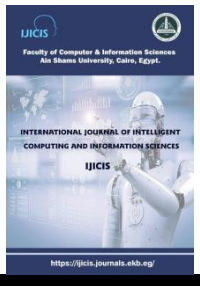
IJICIS, Vol.22, No.2, 112-125
DOI: 10.21608/ijicis.2022.123177.1168

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



UNCOVERING THE EFFECTS OF DATA VARIATION ON PROTEIN SEQUENCE CLASSIFICATION USING DEEP LEARNING

Farida A. Mostafa*

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt
0000-0002-9982-1030
farida.alaaeldin@cis.asu.edu.eg

Yasmine M. Afify

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt
0000-0001-6400-8472
yasmine.afify@cis.asu.edu.eg

Rasha Ismail

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt
0000-0003-3581-8112
rashaismail@cis.asu.edu.eg

Nagwa Badr

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University
Cairo, Egypt
0000-0002-5382-1385
nagwabadr@cis.asu.edu.eg

Received 2022-02-21; Revised 2022-05-01; Accepted 2022-05-06

Abstract: *Bioinformaticians face an issue in analyzing and studying protein similarity as the number of proteins grows. Protein sequence analysis helps in the prediction of protein functions. It is critical for the analysis process to be able to appropriately categorize proteins based on their sequences. The extraction of features from protein sequences is done using a variety of methods. The goal of this study is to investigate the different variations of data on the classification performance of a deep learning model employing 3D data. First, few research questions were formulated regarding the impact of the following criteria: dataset size, IMF importance, feature size, and preprocessing on the proposed deep learning classification process. Second, comprehensive experiments were conducted to answer the research questions. Six feature extraction methods were utilized to create 3D features with two sizes (7x7x7 and 9x9x9), which were then fed into a convolutional neural network. Three datasets different in their sorts, sizes, and balance state were used. Accuracy, precision, recall and F1-score are the standard assessment metrics used. Experimental results draw significant conclusions. First, the 7x7x7 feature matrix has a positive correlation between its dimensions, which improved the results. Second, using the sum of the first three IMF components had better impact than using the first IMF component. Third, the classification process did not benefit from the normalization of features for small datasets unlike the large dataset. Finally, the dataset size had a significant impact on training the CNN model, with a training accuracy reaching 84.03%.*

Keywords: *Deep Learning, Proteins, EMD, IMF, Feature Matrix.*

***Corresponding Author**: Farida A. Mostafa

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

Email address: farida.alaaeldin@cis.asu.edu.eg

1. Introduction

Proteins are essential components of bones, muscles, cartilage, skin, and blood, as well as the building blocks of every cell in the body. They're also necessary for the manufacturing of hormones, enzymes, and other biological substances. Proteins are necessary for the formation and healing of damaged tissues, as well as the production and release of hormones like insulin, serotonin, and calcium.
The three types of structures discovered in proteins are primary, secondary, and tertiary structures [1]. The most fundamental of these structures is the primary structure, which is made up of a series of amino acids connected together by peptide bonds. Any change in a protein's main structure, despite its significance, can lead to different product, which can lead to deadly behavior.

Protein sequences are constantly increasing as a result of sequencing technologies that supply protein sequences to bioinformaticians. Sequencing techniques create data in the form of a sequence of amino acids expressed as various compositions of twenty characters. This data cannot be utilized as an input for machine learning or deep learning models due to its nature. As a result, feature extraction is a critical step in protein analysis.

Recent studies look at how to extract 2D features from proteins based on their physicochemical properties. Statistical approaches are then utilized to discover protein comparisons using the extracted features. According to research, protein sequence similarity correlates with functional similarity [2], indicating that protein primary sequences should be investigated further. Since then, protein sequence alignment techniques have gotten a lot of interest in bioinformatics [2]. One of the fundamental disadvantages of alignment methods is that they favor accuracy above efficiency [3].

In the realm of data analysis, deep learning algorithms are being employed increasingly often and broadly. The Convolutional Neural Network (CNN) has been employed in medical research to examine health informatics, among other things. CNN is also being used by researchers in the field of medical analysis, and the results are promising [4]. Deep learning has received very little attention when it comes to protein analysis.

The motivation behind this study is to address four research questions, namely:
Q1: Does the size of the feature matrix affect deep learning model accuracy?
Q2: How does employing different IMF components affect the results?
Q3: Examine the effect of normalization on accuracy results.
Q4: Study the impact of data size on training and validation accuracies.

