins" from the NCBI non-redundant protein database (https://www.ncbi.nlm. nih.gov /protein/https://www.ncbi.nlm.nih.gov/protein/) was set, and PSPD proteins were collected [25] as shown in Table 1.

A sequence of a pathway-specific protein was suggested as a positive test sample, and the sequence was referred to as a negative sequence with no known site for pathway association. They were randomly chosen to achieve a balance between positive and negative samples for training datasets and independent test datasets. UniProt/Swiss-Prot online database, which contains multiple species, was used, but only human-related proteins specially involved in human pathways were considered in this research study. In step one, 217 PSPD proteins were downloaded and uploaded on the CD-HIT for similarity measurements; after CD-HIT [26], 105 proteins associated with pathway-specific proteins were received. In step two, 283 other proteins were downloaded and uploaded on the CD-HIT for similarity measurements; then, 140 non-PSPDs were received. According to this preprocessing approach, 245 proteins were finalized after the removal of redundancy. In step three, the query "cancer pathway-specific proteins" was set, and 1,104 proteins were found in FASTA format and uploaded on the CD-HIT for similarity. Redundancy was reduced, and 532 proteins containing 224 PSPD proteins and 308 non-PSPD proteins were received.

TABLE 1. Our experimental data collected as PSPDS and non-PSPDS sequences.

	Collected	Non-Redundancy	Cross-Validation	Independent
PSPDs	1763	1612	400	103
Non-PSPDs	2251	1200	1323	448

B. FEATURES EXTRACTION FOR IDENTIFYING PATHWAY-SPECIFIC PROTEIN ASSOCIATION

Another problem of the current hypothesis is that the extraction of features is an important step in the classification process; that is, protein sequence information is translated into numerical data. In this study, knowledge about protein sequences was chosen based on structure, physicochemical characteristics, and evolutionary-related characteristics. They could be further broken down into two subtypes: dipeptide deviation from the expected mean (DDE) and dipeptide composition (DPC). A sparse matrix of two dimensions consisting of 20 $\times$ 20 was obtained and extended into a single-dimensional vector. Instead, for the vector to achieve a compact functionality set via random projection, an effective measuring matrix was chosen. Therefore, new technology has been introduced to the extraction of compressive sensing functionality.

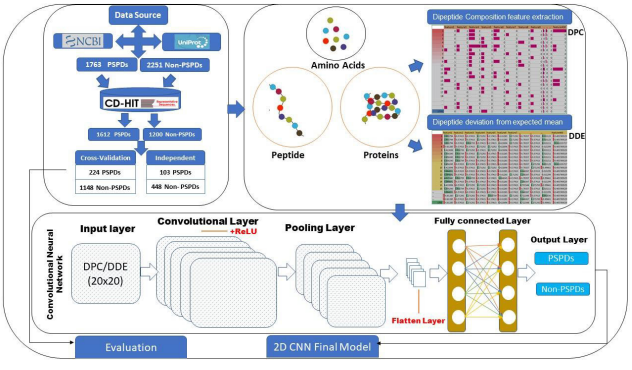

FIGURE 1. Proposed framework model of 2D-CNN PSPD.

The subjects of this study consisted of the 2D CNN and the DDE and DPC feature profiles, and an important method was developed to classify pathway-specific proteins. The system involves four procedures: data collection, feature extraction, CNN generation, and model assessment. Figure 1 shows our system flowchart and explains its specifics as follows. The subjects of this study consisted of the 2D CNN and the PSPD PSSM matrix feature extraction profiles. An important method was developed to identify and classify pathway-specific proteins involved in human pathways. The PSSM matrix feature extraction profile was treated on the basis of encoding based on DDE for physicochemical property-based features. Peptides of equal length were encoded using DPC descriptor for evolutionary-derived features.

1) DIPEPTIDE DEVIATION FROM THE EXPECTED MEAN (DDE)

DDE-PSSM was used to collect physicochemical data, sequence information, and evolutionary information. Therefore, the DDE, a new amino acid composition-based descriptor, was proposed and developed in this study to efficiently recognize PSSPD and PSSM from non-PSPDs. The efficiency of the DDE characteristic vector in enhancing the particular linear proteins associated with pathway prevention was demonstrated and compared with other characteristic representations. In comparison with other amino acid-derived features on different datasets, DDE function vectors had better performance (with accurate differential cross-validation and independent datasets) different datasets. The amino acid frequencies are divergent [27] to extract the features and their protein relation with a feature vector widely employed in various protein function prediction methods DDE of their respective median predicted levels of acid [27].

In this analysis, dipeptide composition aspects were used to measure the dipeptide frequency deviations from the predicted average values in accordance with previous studies [28]. Three important computer parameters were built to create the DDE feature vector: theoretical mean (T_m), theoretical variance (T_v), and dipeptide composition (C_c). The three parameters and DDE are calculated as follows, and $D_{C(i)}$, an indicator of the Cc of dipeptide i in peptide P is given by

$$D_{c(i)} = \frac{n_i}{N} \tag{1}$$

The features with a length of 400 dipeptide properties (20×20 regular amino acids) were extracted, but not all of them were going on in any sequence. Nor is the occurrence of dipeptide I and N is L-1 (i.e., potential quantity in P). $T_{M(i)}$ the theoretical mean

$$T_{M(i)} = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \tag{2}$$

For the first amino acid, C_{i1} is the number of codons, and Ci_2 number and for the second amino acid of C_{i2} codons for the specified dipeptide $'i.'$. CN is the total number of codons available except the three stop codons. $T_{M(i)}$ does not depend on peptide P, so the features with a length of 400 dipeptides were extracted and precomputed. $T_{V(i)}$ is given by dipeptide i theoretical variance as follows:

$$T_{v(i)} = \frac{T_{M(i)}\left(1 - T_{M(i)}\right)}{N} \tag{3}$$

The theoretical average of i is $T_{M(i)}$ determined with Equation (2). The number of dipeptides in peptide P is again and N is L-1. $DDE_{(i)}$ is finally determined as

$$DDE_{(t)} = \frac{D_{c(i)} - T_{m(i)}}{\sqrt{T_{V(i)}}} \tag{4}$$

Finally, DDE was calculated for each feature of the 400 dipeptides, and the 400-dimensional characteristic vector was employed:

$$DDE_p = \left\{DDE_{(i),}, \ldots, \ldots DDE_{(n)}\right\}, \ where, i = 1, 2, \ldots, 400 \tag{5}$$

2) FEATURES EXTRACTION USING DIPEPTIDE COMPOSITION(DPC)

Two consecutive residues consist of dipeptide composition (DPC). The sequence lengths are set to 400. This commonly used representation of the sequence includes details on the amino acid fraction and their local order. We applied to this model with a protocol of feature extraction DPC-PSSM for the optimal feature's foundations. We developed by utilizing the next DPC model of the sequence feature extraction model. The DPC represents the occurrence of an amino acid in two adjacent positions in a protein sequence that represents the number of amino acid incidents. In the series, for example, MALMAC and CC dipeptide frequencies: 2, 1, 1, 1, 1, and 1, respectively, MA, AL, LM, AC, and CC. A total of 400 dipeptides were used, i.e., the number of feature elements. By dividing the frequencies by (N-1), while N is the sequence length, the DPC characteristics were standardized and multiplied by 100 [29]. Dipeptides capture the amino acid composition a new meaning as some local details can be obtained in terms of the frequency of two contiguous amino acids [29]. Thus, for cases that need localized information, such as homologic information, the dipeptide composition is appropriate.

$$f_j = \frac{\#\ of\ dipeptide_j}{N - 1} \times 100 \tag{6}$$

C. PROPOSED 2D CONVOLUTIONAL NEURAL NETWORK FRAMEWORK

TensorFlow structures were introduced and distinguished from matrices. The 2D-CNN PSPDs were commonly used to define images with each input image converted into the input window so that the size of the image contain on window size and the feature-length was the distance. All input features of a scalar and two-dimensional information were converted to two-dimensional features (channels) to create an input window for each protein of any length so that all features, including those already expressed in 2D, were two dimensional and

could be viewed as individual channels. However, each two-dimensional function, such as the solvent accessibility clause, was duplicated across the line and across the column to generate two channels, while scalars such as sequential length were duplicated into a two-dimensional matrix (one channel). The size of the protein features channels was determined on the basis of the window length. Each filter in a convolution layer that converted the entry window to each filter had access to all input functions and could learn the connections across the channels by having all of its properties in separate input channels [30] and feature extraction of the secondary structure of proteins [31]. The underside of Figure 1 demonstrates the simplified framework model of 2D-CNN PSPD.

