SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com
ISSN Online: 3006-4708
ISSN Print: 3006-4694

An explainable phosphorylation peptide associated with SARS-CoV-2 infection employing a 2D Convolutional Neural Network (2DCNN)

Ali Ghulam[1], Tarique Ali[2], Taha Hussain[3], Nida Jabeen[4], Taiyaba Qureshi[5], Mujeeb ur Rehman[6], Rahu Sikander[7], Sultan Ahmed[8]

[1,2] Information Technology Centre, Sindh Agriculture University, Tandojam, Sindh, Pakistan, Correspondence Author, garahu@sau.edu.pk
[3] Electrical and Electronics Engineering, Uskudar University, Istanbul, Turkey
[4] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[5] School of Computer Science, University of Science and Technology of China, Hefei, Anhui, China
[6] School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China
[7] Computer Science and Software Engineering Jinnah University for women, Karachi, Pakistan
[8] Changshu Institute of Technology, Suzhou. P.R China

DOI: https://doi.org/10.63163/jpehss.v3i1.157

Abstract

Phosphorylation is a post-translational modification process plays a critical role in the regulation of many cellular processes, including viral infection for example SARS-CoV-2. The SARS-CoV-2 is the virus responsible for causing the COVID-19 pandemic. The identification and characterization of phosphorylation sites on SARS-CoV-2 proteins could provide valuable insights into the mechanisms underlying the virus's pathogenesis and may lead to the development of new therapeutic strategies for COVID-19. The development of computational predictors for phosphorylation site identification has received remarkable attention recently, however these methods limited to find phosphorylation sites in SARS-CoV-2-infected host cells. Viral-host protein-protein interactions cause alterations in phosphorylation and may influence host protein subcellular localization. In this work we proposed a predictor called 2Deep-IPs using two-dimensional convolutional deep neural network (2D-CNN) for identification of particular phosphorylation sites. We extracted the amino acid composition-based features from protein sequence by using dipeptide deviation from expected mean (DDE) descriptor. Further, we used shapely additive explanation's (SHAP's) algorithm to rank the effective attributes that adequately contain crucial biological information. The proposed model outperformed on top 15 high ranked features. The empirical outcomes of 2Deep-IPs based on 10- fold cross-validation achieved accuracy score **96.71**, Sen score obtained **94.46** and Spec score obtain is **99.69** and MCC score obtain **0.939**. The results analysis based on independent datasets achieved accuracy score **95.70**, Sen score obtained **97.83** and Spec score obtain is **91.89** and MCC score obtain **0.782**, respectively. Thus, the anticipated results reveal that 2Deep-IPs outperforms other phosphorylation sites predictors both on cross-validation and independent test respectively. We hope that the proposed Deep-IPs will provide in-depth knowledge to other methods that can be used to predict general phosphorylation sites.

Keywords: Phosphorylation sites, SARS-COV-2, 2DCNN, DDE, 2DEEP_IPs.

Introduction

A highly contagious and pathogenic coronavirus that developed in late 2019 has led to a pandemic called "COVID-19 2019", an acute respiratory disease, which is a huge health and socioeconomic problem [1, 2]. Similar to the severe disease brought on by SARS-CoV-2, severe coronavirus disease 2019 (COVID-19) is characterized by immediate respiratory distress and excessive inflammation that can result in respiratory failure, multi-organ failure, and death. The coronavirus pandemic is a highly contagious and highly pathogenic disease. Research on phosphorylation has increased dramatically over the past few decades as a result of the importance of phosphorylation in comprehending biological systems of proteins and providing advice for the design of fundamental therapeutic drugs. And numerous attempts have been made to locate phosphorylation sites using experimental techniques and computational prediction tools [3]. Although there is a lot that can be inferred about SARS- CoV-2 based on similarities to SARS-CoV-2, SARS-CoV-2 is a novel coronavirus with special traits that aid in its pandemic-scale propagation. SARS-CoV-2 infection is typically asymptomatic, unlike SARS-CoV, especially in the younger population [4-5]. By assessing changes in protein abundance and phosphorylation, the pathway of pathogenesis can be elucidated [6], providing a powerful tool for proteomic approaches. To explore the relationship between SARS-COV-2 and host cells, [7] described a system of interaction groups, proteomes, and signaling processes. To demonstrate a significant response of phosphorylation in host and viral proteins, Miao, M., et al. [8] presented a quantitative, mass spectrometric, phosphoproteomic study of SARSCOV-2 infection in Vero E6 cells. Klann et al. [9] studied phosphorylated proteome change signals using the CACO-2 human cell system for SARS-CoV-2 infection. Hekman-metal [10] conducted quantitative SARS- CoV-2 infection phosphoproteomic studies on iAT2 cells to utilize the infection and pathogenesis process. And rapid output Mass spectrometry methods can precisely identify phosphorylation sites, yielding a huge number of examples of phosphorylation. Traditional experimental techniques, however, require a lot of work and time, especially when used to verify a large number of potential phosphorylation sites. Alternatively, computational methods are gaining popularity as a means of overcoming the limitations of experimental tactics. Few attempts have been made to investigate the 40 computational techniques that have been developed to date for locating phosphorylation sites, and a sizable portion of them are based on machine learning algorithms, such as Support Vector Machine (SVM),, Bayesian decision theory, logistic regression and Random Forest [11]. In combination with enhanced logistic regression model, Quokka used various sequence measurement functions to predict phosphorylation sites [12]. In GPS 5.0, after identifying phosphorylation weights, two new techniques, namely position weights determination and matrix scoring optimization, were used to adopt logistic regression method [13]. Although these methods can accurately predict the performance of phosphorylation sites, "feature engineering" involves manual patterns, which may result in skewed features [14] being limited. Deep learning is a feasible and attractive way to solve this problem. Deep learning has obvious benefits over the laughable "feature engineering" of traditional engineering methods. It can automatically construct complex patterns and obtain highly abstract information from training information. For example, MusiteDeep uses the input of raw sequence data and the use of CNNs to predict phosphorylation sites with new two-dimensional attention mechanisms [15]. CapsNet has created a multi-layer NETWORK of CNN capsules to identify proteins after altered sites and has provided several excellent capsules that describe biologically relevant properties [16]. DeepPSP has constructed a deep neural network based on global and local information to predict phosphorylation sites [17]. These methods are superior to typical machine learning methods that use only raw sequences. However, in host cells infected with SARS-CoV-2, there is no special deep learning structure to recognize phosphorylation sites. When SARS-CoV-2 vaccination and infection were combined with past exposure to other human coronaviruses, a 15-amino acid stretch of the SARS-CoV-2 spike glycoprotein was found to elicit significant T-cell reaction [18]. An immunodominant coronavirus specific peptide epitope may be generated from the corresponding peptide sequence, which is found in the spike glycoprotein fusion domain and may be present in various HLA alleles. Previous research [19] has supported the hypothesis that Follow-up experiments confirm the findings. The DEE deploy features extraction algorithm, the amino acid composition AAC algorithm, and the pseudo amino acid composition PseAAC algorithm have all received significant attention from researchers. To better predict anticancer peptides, we introduce a new deep learning-based method here called ACP-2DCNN. Dipeptide deviation from expected mean (DDE) is used to extract relevant features, and a two-dimensional convolutional neural network is used for model training and prediction (2D CNN) [20]. Here, we provide a new proposed method structure, called DeepSARS-CoV-2-IPs 2D-CNN, which includes the Two-Dimensional Convolution Neural Network (CNN) that reliably predicts the phosphorylation sites of SARS-COV-2 host cells (Figure 1). We have generated a number of independent data sets to evaluate IPs 2D-CNN performance. The results show that the recognition ability of general phosphating sites is very strong through word combination and IPs 2D-CNN structure. We believe that the proposed 2Deep-IPs based on 2D-CNN architecture can also better address the additional difficulties in bioinformatics than earlier solutions. In addition, we present an early example of popular word 2D-CNN methods in biological sequence analysis and may highlight the challenges of other biological predictions.

Materials and methods

Data collection and preparation

In this study, literature collection was conducted at the experimental test site for human A549 SARS-CoV-2 phosphorylation [21] including SARS-CoV-2 10474 phosphorylation sites. The experimentally confirmed phosphorylation sites of SARS-CoV-2-infected human A549 cells were taken from the literature for this work [21]. A considerable datasets similarity redundancy [22] of research has employed the CD-HIT program that removed the similarity [22] with a ratio of 30% sequence consistency standard to reduce redundant sequences for protein phosphorylation and prevent model overplay. The processed protein sequences were cutoff 33-residue-long sequence pieces with S/T or Y in the middle to make comparisons with other phosphorylation site prediction algorithms easier. The fragment is defined as positive if the central S/T or Y is phosphorylated, otherwise it is defined as a negative sample [23-24]. We used in cross-validation that means (Training sets) datasets 4308 positive samples, 4308 negative samples, S/T sites were obtained and then independent tests datasets 1079 positive samples, 1079 negative samples, S/T sites were obtained [25-26]. Meanwhile, in the deep learning scenario of sequence analysis, a general performance evaluation technique is implemented to separate the data set into strictly deployed training sets and conduct on independent tests at a random ratio of 8:2. In the Table 1 provides a detailed description of the datasets. The sample was sequenced positive for the SARS-COV-2 specific protein, and one was sequenced negative, but no definitive link was identified. Positive and negative samples were randomly selected and balanced with different test datasets. In this study, only proteins relevant to humans were studied, but the unified Swiss-Prot online database containing different species was used.

Table 1. The training set and independent testing set of the benchmark datasets of the S/T sites were partitioned at random in the ratios of 8:2 respectively.