The objective of this paper is to study unattended research area of protein sequence classification using deep learning employing 3D data. Investigating the different variations of data on the classification performance of a deep learning model employing 3D data is crucial. To achieve this purpose, three datasets were used, each with its own kind, size, and balancing state. Accuracy, precision, recall and F1-

score are the standard assessment metrics used. The results of the studies show that when employing the suggested deep learning model, the size, preprocessing, and dimensionality of the features along with dataset size play a vital influence in the protein categorization process.

The content of this study is organized as follows. First, some of the relevant work that's been done in the field is presented in section 2. Details of the investigation are described in section 3. Section 4 represents the results and discussion. Finally, in the conclusion, section 5 puts the study to a close.

2. Related Work

In the recent several decades, the area of protein sequence analysis has changed dramatically, with the appearance of machine learning and neural networks as important participants. This section will be covering some of the alterations proposed by many researchers.

Authors in [5] explain how to use 2D data to create 3D information by utilizing the amino acids evolutionary index as first component and the class of amino acid values as the second. The sequence signal is then altered to the frequency domain by applying Discrete Fourier Transform (DFT). The distance between sequences is then calculated using the new numerical sequences to determine how alike protein sequences are.

In [3], the physicochemical characteristics of amino acids have been shown to be strongly linked to protein structure and function. As a result, a slew of approaches based on these characteristics have emerged. By integrating amino acid properties with RQA, SVM-Recurrence Quantification Analysis (RQA) offers a method for finding distant homology.

The non-integer idea of fractal geometry [2] can be utilized to characterize the dynamical structure of a signal. It can also be used to indicate amplitude and frequency variations. Based on the notion of fractal geometry and the physicochemical characteristics of amino acids, the study team proposed a hybrid technique based on discrete wavelet transform (DWT) and fractal dimension to explore and assess protein similarities.

A support vector machine was used to forecast Phage Virion Proteins (PVPs) based on a set of optimal qualities, according to [6]. Amino acid composition, dipeptide composition, atomic composition, physicochemical properties, and chain transition distribution were among the features chosen using a feature selection approach from a large number of alternatives.

Using a benchmark dataset, PVPred-SCM, a predictor developed by [7], is used to predict and evaluate PVPs. PVPred-SCM was created by estimating the propensity scores of 400 dipeptides using the scoring card technique (SCM) in conjunction with just DPC for predicting and evaluating PVPs. The gene scores were adjusted using their unique genetic algorithm to improve prediction performance.

Research presented by [8] employs a unique sequence-based meta-predictor dubbed the Meta-iPVP to distinguish PVPs from non-PVPs. They supplied a new balanced dataset with 313 empirically confirmed PVPs and non-PVPs. They use a feature representation system that combines probabilistic data from four machine learning techniques with seven different feature encodings. The probabilistic characteristics were employed as input features in the SVM model.

A collection of research for the recognition of the HAR in terms of machine and deep learning approaches are described in [9]. The data, characteristics, and classification techniques from the investigations were summarized. Deep learning outperformed machine learning in terms of accuracy and the number of activities that were employed and identified. It is recommended that more contemporary deep learning algorithms be used.

The authors of [10] propose a 3D-CNN based on an upgraded sparse autoencoder for Alzheimer's disease prediction. The 3D-CNN was used to produce predictions based on the learnt records, and the sparse autoencoder was used to learn the best data representation. With the Adam approach and batch normalization, the SAE was improved. The data used was MRI imaging data from Alzheimer's disease patients.

In [11], authors employed deep learning to categorize Melanoma. Melanoma is a fatal, metastatic cancer that may spread to other organs and tissues. Early identification is critical to recovering from melanoma and reducing death. They applied a variety of original picture alterations to expand the number of photos in the training set, including grid distortion, horizontal flip, and vertical flip. The findings of the skin lesion picture segmentation trials demonstrate that the five structures worked effectively and closely.

The common between available feature extraction approaches is that they only extract 2D or 1D data, according on the prior literature review. Model development with 3D data has received very little attention. Machine learning models, such as SVM, are the most common categorization models. It's worth noting that the research of protein diseases didn't get the attention it deserved. Deep learning was used in other fields having a positive impact on the results encouraging applying it to protein studies.

3. Materials and Methods

3.1. Protein Datasets

In this section, the methodology for selecting the datasets is explained in depth. This study employed a number of different datasets. There are two PVP databases, that differ in size and balancing state, in addition to the disease dataset. The features will be deliberated in this section.

