Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

More-SPEED: Enhancing Protein Activity Prediction from DNA Sequences

Samaher Al-Janabi* [1], **Zena Kadhuim**[2]

[1] Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq
[2] Department of Software/College of Information Technology, University of Babylon, Babylon, Iraq
*Corresponding Author

DOI: https://doi.org/10.52866/ijcsm.2023.04.04.005
Received July 2023; Accepted September 2023; Available online October 2023

ABSTRACT: This work presents More-SPEED, a novel model for accurately predicting protein activity while minimizing computational demands. Leveraging optimized structures and data preprocessing techniques, More-SPEED achieves high accuracy in protein activity prediction. The model incorporates the data compression three dimension (DC-3D) layer, utilizing the graph mining pattern-fist frequency graph mining (GMP-FFGM) algorithm for efficient preprocessing of complex Deoxyribonucleic acid (DNA) sequence datasets. Additionally, the deterministic structure network using the natural-inspired optimization algorithm called Whale Optimization Algorithm (DSN-WOA) structure optimizes parameters of the Biological dynamic long short term memory (BDLSTM) model, reducing processing time and eliminating manual parameter selection. The BDLSTM layer plays a crucial role in matching codons and predicting protein names, reducing computational complexity without compromising accuracy. The Bi-Rule layer efficiently determines protein activity, especially in disease contexts, providing valuable insights in a shorter time compared to alternative approaches. Evaluation metrics validate the effectiveness of More-SPEED in accurately predicting protein activity, making it a promising solution for advancing protein research.

Keywords: Intelligent Data Analysis; Deep Learning; GMP-FFGM; DSN-WOA, BDLSTM, Bi-Rule; Active Proteins.

1. INTRODUCTION

Technological advancements have significantly improved various aspects of our daily lives through the invention of tools and devices that make life easier and faster. Communication technologies, for instance, have enabled individuals from around the world to communicate with ease. The recent outbreak of epidemics such as COVID-19 has further highlighted the importance of internet-based services, which have become essential for limiting physical interactions and reducing the spread of the virus [1]. Diagnosing diseases associated with an organism requires an understanding of its DNA structure and the genes that generate proteins. However, predicting the proteins that promote or inhibit the presence of diseases is a complex process that necessitates studying the organism's DNA structure [2].

Intelligent Data Analysis (IDA) is a multidisciplinary field that employs techniques from artificial intelligence, high-performance computing, pattern recognition, and statistics to extract meaningful knowledge from data (Sarker, 2021). The IDA process involves three primary steps: problem identification and parameter understanding, model building using techniques such as clustering, classification, prediction, optimization, etc., and evaluation of the results. Finally, the results must be interpreted in a way that is understandable to both specialists and non-specialists [3]. The main benefits of IDA can be summarized as follows ([1]: (a) Extraction of meaningful knowledge from data. (b) Multidisciplinary approach that combines techniques from various fields. (c) Identification and understanding of real-world problems. (d) Effective model building and evaluation techniques. (e) Interpretation of results in a way that is understandable to all specialists and non-specialists.

Data analysis[16][19] is divided into four types: Descriptive Analysis, Diagnostic Analysis, Predictive Analysis, and Prescriptive Analysis. The goal of intelligent data analysis is to extract knowledge from data, which can be used to inform decision-making and improve outcomes. Prediction is a crucial data analysis task that involves estimating the value of a target feature that is not known. Prediction techniques can be classified into two main fields based on the scientific area: prediction techniques related to data mining and prediction techniques related to deep

*Corresponding author: samaher@itnet.uobabylon.edu.iq

learning, such as neuron computing techniques [4]. Various types of data analysis have been introduced, each with their unique advantages and applications in different fields. The aim of prediction is to analyze trends by making estimations for future events based on the impact of past and present data.

Bioinformatics[23][19] is a sub-discipline that lies at the intersection of biology and computer science, dealing with the extraction, storage, analysis, and dissemination of biological data. The primary objective of bioinformatics is to manage data in a way that allows easy and efficient access to information and to submit new entries as they are produced. Moreover, bioinformatics involves the development of technological tools that help analyze biological data. As a field, bioinformatics encompasses a wide range of disciplines, including drug designing, genomics, proteomics, systems biology, machine learning, advanced algorithms for bioinformatics, structural biology, computational biology, and many others. Bioinformatics [27][28] deals with complex DNA and amino acid sequences called proteins, which are extracted from DNA. Due to its interdisciplinary nature, bioinformatics has become a significant area of research, with numerous applications in various fields such as medicine, agriculture, and environmental science. Bioinformatics poses a significant challenge due to the complexity involved in extracting accurate and authentic protein from a complex network that represents the DNA sequence. To address this challenge, Intelligent Data Analysis (IDA) techniques are used. However, these methods can suffer from time complexity and require large computational resources.