Kera's library with the TensorFlow backend [32], [33] was used to implement our deep learning architecture. The 2D-CNN PSPDs is generally composed of numerous layers with a particular function, causing each layer to transform its input into useful representation. The architecture of our 2D-CNN PSPDs model was coupled with a particular order. Optimization should be applied to find the correct architecture and hyperparameter and to construct an effective model, as revealed by several studies in this field [34], [35]. A various set of layers and hyperparameters was required for various problems and datasets. In this review, this procedure was carried out and described as follows in accordance with this law.

1) CNN PREDICTION MODEL

A best computational model and protein feature representation can quickly annotate the functions of the enzymes in chemical reactions in the prediction of enzyme proteins, which is a specific pathway function. CNN module 2D structure information into the window as a figure convenient for convolutional neural networks (CNNs) and discard a large amount of related information. We, therefore, proposed a method that would directly predict the function of the pathway-specific enzyme proteins using the relation between amino acids. First, we have introduced a new structural feature, the relative angle of amino acid, in addition to standard structural features. A variety of applications were undertaken to identify the type of protein, predict binding sites, prediction of protein-protein interactions based on knowledge from sequences in the bioinformatics area of the CNN model. For example, classification of pathway-specific proteins and transportation proteins, prediction of electron transport proteins, secondary protein structure prevention, DNA-protein binding site prediction, and protein-protein interaction prevision are used by many researchers in the field of bioinformatics. The main advantage of this method is that data will be processed in an appropriate image format after the automatic use of features. 1D convolution is used on features associated with the sequence of amino acids, whereas 2D convolution is associated with the specific position marking matrix or any additional map function.

CNNs are ideally suited for these problems because the main concept in convolutional layers, regardless of the spatial location of their input, is to identify local patterns. If this concept is taken into account in enzyme related pathway protein predictions, using convolutional filters for an amino acid covariance matrix, say, the pattern allows to detection interactions between locally separated sequence patterns by an arbitrary amount of residues that match well with observed structural patterns.

2) 2D CNN OPTIMIZATION PROCESS

The advantage of this 2D-CNN method is end-to-end differentiability, which means that all parts of the organization can be optimized simultaneously through independent and cross-validation, from acquiring input features to predicting two-dimensional coordinates. We have optimized our method based on deep learning (DL) models.

3) INPUT LAYER

Throughout the analysis, the parameters of the input layer were translated to 20×20 matrices throughout the DPC model as for dipeptide feature profiles. With our input data, these matrices could be applied as a method to distinguish PSPD proteins in the binding pathways. Furthermore, the dipeptide composition PSSM was used as an input in the 2D-CNN model. The same points were used as a pathway-specific protein family and inserted into independent sets. Then, the training performance was assessed with a 10-fold cross-validation process. This research was carried out using 2D-CNN, the largest deep neural network. CNN has been used in many fields, and impressive results have been obtained through computational vision, especially if the input is normally a 2D image pixel density matrix. A 2D structure of CNN architecture input image was utilized on the basis of these results, and 2D inputs of 20×20 size window PSSM matrices were conveniently generated. The 2D CNN models rather than 1D models were preferred to capture the hidden figures confidential the PSSM matrix profiles. PSSM profiles were connected from the input layer to the output layer via the 2D CNN design architecture.

4) ZERO PADDING LAYER

The block of a CNN was a pooling layer that could slowly decrease the representation's spatial size, the number of network parameters, and measurements. In each function diagram, the pooling layer operated separately. The function of this layer is also known as "down-sampling" because it eliminates certain values that lead to fewer systems and overfit operations while preserving essential characteristics. We set datapoints window, or any region that moves through the input matrix is often required for the grouping layer to become a representative of all values. In the top, bottom left, and right of the features profile matrix, you can add columns and rows of zero values. When 2×2 strokes were used, the frequency of the production was 22×22 strokes in a 20×20 matrix. After the filters were applied to the input data, our model did not have different output dimensions.

5) CONVOLUTIONAL LAYER

The features in the 2D input matrix were extracted through convolution by using a coding layer. A sliding window was used to transform the values into representative values and moved in step across the input. The convolution activity maintained the spatial relationship between numerical inputs in hybrid feature profiles by learning useful functionality through small input squares. When our model was designed by using a 3×3 sliding window. Each neuron was obtained, and inputs from the previous layer were trained with weights and biases.

6) ACTIVATION LAYER

The important contextual information about the carrying function as the activation mechanism used in the creation of 2D-CNN for the classification of PSPD proteins was performed with a rectified linear unit (ReLU). ReLU has been commonly used as the most important triggering function of all deep neural networks. The ReLU function is defined by the following formula, where x is the input number of the neural network.

$$f(x) = max(0, x) \tag{7}$$

7) POOLING LAYER

A pooling layer is normally placed in convolution layers to reduce the size of the matrix measurement for the next convolutional layer. The block of a CNN is a pooling layer. This has a feature of slowly decreasing the representation's spatial size and reducing the number of network parameters and measurements. On each function diagram, the pooling layer operates separately. The function of this layer is also known as the ''down-sampling'' because it eliminates certain values that lead to fewer systems and overfit operations, while still preserving essential characteristics. When we set a sliding window or any region that moves through the input matrix is often required for a grouping layer to become representative of the values. Transformation either takes the maximum value (max pooling) or the mean of the values (average pooling). In this analysis, two pooling phases with three or three filters were planned with a commonly recognized method.

8) DROPOUT LAYER

In this step, the key factors of the dropout layer were identified and introduced to strengthen the current model's predictive performance and avoid overfitting [36], [37]. The model was randomly deactivated in the dropout layer with a certain probability P. The neural network ignored the selected neurons in the training if the dropout value was introduced to a layer and if the training time was extended. Dropout is often used to regularize deep neural networks; however, applying dropout to fully connected layers and convolutional layers is radically different. As well as being dropout in the deep learning community. As such, the dropout function with only 0.02 value was applied to the fully connected layers.

9) FLATTEN LAYER

Data are transformed into a one-dimensional array to the next level through flattening. The contribution from convolutional layers is flattened to create a single long vector. The final classification model, called a completely connected layer, is related to the model. All classes should be distributed to evaluate the output layers, and the input matrix should be converted to a vector by using the flattened layers.

10) FULLY CONNECTED LAYER

The layers in which all inputs of one layer are linked to each activation unit of the next layer are fully connected into neural networks. Neurons in a completely linked system, as seen in normal neural networks, are completely connected to all activations in the previous layer. Their activations can thus be determined by multiplying the matrix by an offset of the bias. For more details, see the Neural Network (NN) section of the notes. Implementation of dense layer, fixed that is a standard and completely connected NN, can be seen [38]. The characteristics of convolution and pooling layers are described in this section. The use of a completely connected layer is a popular approach to nonlinear hybrids.

11) LOSS FUNCTION

Binary cross-entropy was used to train a model and simultaneously overcome many classification problems if any classification could be reduced to a binary choice (e.g., yes or no, A or B, and 0 or 1). Binary cross-entropy is a loss function used in binary tasks. Tasks that answer a question by two options alone (e.g., yes or no, A or B, 0 or 1, and right or left). For a variety of binary classification problems, the loss function has been demonstrated [39]. As described above, the SoftMax output can be compared with the target value and minimization of (produced by one-hot encoding). We use cross-entropy to distinguish between them. Entropy is a loss function that maximizes the likelihood value as a target of the appropriate class mark. It is easy to see that an overtrained model will be very small to zero and could be accomplished by minimizing the loss function in a relatively simple manner. A variety of regularization strategies may be employed to prevent overfitting (e.g. protein 1 or protein 2 penalties, typically employed in proposed models), such as the inclusion of penalties in the loss function.

12) SOFTMAX UTILIZATION

The model output was assessed in terms of a SoftMax function, which reduces the probability of any output [40]. This function is a formula-defined logistic function and feature form used in ANN in the output layer and multiclass categorization problems. Inactivation, the value production is converted to values between 0 and 1 (distribution of the probability), where z in the formula above is a K-dimensional vector $\sigma(z)j$ entry, and *j-th* is the expected probability of sample vector x. Then, $(0, 1)$, *j-th* is a true value of the range $(0, 1)$ and *j-th*. In the model, trainable params with

TABLE 2. Parameters used as a trainable in 2d CNN model.

