Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2304.2793

2023, 42(4) 28-40

Identification of cancerlectin proteins using hyperparameter optimization in deep learning and DDE profiles

Rahu Sikander [a], Ali Ghulam [b, *], Jawad Hassan [c], Laiba Rehman [d], Nida Jabeen [e], Natasha Iqbal [f]

[a] *Center for Computing Research, Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi Pakistan*

[b] *Information Technology Centre, Sindh Agriculture University, Tandojam, Sindh Pakistan*

[c] *Department of Computer Science and Engineering, Air University Multan Campus, Multan, Punjab Pakistan*

[d] *Women university Multan*

[e] *College of information and compute, Taiyuan university of Technology 030024, Shanxi Taiyuan China*

[f] *Department of Botany, Government College University of Faisalabad, 37000, Faisalabad Pakistan*

[*] Corresponding author: Ali Ghulam, Email: garahu@sau.edu.pk

Received: 17 May 2023, Accepted: 25 September 2023, Published: 01 October 2023

K E Y W O R D S

Cancerlectin

Proteins Sequence

Proteins Features Encoding

2D-CNN

Machine Learning

Deep Learning

Bio- Engineering

A B S T R A C T

This study focuses on the development, metastasis, and spread of cancer diseases. It is therefore very desirable to establish deep learning method that classify cancerlectin proteins function efficiently and effectively. We used feature extraction model for physicochemical properties, such as Cancerlectins protein structure, functions, and other compounds. We propose a computational method, namely, cancerlectin two-dimensional convolutional neural networks (Lectin2D-CNN), for predicting cancerlectin proteins. Additionally, we conduct the cross-validation experiments. In addition to this approach, our paper proposes using cancerlectin two-dimensional convolutional neural networks (Lectin2D-CNN) to do image-based classification. The results indicate the proposed method Lectin2D-CNN achieved high accuracy and satisfactory specificity for comparison data sets and was superior to the compared methods. Various classifiers were used to predict cancerlectin protein functions. We developed a prediction model based on the 2D-CNN architecture to increase the recognition sensitivity and accuracy for cancerlectins. Results provide a basis for the estimation of cancer lectins and demonstrate deep learning approaches in in computational proteomics. When the Cross-validation using 2D-CNN random number generator has produced accuracy score obtain 0.7169%, Sensitivity score obtain 0.7012%, Specificity score obtain 0.7326%, MCC score obtain 0.4428%, ROC-AUC score is 0.76%, respectively, then we know we've attained a reliable result. When the Independent datasets using 2D-CNN random number generator has produced accuracy score obtain 0.6375%, Sensitivity score obtain 0.6160%, Specificity score obtain 0.6589%, MCC score obtain 0.2851%, and ROC (auc) score is 0.76%, respectively.

1. Introduction

Researchers have suggested several computational strategies for diverse classification problems, including machine learning, to achieve superior results [1]. Cancerlectins are involved in a wide range of biological processes, including host-pathogen interactions, cell identification, activation pathways, cell cycle regulation, and apoptosis. Scholars have recognized the moisture of sugar in certain proteins and found that the majority of lectins are highly sensitive and selective and bind to the sugars reversibly and without any alteration in the bound carbon moisture [2]. These glycoproteins are commonly divided into five classes based on the monosaccharides for which they have the highest affinity [3]. Cancerlectins differ significantly from one another not only in terms of its functional activities but also in terms of their sequences, structures, site binding systems, quaternary structures, affinities, specificities with carbohydrates, and potential uses [4]. Studies provide support for the validity of cancer-related processes, such as tumour cell discrepancy, control of cancer cells, cell labelling of tumour tissues, and metastasis of cancer [5, 6]. This discrepancy might be due to the fact that cancerlectins are usually detected by biological experiments, but the results are unreliable and expensive. Algorithms and classifiers are also a problem in the current hypothesis to construct a new combination of machine learning strategy to obtain increased prediction accuracy. Computational prediction methods are used to search for new sequences of cancerlectin protein and obtain candidates for cancer markers [7]. However, a number of biological processes may be regulated, including host–pathogen interactions, cells, complementary activation mechanisms, cell cycle regulation, and apoptosis. This usually contains two or more carbohydrate sites as a lectin molecule. Additionally, interaction with cells would cause them to cross-link and subsequently agglutinate in addition to mixing them with the sugars on their surfaces [8]. Tumour cells interact with one another, and adhesion to cell, cell growth, differentiation of tumour cells, metastasis, and cellular infection, are the key to cancerlectin a kind of lectin [9, 10].

We have been motivated to integrate various classifiers and extractors for improving prediction accuracy by sophisticated applications of machine learning that demonstrate excellent results for cancer identification [11, 12]. After contrasting their effectiveness and acceptance, classification algorithms are chosen to show how machine learning affects cancer detection [13]. Tumour tissue-derived cells can be characterized using cancerlectins. For instance, the agglutinin of Helix pomatia is a useful forecasting indicator of colorectal carcinoma. Cancerlectins may also be used for cancer treatment because they bind to receptors on the surface of cancer cells to induce cytotoxicity, inhibit cancer growth, or promote apoptosis [14, 15]. Most of cancerlectins proteins have therapeutic function by binding to membranes of cancer cells or receptors, resulting in cytotoxicity, apoptosis, and cancer growth inhibition [16]. Galectin is a cancer modulator and play a role in infections [17]. Cancerlectin diagnostic and molecular correlations between therapeutic markers and molecular cancer were negative and statistically significant [18, 19]. Therefore, screening specific cancerlectins from multiple cancerlectins is critical for better understanding and overcoming cancer.