Table 1 Detailed description of protein dataset sizes.

	PVP-Benchmark		PVP-Balanced		Disease		
	PVP	Non-PVP	PVP	Non-PVP	AIDS	Tumor Suppressor	Proto-oncogene
Training	99	208	250	63	388	383	425
Independent	30	64	250	63	130	129	142
Total	129	272	500	126	518	512	567

3.1.1. Phage Virion Proteins datasets

Two datasets of phage virion proteins are used. The main difference between the two datasets is that the first is a benchmark dataset that has been utilized in multiple research papers [6]–[8] and will be referred to as the PVP-Benchmark dataset from now on. The second dataset is the PVP-Balanced dataset, which was recommended by [7] as a unique balanced dataset. Each dataset comes with both training and independent datasets.

3.1.2. Disease Dataset

Uniprot.org [12], a global database of protein sequences and functions, provided the disease dataset. were Three were chosen for this study: AIDS, tumor suppressor, and proto-oncogene. These three diseases were considered because they had almost similar numbers of accessible protein sequences, which helps to reduce classification bias. For each of these disorders, there are 518, 512, and 567 protein sequences accessible, respectively.

A filtration technique was necessary. When comparing protein sequences of the three diseases, it was discovered that tumor suppressor and proto-oncogene proteins had 17 identical protein sequences, leading to their removal. Table 1 contains further information about the datasets.

3.2. Proposed Protein Classification Deep Learning Model

This section describes the suggested strategy in depth, including a workflow diagram and a step-by-step explanation. The suggested technique, as illustrated in Figure 1, involves five phases: feature extraction, feature processing, model construction, model training, and protein classification and validation.

3.2.1. Feature Extraction Phase

In this phase, feature extraction, the most important step in protein analysis, is introduced. Because of the amino acid representation, protein sequences can't be fed to the CNN model. As a result, the focus of this stage is on transforming the amino acid representation of a protein sequence into a feature descriptor that can be examined using any machine learning approach. Using amino acid properties as a predictor assures the construction of a powerful predictor. [3].

A total of twelve feature descriptor categories are available [13]. This study employed the amino acid composition group, as well as the C/T/D, conjoint triad, and quasi-sequence-order groups. To extract features, the iFeature python package was used. The five types of feature extraction techniques used are listed in Table 2.

Two feature matrices are provided. The protein sequences were initially subjected to feature extraction methods in order to produce those matrices. The protein sequence was subjected to the conjoint triad approach in order to obtain the 1x343 feature vector. The 1x729 feature vector, on the other hand, was obtained using a variety of feature extraction methods, including CTriad, AAC, GDPC, CTDC, CTDT, CTDD, and SOC number.

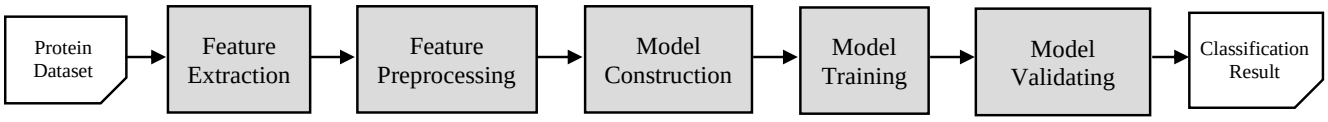


Figure 1: Proposed workflow diagram for protein classification.

Table 2 The five types of amino acid feature extraction procedures are described in depth.

Group	Descriptor	Equation	Number of Features
Amino Acid Composition	Amino Acid Composition	$\boldsymbol{f(t) = \frac{N(t)}{N}, t \in \{A, C, D, \ldots, Y\}}$	20
Group Amino Acid Composition	Grouped Dipeptide Composition	$\boldsymbol{f(r,s) = \frac{N_{rs}}{N-1}, r, s \in \{g1, g2, g3, g4, g5\}}$	25
C/T/D	C/T/D Composition	$\boldsymbol{C(r) = \frac{N(r)}{N}, r \in \{polar, neutral, hydrophobic\}}$	39
	C/T/D Transition	$\boldsymbol{T(r,s) = \frac{N(r,s) + N(s,r)}{N-1}, r, s \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\}}$	39
	C/T/D Distribution	The Distribution descriptor consists of five values for each of the three groups (polar, neutral and hydrophobic)	195
Conjoint Triad	CTriad	$\boldsymbol{d_i = \frac{f_i - min\{f_1, f_2, \ldots, f_{343}\}}{max\{f_1, f_2, \ldots, f_{343}\}}}$	343
Quasi-Sequence-Order	SOC Number	$\boldsymbol{\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2, d = 1, 2, 3, \ldots, nlag}$	68