Data Types	Residue type	Positive samples	Negative samples	Total Data Points
Cross-Validation (Training)	S/T	4308	4308	8616
Independent (Testing)	S/T	1079	1079	2158

Feature extraction for phosphorylation sites

The current perception also involves extracting features as a key step in the method configuration; That is, information about protein sequences is converted into integer data as with probability matrix. There [27] is limited research investigating distributed into two subtypes: dipeptide deviation extracted from the predicted mean (DDE). A two-dimensional vector score matrix consisting of 20 × 20 features length image matrix is obtained and extended to a vector. Instead, an efficient measurement matrix is designed, and a compact function is obtained by random projection. Therefore, a new method to extract the compressed sensing function is proposed.

Dipeptide deviation from the expected mean (DDE)

Therefore, this study proposes and develops a new amino acid descriptor based on sarS-CoV-2 composition and non-SARS-CoV-2 with DDE vector score. The results showed that the effectiveness of DDE feature vectors in improving the prevention of specific linear protein sequences was compared with other feature expressions. DDE function carriers perform better (with accurate differential cross validation and separate data sets) than other amino acid derived features in various data sets. For different protein functional methods, amino acid frequencies vary [24] to extract characteristics and protein relationships with vectors commonly used to predict the DDE of their respective acids [28]. In order to study dipeptide composition characteristics, past studies have been used to measure dipeptide frequency changes, which are consistent with earlier results in dipeptide composition. Theoretical mean value (Tm), theoretical variance (Tv) and dipeptide composition (Cc). 3. The indicator DDE and DC_i of dipeptide C_c in peptide P were calculated as follows, and the indicator DC_i of dipeptide i *Cc* in peptide *P* was calculated as follows.

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

The functional length (**20 × 20** common amino acids) is characterized by 400 dipeptides, but they are not all extracted in any sequence. Dipeptides ***I*** *and* ***N*** do not appear as ***L-1*** (i.e., potential amounts in ***P***). Theoretical mean $\boldsymbol{TM_{(I)}}$

$$T_{mi}(r,s) = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \tag{2}$$

The number of $\boldsymbol{C_{i1}}$ codons and the quantity of $\boldsymbol{C_{i2}}$ and the quantity of the additional amino acid of the $\mathbf{C}_{i2}$ codon given dipeptide ***i*** is the quantity of the first amino acid ***CN*** is the total quantity of accessible codons except for the three stages. $\boldsymbol{T_{mi}}$ was independent of peptide P, so 400 dipeptide lengths were extracted and pre calculated. The theoretical variance of $\boldsymbol{T_{vi}}$ is: dipeptide ***i***

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

Equation $\boldsymbol{i}$ (theoretical average value) is estimated as $\boldsymbol{T_{m\,(i)}}$ by Equation (2). $\boldsymbol{N}$ is one less than the quantity of dipeptides in peptide $\boldsymbol{P}$; Thus, the quantity of dipeptides in peptide $\boldsymbol{P}$ is increased again, and N is L-1. $\boldsymbol{DDE_{(i)}}$ is now determined to be

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

The 400-dimensional feature vector is used to calculate the DDE for each property of the 400 dipeptides.

$$DDE_P = \{DDE_{(i),} \dots, \dots DDE_{(n)}\}, where\ i = 1,2, \dots, 400 \tag{5}$$

2D-CNN based framework

The study included two-dimensional 2D-CNN and DDE vector profiles feature maps and designed an important method for the classification of individual proteins. It includes four methods: data collection, feature extraction, 2D-CNN model construction algorithm configuration and then model validation and evaluation. Figure 1. shows our proposed framework model and provides the following details. Topics of this work include the DDE features extraction of profile used with in 2DCNN method. A key method for identifying and classifying individual proteins associated with the human SARS-CoV-2 has been developed. DDE encoding was used to process the features vector score matrix extraction profile of physical and chemical qualities. The DDE features vector score matrix extraction profile of physical and chemical properties was processed by 2D-CNN model. Dipeptide Deviation from Expected Mean (DDE) is used to extract the key features, and a two-dimensional convolutional neural network is used for model training and prediction (2D CNN).

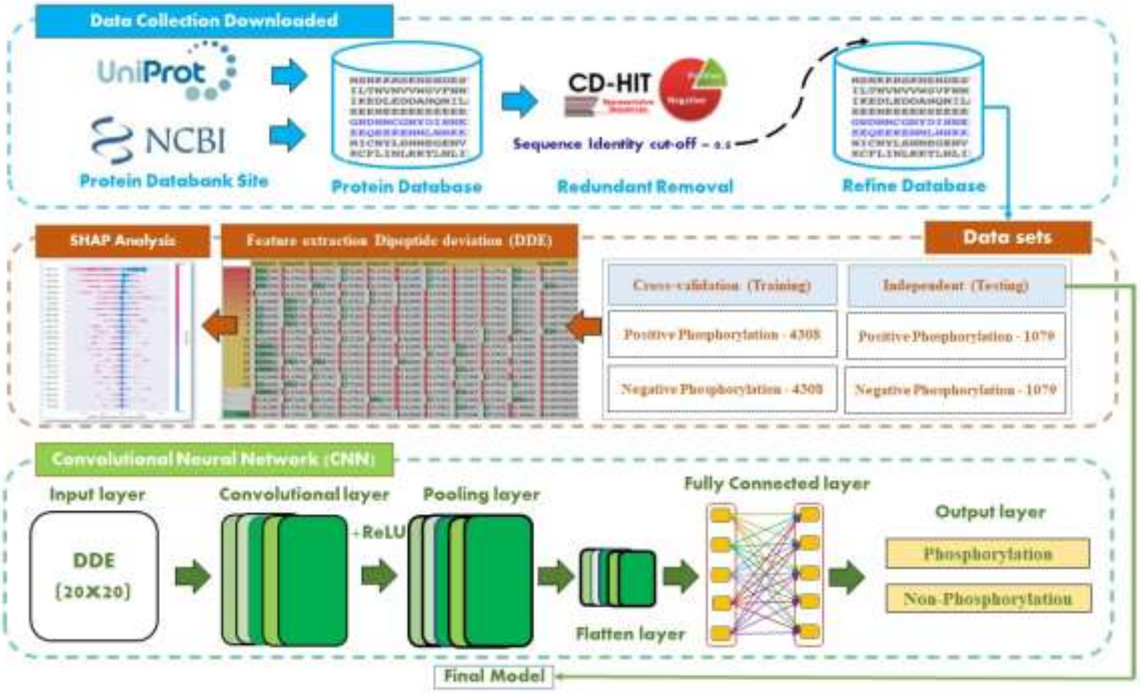
Data Collection Downloaded
UniProt
NCBI
Protein Databank Site
Protein Database
CD-HIT
Sequence Identity cut-off – 0.5
Redundant Removal
Refine Database
Data sets
SHAP Analysis
Feature extraction Dipeptide deviation (DDE)
Cross-validation (Training)
Independent (Testing)
Positive Phosphorylation - 4308
Negative Phosphorylation - 4308
Positive Phosphorylation - 1079
Negative Phosphorylation - 1079
Convolutional Neural Network (CNN)
Input layer
DDE
(20X20)
Convolutional layer
+ReLU
Pooling layer
Flatten layer
Fully Connected layer
Output layer
Phosphorylation
Non-Phosphorylation
Final Model

Figure 1. Proposed framework model of 2D-CNN SARS-CoV-2

2D-CNN Framework model which construct the basic structure of a convolutional neural network (CNN). Each CNN had three hidden layers, then an input layer (which was included in a fully connected layer and a pooling layer), and finally a convolutional layer. The 2D-CNN phosphorylation typically consisted of multiple layers, each of which performed a unique function in order to convert the input into an actionable representation. Our 2D convolutional neural network

SARS-CoV-2 architecture was then constructed with a ranked architecture. Hyperparameters can be estimated by analyzing a model's predictive accuracy and power, as has been demonstrated by numerous prior studies in this field. First identify the image-based matrix and tensor flow structure, and then make a judgment. SARS-CoV-2 is widely used in image transformation so that each input image is contained at the same window size and feature distance. The streamlined framework model of 2D-CNN is shown in the figure 1. In addition to this approach, our paper proposes using two-dimensional convolutional neural networks (2D-CNNs) to do image-based classification [29, 30]. A 2D-CNN convolutional neural network, a common neural network for deep learning, was used in this investigation. The simplified convolutional neural network's layer structure is shown in the lower portion of Fig. 1. The Keras library and (ReLU) function backend was used in our deep learning architecture. Additionally used to more effectively improve the performance were GPU computing and the CUDA kernel. The input layer parameters for this investigation were taken from DDE vector profiles and transformed into 20x20 matrices. We intend to offer a method to categorize SARS-CoV-2 proteins into different molecular activities utilizing these matrices as the input data. To train the 2D-CNN model with various weights and biases to improve its predictive performance, we assumed the 20x20 matrix to be an image with 20x 20 pixels. Instead of a 1D structure, a 2D CNN model is used to capture the hidden features within DDE profiles. Our deep learning system is implemented using Kera's package, which uses Tensor Flow as the back end. 2D-CNN SARS-CoV-2 consists of multiple layers, each with a specific purpose, so that each layer transforms its input into valuable information. We designed the 2D-CNN SARS-CoV-2 model using specific sorting rules. Various studies have found that it is important to apply optimizations to find the optimal architecture and hyper parameters when creating an effective model, as shown in [31].