Layer(type)	Remarks	Output Shape	Parameters#
Zeropadding2d_1	Padding = (2,2)	(None, 22, 22, 1)	0
conv2d_329 (Conv2D)	Filters=32, kernels= (3,3)	(None, 1, 20, 32)	5792
max_pooling2d_346 (MaxPooling)	Pool Size= (2,2)	(None, 1, 10, 16)	0
zero_padding2d_330 (ZeroPadding)	Padding = (2,2)	(None, 3, 12, 16)	0
conv2d_330 (Conv2D)	Filters=64, kernels= (3,3)	(None, 3, 12, 64)	9280
max_pooling2d_330 (MaxPooling)	Pool Size = (2, 2)	(None, 3, 12, 64)	0
Zeropadding2d_3	Padding= (2,2)	(None, 9, 12, 32)	0
conv2d_329 (Conv2D)	Filter=128, kernels	(None, 7, 10, 128)	36992
max_pooling2d_346 (MaxPooling	Pool Size = (2,2)	(None, 7, 5, 64)	0
Zeropadding2d_3	Padding= (2,2)	(None, 160)	0
Flatten_1	Flatten	(None, 160)	0
dropout_213 (Dropout)	P=0.4	(None, 160)	0
dense_227 (Dense)	Units=128	(None, 128)	286848
dense_216 (Dense)	Unit=2	(None, 2)	258
Activation (Activation)	SoftMax	(None, 2)	0
Total params:	339,170	Total params:	339,170

339,170 data points were established, as shown in Table 2.

$$\sigma\,(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_i}} \tag{8}$$

13) HYPERPARAMETER

Deep neural networks are highly responsive and successful in terms of choosing the hyperparameters that characterize a network structure and a learning process. As such, these hyperparameters need to be measured automatically. Derivative-free optimization is an area in which methods are developed to optimize functions without relying on derivatives. The hyperparameters tuning for a deep neural network is a vital process, but it consumes time and computational resources, mostly manually based on expert knowledge. Nevertheless, the growing popularity and use of deep neural networks for various applications called for the automation of the process to adapt to each problem. Two groups may be used to distinguish the hyperparameters forming a deep neural network: the one representing the network architecture and the other influencing the training process optimization. Hyperparameters vary from the parameters of a model trained via backpropagation at an architectural level. In the construction of a profound learning model, the choice of such hyperparameters is decided by a variety of factors. The performance of the model has a remarkable effect. To improve the training and prevent overfitting, many parameters should be chosen, as suggested by Chollet [41]. For instance, the question of HPO can be seen as strengthening learning through which the key difference between each approach relies on the description and care of agents. A neural network can build other neural networks by observing potential settings.

- Set hyperparameters for selection
- Create the appropriate model
- Put on a model the training data and measure on a validation dataset the final formative data.
- Use the hyperparameter range of the next set time.
- Return/repeat
- Quantify or assess the output execution on an independent dataset

14) PERFORMANCE EVALUATION OF MODEL

This analysis mainly aims to identify pathway-specific protein sequence is a PSPD protein or not; thus, the definition of PSPD proteins is ''positive,'' and the definition of the non-PSPD protein is ''negative''. For each dataset, a 10-fold cross-validation technique is first applied to the model in the training dataset. Hyperparameter optimization is used to find the best model for each dataset based on the 10-fold cross-validation findings. Finally, the predictive potential of the current model is tested using an independent data collection. The following results are considered: sensitivity, precision, accuracy, and Mathews' correlation coefficient (MCC) as the measurements used to assess the prediction performance of our proposed model. TP, FP, TN, and FN are referred to as genuine or true positive, false positive, and false negative. The evaluation metrics are then specified accordingly.

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$MCC = \frac{TP^{*}TN - FP^{*}FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{12}$$

III. RESULTS

Our findings can be compared with previous results in terms of the proposed performance and reliability of research modeling techniques that are essential to the analysis. Primarily, experimentation is developed by evaluating data, calculating, and comparing numerous results and consultations. According to our two models, which contain DPC, and then used the DDE model.

A. PSPDS AND NON-PSPDS SEQUENCE FOR THE AMINO ACID COMPOSITION

In PSPD and non-PSPD sequences, the amino acid composition was analyzed by calculating its frequency. A compilation of (ARNDCQEGHILKMFPSTWYV-) 20 numerical values representing the various physicochemical and biological features of amino acids is an index of amino acids that were submitted to content analysis. The 20 amino acids that contribute to two separate datasets at a considerably higher level. The two types of data do not considerably different, but some

exceptions are noted. C and P amino acids are located at the maximum concentration frequencies throughout the proteins. Therefore, the discovery of PSPD proteins in these amino acids is important. Thus, our model can reliably predict PSPD proteins based on the different characteristics of these amino acids.

B. 2D-CNN TRAIN THE MODEL

A related idea may explain the training of the features of the model. In our proposed model, 150 epochs are used as model trains. Features are fit to return an object from a history, which can be used to history accuracy and loss function plots between training and validation by history the results of this function in 2d-CNN, which allows the visual measurement of the model's performance. Lastly, the model in 150 epochs with 2D-CNN PSPDs is trained, and the model is good because the precision of the training after 150 epochs is 0.95%, and the loss of training score is 0.17%, which is very small. However, as the validation loss is 0.22% and the validation precision is 0.91%, the model seems overfitted. Overfitting provides an assumption that the network has an excellent memory of the training data but does not see the hidden data; thus, the quality of training and validation varies. We resolved to deal with this possibility. In the following section, our model is developed by incorporating a dropout rate in the network, and the other layers are kept unchanged. Next, the efficiency of the model is analyzed before our conclusion is presented.

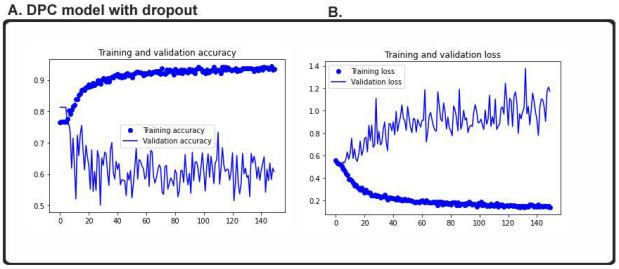

FIGURE 2. DPC model test and loss plot: (a) Validation accuracy and (b) validation loss.

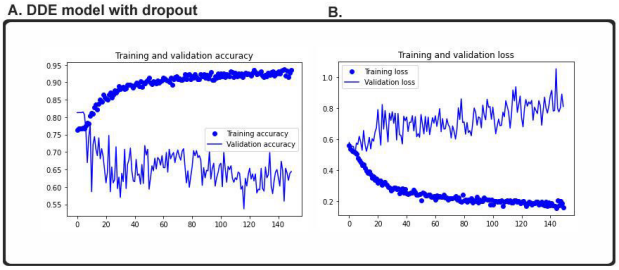

FIGURE 3. DDE model test and loss plot: (a) Validation accuracy and (b) validation loss.

C. TEST SET MODEL EVALUATION

Our proposed DDE model test accuracy of 0.9541 and test loss of 0.1704 are shown in Figure 2. The accuracy of the test is impressive. The model also compares well with the deep learning models. In our other model named DPC model, the predicted test accuracy is 0.9106 and the test loss is 0.2233. This model is examined to evaluate and plot the accuracy and loss between training and validation data as shown in Figure 3. We solved the overfitting issue to some extent; these findings are less surprising if we consider adding a dropout rate as a layer. Dropout turns a fraction of neurons off randomly during the training process, thereby reducing the dependency on the training set by a certain amount. The hyperparameter that can be modified accordingly defines how many fractions of neurons want to dropout. This step prevents the network from memorizing training data by shutting off certain neurons because not all neurons are active at the same time, and inactive neurons learn anything. Then, we develop, compile, and train the network again, but dropout is disregarded at this time. We run the network with a batch size of 10 and 150 epochs.

TABLE 3. DDE model predicted performance of PSPDS with different filters.

Filter numbers	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
32	0.758	0.664	0.711	0.427	0.734	0.825	0.788	0.564
32-64	0.770	0.672	0.721	0.446	0.731	0.851	0.790	0.589

D. PERFORMANCE RESULT FOR IDENTIFYING PSPD PROTEINS WITH 2D-CNN

Previous results indicated that the use of the Tensorflow backend Keras package is consistent with the findings. Our 2D-CNN architecture is implemented. Next, the best configuration for hidden layers is determined with the two separate convolutional layers 32, 64. The DDE model results of the cross-validation data collection of the various filter numbers used are shown in Table 3. We identify PSPDs and detect the sequences with a 10-fold average cross-validation accuracy of 0.7212% and independent set accuracy of 0.7909%. The results are higher than the average with other filter numbers from other metric calculations involving various filters. We achieved the cross-validation set performance of sensitivity of 0.7700%, the specificity of 0.6724%, and MCC of 0.4275%. The results consist of independent datasets by using various filter numbers. We achieved the performance of independent set accuracy of 0.7909%, sensitivity of 0.7310%, the specificity of 0.8511%, and MCC of 0.5894% as shown in Table 3. In this way, we used our model with this evolutionary structure of the layer. We implemented five hyperparameter optimization model to build our concluding model with Adadelta, a robust performance optimizer. Further DPC model results of the cross-validation datasets and Independent sets of the various filter numbers used are shown in Table 4.

Therefore, in these hidden layers, our model was built by using this convolutional layer structure. Afterward, the neural networks were optimized with different optimizers:

TABLE 4. DPC model predicted performance of PSPDS with different filters.

