Current Bioinformatics, 2022, *17*, 245-253

RESEARCH ARTICLE

Deep Learning Model for Protein Disease Classification

Farida Alaaeldin Mostafa[1,*], Yasmine Mohamed Afify[1], Rasha Mohamed Ismail[1] and Nagwa Lotfy Badr[1]

[1]*Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt*

ARTICLE HISTORY

Received: May 05, 2021
Revised: July 12, 2021
Accepted: August 29, 2021

DOI:
10.2174/1574893616666211108094205

Abstract: ***Background*:** Protein sequence analysis helps in the prediction of protein functions. As the number of proteins increases, it gives the bioinformaticians a challenge to analyze and study the similarity between them. Most of the existing protein analysis methods use Support Vector Machine. Deep learning did not receive much attention regarding protein analysis as it is noted that little work focused on studying the protein diseases classification.

***Objective*:** The contribution of this paper is to present a deep learning approach that classifies protein diseases based on protein descriptors.

***Methods*:** Different protein descriptors are used and decomposed into modified feature descriptors. Uniquely, we introduce using the Convolutional Neural Network model to learn and classify protein diseases. The modified feature descriptors are fed to the Convolutional Neural Network model on a dataset of 1563 protein sequences classified into 3 different disease classes: AIDS, Tumor suppressor, and Proto-oncogene.

***Results*:** The usage of the modified feature descriptors shows a significant increase in the performance of the Convolutional Neural Network model over Support Vector Machine using different kernel functions. One modified feature descriptor improved by 19.8%, 27.9%, 17.6%, 21.5%, 17.3%, and 22% for evaluation metrics: Area Under the Curve, Matthews Correlation Coefficient, Accuracy, F1-score, Recall, and Precision, respectively.

***Conclusion*:** Results show that the prediction of the proposed CNN model trained by modified feature descriptors significantly surpasses that of Support Vector Machine model.

Keywords: Protein prediction, disease classification, CNN, EMD, IMF, amino acids.

1. INTRODUCTION

Proteins are an important component of every cell in the body. They are used by the body to build and repair damaged tissues. Furthermore, they are essential for making enzymes, hormones, and other body chemicals. They are an important building block of bones, muscles, cartilage, skin, and blood.

Proteins have three different types of structures: primary, secondary, and tertiary [1]. The simplest form of those structures is the primary structure that is composed of a sequence of amino acids bound by peptide bonds. Despite their importance, any change in the primary structure of the protein may lead to different products, resulting in different behavior, which can be lethal.

Recent research shed light on using protein physicochemical properties in extracting features that are used to detect protein similarities. Based on the extracted features, statistics are made that show the extent of the protein similarities, and thus their classification into families, and the determination of their proximity to each other.

Studies have shown that when the similarities between protein sequences are more, their functionality is also more similar [2], which motivated further analysis of the primary sequence of proteins. The protein classification is a topic worth scrutinizing because of its importance in revealing the function of proteins of unknown function or activity that may lead to death.

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt

Since then, protein sequence alignment methods have gained more intensive attention in bioinformatics [2]. One of the main shortages of alignment methods is that they tend to reduce accuracy in exchange for improving efficiency [3].

Deep learning techniques have started to be used frequently and widely in the field of data analysis. Convolutional Neural Network (CNN) has been used in the medical research field, such as analyzing health informatics. It is also noted that researchers in the medical analysis field are moving into CNN and obtaining desirable results [4]. Deep learn-

*Address correspondence to this author at the Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt; E-mail: farida.alaaeldin@cis.asu.edu.eg

2212-392X/22 $65.00+.00

ing did not receive much attention regarding the study of protein analysis. Moreover, the usage of Empirical Mode Decomposition (EMD) with CNN was used in the classification of cardiovascular diseases but not once on protein diseases.

The related work in the field of protein sequence analysis has a lot of variations. We will cover the variations proposed by many researchers starting from the representation that helps in the analysis phase to the algorithms used for feature extraction and making machine learning and neural network models.

One approach shows that 2D data was used to obtain 3D information by using the amino acids evolution index as the first dimension and the class of amino acid information as the second dimension [5]. Then, the DFT is used to transform the sequence signal to the frequency domain. After that, the distance of sequences is computed based on the new numerical sequences to analyze the similarity of protein sequences.

It has been shown that amino acid physicochemical properties are highly related to protein structure and function [3]. Thus, several methods are developed based on these properties; for example, SVM-RQA proposes a scheme for remote homology detection by using both the amino acid properties and Recurrence Quantification Analysis (RQA). SVM-PCD uses the normalized physicochemical distributions of the 4-mers in protein sequences.