1) CNN prediction model

In this paper, we use a two-dimensional convolutional neural network (2D CNN), a common type of deep neural network, to propose a framework for implementing DL in bioinformatics. When compared to conventional machine learning strategies in bioinformatics, we expect our approach to produce substantial improvements. The CNN module converts the 2D structure information of the window into the image easily understood by CNN, discarding a large amount of relevant information. Amino acid sequences will be associated with phosphorylation sites specific SARS-COV-2 associated proteins and enable prediction of SARS-COV-2 infection. In addition to the basic structural elements, we have introduced a new structural feature, namely the relative angles of the amino acids. Various research projects have been carried out to determine protein types, predict binding sites, and predict protein-protein interactions using sequence knowledge of CNN models. Many studies have applied SARS-COV-2 -specific and transporter protein classification and pathway-specific structure prevention to various problems in bioinformatics. This approach is beneficial because once the features are used automatically, the image will be processed in the appropriate format. One dimensional convolution is used to connect features to amino acid sequences, while two-dimensional convolution is used for other features or to perform additional mapping operations. The convolution layer works well for CNN [32] because the overall goal is to find patterns in the image, even if the input comes from a specific location. In order to discover interactions between patterns of spatially separated sequences, a specified number of residues must map perfectly to observable structural patterns.

2) 2d CNN optimization process

This two-dimensional CNN approach provides end-to-end differentiability, meaning that different regions of the organization can optimize features and then predict two-dimensional coordinates up to the training model [33]. Our technology has been optimized with DL models (deep learning) [34].

3) Input layer

During the analysis, all limitations of the input layer are remapped to a 20x20 matrix that can be used to distinguish SARS-CoV-2 proteins in the binding pathway based on the input data [35]. Similarities are used to identify families of proteins as pathway-specific proteins, which are divided into different subgroups. In order to be trained in the cross-validation program, the machine was 10-fold cross-validated. The study used a deep neural network called 2D-CNN, which is the largest of its kind. In the field of computational vision, especially when the input is a two-dimensional image-based matrix, has shown outstanding effects in various applications of CNN. Based on these results, a CNN architecture for training using input images is constructed, and a 20×20 window PSSM matrix can easily be entered with 2D structures of this size [36, 37]. Two-dimensional CNN models are used instead of one-dimensional models for DDE matrix contours because they can reveal hidden shapes. The 2DCNN design architecture used in DDE profile creation connects them between the input and output layers.

4) Zero padding layer

Zero padding is a symmetric addition of zeros to the input matrix, allowing the input size to be modified. In the model under study, zero values are added at the beginning and end of a window size 20×20 matrix. This allows filters to be applied to matrix boundaries placed adjacent to each other [38].

5) Convolutional layer

The encoding layer is used to compute features from the 2D input matrix, and then the output is fed into the 2d convolution layer. Use a sliding window to get representative values from these values, and then move the values in turn over the input. Using small square inputs, the convolution activity maintains a interaction between the mathematical inputs in the mixed feature profile. We used a 3x3 sliding window to build our model. Neurons are collected and input signals from the existing layer are fine-tuned with weights and biases [39, 40].

6) Activation layer

To investigate the role of two-dimensional convolutional neural networks (2D-CNN) in SARS-CoV-2 protein classification, we employed an activation mechanism to detect contextual information showing the carrying function of 2D-CNN activation (ReLU) [41]. ReLU has been shown to be the most commonly used deep neural network (DNN) activation function. The following formula gives the definition of the ReLU function, which is equal to the number of inputs to the neural network [42].

$$f(x) = max\ (0, x) \qquad (6)$$

7) Pooling layer

We reduced the amount of matrix calculation; the pooling layer is often inserted into the convolution layer. The operation of this layer is called " downsampling " because it removes some data, thereby reducing computation and model fitting, and preserving specific representational characteristics. One of the many techniques for downsampling is to take the maximum (maximum pool) (average pool) in the window. To get the image quality we want, we use the maximum pool and downsample the data to a factor of 2, where the maximum is selected in a window of 2×2[43].

8) Dropout layer

The predictive performance of existing models is enhanced by identifying and including key exit factors, and overfitting is avoided by avoiding overfitting [44]. At the exit layer, the model randomly fails with a specific probability P. If the training time is extended and the dropout is used in the layer, the dropout ignores the selected neuron [45]. To normalize deep neural networks, many training methods use dropout, but applying dropout to fully connected and convolutional layers results in a fundamentally different process. In addition, she is considered a drop-out in the deep learning community. To simplify, apply the dropout function to a fully connected layer and get a value of 0.02.

9) Flatten layer

The flattening layer further reduces the dimension of data by transforming it into a one-dimensional array. The convolution layer flattens the contributions of each layer to produce a long vector. A fully connected layer, called the last classification model, is built on top of the model. In order for the system to properly analyze the output layer, all classes should be assigned to RUN, and the input matrix should be flattened by using the flattening layer [46].

10) Fully connected layer

A fully linked layer is a network in which each node is connected to all previous layers. Also known as the last layer in CNN, the fully connected layer is usually used near the end of the network. To merge the two joined layers, the current model contains two fully joined layers [47]. Our model can grow more knowledge and perform better because the first layer connects all the nodes to produce a flat layer. Connecting the first connection layer to the output layer requires a second layer. There are two nodes at the output layer because predicting ATP binding sites can be seen as a binary classification problem [48].

11) Loss function

In order to overcome a wide range of classification problems simultaneously (yes or no, A or B, 0 or 1), the model is trained with bivariate cross-entropy. In binary tasks, the binary cross entropy is a loss function. Questions that require two answers (for example, yes or no, A or B, 0 or 1, right or left). The loss function has been proven for some binary classification tasks. Minimizes SoftMax (generated by a hot code). To distinguish them, we apply cross-entropy. When it comes to class markers, we use entropy to maximize likelihood as a goal. In connection to the loss function, you can see that if your model is overtrained, it will be very small to zero, and can be done in a simple way by minimizing the loss function. Regularization measures, such as penalties included in loss functions, can be used to combat overfitting [49,50].

12. SoftMax utilization

These findings would suggest that SoftMax function, which reduces the probability of any results, is applied to the results of the model. The characteristic form of the logistic function and the OUTPUT layer of the ANN layer are used as the equation definition logistic function and multi-class classification problem. Deactivation, which I call a probability distribution, represents an equation in which x is a k-dimensional vector with one entry for each possible sample value whose expected probability is defined as ***σ (x)*** j entry. For example, if we take 0 as the lower limit of the range *(0, 1),* then ***j-th*** is the true value of the range *(0, 1).* Trainable parameters in the model as shown in Table 2. [51,52].

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

Layer (type)	Output Shape	Param #
zero_padding2d_113 (ZeroPadd	(None, 3, 22, 20)	0
conv2d_122 (Conv2D)	(None, 1, 20, 32)	5792
activation_197 (Activation)	(None, 1, 20, 32)	0
max_pooling2d_122(MaxPoolin	(None, 1, 10, 16)	0
zero_padding2d_114 (ZeroPadd	(None, 3, 12, 16)	0
conv2d_123 (Conv2D)	(None, 1, 10, 64)	9280
activation_198 (Activation)	(None, 1, 10, 64)	0
max_pooling2d_123(MaxPoolin	(None, 1, 5, 32)	0
flatten_103 (Flatten)	(None, 160)	0
dropout_184 (Dropout)	(None, 160)	0
dense_203 (Dense)	(None, 64)	10304
activation_199 (Activation)	(None, 64)	0
dense_204 (Dense)	(None, 2)	130
activation_200 (Activation)	(None, 2)	0

$$\sigma(Z)_i = \frac{e^{Z_i}}{\sum_{k-1}^{k} e^{Z_i}} \qquad (7)$$

13) Hyperparameter optimization

For hyperparameters, deep neural networks are very sensitive and effective. More specifically, automatic measurement of hyperparameters is very important. While it is necessary to find the best value for the function, methods have been devised that does not rely on derivatives. The time and computational resources needed to manually adjust the hyperparameters of deep neural networks are very valuable, but the process of doing so requires time and a great deal of expertise. However, the use and popularity of deep neural networks have promoted the implementation of automated mechanisms to meet individual needs [53]. Two parameters can be used to identify deep neural network hyperparameters: structure formation group and optimization process influence group. Most of the hyperparameters are orders of magnitude different from those used when training models using backpropagation algorithms at the architecture level. In the design of deep learning model, the selection of hyperparameters is guided by many aspects [54]. The metric performance of this model was excellent. Chollet recommended using several parameters to optimize training and avoid overfitting. For example, the difference between HPO and IPSO can be understood as emphasizing different aspects of learning.

2.6. Assessment of predictive ability

Each protein sequence extracted from the GTP-binding protein set was placed in a different dataset: a separate test dataset and a separate training dataset. Initially, we used a quintuple cross-validation strategy to build our model and maximize its effectiveness. This method suggests that 5-fold cross-validation for the model validation, then select one portion for testing and the other four for training. In addition, the experimental training data sets are tested with independent data sets. We selected criteria such as sensitivity, specificity, accuracy and MCC to evaluate predictive power. Our entry describes true positives, false positives, true negatives, and false negatives as corresponding true positives, false positives, true negatives, and false negatives. To achieve the best balance between sensitivity and specificity, thresholds were selected based on sensitivity, specificity, accuracy and MCC [55].

$$Sensitivity = \frac{TP}{TP+FN} \tag{7}$$

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

Results

The conclusions we draw can be compared with earlier results to analyze the feasibility and integrity of the required research modeling methods. Data evaluation, calculation and comparison are the main methods of the experiment. Based on our model, which includes DDE features extraction method for the prediction is that the model.