Filter numbers	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
32	0.753	0.694	0.724	0.450	0.723	0.824	0.774	0.555
32-64	0.749	0.699	0.724	0.451	0.742	0.824	0.783	0.571

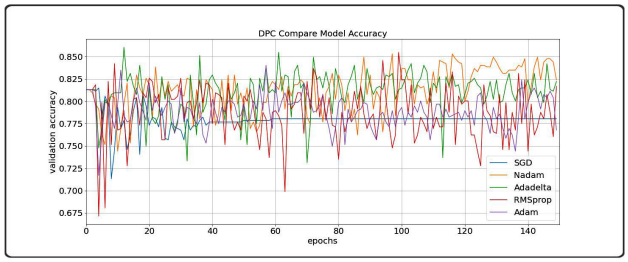

FIGURE 4. DPC model performance of the training and validation accuracy of various optimizers in this analysis (from 0 to 150).

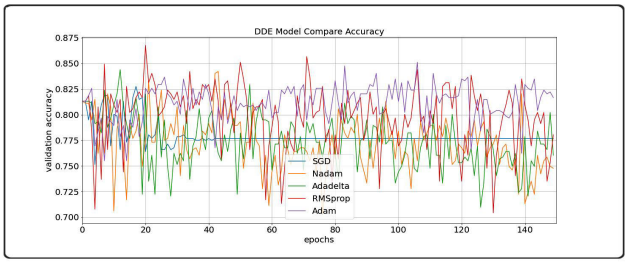

FIGURE 5. DDE model performance of the training and validation accuracy of various optimizers in this analysis (from 0 to 150).

RMSprop, Adam, Nadam, SGD, and Adadelta. After each optimization in each round, the model was reset, that is, a new network was created so that the different optimizers were comparable. The results are displayed in Figure 4. Our final model was created by choosing Adam, an optimizer with a robust performance. The best optimizer for our proposed model was chosen for Adam. During the experiment, the default learning rate (float, default = 0.001 steps), batch size = 10, and dropout rates = 0.2 were used, and the different iterations from 100 to 150 were run. Moreover, the accuracy of our model in terms of predicting new sample data was checked with independent testing data, and the results were compared with the other performance. In Figure 5, our model validation accuracy was improved after the 150th epoch based on training accuracy. Therefore, our training was completed at the 150th level to reduce training time and prevent overfitting were modified (Table 5) to obtain the best result in the performance of the dataset. After this overfitting point, the main problem of all the problems of machine learning is that our classification can only function well in our training method. Still, it can be worse in a different invisible dataset. An independent test was conducted to make sure our model still fit well in a blind dataset.

Our independent dataset included 103 PSPDs and 848 PSPDs, as defined in the previous section. None of these samples occurred in the training set. Two confusing matrices

TABLE 5. Optimal hyperparameters used in our proposed method.

Used Hyperparameter	Vales
Number of epochs	80
Learning Rate	0.001
Batch size	10
Kernel size	3
Dropout rate	0.4
Optimizer	Adam

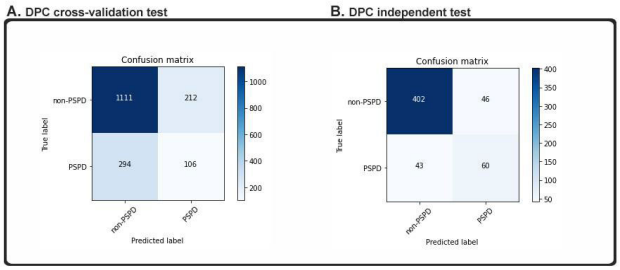

FIGURE 6. DPC model confusing matrices of (a) cross-validation test and (b) independent test.

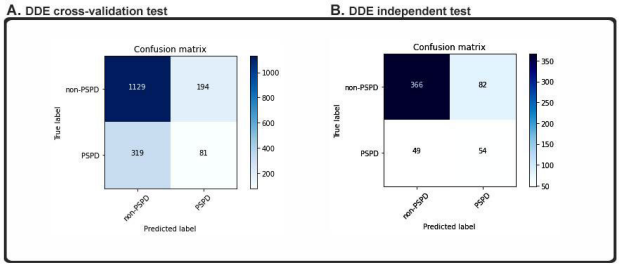

FIGURE 7. DEE model confusing matrices of (a) cross-validation test and (b) independent test.

are shown in Figures 6 and 7, with more detailed results. In Figure 5, which was consistent with the result from cross-validation with our independent test dataset result. In particular, our model achieved 85.8% precision, 82.2% sensitivity, 69.2% specificity, and 0.70% MCC in independent testing. The discrepancies were not too high compared with the cross-validation result and might demonstrate that our model was not overfitted. Another explanation was the use of dropouts, and the duplication of our CNN program was effectively prevented.

E. FURTHER STUDIES ON CNN SIGNIFICANT FUNCTION

The hypothesis that deep learning methods need further support. The extracted features vary from local to abstract hierarchical, so the essential feature of our model of CNN can be difficult to identify. We tried to resolve the issue to provide more valuable knowledge to readers and biologists. Considering that we inserted 20 × 20 hybrid feature profiles into our CNN system, we analyzed the core features of these matrices. To classify the most relevant features in the creation of the problem result, we used the F-score. Our research

aimed to determine which sequences of PSPDs and non-PSPDs would rely on our model to produce better results. All our feature's functionality in F-scores, and variations between the two datasets are observed. In summary, our model could classify amino acids as important hidden features, help us learn the most important protein characteristics, and achieve the best result for each of them.

TABLE 6. DDE model predicted performance of PSPDS with different optimizers.

Optimizers	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
Adam	0.770	0.684	0.727	0.457	0.730	0.865	0.798	0.603
Adadelta	0.755	0.711	0.733	0.469	0.727	0.850	0.789	0.585
RMSprop	0.744	0.681	0.713	0.427	0.715	0.825	0.770	0.547
Nadam	0.752	0.679	0.715	0.433	0.694	0.832	0.763	0.535
SGD	0.729	0.704	0.717	0.435	0.695	0.745	0.720	0.444

F. DDE MODEL RESULT OF IDENTIFICATION PSPD WITH DIFFERENT OPTIMIZERS

In the content analysis, the optimization of hyperparameters was calculated, or it is hard to determine the best hyperparameter optimizer for our model. Most researchers usually aim to optimize their algorithm performance based on an independent dataset. Several findings from this study warrant further discussion, for example, in algorithms for learning. One possible reason for this discrepancy could be that this simplification performance is evaluated through cross-validation. The hypothesis that the optimization of the hyperparameters differences with real learning problems, which are often considered optimization issues, optimizes a loss function alone. In addition, the learning algorithms learn that they can reconstruct their inputs. At the same time, the optimization of the hyperparameter ensures that the model does not overfit its data by tuning, for example, regularization, as shown in table 6 and table 7. Being part of the deep learning and the convolutional networks, we can readily modify and play hundreds of different parameters (although we seek to reduce the number of variables to just a few in practice), each influencing some (possibly unknown) degree of our overall classification. Our results indicate that Adam, Adadelta optimizer gives the best performance to estimate. Our research was establishing pathways and predicting their roles with a certain protein. The analysis of the findings was based on 10-fold cross-validations of cross-validation datasets and independent datasets. When used on the DDE model and calculation based on 5 optimizers, we achieved superior performance of Adam optimizer with the DDE model as shown in table 6 and Figure 8.

G. DPC MODEL RESULT OF IDENTIFICATION PSPD's WITH DIFFERENT OPTIMIZERS

The usage of deep learning technology was analyzed with five different optimization models. The comparison was then performed to determine the most appropriate optimizer. In this

TABLE 7. DPC model predicted performance of PSPDS with different optimizers.