A novel position-feature model for protein sequences is based on physicochemical properties of 20 amino acids and graph energy [6]. According to the specific position of amino acids in the sequence, the position-feature matrices consisting of 0 and 1 were constructed the matrices were mapped to bipartite graphs. By computing the energy E of each graph, a characterizing vector E* for the protein sequence is obtained. Modifying the vector E*, a protein-based characteristic B-vector is used, and relative entropy is applied to analyze the similarity/dissimilarity between sequences.

Another method for analyzing protein sequence similarity [7] calculated the spectral radii of 20 amino acids and put forward a novel 2D graphical representation of protein sequences. To characterize protein sequences numerically, three groups of features were extracted and related to statistical, dynamics measurements, and fluctuation complexity of the sequences. With the obtained feature vector, two models utilizing Gaussian Kernel similarity and Cosine similarity are built to measure the similarity between sequences.

Fractal geometry [2] is a non-integer and useful concept in describing the dynamical structure. It is also a useful method for indicating variations in both amplitude and frequency of a signal. Based on the concept of fractal geometry and the physicochemical properties of amino acids, a hybrid method based on discrete wavelet transform and fractal dimension to study and analyze the similarity of proteins is used.

The main highlights of an algorithm for analyzing ECG signals [4] include feature extraction with no need for using selection techniques. An 11-layer CNN model is implemented and validated using 10-fold cross-validation, hence increasing the robustness of the system. Denoising is not required.

Support Vector Machine (SVM) was used for the prediction of Phage Virion proteins using a set of optimal features [8]. A feature selection protocol is employed to identify the optimal features from a large set that included amino acid composition, dipeptide composition, atomic composition, physicochemical properties, and chain-transition distribution.

Deep learning was introduced in protein sequence analysis to predict protein solubility that is significant in pharmaceutical research [9]. CNN is the model used in this research. Fifty-seven features are used to represent the protein sequence that are sequence-based features and structural features.

Deep learning models are also used on another type of feature as Electrocardiograph signals (ECG). Learning features based on machine learning algorithms and CNNs have added an extra boost to the literature and successful ECG signal analysis. It has been extensively used in heart disease classification [4, 10].

In one study, the CNN is designed to handle one-dimensional ECG data, and all the convolution operations in the convolutional layers are performed on the 1D sequence [10]. The kernel size in each layer is modified to be applied to the 1D sequence. The first five layers of the network are convolutional layers and are followed by three fully connected layers. The final output of the network has a soft-max regressor with a specific number of classes that vary among different ECG databases.

Based on the above literature review, it can be noted that most of the existing protein analysis methods use SVM. Deep learning did not receive much attention regarding protein analysis. It is noted that there is limited work focusing on studying the protein diseases classification.

In this paper, a deep learning approach that classifies protein diseases based on protein descriptors is presented. Protein features are calculated using different feature extraction groups (*i.e*., Amino acid composition, C/T/D). The features are then decomposed into Intrinsic Mode Functions (IMF) using EMD. The higher-order IMF is used as modified feature descriptors. A deep learning model is developed using conventional layers. The proposed approach is applied to the feature descriptors of the 1563 protein sequences that are classified into 3 different disease classes. The evaluation metrics used to evaluate the work are Matthews Correlation Coefficient (MCC), precision, recall, accuracy, F1-score, and Area Under the Curve (AUC).

2. MATERIALS AND METHODS

2.1. Dataset

The Dataset used can be downloaded from uniprot.org [11], the universal resource for sequence and functional information relating to proteins. Three sample diseases were chosen: AIDS, tumor suppressor, and proto-oncogene.

The reason behind choosing those three is that they have almost the same number of available protein sequences, which helps in avoiding bias classification problems. The available protein sequences for those diseases are 518, 512, and 567 sequences, respectively. The variance in the number of protein sequences among the three selected diseases are 1.1%, 8.6%, and 9.7%. On the other hand, the variance is massive when compared to other diseases. For example, AIDS has 74.9% more protein sequences than malaria. Proto-oncogene has 38.6% and 82.8% fewer protein sequences than allergen and disease mutations, respectively. Also, tumor suppressor has 59% more protein sequences than epilepsy. The protein sequences have been reviewed [11], and additional biological information about the diseases is also available [12, 13].

A filtration process is required. The protein sequences of the three diseases were compared against each other, and it was found that tumor suppressor and proto-oncogene proteins share 17 identical protein sequences. Thus they were removed.

2.2. Evaluation Metrics

To quantify the performance of the proposed model, the following measures were calculated as shown in Table **1**. TP, TN, FP, and FN are short-term, denoting the total number of True Positive, True Negative, False Positive, and False Negative of instances, respectively.

2.3. The Proposed System Architecture

In this section, we present the system architecture of the proposed work and explain in detail how each module works. The proposed system architecture is shown in Fig. (**1**). It consists of three modules: feature extraction, feature processing,

Table 1. Details on evaluation metrics.