SHAP Feature Importance

Shapely Additive exPlanaitions (SHAP's) [56, 57] is a well-known algorithm for analyzing the biological attributes of proteins samples. To get an overall sense of significance, we take the mean of the absolute Shapley values for each feature in the dataset by following mathematical expression. The important characteristics are those with high absolute Shapley values.

$$I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\phi_j^{(i)}\right| \tag{11}$$

The features are then ranked in order of decreasing relevance and displayed graphically. The significance of the SHAP feature in the previously trained 2Deep-IPs model to improve the prediction of SARS-Cov-2 for predicting phosphorylation site depicted in the following figure 2. It is possible to replace permutation feature importance with SHAP feature importance. The two metrics of value, however, cannot be compared in any meaningful way: The relevance of permutation features is determined by how much the model's performance suffers. When determining SHAP, feature attribution magnitude plays a key role. Information was retrieved from raw protein sequences using 25 different characteristics. Next, we used the CNN algorithm to obtained the best features scores. For the purpose of checking the model's validity, we devised a 10-fold cross validation test in addition to an independent test. Lastly, we employed both the standard CNN approach and SHAP values to determine the significance of features and understand the model's meaning.

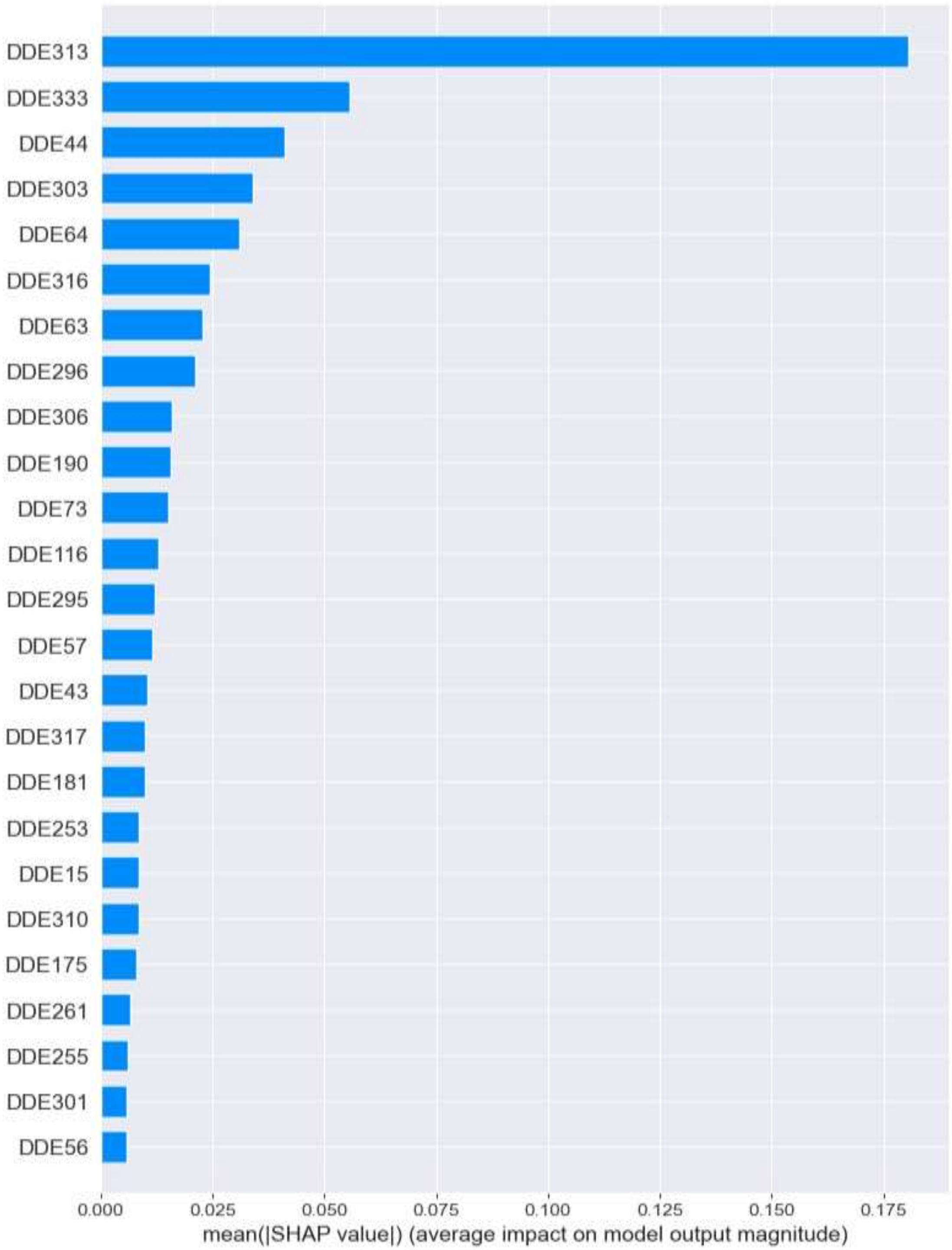

Figure 2. Top ranked features using SHAP algorithm for 2Deep-IPs model

SARS-CoV-2 and Non- SARS-CoV-2 feature analysis using SHAP

In order to analyze the impact of dominant features that help the 2Deep-IPs model to improve the prediction of SARS-Cov-2 samples are evaluated through SHAP [56, 57] method.

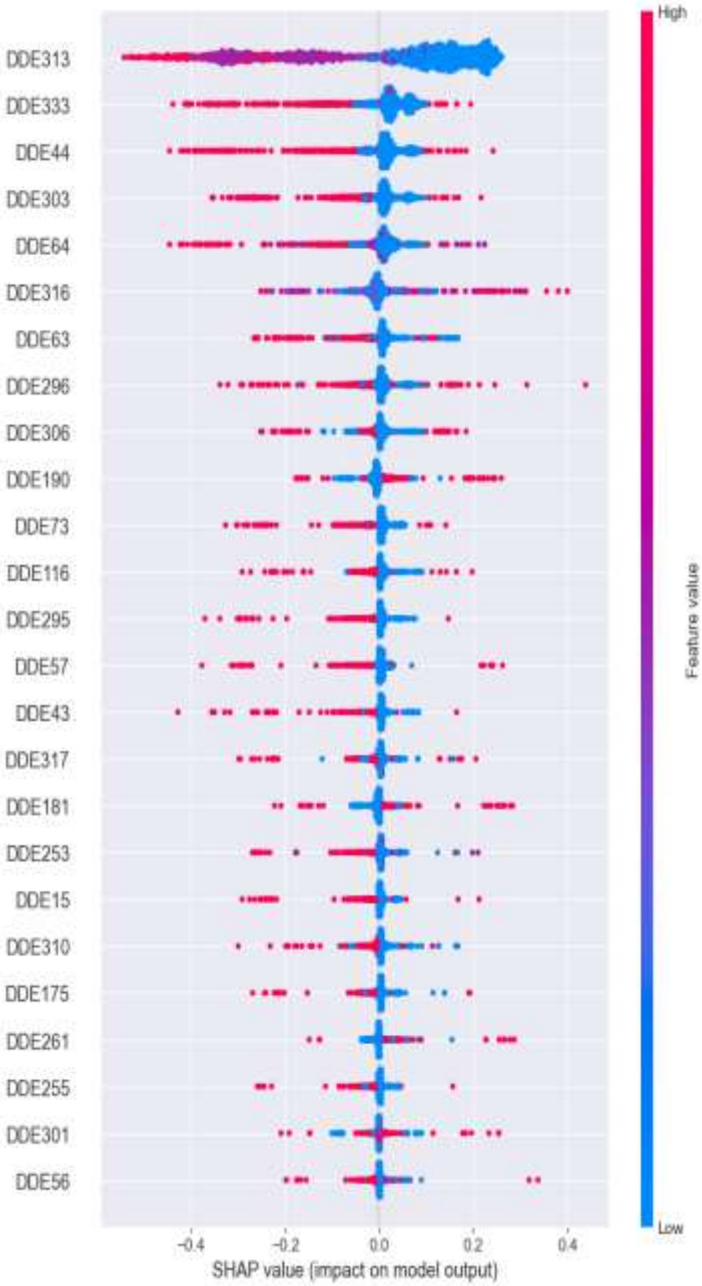

Figure 3. Analysis of ranked features using SHAP algorithm for 2Deep-IPs model

From figure 3, it is apparent that the negative and positive SHAP values for the top ranked attributes contribute the characterization of non-SARS-CoV-2 and SARS-CoV-2 and protein sequences. We also noticed that majority of the features having high rank value especially DDE313, DDE333, and DDE44 extract the key features of SARS-CoV-2 and non- SARS-CoV-2 proteins as shown in Table 3. Furthermore, to calculate the frequency of amino acid composition, we evaluated the both classes. The amino acid index can be found in a compilation of 20 values that describe the various physicochemical and biological properties of amino acids. In two different data sets, 20 amino acids combined to make a significant contribution. The difference between the two data sets is small, but there are some notable outliers. Each protein contains the highest concentrations of the C and P amino acids. This is why it is crucial to identify the SARS-CoV-2 protein among these amino acids. Because of these differences, our model was able to successfully predict SARS-CoV-2 proteins using these amino acids.

Table 3. Analysis of ranked features importance scores.

Features IDs	Features Names	Features importance scores
312	DDE313	95.330980
332	DDE333	95.835080
43	DDE44	71.063100
302	DDE303	58.440474
63	DDE64	53.387209
315	DDE316	42.203436
62	DDE63	39.488891
295	DDE296	36.411014
305	DDE306	27.400451
189	DDE190	27.217468
72	DDE73	25.941024
115	DDE116	22.352317
294	DDE295	20.732134
56	DDE57	19.704612
42	DDE43	17.919594

2D-CNN train the model

It may also help to understand the training of model features. We propose to use 150 eras as model trains in our models. The training data is correct because it is memory mapped to the hidden data; however, validation and validation are unreliable because they do not interpret memory mappings [58]. We acknowledge that this is a possibility, and we plan to address it. In the following steps, the dropout rate is applied to the network model, leaving the other layers unchanged. The next step is to check the efficiency of the model before giving our conclusions.