Different species' cancerlectin effectors were predicted by Zou et al. [20]. PSSM (PSSM) is utilized as an amino acid (AAC), dipeptide composition (DPC), position-specific matrix composition, and auto covariance transformation (PSSM and AC) as a functional vector for classification tasks. These features are characterized based on the primary protein sequence structure. Our proposed deep neural network method has achieved a number of achievements in the classification of images, processing of natural languages, and many other fields [21]. A brand-new deep learning-based approach called TPSO-DBP is proposed and is based on TPSO. It uses fully connected (FC) neural networks, bidirectional long short-term memory (BiLSTM), and sequence-based single-view features to learn the DBP prediction model [22]. In the current study, Ge, Fang, (2022) et al., developed a novel feature extraction technique that takes into account the influence of residues in the vicinity of the mutation site. To assess the effectiveness of the suggested feature extracting method, rigorous cross-validation and independent tests were run on benchmark datasets [23]. The solution is to measure a wide range of features and apply deep learning techniques to the classification function; however, this method still relies on a collection of features that are initially developed. We go one step forward, directly from the raw protein sequences, to learn low-level features [24]. The study carried out by Ali Ghulam, (2020) et al., revealed that pathway-specific study prediction built on a two-dimensional neural network (2D-CNN-PSPD) [25]. In this article, Ali Ghulam, (2022) et al., introduce the ACP-2DCNN, a revolutionary deep learning-based technique for enhancing anticancer peptide prediction. Dipeptide Deviation from Expected Mean (DDE) is used to extract

the key characteristics based on (2D CNN) is used to train the model and provide predictions [26]. The theoretical framework underpinning this study was noncoding RNAs (ncRNAs), they perform numerous essential roles in biology. The aim of this study was to extend this area of investigation of ncRNA–protein interactions (NPIs) are always costly and time-consuming to discover with experimental approaches. A number of/numerous scholars have contended [27] the implications of machine-learning techniques for prediction. It is generally agreed that GNN is a new and cutting-edge technique for finding strong link predictions on complicated networks. Empirical evidence appears to confirm the notion that noncoding RNA–protein interaction prediction using Graph Neural Networks (NPI-GNN) such as GNN-based technique for predicting NPIs. A widely accepted hypothesis [27] is that predictions of fresh interactions that stem from information about the network and sequence.

The algorithms of the two-dimensional conventional neural network (2D-CNN), a subset of ANNs, have multiple hidden layers. 2D-CNN take as input low-level characteristics, creating more advanced functionality for each layer. In this research, we lead to the creation and use of a stack of 1,101 multi-task feed 2DCNN for large-scale deep learning. Our proposed Lectin2DCNN framework can predict thousands of functional concepts based on specific cancerlectin proteins. Furthermore, the prediction of cancerlectin protein sequences association terms with very few training instances, an important problem in the field of the prediction of automated protein functions, has been tackled by proposing a realistic data increase solution by incorporating automated functional predictions previously manufactured into the system.

In this study, a computational computing strategy for predicting the function of cancerlectins proteins based on protein sequence is proposed. We provide DDE features extraction computational model were employed. We suggest the use of the Lectin2D-CNN, a technique for encoding the sequence of cancerlectins protein domains. The cancerlectins protein controls almost all cellular activities, making it challenging to tune deep learning algorithms on hyper parameters.

2. Materials and Methods

2.1 Data Source

The following section shows valuable contextual details about the collection of objective benchmarks the development of a power classifier is based on a valid and objective benchmarking dataset. The original data set has been collected which have been extracted from (Kumar and Panwar, 2011) [28]. The details on annotation proteins and Lectin2D-CNN sequences 385 proteins were developed to form a positive dataset following removal of duplicated sequences and sequences without experimental evidence. In general, if the constructed dataset has very similar sequences, false results are obtained with overestimated accuracy and the potential for generalization of the proposed model is decreased.

The program of CD-HIT has thus been used to delete redundant sequences by 30% as the sequence identity cut off [29]. In total, 178 cancerlectin sequences were obtained and 226 non-cancerlectin sequences were formulated as shown in Table 1. The study of 404 samples has become a key aspect of correspondingly: by searching the UniProt database. We would encourage research to examine 178 proteins associated with cancerlectins and 226 proteins that are not associated with cancerlectins. We set the 30% identical similarity sequences were deleted with the famous CD-HIT tool by setting similarity equivalent to previous studies. CD-HIT. No duplication was conducted non-cancerlectin between the cancer dataset and the non-cancerlectin dataset, all of which are non-redundant.

Table 1

Collected benchmark datasets divided cross-validation and independent dataset.

Lectins	Benchmark Dataset	Cross-Validation	Independent dataset
Cacnerlectin	178	126	52
Non-Cancerlectins	226	126	100

2.2 Feature Extraction of Cancerlectin Protein Using With DDE Approach

The protein sequence characteristics of the cancerlectin were provided for content analysis. The authors identified the physiochemical characteristics of proteins based on their structure [30]. This leads to the discovery of a new compressive functionality. The whole DDE sequence with all of the information is the suggested sequential representation for protein samples. DDE is an important approach for categorizing proteins involved in the cancerlectin was created in this study. We used experimental work to undertake four analyses: data collection, feature extraction, 2D-CNN creation, and model evaluation. Using our method's flowchart as shown in Fig. 1. 2D-CNNs and DDE extraction feature profile matrices were used in this investigation. An

essential method for identifying and categorizing human routes was created using data on a number of different characteristics.

2.3 Feature Extraction Approach

After being standardized by DDE, the data were processed to yield physicochemical characteristics, attributes obtained from the protein sequences implicated in the cancerlectins, and information on evolution [31]. DDE gives end-users the ability to visualize interrelationships within a set of proteins without requiring them to have any prior assumptions about those proteins; it does not concentrate on the alignment of protein interactions.

We refer to this behaviour as DDE when referring to the projected mean variance ***(Tv).*** The theoretical mean value (T_m), theoretical difference (T_v), and dipeptide composition make up the three basic parameters that make up the DDE function vector (C_c). The dipeptide i's (C_c) on peptide p is represented by $(D_{C(i)})$. The following is a list that includes the three components that were discussed earlier, as well as the results of the DDE calculation. The amount of dipeptide I that is contained in peptide P is referred to as the $\boldsymbol{DC_{(i)}}$.

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

Numerous research has looked into the cancerlectin protein features extracted using with the 400 dipeptide features length as calculation of (20× 20). DDE features length is defined as the link between the samples' shape and physical characteristics, which is a dipeptide, although the not useful samples were excluded. Dipeptide 1 and N, on the other hand, are L-1, in my opinion. As a result, it differs from ***L-1*** (i.e., probable quantity in P). The theoretical mean is represented by TM (i).

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

Additionally, those three stop codons were eliminated, and the total number of codons that were present in the amino acid sequence was determined. It gets increased by the dipeptide ***Ci2***, which has the number of codons that are specified, and the specified dipeptide ***Ci1*** is the quantity of codons.