Metric	Equation	Refs.
Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$	[14]
Recall	$\frac{TP}{TP+FN}$	[15]
Precision	$\frac{TP}{TP+FP}$	[16]
F1-score	$2\times\frac{Precision\times Recall}{Precision+\ Recall}$	[17]
Matthews Correlation Coefficient	$\frac{TP\times TN\ -\ FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	[17]
Area Under the Curve	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	[8]

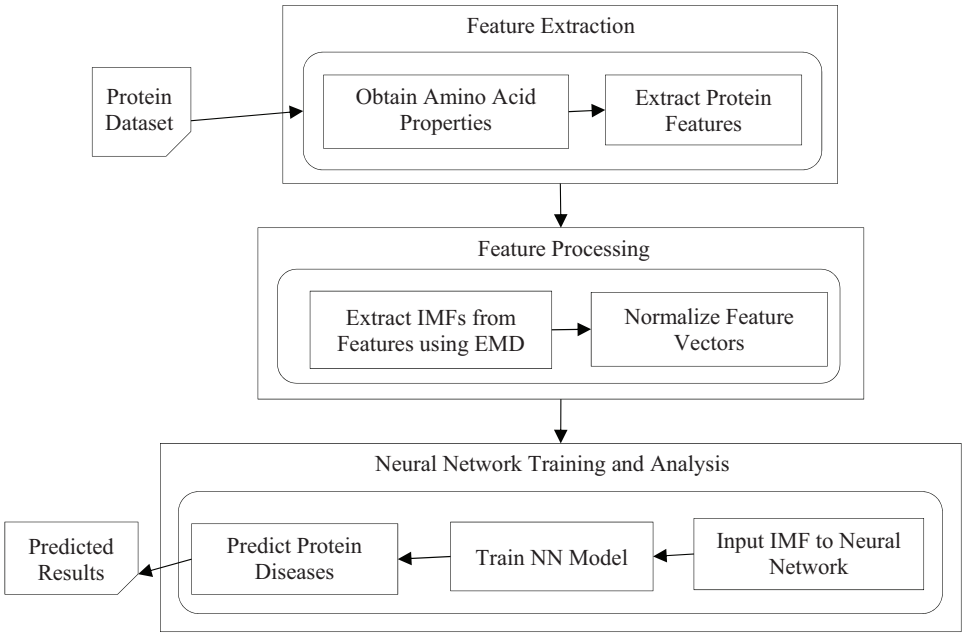

Fig. (1). Protein disease classifier system architecture.

Table 2. A detailed description of the three groups of amino acid feature extraction techniques.

Group	Descriptor	Equation	Number of Features
Amino Acid Composition	Amino Acid Composition	$f(t) = \frac{N(t)}{N}, t \in \{A, C, D, \ldots, Y\}$	20
Group Amino Acid Composition	Grouped Dipeptide Composition	$f(r,s) = \frac{N_{rs}}{N-1}, r, s \in \{g1, g2, g3, g4, g5\}$	25
C/T/D	C/T/D Composition	$C(r) = \frac{N(r)}{N}, r \in \{polar, neutral, hydrophobic\}$	39
	C/T/D Transition	$T(r,s) = \frac{N(r,s) + N(s,r)}{N-1}, r, s \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\}$	39

and neural network training and analysis. Module details are presented in the following subsections.

2.3.1. Feature Extraction Module

In this module, the main step in protein analysis is conducted, which is feature extraction. Because of the amino acid representation, protein sequences cannot be fed to the CNN model. Therefore, this step focuses on the digital representation of a protein sequence into a feature descriptor that can be analyzed using any machine learning methodology. Using features extracted from amino acids guarantees building a powerful predictor [3].

There are twelve groups of feature descriptors [18]. In this paper, we have used the amino acid composition group, grouped amino acid composition, and C/T/D groups. Feature extraction was conducted using the iFeature website [18]. Three groups of feature extraction methods were used, as shown in Table **2**.

2.3.2. Feature Processing Module

In this module, the feature descriptor is decomposed into IMF using EMD [2]. IMF forms the multi-scale feature descriptor from higher frequency to lower frequency. The high-frequency features are then used as they are purified of any noise [10]. The IMFs are then normalized and ready to be used for the next step, which is the model training and analysis.

The feature descriptors are decomposed into IMFs that are ordered from high significance to lower significance (we use significance to refer to frequency). The higher-order IMF represents a modified feature descriptor without most of the noise [10]. The lower order IMFs are neglected as they are of lower frequencies, thus considered as noise. The higher-order IMF is then normalized using standard scaler that follows Eq. (1):

$$z=(x-u)/s \quad (1)$$

Where x is the feature, u is the mean, and s is the standard deviation.