Optimizes	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
Adam	0.754	0.688	0.721	0.446	0.712	0.804	0.758	0.522
Adadelta	0.764	0.668	0.716	0.436	0.734	0.836	0.785	0.577
RMSprop	0.738	0.681	0.710	0.421	0.669	0.854	0.761	0.536
Nadam	0.738	0.681	0.710	0.421	0.717	0.832	0.774	0.558
SGD	0.758	0.680	0.719	0.441	0.688	0.731	0.709	0.421

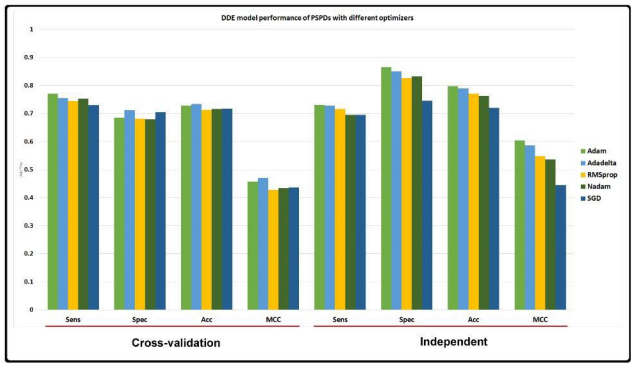

FIGURE 8. Comparison of performance between 5 optimizers with a DDE model based on 10-fold cross-validation on cross-validation datasets and Independent datasets.

situation, it is a difficult job to select an optimizer for training the network of CNN. To compare and classify the best optimizer for estimating PSPD functions 5 best optimizers were selected. Based on their processing times, prediction accuracy, and error, the five optimizers, Adadelta, RMSprop, Adam, Nadam, and SGD were compared. When we measured the prediction results of the DPC model. We also measure the sensitivity, specificity, accuracy values, F-score, and Matthews correlation coefficient values, which represent the best overall performance of the Adadelta optimizer with the DPC model as shown in Table 7 and Figure 9,

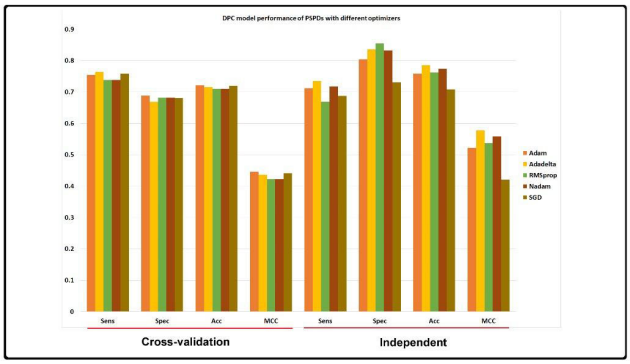

FIGURE 9. Comparison of performance between 5 optimizers with the DPC model based on 10-fold cross-validation on cross-validation datasets and Independent datasets.

TABLE 8. DPC model performance results of identifying PSPDS with other ML classifiers.

ML classifiers	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
AdaBoost	0.755	0.848	0.802	0.607	0.599	0.832	0.711	0.453
Random Forest	0.580	0.898	0.739	0.493	0.731	0.843	0.780	0.584
LSTM	0.704	0.758	0.731	0.466	0.706	0.881	0.762	0.547
CNN	0.802	0.755	0.779	0.558	0.785	0.925	0.855	0.600

to better investigate the predictive capacity of the hyperparameter optimizer.

H. PSPD's BETWEEN 2D CNN AND SHALLOW NEURAL NETWORKS WITH A COMPAREABLE EFFICIENCY

A possible interpretation of this finding is that it examined the performance of various machine learning techniques for the identification of proteins from PSPDs. We used four machine learning classifiers (e.g., AdaBoost, Random -Forest, and LSTM). To test the model, CNN Long Short-Term Memory Networks architecture implemented convolutional neural network (LSTM) perceptions, and 1D CNNs compared the effects of our 2D CNNs to those of them. We used the optimum parameters in all the experiments for equal comparisons with all the classifiers, as shown in Table 8. We demonstrated that the performance of our 2D CNN with the same experimental structure was better than that of other conventional machine learning techniques. In particular, by using a separate dataset, our 2D CNN implemented specific algorithms.

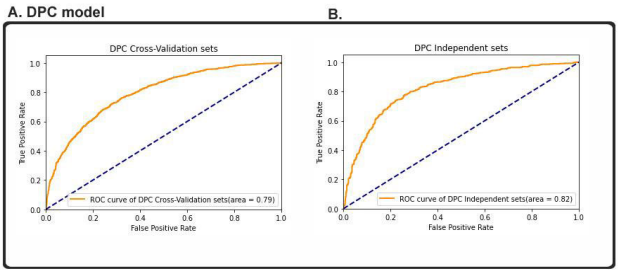

FIGURE 10. DPC ROC-AUC model of (a) cross-validation test and (b) independent test.

I. COMPARATIVE PERFORMANCE OF THE IDENTIFICATION OF PSPD's BY USING ROC-AUC CALCULATION

In this section, our findings could be compared with the results of earlier studies that compared the performance of the binary classification problem of this study. Our data were consistent with most machine learning classification models used. Researchers deploy the ROC curve plot and the AUC, along with other metrics, such as the accuracy of the algorithm or the confusion matrix. In this section, the ROC curve and AUC were used to analyze the 2D CNN output via multiple classifications, as seen in Figures 10 and 11. Displays the 2D CNN PSPD Multilink ROC curve. The results are somewhat but slightly similar to those of binary classification, indicating that our deep neural network architecture could perform highly even with the multiclassification method, but more data were needed to explore this finding further. Therefore, our proposed 2D-CNN model showed the best performance and had no overfitting because the 2D-CNN model cross-validation accuracy score was 0.87% and the independent accuracy score was 0.86%. The comparison of the same data points revealed that the DPC model cross-validation datasets had ROC and ACU score of 0.79%, and independent datasets achieved RCO-AUC score of 0.82%.

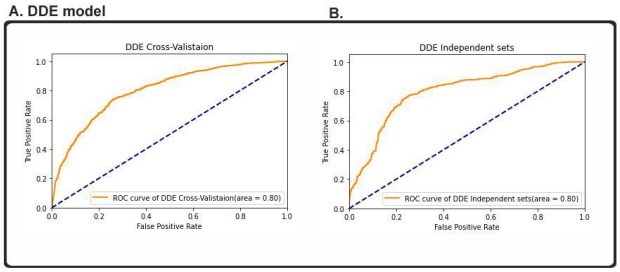

FIGURE 11. DDE ROC-AUC model of (a) cross-validation test and (b) independent test.

The ROC curves were derived from cross-validating results and used to further evaluate the CNN model's efficiency. Figures. 11 display the value for each protein association with the pathway class of the roc curves and the area under the curve (AUC). The ordinate is the true positive (TPR), and the abscissa is the false positive rate (FPR). The comparison of the DDE with the same data points revealed that the DPC model cross-validation datasets had an ROC and AUC score of 0.80%, and independent datasets achieved RCO-AUC score of 0.80%.

Our findings could be compared with the results of earlier studies with 10-fold cross-validation checks, and ROC (AUC) of 0.79% and 0.82% were achieved, and they were similar to our proposed model of 2D-CNN for both DDE and DPC composition. This result suggested the efficacy of the functionality protocol. The output of 2D CNN PSPD was also tested with a different dataset, and the results showed the 2D-CNN PSPD relation. Additionally, three machine learning classifiers were deployed to compare the results with AdaBoost ROC-AUC values of 0.76%, which was closely related to our method. The ROC-AUC value is the Random forest of 0.79%, and the LSTM classifier achieved a score of 0.81%.

IV. CASE STUDY

A. PROTEIN-PATHWAY ASSOCIATION

G protein pathway suppressor 2 (GPS2): In two human DNAs, including an *Arabidopsis* FUS6 homolog (COP11), the G protein- and mitogen-activated protein-kinase-mediated signal transduction is blocked [42], [43]. ''G-protein pathway suppressor arginine differential methylation recognized in melanoma by tumor-specific T cells'' [44].

The data provide convincing evidence of a strong association between proteins and pathways, such as the Uni-ProtKB database-associated protein ID of Q13227 (GPS2_HUMAN), which is involved in GPS2. The protein Q13227 acts as a B-cell production regulator that inhibits UBE2N/Ubc13, thus limiting the activation of the signaling pathway of Toll-like receptors (TLRs) and B-cell antigen receptors (BCRs). In response to depolarization, the role of a central mediator in the mitochondrial stress response relocates from the mitochondria to the nucleus. It disinfects the expression of mitochondrial-encoded genes in the nuclear field [45]. GPS2 is identified as a mediating retrograde mitochondrial signal and transcription activator encoded in the nuclear mitochondrial gene. These findings indicate an additional mitochondrial gene transcription regulation, a guided mitochondrial–nuclear communication pathway, and suggest that the key part of the mammalian mitochondrial stress response is retrograde GPS2 signalization shown in Table 9.

TABLE 9. Provided protein-pathway association evidence.