The peptides that were recovered from *TM (i)* - were not related to *TM (i);* as a result, they were excluded. In order to create compact dipeptide P, 400 dipeptide characteristics were kept, processed, and precalculated. Dipeptide was used to derive *TV (i)* theoretical variance, which was then computed as shown below:

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

Where we used *TM (i)* is given by *i (2).* Peptide P contains L-1 dipeptides, which is equal to the total of N. $^{DDE_{(i)}}$. $^{DDE_{(i)}}$Was provided after being expanded as follows:

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

In order to obtain the 400-dipeptide features which extracted trough DDE approach. The features extraction process used a 400-dimensional vector, which produced a collection of four hundred- feature length.

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

2.4 The Proposed CNN Predictive Model Architecture

Deep learning shown outstanding performance in the study disciplines of image processing [32], computer vision, image processing [33, 34], and bioinformatics during the last decade. Accurate predictions for a variety of biological problems, including cancerlectin protein prediction [36], cancerlectin protein identification [37], and cytoskeleton motor protein identification [35]. We used 2D-CNN in this work, which Le et al. had used in their research [36] to test the effectiveness of 2D CNN for DBP prediction. The research design involved the approach used in the current study used Tensor Flow objects. As shown in Fig. 1 presented the two-dimensional convolutional neural network (CNN) architecture. Three thinned-out layers, followed by hidden layers, made up each CNN. CNN has a wide range of applications and can produce amazing outcomes [25] as well as boost protein synthesis [26]. In addition, various datasets or challenges required varying quantities of feature, hyperparameter, or parameters. Convolution layers, pooling layers, and efficient computing are just a few of CNN's essential traits that set it apart from other algorithms. A typical CNN architecture has three primary layers: the input layer, the hidden layer, and the output layer. These layers are referred to as the input, hidden, and output layers, respectively.

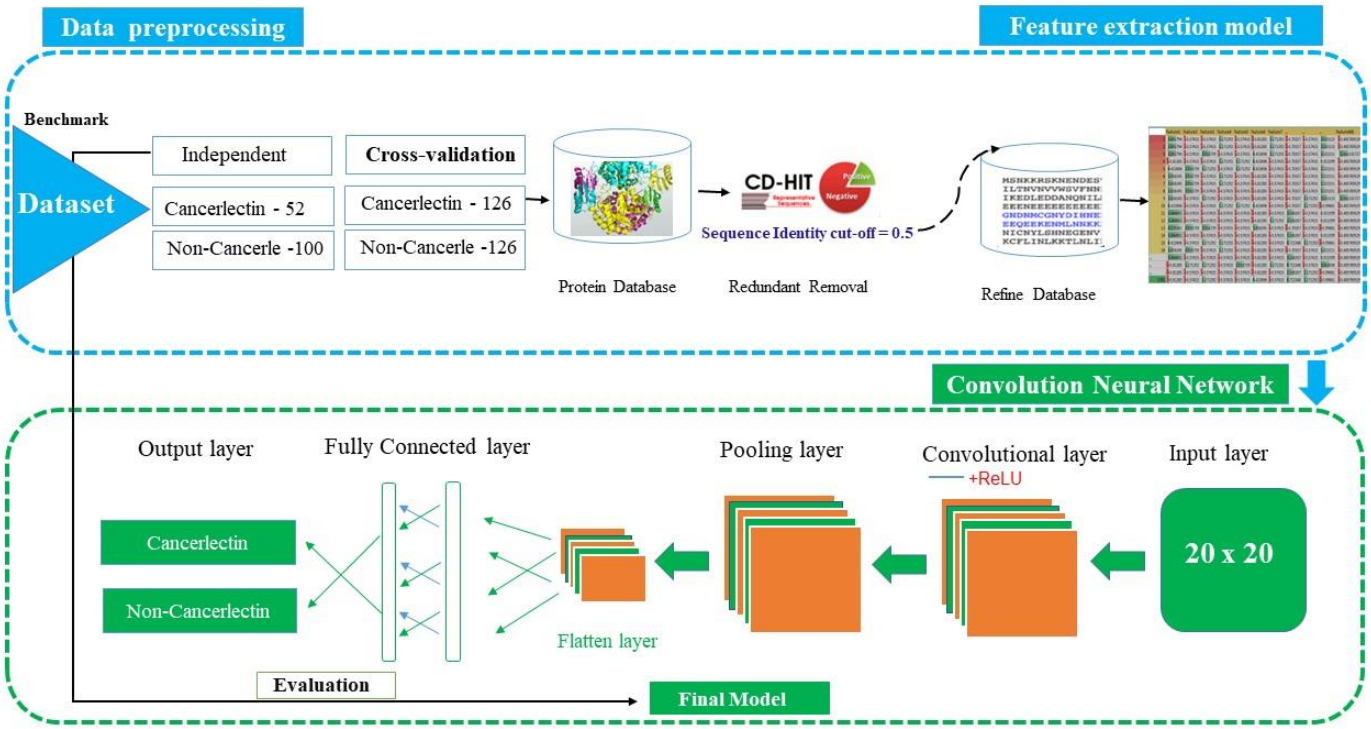

Fig 1. Proposed Framework model for 2D-CNN

We made use of 2D-CNN to extract as many of the DDE score matrix's hidden spatial properties as we possibly could. The proposed technique insured that disorder problems would be produced, and it ensured that the features formed in amino acid sequences would be accurate. The development of the hidden layers will make it possible for CNN's hidden characteristics to quickly recognize cancerlectin proteins once they have been generated. For this approach, we employed three different kernel sizes in each of the four filter layers (32, 64, 128, and 256). This filter is employed and mentioned in relation to convolutional neural networks as a windows or kernel. We can verify the image using a number of filters to generate different feature mappings of the image. Each convolution layer is made up of several filters. They actually appear as a number, such as 32, 64, 128, 256, 512, etc. The same points were added to separate sets as a protein family that is particular to the cancerlectin peptides. The input layer of a CNN was composed of 2D-CNN image-based data points. A two-dimensional matrix was used to present the images in the conversation that came before it. Before entering the input, we further made an image with the dimensions 20x20 = 400 features length.