To overcome this issue, different hardware solutions, such as Field Programmable Gate Arrays (FPGAs)[30][31] have been introduced. While FPGAs offer faster processing times, they require significant investment costs. Furthermore, the use of FPGAs has several advantages. They are highly customizable and can be tailored to suit specific bioinformatics applications. Additionally, they offer high levels of parallelism, which can reduce processing times significantly. Despite the cost associated with FPGA technology [33], it has proven to be a valuable asset in the field of bioinformatics and has led to advancements in the extraction and analysis of biological data.

2. MAIN CHALLENGES

The problem of finding any protein effect in specific disease based on intelligent data analysis techniques can divided into parts: The first related to programming challenges while the second related to application challenges. As we know the prediction techniques split based on the scientific field into two fields; prediction techniques related to data mining and prediction related to deep learning. In our work deal with the second type:

The profoundness is one of the main characteristics of DNA sequence and as programmer it is very difficult to extraction useful knowledge from it or work directory on that sequence; therefore, to avoid those limitations need efficiency technique to split that sequence into multi subsequence automatically effect to each disease

Long short term memory, (LSTM) is the deeplearning predictions techniques that contain many features make it as a best (i.e., high accuracy and memory) but on other hand as any other Nerocomputing techniques; it based, on the (try and error), principle for determined the structure of network (i.e., number of Nodes; number of Layers; weights for(input , cell, forget and output gate) also objective function is consider as one of the firstly parameter to successful in reach of the goal. Therefore, need to avoid these limitations (i.e., high computation and time implementation).

Time implementation is so important when we dealing to medical diagnosis, for this proposed we need to implement our work in speed processor that depend on multithreaded technologies as shown in Figure (1).

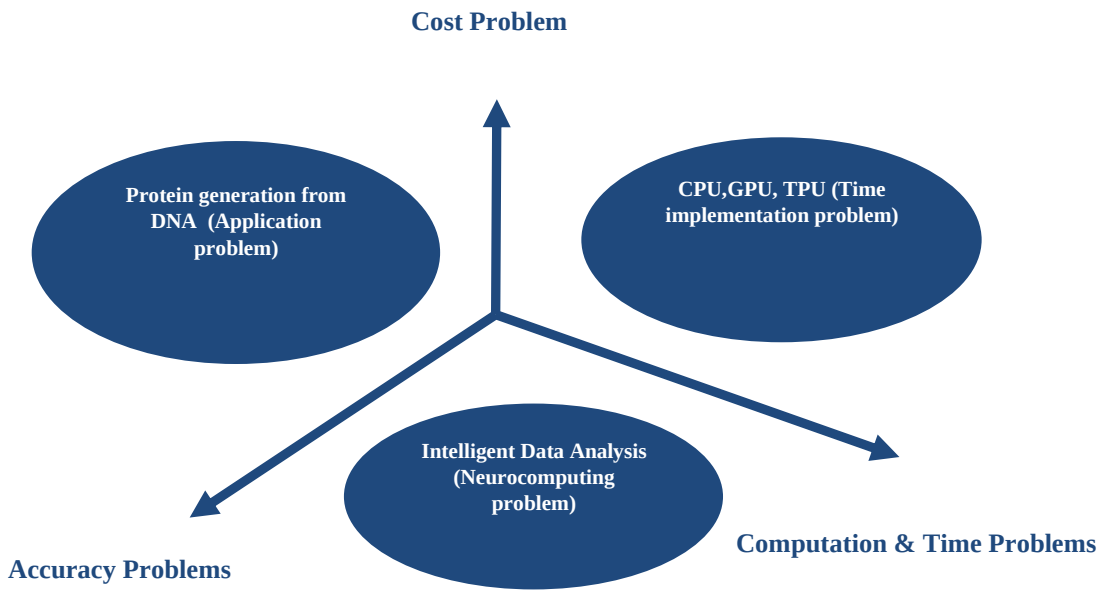

Figure (1): Relationship of basic element of problem

Protein prediction is a highly important concern that directly affects people's lives and their overall health. Since the goal of this study is to establish a modern prediction method for dealing with large amounts of data, it falls under the category of data series. In this section, we will present past research in the same field and try to solve three challenges in Figure (1) and compare these works based on five essential characteristics. These characteristics include the database used, the methods employed, as well as the benefits and limitations of each method. Additionally, we will evaluate each method thoroughly to provide a comprehensive understanding of the strengths and weaknesses of each approach.