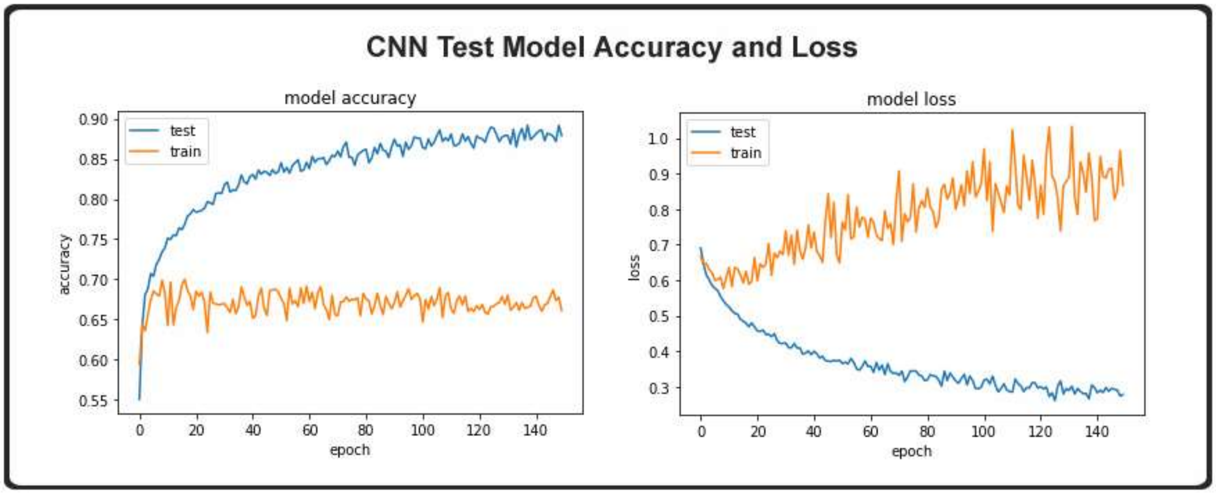

Figure 4. CNN model accuracy and model loss

Test set model evaluation

In Figure 4, we presented the proposed CNN model has a test accuracy of 97.59 and a test loss of 10.62. The accuracy of this test is amazing. Although we have tried to solve the problem of overfitting before, these findings are not as surprising as they seem [59]. In fact, we might expect the study to proceed better if we had the dropout rate as the second tier. Dropping out can reduce the need for training by a certain amount. In the case of a neuron displaying more than one part that wants to exit, the adjustable hyperparameter is defined as the neuron displaying more than one part that wants to exit. Disabling inactive neurons helps prevent the network from remembering training data. Once the network is trained, we redevelop, compile, and train the network, but we ignore the dropouts at this point. Our batch size is set to 10 and we use 150 epochs.

Performance result for identifying sars-cov-2 proteins with 2D-CNN

Previously found Keras backend in Tensorflow is consistent with the results. We've built a 2D-CNN system. Then, the optimal hidden layer configuration is selected, which includes a convolution layer 32 filter numbers, 64 filter numbers and 128 filter numbers long as shown in Table 4 [60]. An independent sample was used to calculate the cross-validation error rate, SARS-COV-2 was identified, and sequences were found with an average accuracy of 10 times, with an independent set accuracy of **92.11** score achieved and then compared to other filters, the results are higher than other filters. The accuracy score on the basis of cross validation that means (Training sets) was **96.24**, specificity cross validation set was **98.00**, and MCC cross validation set was **0.924**. In this project, several filter numbers were used to produce separate data sets. The metric performance shown in Table 3 which mentioned accuracy of **96.24**, sensitivity of **93.41**, specificity of **98.00**, and MCC of **0.924** score achieved. The accuracy score on the basis of independent sets that means (Testing sets) obtained **92.11**, accuracy, sensitivity of **90.81**, specificity of **93.41** and MCC score obtained **0.843.** Then we apply our model to simulate this evolutionary layer. To summarize our model, we construct five hyperparametric optimization models.

Table 4. S/T Sites Datasets predicted performance of SARS-COV-2 with different filters.

Filter numbers	**Cross-Validation (Training sets)**				**Independent sets**			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
32	90.41	90.62	90.53	0.810	91.61	93.21	92.31	0.849
32-64	94.38	97.98	96.18	0.924	92.39	93.13	92.76	0.855
32-64-128	**94.44**	**98.00**	**96.24**	**0.924**	**90.81**	**93.41**	**92.11**	**0.843**

Therefore, the model is developed by deep convolution layer structure. The models then went through an optimization process that included RMSprop, Adam, Nadam, SGD and Adadelta. To ensure that different optimizers are comparable, the model is reset after each optimization. See Figure 5. We built our final model with Adam, which is an intelligent optimizer with excellent performance. For the model we proposed, the optimal optimization was selected for Adam. The experiment used a default learning rate (0.001 step) and a batch size (10 steps) with a dropout rate of 0.2, and the number of iterations ranged from 100 to 150. In addition, we used other independent test sets data to ensure the accuracy of the model and compared the results with those of our competitors. After 150 thresholds, due to the improvement of training accuracy, the accuracy of our model verification is improved. Therefore, since we finished training around level 150, we shortened the training time to minimize over-fitting of the final results (see Table 5). The problem with all machine learning problems ultimately stems from overfitting points, which occur when our training methods fail to adequately capture the training data. Even so, in another hidden data set, the situation could be even worse. To re-check the accuracy of our model in a blind data set, we ran an independent test.

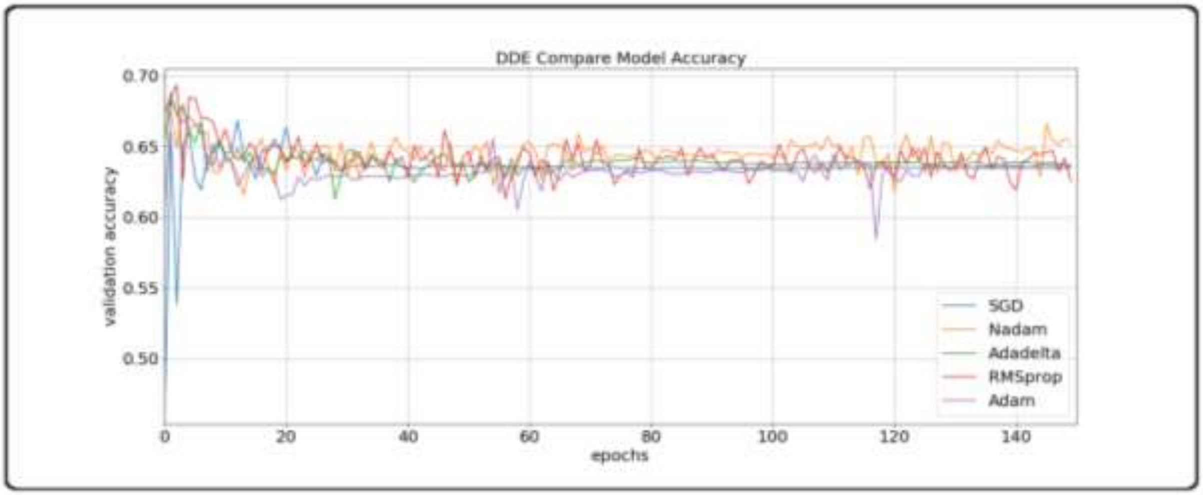

Figure 5. S/T Sites Datasets performance based on optimizers in this analysis (from 0 to 150).

Table 5. Hyperparameters optimization used in our proposed method.

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

The data were normalized by cross-validation that means (training sets) consisted of 4308 data points as positive SARS-COV-2 and 4308 data points as negative non- SARS-COV-2. Our dataset consisted of 1079 data points as independent sets positive SARS-COV-2 and 1079 data points as negative non- SARS-COV-2. In the results of cross-validation are consistent with the results of our separate test data sets. The cross-validation results did not differ much, which probably meant that our model did not overfit. One theory is that by using dropouts, replication of CNN's programming was stopped.

CNN important functions

The assumption is that deep learning models need more help. The hierarchical structure of the attributes we extract varies from local to abstract, which makes it difficult for us to accurately locate the basic features of the CNN model. In order to provide readers and biologists with more relevant knowledge, we have done our best to solve this problem. Then, after inserting 20×20 mixed features into the CNN system, the main features of these matrices are analyzed. To create a useful question result, we use an F-score to help us categorize the most relevant features. To test the model, we measured the effectiveness of several SARS-CoV-2 and non-SARS-CoV-2 sequences and how dependent they were on the model. There is no significant change in the usefulness of our features between the F-values of the two data sets. Our model can learn the importance of hidden features, support us to understand the most essential protein qualities and even select the most effective methods to achieve our desired results.

Result of identification SARS-CoV-2 with different optimizers based on barchart

For most scholars, the goal of algorithmic performance is to optimize performance based on an independent data set. It is worth noting that some of the conclusions of this study deserve additional consideration, especially in terms of algorithm design. This could be a possible explanation for the mismatch, since cross validation is a model assessment that includes out-of-sample predictions. Although it is believed that hyperparametric optimization optimizes only the loss function, the hypothesis claims that the hyperparametric optimization technique we provide optimizes both the loss function and different loss functions. The algorithm found that it could recreate the input. This is achieved through hyperparameter optimizations, such as regularization, as shown in Tables 6. Using Adam, the Adadelta optimizer, we have determined that it is the best estimate. Our research is focused on finding phosphorylation and figuring out how this phosphorylation connect to particular SARS-CoV-2 proteins. The cross-validation data set and independent data set were used for analysis.

Table 6. DDE model predicted performance of SARS-CoV-2 with different optimizers.