Entry	Entry name	Gene Name	Organism	Pathway Ids
Q13227	GPS2_HUMAN	GPS2	Homo sapiens	R-HSA-1989781
Q13098	CSN1_HUMAN	GPS1, COPS1, CSN1	Homo sapiens	R-HSA-5697010
Q9Y618	NCOR2_HUMAN	NCOR2	Homo sapiens	R-HSA-383280
Q9NQS5	GPR84_HUMAN	GPR84, EX33	Homo sapiens	R-HSA-418555
C9JFE4	C9JFE4_HUMAN	GPS1,1987516	Homo sapiens	PTHR14145:SF2
A8K070	A8K070_HUMAN	GPS1	Homo sapiens	PTHR14145:SF2
I3L3Y9	I3L3Y9_HUMAN	GPS2	Homo sapiens	PTHR22654
I3L1H4	I3L1H4_HUMAN	GPS2	Homo sapiens	PTHR22654
I3L242	I3L242_HUMAN	GPS2	Homo sapiens	PTHR22654
I3L4X7	I3L4X7_HUMAN	GPS2	Homo sapiens	PTHR22654

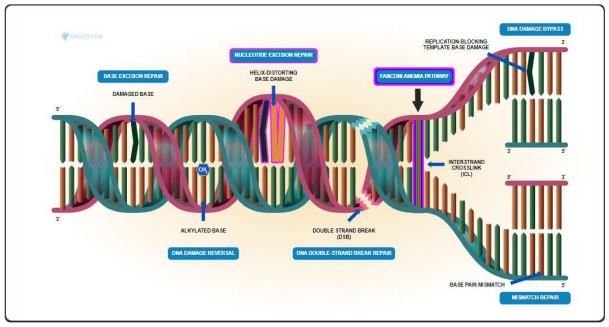

FIGURE 12. Protein–pathway involvement in the Fanconi anemia pathway with DNA damage repair [48].

B. PROTEIN-PATHWAY FUNCTIONS

The Uni-ProtKB database protein entry is Q13098 (CSN1_HUMAN), as shown in Table 3, provide convincing evidence about the COP9 signalosome complex subunit 1, which is associated with R-HSA-5697010 reactome pathway, such as the one associated with the inner part of two sub-pathways. In Figure 12, the Fanconi Pathway for Anemia and Nucleotide Excision Repair. The main component of the COP9 signalosome complex (CSN), a multi-cellular and developmental complex [46]. CSN is a significant regulator of the ubiquitin conjugation route (Ubl) via deneddylation in E-3 complex subunits, resulting in a reduction in ligase-like Ubl activities of SCF complexes, such as SCF, CSA, and DDB2. SCN is an essential regulator in this field. A complicated external ICL repair system, along with other repair processes, such as homologous recombination, nuclear cure, translational synthesis, and alternative terminal joints, is designed to repair different DNA injuries by Fanconi anemia pathways. Proteins of Fanconi anemia (FA) are used for retaining genomic constancy. Their primary function is to repair the interconnected DNA strands, which impede replication and transcription due to the covalent bonding of the Watson and Crick strands of DNA [47]. We have introduced a graph database, enhanced performance of data analysis tools, and developed new data structures and strategies to enhance diagram viewer performance.

C. FANCONI ANEMIA PATHWAY ASSOCIATED WITH THE OCCURRENCE OF CANCER DISEASE

DNA repair, an active cellular system reacting to constant damage to DNAs, is important in the preservation of genome integrity. Inherited DNA gene repair mutations are identified to prevent the carriers of genetic dysfunctional conditions from developing cancer. For example, DNA double-strand breaks recruit and activate ATM serine/threonine kinase, leading to the arrest of cells. ATM mutation causes ataxia-telangiectasia disorder. Bloom-protein syndrome exhibits a BLM mutation causing the Bloom syndrome, as well as DNA-stimulated ATPase and DNA-dependent helicase activities [49]. This section discusses the mechanisms of the Fanconi anemia pathway involved in ICL damage repair and the appropriate mutations causing genomic integrity deficits and supporting tumorigenesis [50], as shown in Figure 13.

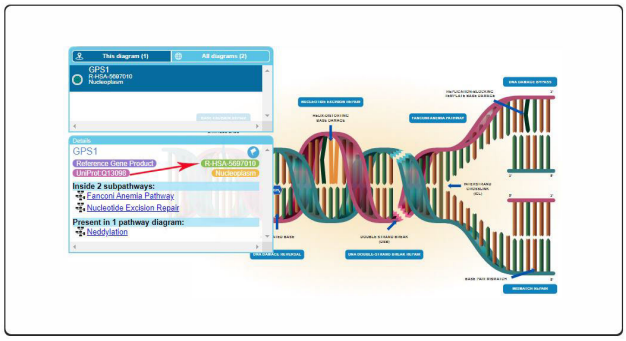

FIGURE 13. Protein–pathway involvement in Fanconi anaemia pathway with cancer disease.

D. PATHWAY ASSOCIATED WITH PROTEINS TRANSCRIPTION-COUPLED

To react to the interstrand crosslink (ICL) DNA lesions, the Fanconi anemia pathway is connected to many repair procedures. It has many roles other than ICL repair.

Fanconi anemic proteins, especially FANCD2, participate in the defense and cytokinesis of replication. The pathway forms a complicated network outside core ICL repair components to repair diverse DNA injuries, along with other repair processes, such as homologous recombination, nuclear reparation, and translational synthesis. These functions include fork stabilization and cytokinesis regulation. As such, fanconi anemia proteins emerge as master genomic integrity regulators that coordinate various repair processes. Here, a detailed overview of the functions of the Fanconi anemia pathway in ICL repair, its relationship with other repair pathways, and its evolving role in the maintenance of genomes is presented in Figure 14. DNA repair proteins can be used to repair post-replication or control the function of the cell cycle. May be involved in the cross-strand repair of DNA and in preserving the normal stability of the chromosome.

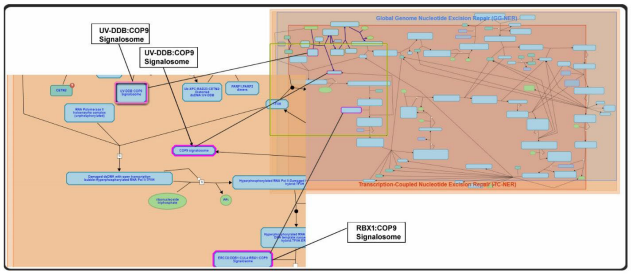

FIGURE 14. Protein–pathway involvement in transcription-coupled nucleotide excision repair.

The tumor suppressor gene candidate. The disorder is caused by gene-related mutations in this section. Description of disease a disorder that affects all the elements of the bone marrow, leading to anaemia, leukopenia, and thrombopenia. It is concerned with malformations of the heart, kidneys, and legs, dermal pigmentation, and malignancies. At a cellular level, hypersensitivity to DNA-damaging substances, chromosomal instability, and deficient DNA repair are associated with this—Fanconi anemia group G protein association with enzyme and pathway databases such as R-HSA-6783310 Fanconi anemia pathway. The essential component of the COP9 signalosome complex (CSN), a complex, and G protein pathway suppressor 1 (GPS1) are involved in various cellular and developmental processes, as shown in Figure 14.

V. IMBALANCE DATA PROBLEM

We consider our datasets to be unbalanced, affecting the classification process and, thus, significantly. The data set for cross-validation against independent is the same with a ratio of positive-negative rating. Mostly two methods are popular to fix training data imbalance. The first approach is data processing, and the second one is algorithmic. For this analysis, we used the method of data processing by sampling in the training data the minority class. Previous investigators have made substantial progress in over-sampling procedures. In selecting the over-sample approach over the under-sampling approach in resolving the imbalance problem, we have two advantages that have been achieved: data are sufficient to construct a robust model, and useful loss value has been avoided. Keeping this in mind, the number of minority class instances was slowly increased during the experiment, and the performance was reported at every move. With consideration for a balance between flexibility and specificity, the final selected model achieves the best efficiency.

VI. DISCUSSION

Our findings suggest that the computational method can be utilized to classify the biological functions of PSPD proteins. Furthermore, our research is necessary so that our molecular-based studies on functions in signaling pathways, G-protein pathways, and metabolic pathways can be better understood. Our research fills the gap with deep-learning techniques to complete PSPD sequences. This research is also the first to establish a computational method that provides biologists with much useful knowledge for understanding 2D-CNN-PSPD molecular functions and for creating a complex disease pathway based on their application in human diseases. For protein sequences, we also develop a broad and high-performance deep learning architecture. We validate the results with tuned hyperparameters to select the best parameters for efficient optimization. We use the extracted hybrid feature profiles as a vector only when they come into a network, and our findings are a different way of treating and adapting feature profiles to CNN networks. In addition, our two-dimensional CNN models employ many measuring methods to outstrip other approaches at the same level and collect data.

Our approach involving real-time systems is suitable. We can build a retrieval and analytical, biological information system based on computational model protein sequences. This intelligent device is more capable of finding variants or mutations of human diseases based on protein functions. This knowledge is used by biologists to develop drug targets in pharmaceutical studies. Our efforts contribute to our progress with this work, and this success is the key to treating descriptors for evolutionarily derived features as images. However, the proposed approach still has some limitations, and alternative methods are available to enhance the proposed technique in the future. First, a large number of datasets will increase profound learning efficiency, so future research and further information are needed to improve performance. Second, further studies should explore how all descriptors for evolutionarily derived feature information can be entered in CNN networks. We have also encouraged biological researchers to use our model and to suggest interactive experiences in addition to the showing of experimental precision findings. They thought that the model for machine learning plays an important role in understanding proteins with unknown functions and that our deep understanding of the model of amino acid interaction is a groundbreaking approach for future research using structural protein knowledge.