2.4.1 Input layer for 2D convolutional neural networks

The present study employed each feature vector as input into the first layer of CNN at this point. We presented the DDE feature space in a 20 × 20 shape taking into account a matrix image of 20 x 20 pixels [29]. Therefore, DDE vector matrix as input as image, we suggested a method for forecasting cancerlectin proteins. A matrix as image format of 20x20 pixels representing the DDE profile vector matrix was displayed. Then, using this kind of dataset and the two-dimensional CNN technique, we trained the model. The 2D-CNN was then coupled with the input DDE profile, and various parameters were adjusted to enhance the model's performance. With this method, the created features are certain to be accurate, and the disorder issue inside the amino acid sequences is avoided. The more hidden characteristics that are generated in CNN, the easier it is to identify cancerlectin proteins. Three different kernel sizes were employed in each of the four filter layers (with 32, 64, 128, and 256 filters) in this study.

2.4.2 Multiple Hidden Layer

The hidden layer also includes additional layers with various activation functions and hyper parameters, such as (a) zero padding layer: used valid name categories of cushioning were examined in this study using a research approach. The 2D-CNN framework input matrix 20 × 20 windows size were incorporated into the zero padding 2D architecture at the conclusion of the chain. (b) The convolutional layer served as a CNN's fundamental building piece. (c) max-pooling layer: A max pooling layer was placed after each overall layer, selecting and filtering features from the results. A pooling layer is

typically periodically further to a CNN framework. The pooling layer operated independently on each input depth section and could be resized using the maximum pooling layer operating. (d) Fully connected layer: A max pooling layer was placed after each overall layer, selecting and filtering features from the results. A pooling layer is typically periodically added to a CNN framework. The pooling layer operated independently on each input depth section and could be resized using the maximum pooling layer operating. We used the following equation to add 2 x 2 stride to each 20 x 20 vector:

$$zp = \frac{K-1}{2} \tag{6}$$

where, k = 3 present the filters and kernel scale size

A related idea which might explain the convolution layer with filters of predetermined sizes are used for the majority of computations. The filter moves across the input data to produce a feature map for each place. We would encourage researchers to examine output layer is then taken to be the combination of feature maps. In experiment we used the five convolution layers, and each convolution layer makes several convolution operations using various filters. The most popular activation functions are sigmoid, SoftMax, linear, tanh (tangent), ReLU, and reaky ReLU. We discovered the best outcomes for ReLU activation function out of everyone. The ReLU function is written as follows:

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

In the formula 7. represented by n, total number of features. Empirical evidence appears to confirm the notion that max-pooling layer with a 2 x 2 stride is placed after the convolutional layer. The attempted to identify max-pooling layer is to decrease computation, parameter count, and dimensionality. Additionally, Maxpooling reduces the length of time needed for model training and manages the overfitting problem.

2.4.3 Output Layer

A possible interpretation of this finding is that, quantitative methods were employed to analyse the flatten layer, which is the layer that sits directly above the output layer. Following the analysis of the fully linked layers, we found that there is always a flatten layer present, which functions to convert the input matrix into a vector. The fully linked layers that we applied ensure that every node in the layer that comes before it is fully connected to all of the nodes in the layer that comes before it. The last phases of CNNs frequently make use of layers that are fully connected. All of the nodes on the first layer are connected to the flatten layer so that the model may learn more and perform its functions more effectively. Through the use of the second layer, the first fully linked layer can communicate with the output layer. In addition, we included the succeeding layer, which we referred to as dropout, in order to enhance the model's performance outcomes and prevent it from overfitting. In the dropout layer, the model will inactivate neurons at random with a certain probability p, and this layer will be called the dropout layer. By fine-tuning the dropout value (from 0 to 1), the training time may be cut in half, and the amount of time needed to compute for the subsequent layers can be cut in half as well. Following each iteration of the convolution process, a supplementary non-linear procedure called ReLU (Rectified Linear Unit) was executed. For the purpose of defining the ReLU output, we utilized the following formula:

$$f(x) = max(0, x) \tag{8}$$

where x denotes the total number of inputs that a neural network received. The output of the model was computed with the assistance of a SoftMax function, which determined the probabilities associated with each possible outcome. The logistic function referred to as the SoftMax can be described using the following formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{k} e^{z_i}} \tag{9}$$

where ***jth*** class is the projected probability based on sample vector ***x, z*** is the input vector with a ***K***-dimensional vector, ***(z)*** is a K-dimensional vector with real values that fall between 0 and 1, and so on. For the purpose of determining whether or not the model is appropriate for the empirical data, we utilized the binary cross-entropy as the loss function. This function can be explained as follows:

$$loss(T, 0) = -(T(\log(0) + (1-T)\log(1-0) \tag{10}$$

where T and O represent the objective and output, correspondingly. The lectin2D-CNN model is constructed using the Keras package with the Tensor flow backend. Fig. 1 depicts a schematic view of the proposed model.

2.5 Evaluation Performance of The Model

Most scholars seem to agree that metric performance. However, there continues to be debate about empirical research on how to select the correct metric for a given problem shows that although each metric has unique

characteristics, all of them have characteristics that measure various parts of the algorithms being examined [38]. Existing research has focused on actual values but has failed to explore evaluating algorithms and actual values in clinical medicine, which can make it difficult to determine which metrics are most suitable for evaluating algorithms.

The evidence points to prediction accuracy is generally inappropriate in the presence of imbalanced data, error costs vary considerably. Critics of the true positive argue that true negative rates, as well as between recall and precision, are always a part of machine learning performance evaluations. There are three generally used performance measurements in information retrieval: Precision, Recall, and F-Measure [39]. An ROC curve plots false positive and false negative rates on a graph, illustrating the trade-off between these values for every cutoff.