3.2.2. Feature Preprocessing Phase

In this phase, the feature vectors are processed using EMD and decomposition into six IMFs. From higher frequency to lower frequency, IMF creates a multi-scale feature description. The noise is removed from the high frequency characteristics, which are then used [14]. Higher order IMFs are

normalized using the usual scaler. The feature matrices are then constructed by resizing the normalized feature vectors, with the 1x343 being resized to a 7x7x7 feature matrix and the 1x729 to a 9x9x9 feature matrix.

3.2.3. Model Construction Phase

In this phase, a detailed explanation of the layers and dimensions of the deep learning model will be presented. The core component of a CNN is the convolutional layer, which performs computationally demanding lifting. Its goal is to learn to anticipate by extracting characteristics from the protein description. The Dropout layer is a mask that removes certain neurons' contributions to the next layer while leaving the rest alone.
Dropout layers are important in CNN training because they prevent the training data from being overfit. The dense layer is a simple layer of neurons in which each neuron receives input from all of the neurons in the previous layer, thus the name. The Dense Layer classifies characteristics based on the output of convolutional layers. The proposed model has one input layer, three 3D convolutional layers, one flatten layer, four dense layers, and dropouts make up the CNN model as shown in Table 3.

The input layer of the model has dimensions of NxNxN, which correspond to the dimension of the feature matrix. The feature matrix is then transferred to the convolutional 3D layer, which is the second layer. The output of the dimensions 5x5x7 is then sent via a dropout layer, which helps the CNN model maintain consistency throughout training. The second Conv3D layer receives the same-dimensional output and creates a 4x4x6 feature matrix, which is passed to the second dropout layer. The third Conv3D layer takes the dropout layer's output and generates a 4x4x6 feature matrix.

To transform a 3D feature matrix with dimensions of 4x4x6 into a 1D feature vector with dimensions of 49152, the flatten layer uses the 512 filters. Data is sent over a series of dense and dropout layers. A 64x1 dense layer feeds features to a 64x1 dropout layer, resulting in identical-dimensional features. The 64x1 dimension is received by the second dense layer, which transforms it to a 32x1 dimension feature. The 32x1 feature is received by a dropout layer and sent to the third dense layer, which transforms it to a 16x1 feature. Finally, the output of the dense layer is sent to the final dropout layer, which passes it on to the 1x1 output layer, which outputs the classification result.

3.2.4. Model Training Phase

In this phase, the CNN model is trained using the 3D features extracted and processed. The model is trained using 10-fold cross-validation. Also, a validation set is used employing a validation split equal to 0.01. The factors that govern the network topology and how the network is trained are referred to as hyperparameters. Before training, hyperparameters are established. Regarding hyper parameter optimization, the number of epochs used is 40 epochs for each fold, batch size is 5, activation function used is ReLU, and the optimizer used is Adam with learning rate 0.01.

Table 3 Complete explanation of the 12 layers of the CNN model used for protein classification.

Layers	Type	Number of Filters	Activation Function	Kernel Size
1	Convolution 3D	128	ReLU	8
2	Dropout	128	ReLU	-
3	Convolution 3D	256	ReLU	4
4	Dropout	256	ReLU	-
5	Convolution 3D	521	ReLU	1
6	Flatten	49152	ReLU	-
7	Dense	64	ReLU	-
8	Dropout	64	ReLU	-
9	Dense	32	ReLU	-
10	Dropout	32	ReLU	-
11	Dense	16	ReLU	-
12	Dropout	16	SoftMax	-

3.2.5. Model Validating Phase

In this phase, the evaluation metrics that were utilized to evaluate the CNN model are discussed. To quantify the performance of the proposed CNN model, popular metrics were used calculated by Eqs. (1)-(4).

True Positive refers to a situation in which the model accurately predicts the positive class (TP). A model that accurately predicts the negative class is called a True Negative (TN) result. A False Positive (FP) occurs when a model forecasts the positive class wrongly. When the model inaccurately predicts the negative class, it is called a False Negative (FN).