2.3.3. Neural Network Training and Analysis Module

In this module, the normalized IMFs are fed to the CNN constructed of 8 layers (3 convolutional, 1 flatten, and 4 fully connected layers). The model is then trained, validated, and finally tested using the testing set. After the model is trained, it is ready to be used for prediction. The model is built using three components: (I) convolutional layers, (II) activation functions, and (III) dense layers.

2.3.3.1. Convolutional Layer

The convolutional layer serves as the main block of a CNN as it conducts computationally intensive lifting [4]. It aims to extract features from the protein descriptor and learns to predict. In our model, the type of the first three layers of the model is convolutional layers.

2.3.3.2. Activation Functions

The activation functions help in the learning process. The Rectified Linear activation function (ReLU) [10] is a piecewise linear function that will output the input directly if it is positive; otherwise, it will output zero. It allows the model to learn faster and perform better. It is used in all layers except the output layer (layer 8) that has a SoftMax activation function.

2.3.3.3. Dense Layers

The dense layers form fully connected networks. The last dense layer has an output of three neurons as it represents the three labels of the three diseases used in the training process.

2.4. The Proposed Deep Learning Model

In this section, we present the detailed structure of the proposed deep learning model. The deep learning model proposed is built using CNN. The CNN model is designed to handle 1D data, and all the convolution operations in the convolutional layers are performed on the 1D sequence. The kernel size in each layer is different. The first three layers of the network are convolutional layers and are followed by one flatten layer and four fully connected layers. The final output of the network has a SoftMax regressor with a specific num-

Table 3. Detailed description of the 8 layers of the CNN model used for disease classification. (~) differs based on the input size of the feature descriptor.

Layers	Type	Number of Filters	Activation Function	Kernel Size
1	Convolution	512	ReLU	8
2	Convolution	256	ReLU	4
3	Convolution	128	ReLU	1
4	Flatten	~	ReLU	-
5	Dense	64	ReLU	-
6	Dense	32	ReLU	-
7	Dense	16	ReLU	-
8	Dense	3	SoftMax	-

ber of classes that vary among different databases. In our work, there are three classes, so the output has three neurons.

Based on the numbers from Table **2**, the sizes of the feature descriptors used are different, which causes the difference in the number of neurons in each layer of the first four layers. The first convolutional layer using the AAC feature descriptor is fed by a 20 × 1 sequence and is modified with 512 filters of size 8 × 1. Consequently, an output of size 13 × 512 is produced. Similarly, the other three feature descriptors, *i.e.*, GDPC, CTDC, and CTDT, are used. The input size will be 25 × 1, 39 × 1, and 39 × 1, respectively. The output resulted is 18 × 512 using the GDPC feature vector and 32 × 512 using CTDT and CTDC feature descriptors.

The second convolutional layer uses 256 filters of shape 4 × 1 that converts the output resulting from AAC feature descriptor to a shape of 10 × 256. The output resulting from the GDPC feature descriptor is 18 × 256, and from CTDT and CTDC, the feature descriptor is 29 × 256.

The third convolutional layer converts the respective input feature vector space to a shape of 10 × 128, 15 ×128, 29 × 128, and 29 × 128 using AAC, GDPC, CTDT, and CTDC feature descriptors, respectively. The corresponding layer type, number of filters, activation function, and kernel size are shown in Table **3**.

Throughout the CNN model, ReLU acts as the nonlinear activation function used in all layers except the eighth layer that uses the SoftMax activation function to classify the descriptor to the desired class. The flatten layer converts the output of the previous convolutional layer to feed layer number 5 that is the first fully connected layer. The model is trained using 10-fold cross-validation. Also, a validation set is used using a validation split equal to 0.01. The number of epochs used is 40 epochs for each fold and a batch size of 5.

3. RESULTS AND DISCUSSION

In order to assess the proposed deep learning model, a comprehensive set of experiments was conducted with respect to the model accuracy. The objectives of the experiments are: (I) compare SVM using different kernels to find the best SVM kernels to be used in further experiments, (II) compare the usage of CNN with popular SVM in predicting diseases, and (III) show the impact of IMFs to the performance of CNN and SVMs.

3.1. Experiment I

The objective of this experiment is to compare the performance of the SVM algorithm with different kernel functions on the four features without being modified. The kernel functions used are linear, polynomial, Radial Basis Function (RBF), and sigmoid. We will be using them as SVM-linear, SVM-poly, SVM-rbf, and SVM-sigmoid, respectively. It is noted that SVM performs well when it comes to a balanced dataset [17]. Therefore, it is suitable for comparison with our proposed model (with its balanced dataset), and it stands as a good competitor to our CNN model.

As shown in Fig. (**2**), it is observed that SVM-poly outperforms in accuracy for all features used. SVM-poly using CTDT features shows the highest accuracy and outperforms SVM-rbf, SVM-linear, and SVM-sigmoid by 19.8%, 21.9%, and 56.7%, respectively. Similarly, SVM-rbf comes second in place, outperforming SVM-linear and SVM-sigmoid. Since the highest accuracies reached have been obtained by SVM-poly and SVM-rbf, they will be used in further experiments.