We measured the performance of three indices including accuracy (Acc), Sensitivity (Sn), and Specificity (Sp), assessed the rating between cancerlectins and non-cancerlectin. Calculated accuracy is the absolute precision of cancerous and non-cancerlectin classification. Sensitivity and specificity represent the sensitivity and specificities of the 2D-CNN model, such steps are usually worded as follows:

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

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

$$F\text{-}measure = 2 \times (Recall \times Precision / Recall + Precision) \tag{14}$$

$$Precision = \frac{TP}{TP} + FP \tag{15}$$

$$Recall = \frac{TP}{TP} + FN \tag{16}$$

$$MCC = \frac{(TN * TP) - (FN * FP)}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{17}$$

For the above formulas, the numbers of cancerlectins and non-cancerlectins, respectively (True positive), and TN (True negative) are stated in the above formulas. The number of identified non-cancerlectin yet forecasted as cancerlectins and the number of cancerlectin are also FP (false positive) and FN (false negative), respectively, predicted as non-cancerlectins. We also plotted an operating curve of the receiver (ROC) by using X-axis sensitivity and Y-axis specificity. The predictability of cancerlectin proteins features has to be fully explored in applied investigations deep learning classifiers with performance measures. Measures for evaluating various aspects of performance exist in multiple studies, as well as diverse evaluation methods. We have found that the Sensitivity, Specificity, MCC and F1 are equally appropriate measures to review the 2D-CNN models, as they have a possibility of incorrect values when specific circumstances arise. Different words, nomenclatures, or notations may emerge in different contexts even for the same performance metric. Results to conduct a thorough evaluation of the typical measurements and evaluation methodologies for bioinformatics predictors, we performed a review [40, 41]. The ability to accurately comprehend and interpret the results is crucial in bioinformatics, as it dictates which predictors to use and which one to discard.

3. Results

The sample for this study consisted of feature encoding approaches were utilized in this work, which studied several feature extraction strategies for extracting numerical descriptors from biological sequences. Each sampling unit consisted of features that were redundant and irrelevant reduced the learning algorithms' performance. The present study employed DDE features extraction approach that involved using a two-level subset feature selection approach. We reduced the feature space dimensionality while increasing the predictive capabilities of the learning parameter name configuration values learning rate 00.01, Activation function, ReLU function, Cross-entropy, Weight initialization, Optimizer Adam and Dropout rate 0.5.

The research design involved several learning algorithms were given feature spaces from which to choose an operational engine that is best suited for cancerlectins. We conducted in-depth 2D-CNN with DDE, provided the greatest overall performance among the algorithms tested. When the Cross-validation using 2D-CNN random number generator has produced accuracy score obtain 0.7169%, Sensitivity score obtain 0.7012%, Specificity score obtain 0.7326%, MCC score obtain 0.4428%, ROC-AUC score is 0.76%, respectively, then we know we've attained a reliable result. When the Independent datasets using 2D-CNN random number generator has produced accuracy score obtain 0.6375%, Sensitivity score obtain 0.6160%, Specificity score obtain 0.6589%, MCC score obtain 0.2851%, and ROC-AUC score is 0.76%, respectively.

3.1 Test Model Evaluation

In our proposed approach, a model of "train" consisting of 150 epochs was adopted. An object was returned by a feature function [42]. As a result, 2D-CNN has been trained and has maintained a value of 0.8023; however, an increase in duration and a commensurate decrease in accuracy loss (significance) were only 0.7558. In this study, Fig. 2. Presented our suggested model DDE features extraction, which has test loss of 42.14% and accuracy of 55.42%. During the workout, spills (training) intermittently turn off a portion of the synapses in the network, decreasing the network's dependence on that specific activity [43]. For this phase, the number of neuron fractions to decline was specified using the expandable parameter. No data was preserved since we only kept the active node and decreased the number of inactive nodes. As a result, we did not investigate, gather, or retrain the network, but we also did not carry out any more expansion in other areas. We operate with a network that has more than 150 epochs and an input batch size of 10.

We would encourage researchers to examine improved using a number of optimizers, including rmsprop, Adam, nadam, sgd, and adadelta. For a fair comparison of the various optimizers, the model was reinitialized, i.e., a new network was constructed, after each round of optimization. Fig.. 3 displays the overall performance results, and we chose nadam, an optimizer with reliable performance, to build our final model.

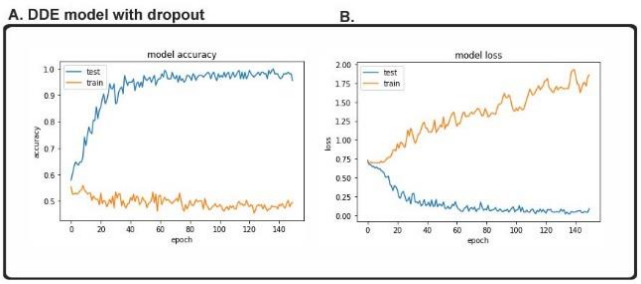

Fig 2. (A) represent the model accuracy (B) represent the model loss

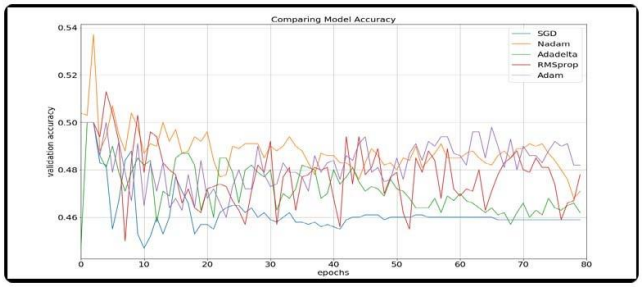

Fig. 3. Optimizer's performance of validation accuracy

3.2 Confusion Matrices Predicted Labels

The main issue with deep learning is over fitting, which means that even if a defined dataset gets worse, our classification will only be accurate in training sets. So that we could ensure that our approach works in a blind dataset as well, we also used an autonomous test. 178 Cacnerlectin and 226 non- Cacnerlectin were present in our independent dataset, as was mentioned in the part above. No samples were included in the workout set. Two uncertainty matrices are displayed as instructive outcomes in the Fig. 4. To be more specific, the results from separate datasets showed that our model had an accuracy rate of 80.5%. There aren't enough variances when compared to the cross-validation result, which should show that our model wasn't over fit. Over fitting and the usage of dropouts in our CNN structure were found to be successful (N. Srivastava et al., 2014) [44].

3.3 Performance Metric Based on Cancerlectin with Comparable Efficiency Of Shallow Neural Networks

We used the same dataset and compared our method to the most popular classifiers; these results are less shocking.