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\textbf{F1score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Results and Discussion

This section presents the results of comprehensive experiments conducted on the proposed deep learning model. The experimental results will address the four research questions. Experiment I goal is to answer the first research question regarding the impact of size of the feature matrix on the learning process. Experiment II goal is to answer the second research question regarding the effect of usage of different IMF components. Experiment III goal is to answer the third research question regarding the normalization effect. Experiment VI goal is to answer the fourth and last research question regarding the impact of size of the dataset used.

4.1. Experiment I: Studying the effect of the size of feature matrix

The objective of this experiment is to investigate the effect of the size of the feature matrix on the classification process. Two sizes of feature matrices were used in this experiment: 7x7x7 and 9x9x9. The 7x7x7 matrix is extracted by using the conjoint triad method while the 9x9x9 is constructed by merging the 7x7x7 matrix extracted using the conjoint triad method with AAC, GDPC, CTDC, CTDT, CTDD, and SOC number.

As shown in Figure 2, the 9x9x9 merged feature matrix did not have a positive impact on the learning and classification process compared to the 7x7x7 conjoint feature matrix. The loss curve of the 9x9x9 feature matrix shows that the model is not training well when using the 9x9x9 feature matrix. Also, there is a significant difference between the accuracies of the 7x7x7 and 9x9x9 feature matrices as shown in Table 4 and Table 5.

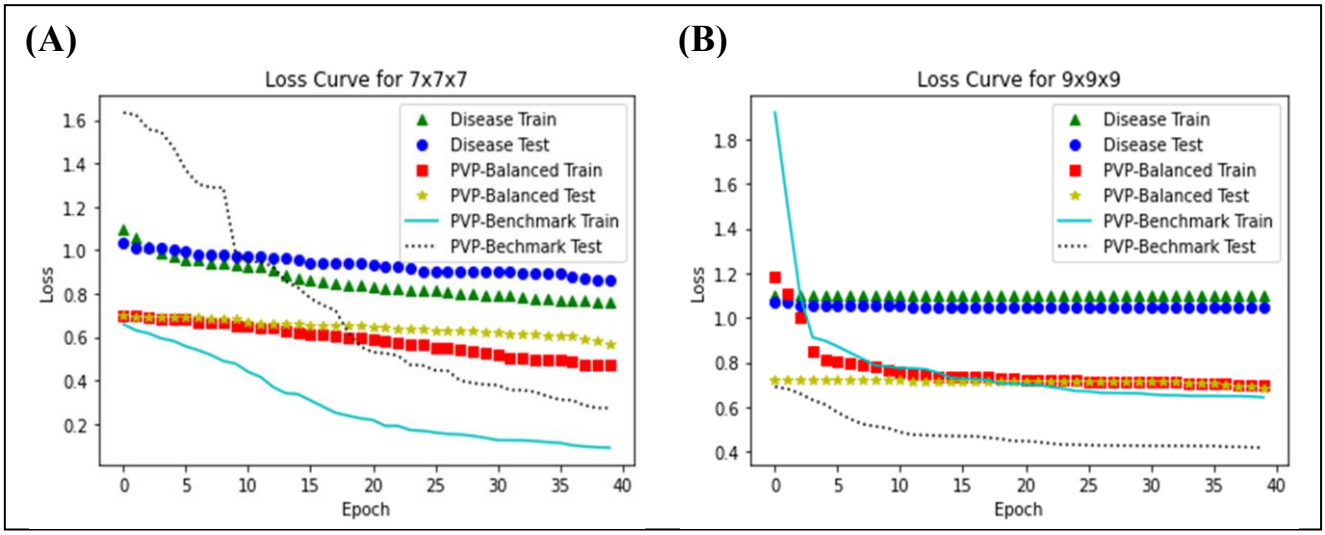


Figure 2: Loss curves for different feature matrix sizes. (A) Loss curve for 7x7x7, (B) Loss curve for 9x9x9

Table 4 Performance evaluation on training set on three datasets with different feature matrix sizes

	Precision		Recall		F1-Score		Accuracy	
	7x7x7	9x9x9	7x7x7	9x9x9	7x7x7	9x9x9	7x7x7	9x9x9
PVP-Balanced	**0.6540**	0.2389	**0.6542**	0.5000	**0.6514**	0.3225	**0.6560**	0.4778
PVP-Benchmark	**0.6635**	0.3422	**0.6533**	0.5000	**0.6353**	0.4050	**0.6809**	0.6843
Disease	**0.8408**	0.2054	**0.8410**	0.3672	**0.8391**	0.2390	**0.8403**	0.3782

Table 5 Performance evaluation on independent set on three datasets with different feature matrix sizes