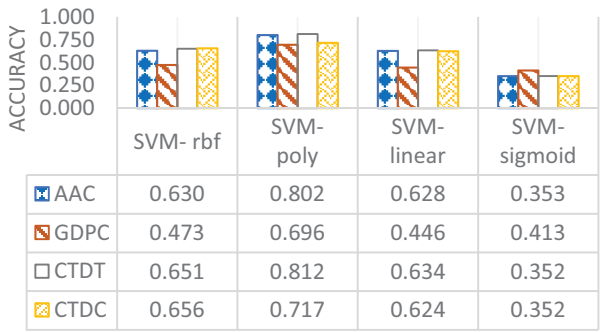

Fig. (2). Accuracy comparison between SVM with four different kernel functions. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

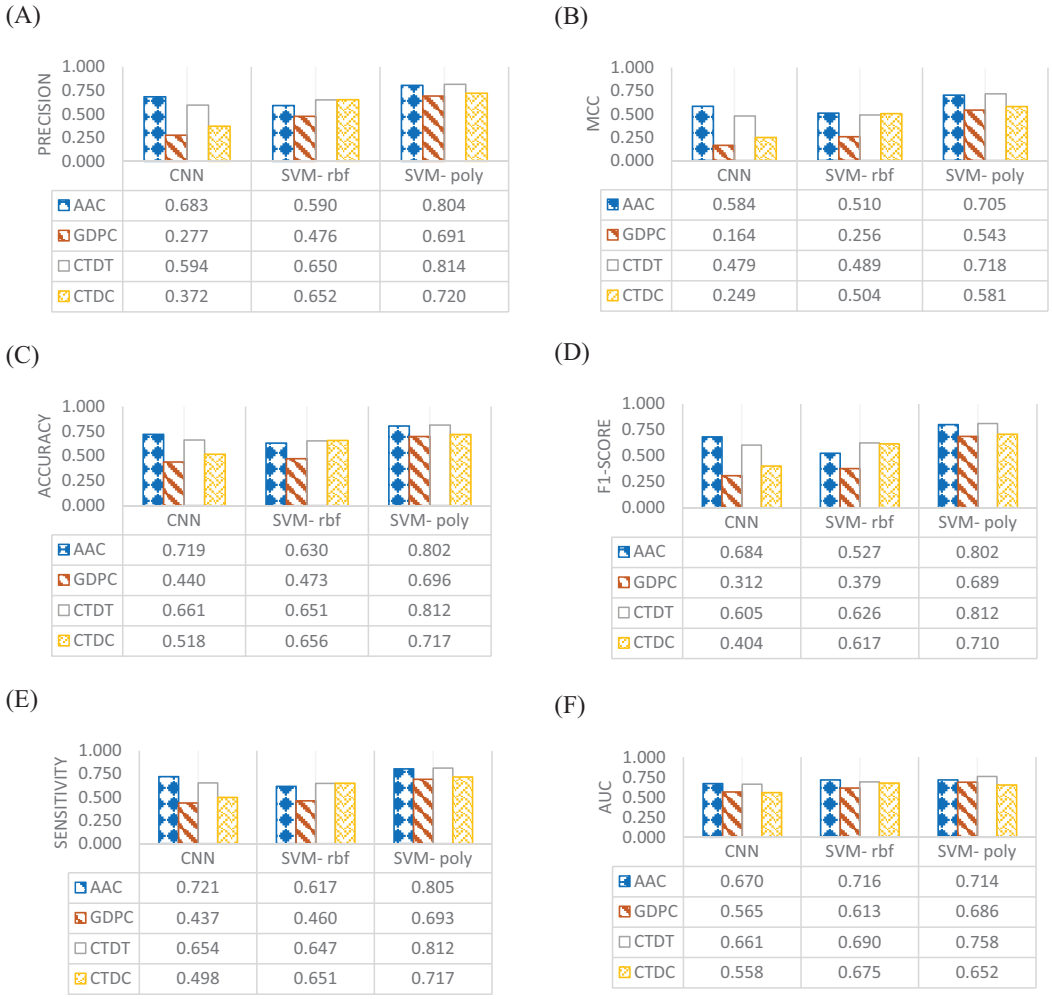

Fig. (3). Performance comparison of CNN, SVM-rbf, and SVM poly using normal feature vectors with several evaluation metrics: (**A**) Precision, (**B**) MCC, (**C**) Accuracy, (**D**) F1-score, (**E**) Recall, and (**F**) AUC. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

3.2. Experiment II

The objective of this experiment is to compare the performance of the CNN model with the SVM algorithm using the two superior kernel functions from experiment I (SVM-poly and SVM-rbf). The four normal feature descriptors are fed to the CNN, SVM-poly, and SVM-rbf.