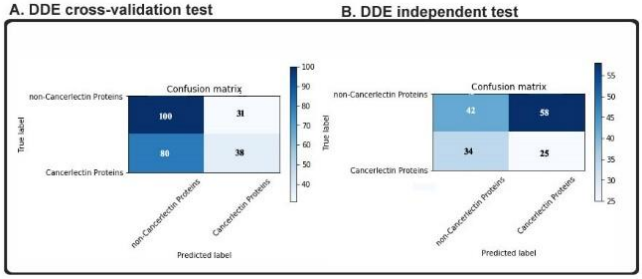

Fig 4. Confusion matrices of: (A) cross-validation test (B) independent test

We employed 2D-CNN and DDE [45]. Table 2 provides a summary of the performance of the proposed technique and the other classifiers at the same stage. Results reveal that independent data sets and cross-validation both outperform the other strategies. It is necessary to go into detail about a number of the comparison's conclusions, including the cross-validation and independent data set outcomes. We obtained the metric performance with comparison, cross-validation datasets accuracy 0.7169%, and independent datasets accuracy score achieved 0.6375%. Our methods results are similar with earlier findings that demonstrated ROC AUC based on a comparison of 2D-CNN and DDE deep learning classifiers [46]. The results of our comparison show that our proposed strategy performs significantly better than other ways as shown in Fig. 5.

Table 2

DDE model performance results of identifying Cancerlectin with filter numbers.

ML Classifiers	Cross-Validation				Independent			
	Sens	Spec	Acc	MCC	Sens	Spec	Acc	MCC
LSTM	0.6935	0.7326	0.7131	0.4298	0.5464	0.5857	0.5660	0.1331
DNN	0.6406	0.6786	0.6596	0.3185	0.5916	0.570	0.5808	0.1726
CNN	0.7012	0.7326	0.7169	0.4428	0.6160	0.6589	0.6375	0.2851

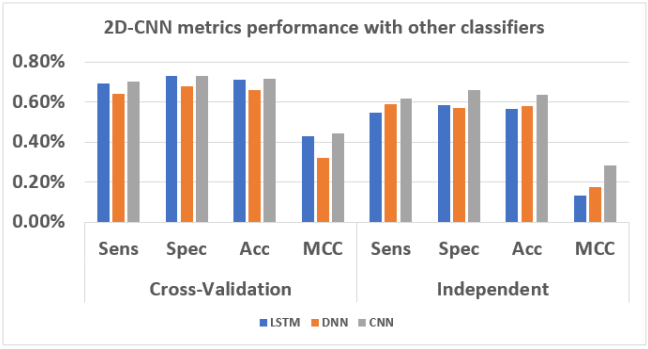

Fig 5. 2D-CNN metrics performance with other classifiers

3.4 ROC (AUC) Comparative Performance of The Proposed Model

The researchers plotted a graphical depiction of their predicted output using the ROC curve in addition to other metrics. In order to investigate how different classifications affected the output of the 2D CNNs, the findings from Fig. 6. To display the components of the up and down double down CURVE, the multilink ROC curve graphic was enlarged. Additionally, adopting twofold and triple cross model yields severely reduced generalization due to cross-validation and non-overfitting. Cross-validation thereby helped the zero overfitting 2D-CNN result, but it prevented cross model overfitting in the other outcomes. The DDE model validation datasets achieved area under the ROC curve (AUC) score 0.76%, according to the results using identical data points, however the dataset comparison revealed that the same dataset's ACU was 0.66%.

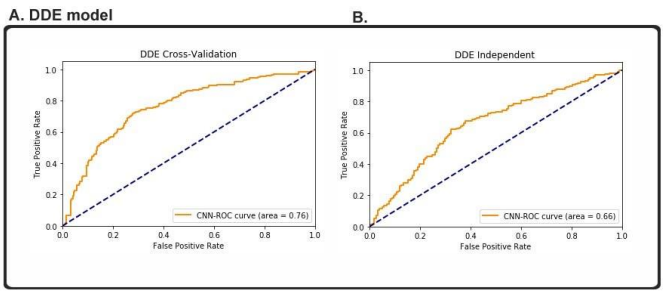

Fig 6. ROC–AUC (A) cross-validation (b) independent test

3.5 Comparison Performance Our Proposed Method with Other Classifiers

In all, these studies show intuitively the validity of results as shown in Fig. 7. Our proposed 2D-CNN method one the basis of independent datasets obtained CNN ROC (AUC) score received is 66.5%, LSTM ROC (AUC) score received is 60.5% and DNN ROC (AUC), score received is 63.7%. Our proposed 2D-CNN method one the basis of cross-validation datasets obtained CNN ROC (AUC) score received is 76.5%, LSTM ROC (AUC) score received is 72.5% and DNN ROC (AUC), score received is 74.7%. We received all ROC (AUC) scores on the basis of 10-fold cross-validation results based on independent. The results analysis of our proposed method is excellent better than other classification models.

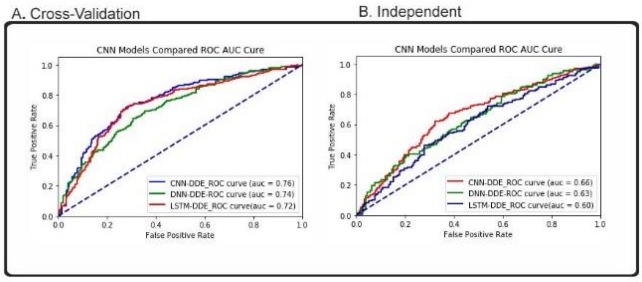

Fig 7. CNN performance comparison with LSTM and DNN

4. Future Direction

Several findings of this study warrant a further discussion about future trends that's way to deal with this gap by modelling the relationship between the protein structure and characteristics derived from human cancerlectins. For future inventions, the algorithmic feasibility methods that concern the single task networks are useful. With these solutions, we can build and test a specific 2D-CNN -based CNN task protein prediction process. Although Lectin2D-CNN increases the precise and precise cancerlectin protein prediction, they still have to reach predictability and efficiency of algorithms. Furthermore, it will aim to define key information about proteins in order to generate more efficient secret characteristics, take into account biological significance

and integrate unique effective algorithms such as 2D-CNN, capsule networks, and generative opponent networks. We will search for and collect more cancerlectin information and update the sample collection in the future [47] can motivate us to develop another approach. It shall also take into account the physicochemical properties, subsequent structures, and other properties of lectins that can help to make a better distinction between cancer and non-cancerlectins. Therefore, we are also trying to provide this method with a robust web server, which will provide the vast majority of researchers and medical scientists with some convenience.