VII. CONCLUSION

In this study, the relationship between human proteins and the human pathway was analyzed on the basis of 2D-CNN-PSPDs architectures. An effective deep learning model was developed to classify PSPD proteins by turning the DDE and DPC descriptor physicochemical characteristics for derived features into matrices for evolutionary features. These matrices were then used as an optimized framework for 2D-CNN-PSPDs. The proposed 2D-CNN-PSPD model with PSSM matrix feature profile prediction based on 10-fold cross-validation and a separate research dataset was used to investigate our model. In contrast to other state-of-the-art neural networks, our approach provided superior efficiency and major improvements in all traditional measuring methods. Over the past decade, traditional methods have not been able to understand better the function of newly discovered DNA damage replication proteins associated with pathways. New PSPD proteins could be precisely defined and used to produce human disease pathways, drugs pathway, DNA repairing pathways via our model.

This study also promoted the use of 2D-CNN-PSPD in biochemical research and bioinformatics, especially in related proteomic and genomic directions for predicting protein sequence functions associated with human pathways. However, our hypothesis was complicated by the approach for mapping the human proteins UniProtKB/Swiss-Prot on four pathway databases. We verified the cross-reference knowledge route via a preliminary web interface. Future implementation will promote research on various biological pathways.

The conclusion of our proposed method for optimizing hyperparameters is provided to improve the prediction efficiency. In order to identify pathway association proteins within repair DNA, we carried out all the analyses that were proposed using 2D-CNN approaches constructed from PSSM matrix profiles. The output was analyzed using a 10-fold cross-validation method and separate radial network data sets. Our method demonstrated the precision of 10-fold cross-validation of 92.5% and 82.26%, respectively for the detection of DNA damage pathway proteins. We provided protein sequence model-independent sets on unlabeled Swiss-Prot protein sequences and finalized fine-tuned in the tasks of protein hyperparameter optimization. New pathway proteins can be reliably identified using our model and are used for DNA-based pathways, such as repair of DNA or production of replication. The contribution of this study may also lead to further work to encourage the use of 2D-CNN in the field of bioinformatics, especially in the prediction of protein functions.

This research focuses on the design of successful and deep learning models for the classification of PSPD/non-PSPD. In the future, we will discuss this concept of pathway adaptive weighting. Pathway-specific proteins are associated with disease, chemicals, and proteins (*e.g.*, genes, drugs, and enzymes). Some pathways are believed to be directly related to diseases. Their incorporation helps increase access to highlight the available pathway tools and provides a context for a particular chemical or target.

DATA AVAILABILITY

Data availability statement: The datasets were analyzed in this study are publicly available. The used all datasets can be found here: NCBI database (https://www.ncbi.nlm.nih.gov/), and https://www.ncbi.nlm.nih.gov/protein/ database and then also verified and compared with (https://www.uniprot.org/uniprot/) Uniport database and then we used proteins datasets in fasta format for removing the similarity (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit), CD-HIT web tools. After then all preprocess our data, we attached in data-availability folder (1- Sample of fasta format data, 2- Extracted PSSM features data). Further, the code used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGMENT

(Ali Ghulam, Xiujuan Lei, and Yuchen Zhang contributed equally to this work.) Xiujuan Lei conceptualized and finalized the manuscript. Ali Ghulam wrote the initial manuscript. Yuchen Zhang helped design the method and the code. Shi Cheng and Min Guo revised the manuscript and polished the expression of English. All of the authors have read and approved the final manuscript.

REFERENCES

[1] T. Nugent and D. T. Jones, "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 24, pp. 1540–1547, Jun. 2012, doi: 10.1073/pnas.1120036109.

[2] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins," *Proteins, Struct., Function, Genet.*, vol. 18, no. 4, pp. 309–317, Apr. 1994, doi: 10.1002/prot.340180402.

[3] J. E. Shim, J. H. Kim, J. Shin, J. E. Lee, and I. Lee, "Pathway-specific protein domains are predictive for human diseases," *PLOS Comput. Biol.*, vol. 15, no. 5, May 2019, Art. no. e1007052, doi: 10.1371/journal.pcbi.1007052.

[4] S. W. Englander and L. Mayne, "The nature of protein folding pathways," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 45, pp. 15873–15880, Nov. 2014, doi: 10.1073/pnas.1411798111.

[5] D. J. Pitman, S. Banerjee, S. J. Macari, C. A. Castaldi, D. E. Crone, and C. Bystroff, "Exploring the folding pathway of green fluorescent protein through disulfide engineering," *Protein Sci.*, vol. 24, no. 3, pp. 341–353, Mar. 2015, doi: 10.1002/pro.2621.

[6] R. R. Copley and P. Bork, "Homology among 8 barrels: Implications for the evolution of metabolic pathways 1 1Edited by G. Von heijne," *J. Mol. Biol.*, vol. 303, no. 4, pp. 627–641, Nov. 2000, doi: 10.1006/jmbi.2000.4152.

[7] Z. Garaiova, S. P. Strand, N. K. Reitan, S. Lélu, S. Ø. Størset, K. Berg, J. Malmo, O. Folasire, A. Bjørkøy, and C. de L. Davies, "Cellular uptake of DNA–chitosan nanoparticles: The role of clathrin- and caveolae-mediated pathways," *Int. J. Biol. Macromolecules*, vol. 51, no. 5, pp. 1043–1051, Dec. 2012, doi: 10.1016/j.ijbiomac.2012.08.016.

[8] F. Wu and P. J. Yao, "Clathrin-mediated endocytosis and Alzheimer's disease: An update," *Ageing Res. Rev.*, vol. 8, no. 3, pp. 147–149, Jul. 2009, doi: 10.1016/j.arr.2009.03.002.

[9] T. Jordan, C. Barcellona, D. Basore, C. Clark, Z. Guo, S. Isern, K. Nand, G. Rabasa, T. Shoemaker, G. Werner, K. Xia, X. Yuan, R. J. Linhardt, S. Michael, and C. Bystroff, "HPV VLPs as scaffolds for vaccine design," *Biophys. J.*, vol. 116, no. 3, p. 58a, Feb. 2019, doi: 10.1016/j.bpj.2018.11.360.

[10] V. Ramakrishnan et al., "Geofold: Topology-based protein unfolding pathways capture the effects of engineered disulfides on kinetic stability," *Proteins*, vol. 80, no. 3, pp. 920–934, 2012, doi: 10.1002/prot.23249.

[11] Y. Yu, Y. Cai, C. Zhang, W. Cai, L.-T. Da, G. He, and Z.-G. Han, "DeepAntigen: A novel method for neoantigen prioritization via 3D genome and deep sparse learning," *Bioinformatics*, vol. 5, p. btaa596, Sep. 2020, doi: 10.1093/bioinformatics/btaa596.

[12] Y. Yuan, Y. Shi, C. Li, J. Kim, W. Cai, Z. Han, and D. D. Feng, "DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations," *BMC Bioinf.*, vol. 17, no. 17, p. 476, Dec. 2016, doi: 10.1186/s12859-016-1334-9.

[13] Y. Shi, X.-B. Su, K.-Y. He, B.-H. Wu, B.-Y. Zhang, and Z.-G. Han, "Chromatin accessibility contributes to simultaneous mutations of cancer genes," *Sci. Rep.*, vol. 6, no. 1, Dec. 2016, Art. no. 35270, doi: 10.1038/srep35270.

[14] V. Ramakrishnan, S. P. Srinivasan, S. M. Salem, S. J. Matthews, W. Colón, M. Zaki, and C. Bystroff, "Geofold: Topology-based protein unfolding pathways capture the effects of engineered disulfides on kinetic stability," *Proteins, Struct., Function, Bioinf.*, vol. 80, no. 3, pp. 920–934, Mar. 2012, doi: 10.1002/prot.23249.

[15] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 29–39, Feb. 2018, doi: 10.1007/s13534-017-0051-2.

[16] V. B. Kolachalama, P. Singh, C. Q. Lin, D. Mun, M. E. Belghasem, J. M. Henderson, J. M. Francis, D. J. Salant, and V. C. Chitalia, "Association of pathological fibrosis with renal survival using deep neural networks," *Kidney Int. Rep.*, vol. 3, no. 2, pp. 464–475, Mar. 2018, doi: 10.1016/j.ekir.2017.11.002.

[17] N. Q. Khanh Le, Q. H. Nguyen, X. Chen, S. Rahardja, and B. P. Nguyen, "Classification of adaptor proteins using recurrent neural networks and PSSM profiles," *BMC Genomics*, vol. 20, no. 9, pp. 1–9, Dec. 2019.