2.1 Application Challenge

Many researchers work on extracting knowledge from DNA especially protein that effect in specific or one disease depend on several preprocessing technologies to finally extract the perfect protein:

In 2020 [22], D. Narmadha and A. Pravin1 used various datasets for different diseases such as cancer, diabetes, asthma, and HPV viral infection. The results showed an accuracy of 92.328% for predicting cancer, with precision, recall, and F-measurement at 93.121%, 92.874%, and 91.102%, respectively. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2021, [4] implemented a new model based on artificial intelligence to perform genome sequence analysis of humans infected by COVID-19 and other viruses such as SARS, MERS, Ebola, and Middle East respiratory syndrome. The proposed work used accuracy for measuring algorithm performance and achieved a high accuracy level of 97% for COVID-19 and 95% for other viruses. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2021, [5] discovered a novel NN-according Parkinson's disease protein cause prediction using the ensemble (n-semble) approach. The study first identified disease-related traits within protein sequences and then identified the most pertinent and important traits from highly dimensional data. Precision, recall, and F Score for this approach were 88.9%, 90.9%, and 89.8%, respectively. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022 [6] designed a method to identify immune-related Retinol-binding protein (RBP), a family of proteins with diverse functions, to predict prognosis and therapy response in prostate cancer. The study employed mRNA dataset and used Pearson Correlation analysis to select immune-related RBP. The method was found to distinguish between high-risk and low-risk groups of prostate cancer patients. The high-risk group showed higher rates of genomic alterations and were more sensitive to targeted and immunotherapy than the low-risk group, as measured by sensitivity. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022, [7] developed a bioinformatics approach called PEPPER to identify PE_PGRS proteins, which are believed to be involved in host response and disease pathogenicity in the Mycobacterium tuberculosis genome. The study employed accuracy and speed to measure the performance of the algorithm, and the results demonstrated good performance. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

TABLE (1) **Description of different application challenges related with DNA applications**

Author(s)	Data set/ Database	Preprocessing	Evaluation tools	Advantage	Disadvantage
[22]	Protein sequence Collection of PPI (string DB, IntAct, DIP) database	Segmentation	Accuracy	predict the, proteins for the respective diseases	Time complexity
[21]	DNA sequences collected in file extension of .fasta and .gb from GenBank National Center for Biotechnology Information (nih.gov)	Feature extraction	Accuracy	understand genetic variants of (COVID-19 , SARS-CoV-2 SARS, MERS, Ebola) to perform genome sequence analysis of human that infected by (COVID-19 , SARS , MERS and Ebola) better classification different gene sequences	Classify only specific type of genome sequences
[23].	RNA Sequence(https://github.com/zhaoxj-tech/DFpin.git.)	remove feature redundancy	Accuracy	predicting protein interaction related to different type of human Disease	Time complexity
[24]	RNA sequence (https://www.uniprot.org/downloads)	Segmentation od RNA	Accuracy	predict protein function based on information of sequence of RNA only	Time complexity
[25]	Set of protein NCBI DB UniProt/SwissProt DB	Clustering	Accuracy	identify PE_PGRS proteins is thought to be involved in host response and disease pathogenicity	Identify specific type of protein

[32]	***Extract protein related to Disease https://archive.ics.uci.edu/ml/datasets/Codon+usage# https://www.kaggle.com/code/peterwu19881230/ml-alzheimer/data***	***Dynamic Compression – three demission (DC-3D)***	▪ ***Accuracy*** ▪ ***Loss*** ▪ ***CM***	▪ ***Identify all type of protein*** ▪ ***Accurate feature extraction*** ▪ ***Less time extraction*** ▪ ***Less computation*** ▪ ***Less parameter used***	-

2.2 Neurocomputing Challenge

Many researchers work on extracting protein from DNA related to protein effect in specific disease depend on several Neurocomputing technologies:

In 2019, [8]. designed a deep learning method to discover anomaly-causing genes in mRNA sequences that cause brain disorders such as Alzheimer's and Parkinson's disease. The mean squared error (MSE) was used to measure the performance of the method. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2020, [9] proposed a method called m6A-pred predictor for predicting the presence of m6A in RNA sequences using statistical and chemical characteristics of nucleotides. The accuracy and Mathew correlation coefficient values were used to evaluate the algorithm's performance, with the study demonstrating accuracy levels of 78.58% and 0.5717, respectively. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2020, [10] developed an improved grey wolf optimization (IGWO) meta-heuristic algorithm, which is a new iteration of the base grey wolf optimization (GWO) for feature selection. Four machine learning algorithms were employed in this study to predict protein structure. Accuracy was used to demonstrate the algorithm's performance, and the proposed study utilized artificial neural networks to predict protein structure with an accuracy of 91%. While there are similarities between our approaches to performance measurement and protein prediction, there are also differences in the techniques used for intelligent data analysis to predict proteins.