5. Discussion

Bioinformatics and computational biology have increasingly used deep learning, a cutting-edge method in several domains. Certain studies were, however, primarily limited to small protein groups and functional groups. In these experiments, Lectin2D-CNN have been extended to predict different types of protein features like amino acid sequences, 3-D structural including feed-forward feed-force feed-architectures, resource-based neural networking, and deep neural auto-encoder networks, deep re-tasking, and re-investigating feed-forward feed-architectures and similar Lectin2D-CNN types. One of the main barriers to developing practical Lectin2D-CNN prediction tools is technological complex investigation methods that restrict the size of input data and the number of types of functional groups that can be implemented in the process. For this reason, previous research mainly focused on a few families of proteins. New analytical methods are also required that can enable in vitro protein tracking research not only to be highly effective but also to be functional in the real world.

In this study, we propose to predict special protein-specific cancerlectin interactions, for the records of protein sequences as well as for poor and non-characterized open readings, a new Lectin2D-CNN, hierarchical, multi-tasking deep learning method in biological data tools, such as UniProtKB. We also provide a robust Lectin2D-CNN -based predictive model analysis for protein sequence and cancer disease knowledge. We conducted in-depth Lectin2D-CNN with DDE; CNN provided the greatest overall performance among the algorithms tested. When the Cross-validation using 2D-CNN random number generator has produced accuracy score obtain 0.7169%, Sensitivity score obtain 0.7012%, Specificity score obtain 0.7326%, MCC score obtain 0.4428%, ROC-AUC score is 0.76%, respectively, then we know we've attained a reliable result. When the Independent datasets using 2D-CNN random number generator has produced accuracy score obtain 0.6375%, Sensitivity score obtain 0.6160%, Specificity score obtain 0.6589%, MCC score obtain 0.2851%, and ROC-AUC score is 0.76%, respectively.

6. Conclusion

This study provides detailed information of cancerlectin is therefore important to examine cancer diseases because it can help recognize tumour markers as well as prevent, treat, and predict the tumour disease. We provide a great deal of data on protein sequences. To the best of my/our knowledge, no study has focused on deep learning approaches used in cancerlectin as our proposed method. Cancerlectins are a number of broadly spread glycoproteins and/or carbohydrate-binding proteins. Cancerlectins comprise a cancer-related group of lectins that play an important role in human tumour initiation, survival, growth, and metastasis. The evolution from simple algorithms, such as deep convolutional neural networks, to more sophisticated methods, such as various classifications and modern deep neural networks, is very little understood. One of the high performances and important challenges during the genomic age is the functional cancerlectin protein Lectin2D-CNN. For feature extraction models used for sequence expressed cancer-associated proteins, we used feature extraction model named DDE is the most commonly used. The complexity of managing external parameters, such as setting residual positions and integrating structure and multiple sequence information. Our proposed method is precise as the latest methods in the literature are listed. The research consisted of evaluating the related data processing for cancerlectin protein data sets. We used 3 model learning machines that were subsequently comparison and tested. Our proposed Lectin2D-CNN method reached accuracy score obtain 78.53%, prediction accuracy at 58.44% of the Matthew's (MCC) correlation coefficient, which is likely to have a better prediction score at Matthew's correlation coefficient (MCC). Our proposed Lectin2D-CNN method has thus achieved ROC-AUCs with DDE scores of 0.762%, which is superior to that of the other machine learning classifications. The aim of this study was to evaluate the effectiveness of Accuracy, Sensitivity, Specificity, MCC and ROC auc appear to be just as discriminating as measures, so we concluded that they are the best to compare classifiers.

7. Acknowledgements

Prof. Dr. Yuping Wang and Rahu Sikander jointly contributed to the study's design. Ali Ghulam and Jawad Hassan conceived of and completed the review and manuscript. The initial copy of the manuscript was drafted by Laiba Rehman and Nida Jabeen. Natasha Iqbal revised the manuscript and refined the English language usage. The final manuscript has been read and authorized by all authors.

Conflict of Interest: none declared.

8. References

[1] K. K. Kandaswamy, et al., "AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties", Journal of theoretical biology 270.1 56-62, 2011.

[2] L. Halina, and N. Sharon, "Lectins: carbohydrate-specific proteins that mediate cellular recognition", Chemical reviews 98.2: 637-674, 1998.

[3] H. Shen, and T. W. David, "Lectin microarray", PROTEOMICS–Clinical Applications 3.2:148-154, 2009.

[4] N. Sharon, and L. Halina, "Lectins as cell recognition molecules", Science 246.4927:227-234, 1989.

[5] S. H. Mody, Rustom, Antaram Joshi, and William Chaney. "Use of lectins as diagnostic and therapeutic tools for cancer", Journal of pharmacological and Toxicological Methods 33.1:1-10, 1995.

[6] L. Halina, and N. Sharon, "Lectins: carbohydrate-specific proteins that mediate cellular recognition", Chemical reviews 98.2: 637-674, 1998.

[7] H. Y. Lai, X. X. Chen, W. Chen, H. Tang and H. Lin, Sequence-based predictive modelling to identify cancerlectins", Oncotarget, 8(17), 28169, 2017.

[8] L. Halina, and N. Sharon, "Lectins: carbohydrate-specific proteins that mediate cellular recognition", Chemical reviews 98.2: 637-674, 1998.

[9] Liu, Fu-Tong, and A. Gabriel Rabinovich. "Galectins as modulators of tumour progression", Nature Reviews Cancer 5.1: 29-41, 2005.

[10] K. Song, R. Young, T. Billiar, and J. Lee Yong, "Role of galectin-3 in breast cancer metastasis: involvement of nitric oxide", The American journal of pathology 160.3: 1069-1075, 2002.

[11] A. F. Sherwani, et al. "Characterization of lectins and their specificity in carcinomas—an appraisal", Indian Journal of Clinical Biochemistry 18.2:169-180, 2003.

[12] R. E. U. B. E. N. Lotan and A. V. R. A. H. A. M. Raz. "Lectins in cancer cells", Annals of the New York Academy of Sciences, 551: 385-96, 1988.

[13] M. Jordinson, J. Calam, and M. Pignatelli, "Lectins: from basic science to clinical application in cancer prevention", Expert Opinion on Investigational Drugs 7.9:1389-1403, 1998.