Optimizers	Cross-Validation (Training sets)				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
Adam	**94.62**	**97.03**	**95.76**	**0.915**	**87.84**	**90.29**	**89.05**	**0.782**
Adadelta	**94.61**	**96.47**	**95.57**	**0.911**	**94.34**	**91.92**	**9318**	**0.863**
RMSprop	94.21	96.70	95.47	0.910	89.14	90.44	89.79	0.796
Nadam	94.63	96.79	95.71	0.914	90.54	90.35	90.44	0.809
SGD	94.38	96.33	95.35	0.907	90.49	90.63	90.53	0.810

2D-CNN and shallow neural networks with a metric comparable performance

Based on this study, we evaluated multiple machine learning algorithms to identify proteins in SARS-CoV-2. We employed four machine learning classifiers (such as AdaBoost, random-Forest, and LSTM, DNN and CNN) as shown in Table 7. We developed an LSTM model and evaluated it using CNN of different sizes (1D and 2D) data points. For uniform comparisons, the independent tests in Table 6 use the best settings parameter tuning optimization for all the classifiers. Using the same experimental setup, we demonstrate that our 2D-CNN performs better than other standard machine learning algorithms. Our 2DCNN specifically makes use of a unique data set, using algorithms customized for that data set.

Table 7. S/T sites datasets metric performance based on ML classifiers.

ML classifiers	Cross-Validation (Training sets)				Independent sets			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
AdaBoost	93.02	98.34	95.66	0.925	96.29	90.74	93.51	0.883
Random Forest	93.02	99.61	96.36	0.932	99.35	90.73	95.04	0.913
LSTM	94.30	98.51	96.28	0.926	95.18	91.58	93.37	0.868
DNN	92.89	97.83	95.38	0.916	92.37	90.18	91.29	0.835
2D-CNN	**94.46**	**99.69**	**96.71**	**0.939**	**97.83**	**91.89**	**95.70**	**0.782**

S/T sites datasets comparative performance of the ROC-AUC calculation

An ROC(AUC) graph, along with other indicators such as the algorithm's accuracy, is used to evaluate the algorithm. As shown in Figure. 6, ROC and AUC curve were used to evaluate the classified 2D-CNN output using multiple classifications indicators. The ROC(AUC) curve of two

- dimensional CNN-SARS-COV-2 multi - link is given. Our deep neural network structure performs well even under multiple classification methods, but additional data is needed to draw clear conclusions. The cross-verification accuracy of 2DCNN model is **0.980**, and the independent verification ROC(AUC) score is **0.950** obtained.

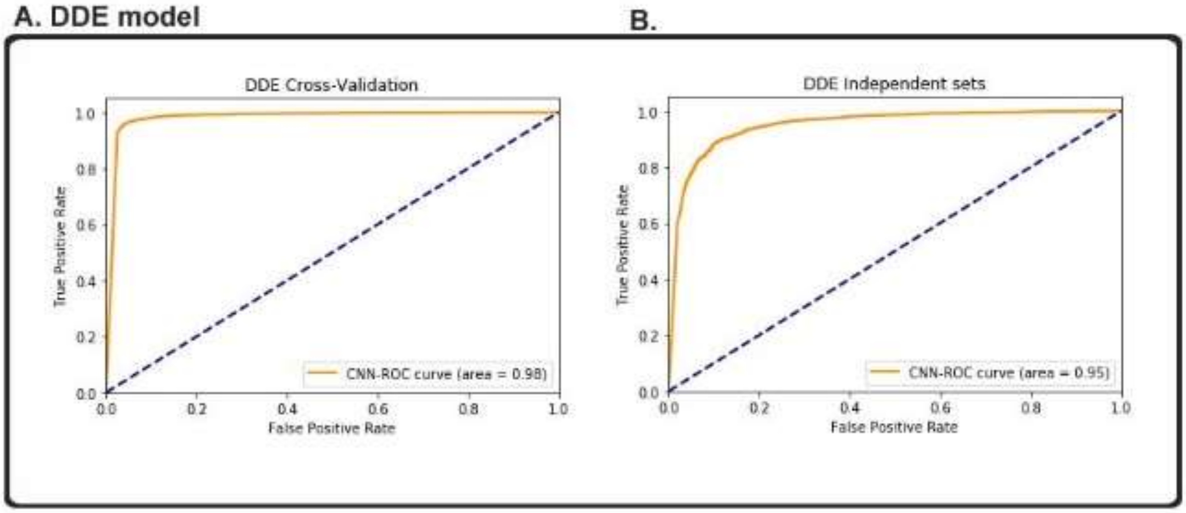

Figure 6. DDE ROC-AUC model of (a) cross-validation test and (b) independent test.

The ROC(AUC) curve is derived from the cross-validation (training sets) results and is used to further verify the validity of the CNN model. Figure 6 show the phosphorylation type of each protein in relation to the ROC curve and the area under the curve (AUC). Results obtained by (TPR) and (FPR) value which consistent with our findings. The RCO-AUC score reached **0.980** by comparing with model cross-validation data set and the independent data set **0.950** score achieved.

Optimizers comparative performance SARS-CoV-2 by using ROC-AUC calculation

Five widely used optimizers, RMSprop, Adam, Nadam, SGD, and Adadelta., were employed to create learning models using individual and feature encoding techniques. The optimization of hyperparameters was estimated during experimental analysis in this study using quantitative methodologies, although it was challenging to identify the appropriate hyperparameter optimizer for our model. Based on integrated feature groups for DDE model as shown in Figure 7. used S/T sites datasets, our proposed model had the highest AUC values (Adadelta obtained 0.97, SGD obtained 0.96, RMSprop obtained 0.95, Adam obtained 0.95 and Nadam obtained 0.95, respectively). 2Deep-IPs models also had comparable AUC values for the phosphorylated S/T. According to our feature extraction model, the 2D-CNN models created utilizing secondary structural information and indicating the biggest contribution of the prediction of S/T sites. The optimization of the hyperparameter, on the other hand, made sure that the model did not overfit its data by setting, for instance, regularization as shown in Figure 6. In contrast, the learning algorithms could recreate their inputs. To compare different optimizers fairly, a fresh network was constructed following each optimization round. The model was given a fresh start. Figure 7 displays the outcomes of the performance. We discovered that the Adadelta optimizer beat all other optimizers when the final model was applied.

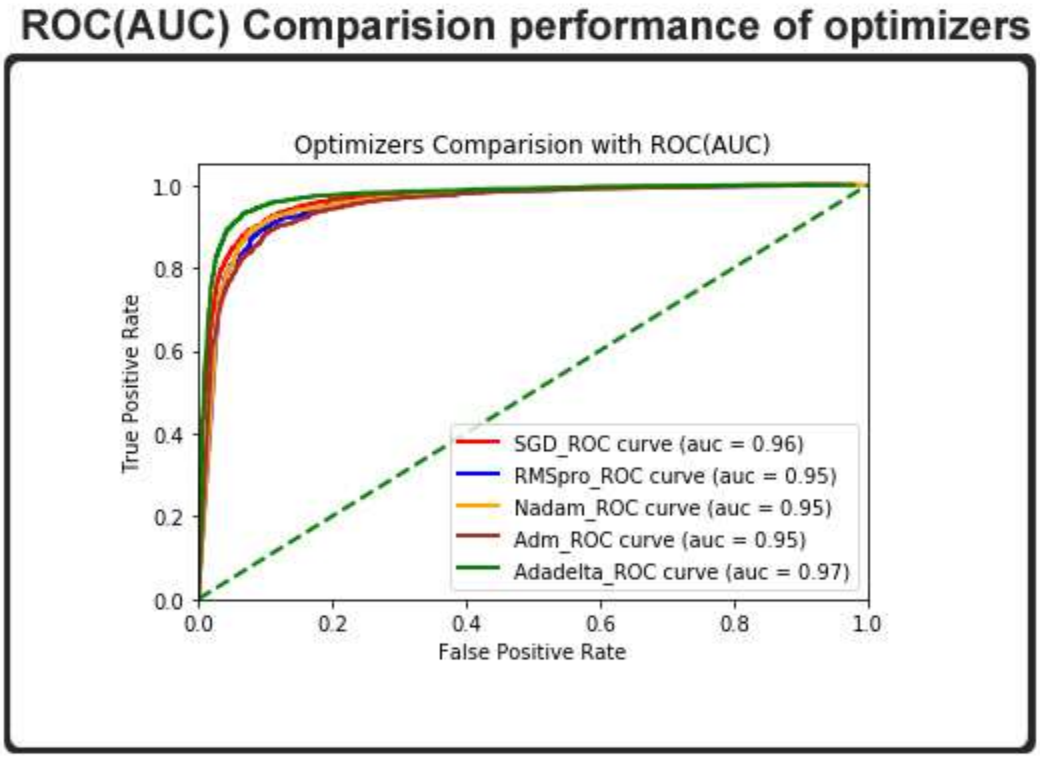

Figure 7. ROC(AUC) Comparison performance of different optimizers

Comparison performance evaluation of different existing methods

Furthermore, we found that on S/T sites, DeepIPs only slightly outperforms MusiteDeep2020 in terms of model performance. The independent dataset may have been incorporated into the training process of these existing methods, resulting in similar results as shown in Figure 6. There [61] is limited research investigating the experimentally verified S/T phosphorylation sites of the SARS-CoV-2-infected Vero E6, Caco-2, and iAT2 cell lines from the literature in order to conduct a fair and unbiased evaluation of the performance [62]. The overlap items between this dataset and the training data for DeepIPs and MusiteDeep2020 were then eliminate using a very strict process. The findings shown that 2D-CNN can accurately identify modification sites, outperforming the outcomes produced better than MusiteDeep2017, MusiteDeep2020, DeepPSP, and DeepIPs as shown in Table.8. based with 5-fold cross-validation (Training Sets). Additional variables were derived from Independent Sets comparison performance with acc of **0.890**,Sens of **0.878**, Spec of **0.902**, MCC of **0.782** and AUC 0.9500 score which is better than all existing methods as shown in Table 9.