Fig. (**3**) shows a detailed comparison between CNN, SVM-poly, and SVM-rbf using the six-evaluation metrics explained earlier in this section. It can be noted that SVM-poly outperforms both CNN and SVM-rbf using normal feature vectors in all evaluation metrics. SVM-poly shows better precision outperforming SVM-rbf and CNN by 20.1% and 27%, respectively, using CTDT normal feature descriptor as an example. MCC is superior in SVM-poly than SVM-rbf and CNN by 31.9% and 33.3%, respectively. SVM-poly surpasses CNN in accuracy, recall, F1-score, and AUC by 18.6%, 19.6%, 25.5%, and 12.8%, respectively. SVM-rbf outperforms CNN by 6.4%, 42.9%, 23.5%, 2%, 21%, and 3.3% in AUC, precision, recall, MCC, accuracy, and F1-score, respectively.

3.3. Experiment III

The objective of this experiment is to compare the performance of the CNN, SVM-poly, and SVM-rbf using the modified feature descriptors using EMD. Fig. (**4**) shows a significant enhancement in all evaluation metrics, ultimately making CNN model superior to both SVM-poly and SVM-rbf. As shown in Fig. (**4**), the performance of CNN and SVM-rbf enhanced when using the four modified feature descriptors in all evaluation metrics.

As shown in Fig. (**3**), it can be noted that SVM-poly did not achieve better results when using the modified feature descriptors than those in experiment II. It can be shown that CNN's performance surpasses the performance of SVM-rbf by 4%, 24.2%, 23.3%, 24%, 29%, and 10.3% in precision, recall, accuracy, F1-score, MCC, and AUC, respectively.