	Precision		Recall		F1-Score		Accuracy	
	7x7x7	9x9x9	7x7x7	9x9x9	7x7x7	9x9x9	7x7x7	9x9x9
PVP-Balanced	**0.6572**	0.2480	**0.6429**	0.5000	**0.6345**	0.3316	**0.6429**	0.4960
PVP-Benchmark	**0.5091**	0.3441	**0.5073**	0.5000	**0.5022**	0.4076	0.6064	**0.6882**
Disease	**0.5454**	0.1180	**0.4706**	0.3333	**0.4660**	0.1743	**0.4788**	0.3541

In Table 4, it can be shown that the 7x7x7 matrix outperforms the 9x9x9 matrix in all evaluation metrics for all three datasets. The precision of 7x7x7 matrix surpasses 9x9x9 matrix by 173.75%, 93.89%, and

309.34% in PVP-Balanced, PVP-Benchmark, and Disease respectively. The recall of 7x7x7 matrix surpasses 9x9x9 matrix by 30.84%, 30.66%, and 129.03% in PVP-Balanced, PVP-Benchmark, and Disease respectively. The F1-score of 7x7x7 matrix surpasses 9x9x9 matrix by 101.98%, 56.86%, and 251.08% in PVP-Balanced, PVP-Benchmark, and Disease respectively. The accuracy of 7x7x7 matrix surpasses 9x9x9 matrix by 37.29%, 0.49%, and 122.18% in PVP-Balanced, PVP-Benchmark, and Disease respectively.

The 7x7x7 matrix outperforms the 9x9x9 matrix in all assessment criteria for all three datasets in the independent set, as shown in Table 5. In PVP-Balanced, PVP-Benchmark, and Disease, the precision of 7x7x7 matrix outweighs that of 9x9x9 matrix by 165%, 47.95%, and 362.20%, respectively. In PVP-Balanced, PVP-Benchmark, and Disease, the recall of 7x7x7 matrix exceeds that of 9x9x9 matrix by 28.58%, 1.46%, and 41.19%, respectively. In PVP-Balanced, PVP-Benchmark, and Disease, the F1-score of 7x7x7 matrix is 91.34%, 23.20%, and 167.35% higher than that of 9x9x9 matrix. In PVP-Balanced, PVP-Benchmark, and Disease, the accuracy of 7x7x7 matrix is 29.61%, 134.89%, and 35.21% higher than that of 9x9x9 matrix.

This experiment was applied using the first IMF when applying EMD to the feature vector before resizing it to the desired feature matrix either 7x7x7 or 9x9x9. This leads us to move to the next experiment regarding the usage of different IMF components by applying EMD. It is shown that the 9x9x9 did not have a positive impact on the classification process. Hence, the rest of the experiments will be conducted on the 7x7x7 matrix.

4.2. Experiment II: Studying the effect of using different IMF components on the feature matrix

The objective of this experiment is to investigate the impact of using different IMF components on the classification process when applied to the feature vector. After the feature extraction EMD is applied to the feature vector to decompose into six IMFs. Two approaches were used in this experiment, either using the first IMF or the sum of the first three IMFs.

As shown in Table 6, the accuracy improved in both PVP-Benchmark and Disease datasets by 4.81% and 3.09%. The PVP-Balanced dataset accuracy on the other hand decreased by 0.0186. The precision recall and F1-score in Disease dataset increased by 2.93%, 3.12% and 3.08%. On the other hand, PVP-Benchmark dataset precision, recall and F1-score increased by 1.22%, 2.01%, and 4.89% respectively.

Table 7 shows a slight improvement in the testing results using the independent set for PVP-Balanced and PVP-Benchmark datasets. For PVP-Balanced dataset, all evaluation metrices improved by approximately 0.05. On the other hand, PVP-Benchmark showed an improvement that varies from 24.12% to 30.11%. Disease dataset did not have much improvement in the independent dataset.