[18] Y.-J. Oyang, S.-C. Hwang, Y.-Y. Ou, C.-Y. Chen, and Z.-W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 225–236, May 2005.

[19] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "NiftyNet: A deep-learning platform for medical imaging," *Comput. Methods Program Iomed.*, vol. 158, pp. 113–122, May 2018.

[20] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data," *Comput. Methods Program Biomed.*, vol. 166, pp. 99–105, Nov. 2018.

[21] S. Babaei, A. Geranmayeh, and S. A. Seyyedsalehi, "Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks," *Comput. Methods Programs Biomed.*, vol. 100, no. 3, pp. 237–247, Dec. 2010.

[22] Y. L. Yip, H. Scheib, A. V. Diemand, A. Gattiker, L. M. Famiglietti, E. Gasteiger, and A. Bairoch, "The swiss-prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants," *Human Mutation*, vol. 23, no. 5, pp. 464–470, May 2004.

[23] N.-Q.-K. Le, Q.-T. Ho, and Y.-Y. Ou, "Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins," *J. Comput. Chem.*, vol. 38, no. 23, pp. 2000–2006, Sep. 2017.

[24] P. Romero, "Computational prediction of human metabolic pathways from the complete human genome," *Genome Biol.*, vol. 6, pp. 1–17, Oct. 2004.

[25] M. H. Arisha, M. Q. Ahmad, W. Tang, Y. Liu, H. Yan, M. Kou, X. Wang, Y. Zhang, and Q. Li, "RNA-sequencing analysis revealed genes associated drought stress responses of different durations in hexaploid sweet potato," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 12573, doi: 10.1038/s41598-020-69232-3.

[26] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: A Web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.

[27] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear B-Cell epitope prediction: A novel amino acid composition-based feature descriptor," *OMICS, A J. Integrative Biol.*, vol. 19, no. 10, pp. 648–658, Oct. 2015, doi: 10.1089/omi.2015.0095.

[28] V. Saravanan and N. Gautham, "BCIgEPRED—A dual-layer approach for predicting linear IgE epitopes," *Mol. Biol.*, vol. 52, no. 2, pp. 285–293, Mar. 2018, doi: 10.1134/S0026893318020127.

[29] L. Zou, C. Nan, and F. Hu, "Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles," *Bioinformatics*, vol. 29, no. 24, pp. 3135–3142, Dec. 2013, doi: 10.1093/bioinformatics/btt554.

[30] N. Q. K. Le, T.-T. Huynh, E. K. Y. Yapp, and H.-Y. Yeh, "Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles," *Comput. Methods Programs Biomed.*, vol. 177, pp. 81–88, Aug. 2019.

[31] Y. Liu, Y. Chen, and J. Cheng, "Feature extraction of protein secondary structure using 2D convolutional neural network," in *Proc. 9th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2016, pp. 1–5.

[32] C. Mirabello and B. Wallner, "Rawmsa: Proper deep learning makes protein sequence profiles and feature extraction obsolete," *Biorxiv*, vol. 1, Jan. 2018, Art. no. 394437, doi: doi:10.1101/394437.

[33] M. Abadi. (2018). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/

[34] Y. Guo, W. Li, B. Wang, H. Liu, and D. Zhou, "DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction," *BMC Bioinf.*, vol. 20, no. 1, p. 341, Dec. 2019, doi: 10.1186/s12859-019-2940-0.

[35] D. Zheng, G. Pang, B. Liu, L. Chen, and J. Yang, "Learning transferable deep convolutional neural networks for the classification of bacterial virulence factors," *Bioinformatics*, vol. 36, no. 12, pp. 3693–3702, Jun. 2020, doi: 10.1093/bioinformatics/btaa230.

[36] B. Adhikari, "DEEPCON: Protein contact prediction using dilated convolutional neural networks with dropout," *Bioinformatics*, vol. 36, no. 2, pp. 470–477, Jan. 2020, doi: 10.1093/bioinformatics/btz593.

[37] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: http://arxiv.org/abs/1605.07146

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[39] R. Shanmugamani, *Deep Learning for Computer Vision: Expert Techniques to Train Advanced Neural Networks Using TensorFlow and Keras*. Birmingham, U.K.: Packt, 2018.

[40] A. Zadeh Shirazi, M. Hatami, M. Yaghoobi, and S. J. Seyyed Mahdavi Chabok, "An intelligent approach to predict vibration rate in a real gas turbine," *Intell. Ind. Syst.*, vol. 2, no. 3, pp. 253–267, Sep. 2016.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] B. H. Spain, K. S. Bowdish, A. Pacal, S. S. Flueckiger, D. Koo, K.-Y. R. Chang, and W. Xie, "Two human cDNAs, including a homolog of arabidopsis FUS6(COP11), suppress G-protein- and mitogen-activated protein kinase-mediated signal transduction in yeastand mammalian cells," *J. Mol. Cell. Biol.*, vol. 16, pp. 6698–6706, May 1996.

[43] H. Bi, "SUMOylation of GPS2 protein regulates its transcription-suppressing function," *Mol. Biol. Cell*, vol. 25, no. 16, pp. 2499–2508, Aug. 2014, doi: 10.1091/mbc.e13-12-0733.

[44] S. Jarmalavicius, U. Trefzer, and P. Walden, "Differential arginine methylation of the G-protein pathway suppressor GPS-2 recognized by tumor-specific T cells in melanoma," *Faseb J.*, vol. 24, no. 3, pp. 937–946, 2009.

[45] M. D. Cardamone, B. Tanasa, C. T. Cederquist, J. Huang, K. Mahdaviani, W. Li, M. G. Rosenfeld, M. Liesa, and V. Perissi, "Mitochondrial retrograde signaling in mammals is mediated by the transcriptional cofactor GPS2 via direct Mitochondria-to-Nucleus translocation," *Mol. Cell*, vol. 69, no. 5, pp. 757–772, Mar. 2018, doi: 10.1016/j.molcel.2018.01.037.

[46] M. Seeger, R. Kraft, K. Ferrell, D. B. Otschir, R. Dumdey, R. Schade, C. Gordon, M. Naumann, and W. Dubiel, "A novel protein complex involved in signal transduction possessing similarities to 26S proteasome subunits," *FASEB J.*, vol. 12, no. 6, pp. 469–478, Apr. 1998, doi: 10.1096/fasebj.12.6.469.

[47] M. Kottemann and A. Smogorzewska, "Fanconi anaemia and the repair of Watson and Crick DNA crosslinks," *Nature*, vol. 493, vol. 7432, pp. 356–363, 2013, doi: 10.1038/nature11863.

[48] A. Fabregat, K. Sidiropoulos, G. Viteri, P. Marin-Garcia, P. Ping, L. Stein, P. D'Eustachio, and H. Hermjakob, "Reactome diagram viewer: Data structures and strategies to boost performance," *Bioinformatics*, vol. 34, no. 7, pp. 1208–1214, Apr. 2018.

[49] H. Kaneko and N. Kondo, "Clinical features of Bloom syndrome and function of the causative gene, BLM helicase," *Expert Rev. Mol. Diag.*, vol. 4, no. 3, pp. 393–401, May 2004, doi: 10.1586/14737159.4.3.393.
[50] H. Joenje and K. J. Patel, "The emerging genetic and molecular basis of Fanconi anaemia," *Nat. Rev. Genet.*, vol. 2, pp. 446–457, May 2001.

ALI GHUALM is currently pursuing the Ph.D. degree with the School of Computer Science, Shaanxi Normal University, Xian, China. His research interests include human disease pathway network modeling and biological pathway databases discovery.

XIUJUAN LEI (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, in 2005. She is currently a Professor and a Ph.D. Supervisor with Shaanxi Normal University. Her research interests include bioinformatics and intelligent computing.

YUCHEN ZHANG (Graduate Student Member, IEEE) received the bachelor's degree from Xi'an Technological University, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Science, Shaanxi Normal University, Xi'an, China. His current research interests include intelligent computing, data mining, pattern recognition, and bio-computing.

SHI CHENG (Member, IEEE) received the Ph.D. degree in electrical engineering and electronics from the University of Liverpool, Liverpool, U.K., and the Ph.D. degree in electrical and electronic engineering from Xi'an Jiaotong-Liverpool University, Suzhou, China. He is currently working with Shaanxi Normal University, Xi'an, China. His current research interests include swarm intelligence, scheduling, and data mining techniques and their applications.

MIN GUO received the Ph.D. degree from Shaanxi Normal University, Shaanxi, China, in 2003. She is currently a Professor and a Ph.D. Supervisor with Shaanxi Normal University. Her research interests include image processing, pattern recognition, and intelligent information processing.