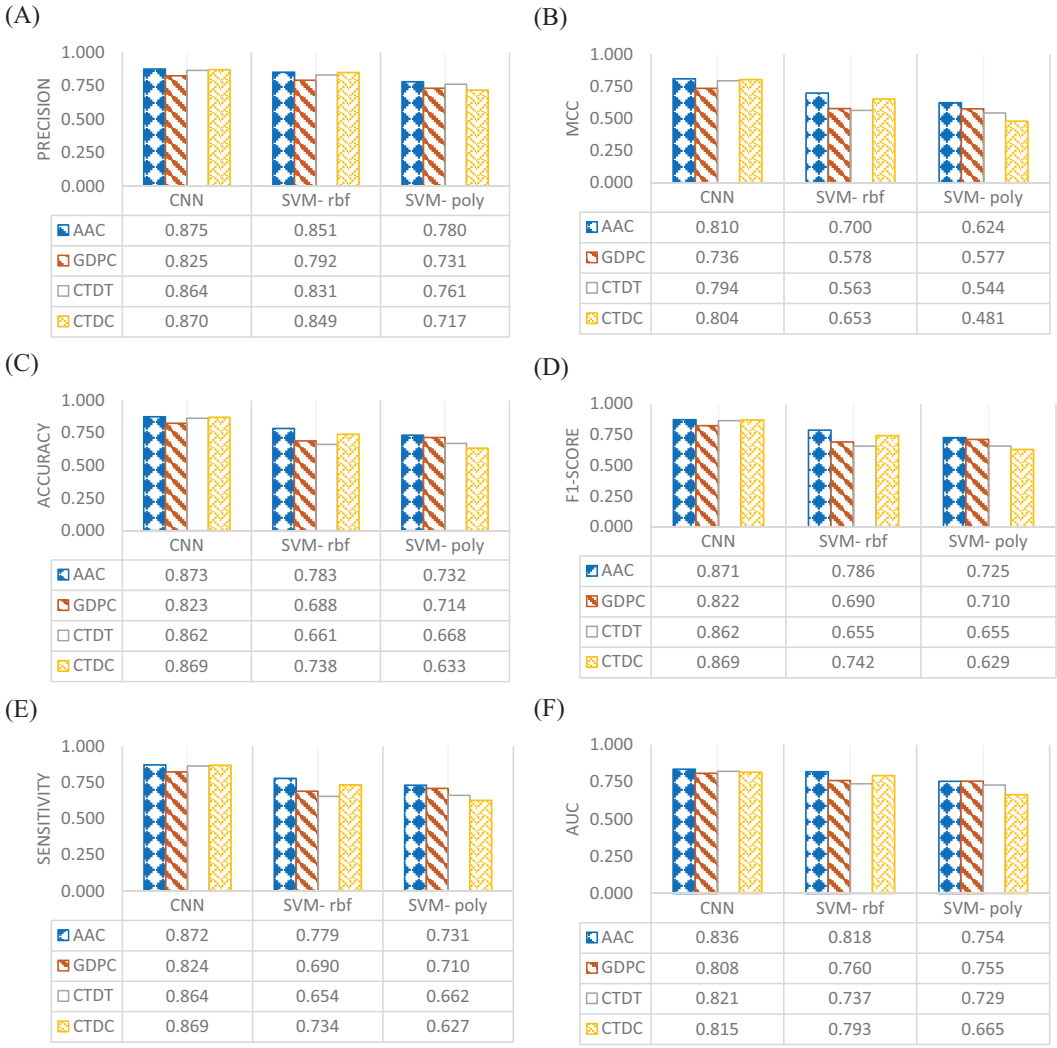

Fig. (4). Performance comparison of CNN, SVM-rbf, and SVM poly using modified feature vectors with several evaluation metrics: (**A**) Precision, (**B**) MCC, (**C**) Accuracy, (**D**) F1-score, (**E**) Recall, and (**F**) AUC. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

CNN also surpasses SVM-poly by 17.6%, 27.8%, 27.2%, 27.6%, 40.1%, and 18.4% in precision, recall, accuracy, F1-score, MCC and AUC, respectively.

To better represent the impact of IMFs, Fig. (**5**) illustrates the improvement of the CNN model predictions using the four modified feature descriptors. AUC has the highest improvement of 31.6% using the modified CTDC feature. MCC, accuracy, F1-score, precision, and recall show significant improvement by 77.7%, 46.6%, 62.1%, 47%, and 66.4%, respectively, when using the modified GDPC modified feature descriptors. It can be noted that using AAC and CTDT modified feature descriptors also have noticeable improvements on the CNN model. AAC modified feature descriptors improved by 19.8%, 27.9%, 17.6%, 21.5%, 17.3%, and 22% in AUC, MCC, accuracy, F1-score, recall, and precision, respectively. CTDT modified feature descriptors also improved by 19.5%, 39.6%, 23.3%, 29.9%, 24.3%, and 31.2% in AUC, MCC, accuracy, F1-score, recall, and precision, respectively.

CONCLUSION

Deep learning techniques such as CNN have not been exploited in protein classification. The objective of this work is to introduce the application of CNN to protein diseases classification. Four types of protein features were extracted from the protein sequences and were modified using EMD and then trained on CNN and SVM models. SVM-poly shows superior results when using the normal feature vectors.

The usage of the modified feature descriptors shows a significant increase in the performance of the CNN model over SVM using both poly and rbf kernel functions. AAC modified feature descriptors improved by 19.8%, 27.9%, 17.6%, 21.5%, 17.3%, and 22% in AUC, MCC, accuracy,