Table 6 Performance evaluation on training set on three datasets with using different components on IMFs on 7x7x7 feature matrix

	Precision		Recall		F1-Score		Accuracy	
	1 IMF	3 IMF	1 IMF	3 IMF	1 IMF	3 IMF	1 IMF	3 IMF
PVP-Balanced	0.6540	0.6455	0.6542	0.6493	0.6514	0.6396	0.6560	0.6440
PVP-Benchmark	0.6635	0.6716	0.6533	0.6664	0.6353	0.6664	0.6809	0.7137
Disease	0.8408	**0.8655**	0.8410	**0.8673**	0.8391	**0.8650**	0.8403	**0.8663**

Table 7 Performance evaluation on independent set on three datasets with using different components on IMFs on 7x7x7 feature matrix

	Precision		Recall		F1-Score		Accuracy	
	1 IMF	3 IMF	1 IMF	3 IMF	1 IMF	3 IMF	1 IMF	3 IMF
PVP-Balanced	0.6572	**0.6824**	0.6429	**0.6746**	0.6345	**0.6711**	0.6429	0.6746
PVP-Benchmark	0.5091	0.6624	0.5073	0.6297	0.5022	0.6368	0.6064	**0.7128**
Disease	0.5454	0.5450	0.4706	0.4205	0.4660	0.3789	0.4788	0.4314

4.3. Experiment III: Studying the effect of normalization

The objective of this experiment is to investigate the effect of normalizing the feature matrices on the classification process. The normalization step is done on the feature vector before resizing it to the 3D feature matrix. After performing EMD and summing the first three IMFs on the 1x343 feature vector in the case of the 7x7x7 feature matrix, normalization was applied.
The identical process, on the other hand, was used on the 1x729 feature vector before scaling it to the 9x9x9 feature matrix. The normalization is done using standard scaler that follows Eq. (5).

$$z = \frac{(x-u)}{s} \tag{5}$$

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

It can be noted from Table 8 that normalization did not improve the values of the evaluation metrics for the three datasets in training set. Table 8 shows that the training accuracy decreased by 21.31%, 4.96%, and 15.01% for precision, recall, and F1-score while accuracy improved by 5.59% for PVP-Balanced dataset. PVP-Benchmark decreased by approximately 2.5% for all evaluation metrics. Regarding disease dataset, normalization did not have much impact on the evaluation metrics.

The independent results shown in

Table 9 imply that not much enhancement occurred. PVP-Balanced, PVP-Benchmark datasets decreased by 0.002 for all evaluation metrices. Disease dataset on the other hand improved by 15.88%, 39.38%, 37.89%, and 36.99% for precision, recall, F1-score, and accuracy respectively. From the above results, it can be noticed that the normalization improved the classification process for large dataset but did not have much impact on the small dataset.

Table 8 Performance evaluation on training set on three datasets with and without normalization on 7x7x7 feature matrix

	Precision		Recall		F1-Score		Accuracy	
	Norm.	No Norm.	Norm.	No Norm.	Norm.	No Norm.	Norm.	No Norm.
PVP-Balanced	0.5321	0.6455	0.6186	0.6493	0.5561	0.6396	0.6800	0.6440
PVP-Benchmark	0.6409	0.6716	0.6473	0.6664	0.6335	0.6664	0.7312	0.7137
Disease	0.8633	**0.8655**	0.8628	**0.8673**	0.8607	**0.8650**	0.5910	**0.8663**

Table 9 Performance evaluation on independent set on three datasets with and without normalization on 7x7x7 feature matrix

	Precision		Recall		F1-Score		Accuracy	
	Norm.	No Norm.	Norm.	No Norm.	Norm.	No Norm.	Norm.	No Norm.
PVP-Balanced	0.6808	**0.6824**	0.6802	**0.6824**	**0.6798**	0.6711	0.6800	0.6746
PVP-Benchmark	0.6948	0.6624	0.6161	0.6297	0.6219	0.6368	**0.7312**	0.7128
Disease	0.6316	0.5450	0.5861	0.4205	0.5784	0.3789	0.5910	0.4314

4.4. Experiment IV: Studying the effect of the size of the dataset

The objective of this experiment is to investigate the impact of the size of the dataset on the training and classification processes. Three datasets were used in this experiment: PVP-Balanced, PVP-Benchmark, and Disease. As shown in Table 1, the dataset size and balance status differ.

As shown in Table 4, it can be noted that the results improvement is positively affected by the increase in the dataset size. The disease dataset shows the highest numbers surpassing PVP-Balanced in all evaluation metrics by approximately 28.01%, and PVP-Benchmark in all evaluation metrics by values between 26.72% and 32.07%. The results indicate that the larger the dataset, the better the model is trained.