Table 8. Comparison performance of existing methods for phosphorylation site prediction with 5-fold cross-validation (Training Sets)

Residue Type	**Method**	**Acc**	**Sens**	**Spec**	**MCC**	**AUC**
S/T Sites	DeepIPs	80.63	79.61	83.50	0.6316	0.8937
	DeepPSP	80.21	76.65	83.78	0.6058	0.8762
	MusiteDeep2020	80.95	82.95	78.96	0.6196	0.8867
	MusiteDeep2017	80.17	78.87	81.46	0.6035	0.8798
	2Deep-IPs	**0.982**	**0.957**	**0.929**	**0.8781**	**0.9641**

Table 9. Comparison performance of existing methods for phosphorylation site prediction with 5-fold cross-validation (Independent Sets)

Residue Type	**Method**	**Acc**	**Sens**	**Spec**	**MCC**	**AUC**
S/T Sites	DeepIPs	80.63	79.61	83.50	0.6316	0.8937
	DeepPSP	80.21	76.65	83.78	0.6058	0.8762
	MusiteDeep2020	80.95	82.95	78.96	0.6196	0.8867
	MusiteDeep2017	80.17	78.87	81.46	0.6035	0.8798
	2Deep-IPs	**0.890**	**0.878**	**0.902**	**0.782**	**0.9500**

Discussion

Computational techniques are a good tool to classify the biological effects of SARS-CoV-2 protein. The work in our lab is critical to our studies of molecule-based signaling networks, G protein pathways, and metabolic pathways. Our deep learning method replaces the missing information to complete the SARS-COV-2 sequence. This calculation is the first to be successfully applied to research and is the first to discover a new computational method to help the scientific community understand how 2Deep-IPs with SARS-CoV-2 molecular function functions and generate complex disease pathways based on this information. We have also developed a generic, but highly optimized, deep learning system for protein sequences. We modify the hyperparameters to optimize the results, and then verify the results with the selected optimal values. When features are retrieved from the network, the mixed feature profile obtained is regarded as a vector only when it enters the network, and as a method to adjust the CNN network feature profile, the results are different. The TWO-DIMENSIONAL CNN model we used several measures to gain an advantage over the competitive model and collect data. Previous research [63] has supported the hypothesis that numerous enzymes and receptors are activated or deactivated by phosphorylation and dephosphorylation processes, which are carried out by kinases and phosphatases, protein phosphorylation is a crucial physiological regulation mechanism. For cellular transduction signaling in particular, protein kinases are responsible, and numerous disorders, mostly cancers, can exhibit their hyperactivity, dysfunction, or overexpression. It follows that it is clear that using kinase inhibitors to treat cancer can be beneficial. we address the mechanism of action of phosphorylation in this review. We also go over the prospect of treating cancers with kinase inhibitors. Our real-time systems strategy works for us. Computational models can help us design a bioinformation system based on protein sequence data retrieval and analysis. It is designed to be intelligent and uses protein function as a basis for detecting disease variations and mutations. This scientific understanding applies to the development of therapeutic targets in drug research. Through hard work, we have helped us advance this project, and the result of our efforts has enabled us to view descriptions of evolutionary features as photographs. However, in terms of further technologies, certain issues remain, and different options can be used to address them in the future. To make the most of this data, the first step is to get more research and data in the future. In addition, it is important to conduct future research to investigate how all descriptor terms involving evolutionarily derived feature information can be fed to CNNs. Biological researchers are also welcomed to use our models, as well as other researchers who can propose to participate in activities that do more than simply demonstrate experimental precision results. In trying to understand proteins whose function is unknown, they believe that machine learning models play a crucial role, which is also considered a pioneering way to apply information about structural proteins in the future. This research suggests a convolutional neural network-based enhanced deep learning method using S-padding. A related 2D-CNN model [64] is created to abstract the comprehensive properties of the phosphorylation site area in protein sequences. The S-padding approach creates a three-dimensional matrix with extension information from the original amino acid sequences. The results of the 10-fold cross-validation trials demonstrate that the suggested technique can perform with an accuracy of accuracy of **95.70** on the human dataset. Furthermore, the accuracy, sensitivity, and specificity of the phosphorylation site prediction task on several organisms achieved score 0.85, suggesting a potential capability. The suggested method improved the metric performance when compared to existing models in terms of accuracy and auc value, and with both cross-validation (Training sets) and independent datasets, the proposed method would perform even better. We proposed 2Deep-IPs, which combines Two Dimensional 2D-CNN architecture with the most used window image-based matrix methods, to predict phosphorylation sites. The analysis was based on cross-validation (Training sets) achieved accuracy score **96.71**, Sen score obtained 94.46 and Spec score obtain is 9969. The results analysis based on independent datasets achieved accuracy score **95.70** reveals that 2Deep-IPs outperforms other phosphorylation sites predictors in terms of performance. The proposed method 2Deep-IPs performance is outstanding when compared to existing methods in terms of Accuracy, Sen, Spec and AUC value, on the basis of both cross-validation (Training sets) and independent datasets, the proposed method would perform even better.

Conclusion

We conducted all analyses using biological processes that link to the occurrence of SARS-CoV-2 infection, phosphorylation plays a significant role. The need for efficient computational approaches to identify phosphorylation in SARS-CoV-2 infection is urgent due to the limitations of experimental site verification, which takes time and money. Accordingly, in this study, we suggest 2Deep-IPs, which combines Two Dimensional 2D-CNN architecture with the most used window image-based matrix methods, to predict phosphorylation sites. Independent testing reveals that 2Deep-IPs outperforms other phosphorylation sites predictors in terms of performance. The aim of this work is to study the human pathway in relation to human proteins by examining the structure of 2D-CNN-SARS-CoV-2. The classification model of SARS-CoV-2 protein takes advantage of the differences between SARS-CoV-2 protein and its derived characteristics and converts them into the characteristic matrix that affects its evolution and development. These matrices were used to construct a two-dimensional CNN-SARS-COV-2 and serve as an effective framework for CNNs. This proposed 2Deep-IPs, SARS-COV-2 model was used to study our model. Our unique neural network exceeds existing state-of-the-art neural networks in efficiency and provides further improvements to all traditional evaluation methods. The failure of standard methods to elucidate new functions of newly discovered DNA damage replication protein pathways has been a persistent problem over the past decade. This allows human disease pathways, drug pathways, and DNA repair pathways to arise based on our models. In this study, the application of 2Deep-IPs SARS-COV-2 in predicting the function of protein sequences associated with human pathways has been accurately annotated from large-scale proteins. However, our assumptions were complicated by the use of four routing databases and the mapping process provided by UniProtKB/Swiss-Prot. The initial online interface allowed us to identify cross-reference knowledge pathways. In the future, this technique will include many different approaches that could help improve biological research especially in the related fields of proteomics and genomics.

Data Availability Statement

The source code and datasets are provided on the author request.

Conflict of Interest

Declare conflicts of interest or state “The authors declare no conflict of interest”.

Author Contributions

GA Conceptualization and designed the experiments, RS performed the experiments and analyzed the data, TA Data collection, writing-original draft preparation. TH and SA editing and reviewed the paper. NJ and TQ Code writing and visualization and conceptualized the review and finalized the manuscript. GA and MR revised the manuscript and polished the expression of English. All of the authors have read and approved the final manuscript.

Funding

This research project is supported by…).

Acknowledgments

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

Reference

[1] Pillai, Dinesh Dhamodhar Mathevan, et al. "Socio-economic impact of coronavirus disease 2019 (COVID-19)–An Indian outlook." Journal of Family Medicine and Primary Care 9.10 (2020): 5103.

[2] Dhama, Kuldeep, et al. "Coronavirus disease 2019–COVID-19." Clinical microbiology reviews 33.4 (2020): e00028-20.

[3] Howell, R., Clarke, M.A., Reuschl, AK. et al. Executable network of SARS-CoV-2-host interaction predicts drug combination treatments. npj Digit. Med. 5, 18 (2022).

[4] Hu, B., Guo, H., Zhou, P. et al. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 19, 141–154 (2021).

[5] Bojkova D. Klann K. Koch B. Widera M. et al., Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. Nature. 2020; 583: 469-472

[6] Wang, G., Pan, J., & Chen, S. D. (2012). Kinases and kinase signaling pathways: potential therapeutic targets in Parkinson's disease. Progress in neurobiology, 98(2), 207-221.

[7] Saud, Zack, et al. "The SARS-CoV2 envelope differs from host cells, exposes pro-coagulant lipids, and is disrupted in vivo by oral rinses." Journal of lipid research (2022): 100208.

[8] Miao, M., Yu, F., Wang, D., Tong, Y., Yang, L., Xu, J., Qiu, Y., Zhou, X. and Zhao, X., 2019. Proteomics profiling of host cell response via protein expression and phosphorylation upon dengue virus infection. Virologica Sinica, 34(5), pp.549-562.

[9] Klann K, Bojkova D, Tascher G, et al. Growth factor receptor signaling inhibition prevents SARS-CoV-2 replication. Mol Cell 2020;80:164–174 e164.

[10] Hekman RM, Hume AJ, Goel RK, et al. Actionable cytopathogenic host responses of human alveolar type 2 cells to SARS-CoV-2. Mol Cell 2020;80:1104–1122 e1109.

[11] Ismail, H. D., Jones, A., Kim, J. H., Newman, R. H., & Dukka, B. K. C. (2015, October). Phosphorylation sites prediction using Random Forest. In 2015 IEEE 5th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) (pp. 1-6). IEEE.

[12] Li F, Li C, Marquez-Lago TT, et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family- specific phosphorylation sites in the human proteome. Bioinformatics 2018;34:4223–31

[13] Huang, S. Y., Shi, S. P., Qiu, J. D., & Liu, M. C. (2015). Using support vector machines to identify protein phosphorylation sites in viruses. Journal of Molecular Graphics and Modelling, 56, 84-90.