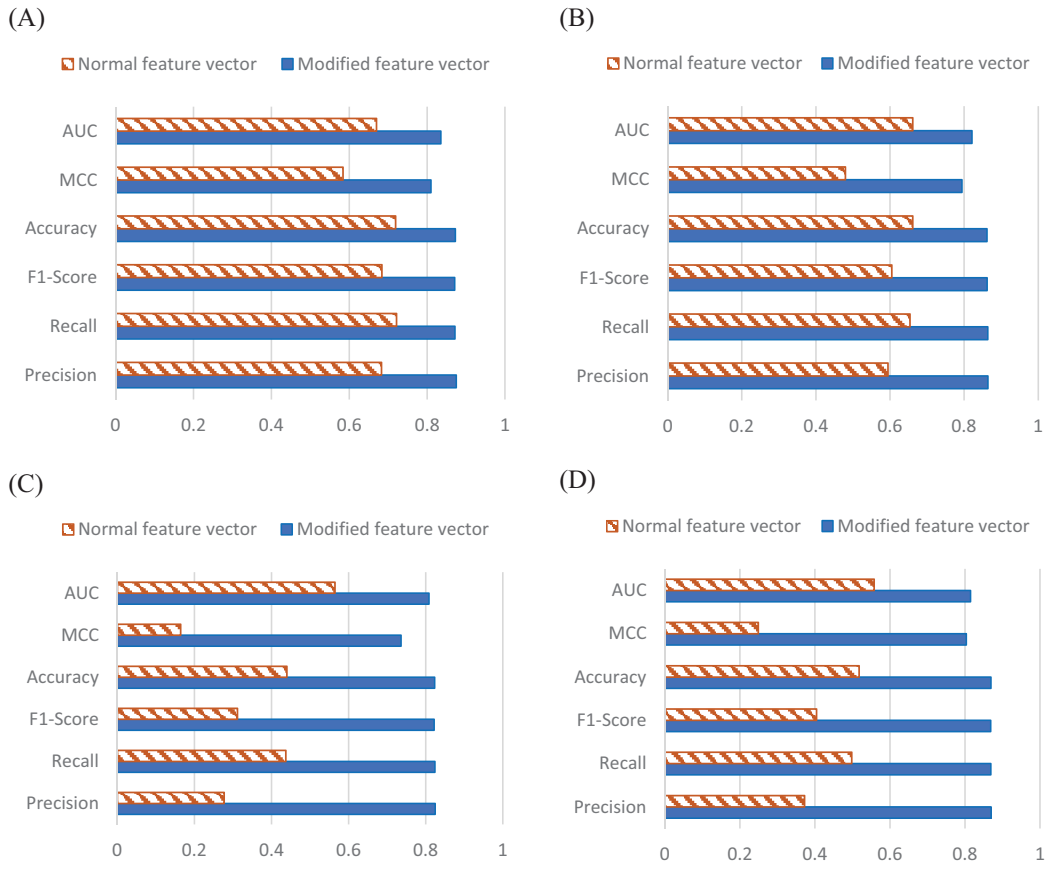

Fig. (5). Evaluation metrics on CNN model using normal and modified feature vectors on: (**A**) AAC, (**B**) CTDT, (**C**) GDPC, and (**D**) CTDC. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

F1-score, recall, and precision, respectively. CTDT modified feature descriptors also improved by 19.5%, 39.6%, 23.3%, 29.9%, 24.3%, and 31.2% in AUC, MCC, accuracy, F1-score, recall, and precision, respectively.

The results show that the CNN model trained by the modified feature descriptors using EMD has the highest performance in comparison to using normal feature descriptors. This encourages us to exert more effort in respect of applying EMD on other sets of features and use different methods for extracting features.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this study are available within the article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Gupta CLP, Bihari A, Tripathi S. Protein classification using machine learning and statistical techniques: A comparative analysis. Recent Adv Comput Sci Commun 2019; 14(5): 16161-32.

[2] Yang L, Wei P, Zhong C, Meng Z, Wang P, Tang YY. A Fractal dimension and empirical mode decomposition-based method for protein sequence analysis. Int J Pattern Recognit Artif Intell 2019; 33(11): 19400202.

http://dx.doi.org/10.1142/S0218001419400202

[3] Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform 2018; 19(2): 231-44. http://dx.doi.org/10.1093/bib/bbw108 PMID: 27881430

[4] Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Inf Sci 2017; 415-416: 190-8. http://dx.doi.org/10.1016/j.ins.2017.06.027

[5] Liao B, Liao B, Lu X, Cao Z. A novel graphical representation of protein sequences and its application. J Comput Chem 2011; 32(12): 2539-44. http://dx.doi.org/10.1002/jcc.21833 PMID: 21638292

[6] Yu L, Zhang Y, Gutman I, Shi Y, Dehmer M. Protein sequence comparison based on physicochemical properties and the position-feature energy Matrix. Sci Rep 2017; 7: 46237. http://dx.doi.org/10.1038/srep46237 PMID: 28393857

[7] Wu C, Gao R, De Marinis Y, Zhang Y. A novel model for protein sequence similarity analysis based on spectral radius. J Theor Biol 2018; 446: 61-70. http://dx.doi.org/10.1016/j.jtbi.2018.03.001 PMID: 29524440

[8] Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. Front Microbiol 2018; 9: 476. http://dx.doi.org/10.3389/fmicb.2018.00476 PMID: 29616000

[9] Khurana S, Rawi R, Kunji K, Chuang GY, Bensmail H, Mall R. DeepSol: A deep learning framework for sequence-based protein solubility prediction. Bioinformatics 2018; 34(15): 2605-13. http://dx.doi.org/10.1093/bioinformatics/bty166 PMID: 29554211

[10] Hasan NI, Bhattacharjee A. Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition. Biomed Signal Process Control 2019; 52: 128-40. http://dx.doi.org/10.1016/j.bspc.2019.04.005

[11] Uniprot. Available from: https://uniprot.org

[12] American Cancer Society team. Oncogenes and tumor suppressor genes. USA: American Cancer Society Inc. 2014.

[13] CDC. HIV Basics. Available from: https: //www.cdc.gov/hiv/basics/whatishiv.html

[14] Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II--a new method for protein solubility prediction. FEBS J 2012; 279(12): 2192-200. http://dx.doi.org/10.1111/j.1742-4658.2012.08603.x PMID: 22536855

[15] Liu L. Combining sequence and network information to enhance protein-protein interaction prediction. BMC Bioinform 2020; 21(16): 537.

[16] Zhou G, Wang J, Zhang X, Guo M, Yu G. Predicting functions of maize proteins using graph convolutional network. BMC Bioinform 2020; 21(16): 420.

[17] Zhang S, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J Theor Biol 2018; 437: 239-50. http://dx.doi.org/10.1016/j.jtbi.2017.10.030 PMID: 29100918

[18] Chen Z, Zhao P, Li F, *et al.* iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018; 34(14): 2499-502. http://dx.doi.org/10.1093/bioinformatics/bty140 PMID: 29528364