[14] U. Schumacher, et al. "Helix pomatia agglutinin binding is a useful prognostic indicator in colorectal carcinoma", Cancer 74.12: 3104-3107, 1994.

[15] D. Mejía, E. González, and V. I. Prisecaru, "Lectins as bioactive plant proteins: a potential in cancer treatment", Critical reviews in food science and nutrition 45.6:425-445, 2005.

[16] C. R. Hill, and G. R. Ter Haar. "High intensity focused ultrasound—potential for cancer treatment", The British journal of radiology 68.816:1296-1303, 1995.

[17] V. R. Gerardo, "Roles of galectins in infection", Nature Reviews Microbiology 7.6: 424-438, 2009.

[18] T. Yau, X. Dan, CC Ng, TB Ng, "Lectins with potential for anti-cancer therapy", Molecules. 20:3791–3810, 2015.

[18] E. Dabelsteen, "Cell surface carbohydrates as prognostic markers in human carcinomas", The Journal of Pathology 179.4:358-369, 1996.

[20] A. Krizhevsky, I. Sutskever, and E. G. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems. 2012.

[21] Z. Lingyun, N. Chonghan, and H. Fuquan, "Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles", Bioinformatics 29.24:3135-3142, 2013.

[22] H. Jun, et al., "Improving DNA-binding protein

prediction using three-part sequence-order feature extraction and a deep neural network algorithm", Journal of Chemical Information and Modeling 63.3, 1044-1057, 2023.

[23] G. Fang, et al., "TargetMM: accurate missense mutation prediction by utilizing local and global sequence information with classifier ensemble", Combinatorial Chemistry and High Throughput Screening 25.1, 38-52, 2022.

[24] B. Ramsundar, et al., "Massively multitask networks for drug discovery", arXiv preprint arXiv: 1502: 02072, 2015.

[25] A. Ghualm, et al. "Identification of pathway-specific protein domain by incorporating hyperparameter optimization based on 2D convolutional neural network", IEEE Access 8, 180140-180155, 2020.

[26] A. Ghulam, 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, 104589, 2022.

[27] Zi-Ang Shen, Tao Luo, Yuan-Ke Zhou, Han Yu, Pu-Feng Du, NPI-GNN, "Predicting ncRNA–protein interactions with deep graph neural networks", Briefings in Bioinformatics, bbab051, 2021. https://doi.org/10.1093/bib/bbab051

[28] D. Damodaran, et al., "CancerLectinDB: a database of lectins relevant to cancer", Glycoconjugate journal 25.3:191-198, 2008.

[29] F. Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data", Bioinformatics 28.23: 3150-3152, 2012.

[30] V. Saravanan, N. Gautham, "Harnessing Computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor", OMICS, 19(10):648-658, 2015. doi:10.1089/omi.2015.0095

[31] V. Saravanan, N. Gautham, "BCIgEPRED—a dual-layer approach for predicting linear IgE epitopes", Mol. Biol. 52 (2), 285–293, 2018.

[32] H.-Y. Lai, C.-Q. Feng, Z.-Y. Zhang, H. Tang, W. Chen, H. Lin, "A brief survey of machine learning application in cancerlectin identification", Curr. Gene Ther. 18 (5), 257267, 2018. https://doi.org/10.2174/1566523218666180913112751.

[33] H. Ding, S.-H. Guo, E.-Z. Deng, L.-F. Yuan, F.-B. Guo, J. Huang, et al., "Prediction of Golgi-resident protein types by using feature selection technique", Chemometr. Intell. Lab. Syst. 124, 9–13, 2013. https://doi.org/10.1016/j.chemolab.2013.03.005.

[34] H. Tang, P. Zou, C. Zhang, et al., "Identification of apolipoprotein using feature selection technique", Sci. Rep. 6, 30441, 2016.

[35] H. Lin, "The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", J. Theor. Biol. 252, 350–356, 2008.

[36] L. Wei, et al., "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides", Bioinformatics 34 (23), 4007–4016, 2018.

[37] D. Gaspar, A.S. Veiga, M.A. Castanho, "From antimicrobial to anticancer peptides", Front. Microbiol. 4, 294, 2013.

[38] X. Li, G. C. Nsofor and L. Song, "A comparative analysis of predictive data mining techniques", International Journal of Rapid Manufacturing, vol. 1, no. 2, pp. 50-172, 2009.

[39] C. Chao, A. Liaw and L. Breiman, "Using random forests to learn imbalanced data", University of California, Berkeley, 2004

[40] Y. Jiao, P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications", Quant. Biol. 4, 320–330. https://doi.org/10.1007/s40484-016-0081-2, 2016.

[41] Z. U. Khan, et al., "iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection", Chemometrics and Intelligent Laboratory Systems 195, 103876, 2019.

[42] S. W. Taju, T. T. D. Nguyen, Nguyen-Quoc-Khanh Le, Rosdyana Mangir Irawan Kusuma, Yu-Yen Ou, "DeepEfflux: a 2D convolutional neural network model for identifying families of efflux proteins in transporters", Bioinformatics, Volume 34, Issue 18, Pages 3111–3117, 2018.

https://doi.org/10.1093/bioinformatics/bty302

[43] C. White, H.D. Ismail, H. Saigo et al. “CNN-BLPred: a Convolutional neural network-based predictor for βLactamases (BL) and their classes”, BMC Bioinformatics 18, 577, 2017. https://doi.org/10.1186/s12859-017-1972-6

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting”, J. Machine Learn. Res. 15 (1), 1929–1958, 2014.

[45] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, C., and E. Elmaghraby, “Ensemble deep learning models for heart disease classification: a case studyfrom mexico”, Information 11 (4), 207, 2020. doi:10.3390/info11040207

[46] L. Pang, J. Wang, L. Zhao, C. Wang and H. Zhan, “A Novel Protein Subcellular Localization Method with CNN-XGBoost Model for Alzheimer’s disease”, Front. Genet. 9, 751, 2019. doi:10.3389/fgene.2018.00751

[47] E. Netto et al. "Herpes simplex and epstein-barr viruses co-infection in early-stage nasopharyngeal carcinoma treated with concurrent chemoradiation: proteomic analysis of formalin-fixed paraffin-embedded samples from a non-endemic region", International Journal of Radiation Oncology• Biology• Physics 105.1: E141, 2019.