[14] Naseer, S., Hussain, W., Khan, Y. D., & Rasool, N. (2021). Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. Analytical Biochemistry, 615, 114069.

[15] Wang D, Zeng S, Xu C, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 2017;33:3909–16.

[16] Wang D, Liang Y, Xu D. Capsule network for protein post-translational modification site prediction. Bioinformatics 2019;35:2386–94.

[17] Guo L, Wang Y, Xu X, et al. DeepPSP: a global-local information-based deep neural network for the prediction of protein phosphorylation sites. J Proteome Res 2021;20:346–56.

[18] Loyal, L., Braun, J., Henze, L., Kruse, B., Dingeldey, M., Reimer, U., ... & Giesecke-Thiel, C. (2021). Cross-reactive CD4+ T cells enhance SARS-CoV-2 immune responses upon infection and vaccination. Science, 374(6564), eabh1823.

[19] Ghulam, A., et al. "Identification of Novel Protein Sequencing SARS CoV-2 Coronavirus Using Machine Learning." Bioscience Research, 2021 volume 18(SI-1): 47-58 48.

[20] Ghulam, Ali, et al. "ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network." Chemometrics and Intelligent Laboratory Systems 226 (2022): 104589.

[21] Stukalov A, Girault V, Grass V, et al. Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV. Nature, 2021;594:246–52. doi: 10.1101/2020.06.17.156455.

[22] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.

[23] Mei S, Li F, Xiang D, et al. Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. Brief Bioinform 2021. doi: 10.1093/bib/bbaa415.

[24] Basith S, Manavalan B, Hwan Shin T, et al. Machine intel- ligence in peptide therapeutics: a next-generation tool for rapid disease screening. Med Res Rev 2020;40:1276–314.

[25] Lv H, Dao FY, Zulfiqar H, Lin H. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. Brief Bioinform. 2021 Nov 5;22(6):bbab244. doi: 10.1093/bib/bbab244. PMID: 34184738; PMCID: PMC8406875.

[26] Zeng, Y., Liu, D. & Wang, Y. Identification of phosphorylation site using S-padding strategy based convolutional neural network. Health Inf Sci Syst 10, 29 (2022). https://doi.org/10.1007/s13755-022-00196-6

[27] Saravanan V, Gautham N. Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. OMICS. 2015; 19(10):648-658. doi:10.1089/omi.2015.0095

[28] Saravanan, V., Gautham, N. BCIgEPRED—a Dual-Layer Approach for Predicting Linear IgE Epitopes. Mol Biol 52, 285–293 (2018).

[29] Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. 2018. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ 6:e4750 DOI 10.7717/peerj.4750.

[30] Palatnik de Sousa I. 2018. Convolutional ensembles for Arabic handwritten character and digit recognition. PeerJ Computer Science 4:e167 DOI 10.7717/peerj-cs.167.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, Curran Associates Inc., Lake Tahoe, Nevada, 2012, pp. 1097–1105.

[32] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Available from: https://www.tensorflow.org/.

[33] Azamfar, Moslem, et al. "Multisensor data fusion for gearbox fault diagnosis using 2-D convolutional neural network and motor current signature analysis." Mechanical Systems and Signal Processing 144 (2020): 106861.

[34] Le, Nguyen Quoc Khanh, et al. "iMotor-CNN: Identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule." *Analytical biochemistry* 575 (2019): 17-26.

[35] Sikander R, Wang Y, Ghulam A, Wu X. Identification of Enzymes-specific Protein Domain Based on DDE, and Convolutional Neural Network. Front Genet. 2021 Nov 30;12:759384. doi: 10.3389/fgene.2021.759384. PMID: 34917128; PMCID: PMC8670239.

[36] Sikander R, Arif M, Ghulam A, Worachartcheewan A, Thafar MA and Habib S (2022) Identification of the ubiquitin–proteasome pathway domain by hyperparameter optimization based on a 2D convolutional neural network. Front. Genet. 13:851688. doi: 10.3389/fgene.2022.851688

[37] Ghulam, A., Ali, F., Sikander, R., Ahmad, A., Ahmed, A., & Patil, S. (2022). ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. Chemometrics and Intelligent Laboratory Systems, 226, 104589.

[38] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312-323.

[39] Cheng, C., & Parhi, K. K. (2020). Fast 2D convolution algorithms for convolutional neural networks. IEEE Transactions on Circuits and Systems I: Regular Papers, 67(5), 1678-1691.

[40] Yan, A., Cheng, S., Kang, W. C., Wan, M., & McAuley, J. (2019, November). CosRec: 2D convolutional neural networks for sequential recommendation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2173-2176).

[41] Wang, Y., Zhang, L., Xia, P., Wang, P., Chen, X., Du, L., ... & Du, M. (2022). EEG-Based Emotion Recognition Using a 2D CNN with Different Kernels. Bioengineering, 9(6), 231.

[42] Jose, J. A., Kumar, C. S., & Sureshkumar, S. (2022). Tuna classification using super learner ensemble of region-based CNN-grouped 2D-LBP models. *Information Processing in Agriculture*, *9*(1), 68-79.

[43] Aziz, A. Z. B., Hasan, M. A. M., Ahmad, S., Al Mamun, M., Shin, J., & Hossain, M. R. (2022). Multi-channel CNN based anticancer peptides identification. Analytical Biochemistry, 114707.

[44] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15(1):1929–1958

[45] Mostafa, F., Afify, Y., Ismail, R., & Badr, N. (2022). UNCOVERING THE EFFECTS OF DATA VARIATION ON PROTEIN SEQUENCE CLASSIFICATION USING DEEP LEARNING. International Journal of Intelligent Computing and Information Sciences, 1-14.

[46] Le, N. Q. K., Yapp, E. K. Y., Ou, Y. Y., & Yeh, H. Y. (2019). iMotor-CNN: identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. Analytical biochemistry, 575, 17-26.

[47] Zhao, L., Zhu, Y., Wang, J., Wen, N., Wang, C., & Cheng, L. (2022). A brief review of protein-ligand interaction prediction. Computational and Structural Biotechnology Journal.

[48] Ren, H., Zhang, Q., Wang, Z., Zhang, G., Liu, H., Guo, W., ... & Jiang, J. (2022). Machine learning recognition of protein secondary structures based on two-dimensional spectroscopic descriptors. Proceedings of the National Academy of Sciences, 119(18), e2202713119.

[49] Jia, S., Despinasse, A., Wang, Z., Delingette, H., Pennec, X., Jaïs, P., & Sermesant, M. (2018, September). Automatically segmenting the left atrium from cardiac images using successive 3D U-nets and a contour loss. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 221-229). Springer, Cham.

[50] Giang Son Tran, Thi Phuong Nghiem, Van Thi Nguyen, Chi Mai Luong, Jean-Christophe Burie, "Improving Accuracy of Lung Nodule Classification Using Deep Learning with Focal Loss", Journal of Healthcare Engineering, vol. 2019, Article ID 5156416, 9 pages, 2019.

[51] Abdel-Hamid, O., Deng, L., & Yu, D. (2013, August). Exploring convolutional neural network structures and optimization techniques for speech recognition. In Interspeech (Vol. 11, pp. 73-5).

[52] Zhang, H., Li, Y., Zhang, Y., & Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. Remote sensing letters, 8(5), 438-447.

[53] Zhang, R., Ghosh, S., & Pal, R. (2022). Predicting binding affinities of emerging variants of SARS-CoV-2 using spike protein sequencing data: observations, caveats and recommendations. Briefings in Bioinformatics, 23(3), bbac128.

[54] Zhang, Tianhao, et al. "Protein Subcellular Localization Prediction Model Based on Graph Convolutional Network." Interdisciplinary Sciences: Computational Life Sciences (2022): 1-10.

[55] Zhang, S.; Hanb, F.; Lianga, Z.; Tane, J.; Caoa, W.; Gaoa, Y.; Pomeroyc, M.; Ng, K.; Hou, W. An investigation of CNN models for differentiating malignant from benign lesions using small pathologically proven datasets. Comput. Med Imaging Graph. 2019, 77.

[56] Lee, YG., Oh, JY., Kim, D. et al. SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. J. Electr. Eng. Technol. 18, 579–588 (2023). https://doi.org/10.1007/s42835-022-01161-9

[57] Sundararajan, Mukund, and Amir Najmi. “The many Shapley values for model explanation.” arXiv preprint arXiv:1908.08474 (2019)

[58] White, C., Ismail, H.D., Saigo, H. et al. CNN-BLPred: a Convolutional neural network based predictor for β-Lactamases (BL) and their classes. BMC Bioinformatics 18, 577 (2017).

[59] White, C., Ismail, H.D., Saigo, H. et al. CNN-BLPred: a Convolutional neural network based predictor for β-Lactamases (BL) and their classes. BMC Bioinformatics 18, 577 (2017).

[60] Rahmati, S., Abovsky, M., Pastrello, C., & Jurisica, I. (2017). pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. Nucleic acids research, 45(D1), D419-D426.

[61] Klann K, Bojkova D, Tascher G, et al. Growth factor receptor signaling inhibition prevents SARS-CoV-2 replication. Mol Cell 2020; 80:164–174 e164.

[62] Bouhaddou M, Memon D, Meyer B, et al. The global phosphorylation landscape of SARS-CoV-2 infection. Cell 2020; 182:685–712 e619.

[63] Ardito, Fatima, et al. "The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy." International journal of molecular medicine 40.2 (2017): 271-280.

[64] Zeng, Y., Liu, D. & Wang, Y. Identification of phosphorylation site using S-padding strategy based convolutional neural network. Health Inf Sci Syst 10, 29 (2022).