5. Conclusion

Protein analysis is essential for determining the cause of mutations and diseases. Using a 3D feature matrix is an unexplored study topic that hasn't gotten the attention it deserves in the classification of proteins. The purpose of this research is to study the effect of different data variations on the proposed deep learning model: size of feature matrix, using different IMF components, preprocessing, and dataset size on the protein classification process. Three datasets were used: two Phage Virion Proteins datasets and a disease dataset. The selection of these datasets was influenced by three factors: various sorts, sizes, and balance state. To validate the proposed model, four evaluation metrics are used to measure the accuracy: Precision, Recall, F1-score, and Accuracy.

The experimental results show that the conjoint triad method feature matrix has a strong positive correlation between its dimensions, which has a favorable influence on the classification process. On the other hand, the loss curve of the 9x9x9 feature matrix shows that the model is not training well when using the 9x9x9 feature matrix. The sum of the first three components of IMF improves the results up to 86.63% in accuracy and 71.28% in independent test. Normalizing features did not have a positive impact on the classification process of small datasets but improved the results of the large dataset. Finally, the size of the dataset played a significant role on training the CNN model were the training accuracy reached 84.03%. The promising findings of this study guide the researchers in their future investigations regarding protein classification. Moreover, encouraging the use of 3D features in further studies.

6. References

[1] C. L. P. Gupta, A. Bihari, and S. Tripathi, "Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis", arXiv:1901.06152v1, 2019.

[2] L. Yang, P. Wei, C. Zhong, Z. Meng, P. Wang, and Y. Y. Tang, "A Fractal Dimension and Empirical Mode Decomposition-Based Method for Protein Sequence Analysis," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 11, 2019, doi: 10.1142/S0218001419400202.

[3] J. Chen, M. Guo, X. Wang, and B. Liu, "A comprehensive review and comparison of different computational methods for protein remote homology detection," no. September, pp. 1–14, 2016, doi: 10.1093/bib/bbw108.

[4] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Inf. Sci. (Ny).*, vol. 415–416, pp. 190–198, 2017, doi: 10.1016/j.ins.2017.06.027.

[5] M. Science, "First Principles Studies on the Interaction of O 2 with X @ Al 12 (X 5 Al 2 , P 1 , C , Si) Clusters," vol. 12, 2010, doi: 10.1002/jcc.

[6] B. Manavalan, T. H. Shin, and G. Lee, "PVP-SVM : Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine," vol. 9, no. March, pp. 1–10, 2018, doi: 10.3389/fmicb.2018.00476.

[7] P. Charoenkwan, S. Kanthawong, N. Schaduangrat, J. Yana, and W. Shoombuatong, "PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method," *Cells*, vol. 9, no. 2, pp. 1–22, 2020, doi: 10.3390/cells9020353.

[8] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, and W. Shoombuatong, "Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation," *J. Comput. Aided. Mol. Des.*, vol. 34, no. 10, pp. 1105–1116, 2020, doi: 10.1007/s10822-020-00323-z.

[9] maha alhumayani, M. Monir, and rasha ismail, "Machine and Deep Learning Approaches for Human Activity Recognition," *Int. J. Intell. Comput. Inf. Sci.*, vol. 0, no. 0, pp. 1–9, 2021, doi: 10.21608/ijicis.2021.82008.1106.

[10] S. A. Soliman, E.-S. A. El-Dahshan, and A.-B. M. Salem, "Diagnosis of Alzheimer'S Disease By Three-Dimensional Convolutional Neural Network Using Unsupervised Feature Learning Method," *Int. J. Intell. Comput. Inf. Sci.*, vol. 0, no. 0, pp. 1–15, 2021, doi: 10.21608/ijicis.2021.80596.1103.

[11] Z. Diame, M. ElBery, M. Salem, and M. Roushdy, "Experimental Comparative Study on Autoencoder Performance for Aided Melanoma Skin Disease Recognition," *Int. J. Intell. Comput. Inf. Sci.*, vol. 22, no. 1, pp. 88–97, 2022, doi: 10.21608/ijicis.2022.104799.1136.

[12] "UniProt." http://www.uniprot.org. Last Access (1 December 2021).

[13] Z. Chen *et al.*, "Sequence analysis iFeature : a Python package and web server for features extraction and selection from protein and peptide sequences," vol. 34, no. March, pp. 2499–2502, 2018, doi: 10.1093/bioinformatics/bty140.

[14] N. I. Hasan and A. Bhattacharjee, "Deep Learning Approach to Cardiovascular Disease Classification Employing Modified ECG Signal from Empirical Mode Decomposition," *Biomed. Signal Process. Control*, vol. 52, pp. 128–140, 2019, doi: 10.1016/j.bspc.2019.04.005.