In 2020, [11]. created a hybrid algorithm that combines a Fuzzy-neural network to categorize proteins based on their unknown sequence into different families. This study employed 497 various test sequences and demonstrated an accuracy level of 90% and execution time of 192ms, compared to the base time execution of 1704ms. Our work shares a similar idea of predicting protein and evaluation methods with Suprativ et al.'s study. However, our approach differs in the techniques used for intelligent data analysis.

In 2020, [12] designed a deep learning platform for predicting the family of proteins (PPF) extracted from DNA sequencing. The study employed rich features to distinguish proteins by capturing disturbances in RNA bases using word-to-vector techniques. Mathew Correlation Coefficient (MCC) and accuracy were used to measure the performance, and different percentages of MCC were found depending on the level of familiarity. The results showed an MCC of about 97.62% for family, 88.45% for subfamily, and 83.09% for sub-family level. Our work shares a similar idea of predicting proteins, but differs in terms of performance measures and techniques used for intelligent data analysis to predict proteins.

In 2020, [13] predicted the protein position of S-sulfenylation using a new method called SulSite-GTB. This protein is involved in different biological processes important for life, such as signaling of cells and increasing stress. The study achieved prediction accuracy of 92.86% and 88.53%, respectively, with AUC values of 0.9706 and 0.9425, respectively, on the training set and the independent test set. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2020, [14] developed a machine learning (ML) model to predict protein-protein interactions between human proteins and viruses, aiming to discover anti-COVID drugs. The first step was to prepare the model to accept human protein sequences of various lengths. For data preprocessing, LVQ was used for feature subset selection, followed by different supervised learning algorithms (SVM, NB, RF, KNN) with multilayer perceptron used for prediction and classification. Confusion matrix was used for evaluating the performance of prediction. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022[15] designed a new method called Deep Learning-based Protein-binding Prediction with Feature-based Non-redundancy from the level of RNA (DFpin) to predict protein interactions related to different types of human diseases. The proposed method used accuracy as a performance measure and achieved a classification accuracy of 93.3%. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022[16] designed a new method called Deep Learning-based Protein-binding Prediction with Feature-based Non-redundancy from the level of RNA (DFpin) to predict protein interactions related to different types of human diseases. The proposed method used accuracy as a performance measure and achieved a classification accuracy of 93.3%. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022, [17] developed a machine learning-based bioinformatics approach called PEPPER to identify PE_PGRS proteins, which are believed to be involved in host response and disease pathogenicity in the Mycobacterium tuberculosis genome. The study employed accuracy and speed to measure the performance of the algorithm, and the results demonstrated good performance. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022, [18] developed a deep learning mechanism named F UTUSA to predict protein function based solely on RNA sequence information. Accuracy was used to evaluate the performance of prediction. Our work shares a similar idea of predicting proteins and utilizes similar performance measures, but differs in the techniques used for intelligent data analysis to predict proteins.

In 2022[19]. developed a design technique called Malsite-Deep to anticipate protein malonylation sites in response to lipopolysaccharide (LPS). The evaluation of the model using independent test sets and 10-fold cross-validation showed an AUC value of 0.99 on the training dataset, with all four independent test datasets achieving AUC values greater than 0.95. While there are similarities between our approaches to performance measurement and protein prediction, there are also differences in the techniques used for intelligent data analysis to predict proteins.

TABLE (2) Description of different Neurocomputing techniques related with DNA applications

Author(s)	Data set/ Database	Preprocessing	Evaluation tools	Method	Advantage	Disadvantage
[8]	Gene Sets of Alzheimer's and Parkinson's www.genecards.org	Feature encoding	MSE	DL based Anomaly Detection	predicting protein interaction related to Alzheimer disease	Time complexity
[9]	RNA sequences	Feature extraction	Accuracy , MCC	m6A-predection	better classification different gene sequences identify features that was discriminative	Classify only specific type of genome sequences
[10]	independent test set (protein sequence) https://github.com/QUST-AIBBDRC/SulSite-GTB/.	Feature encoding	Accuracy	SulSite-GTB	predict the proteins for the respective diseases	Time complexity
[11]	RNA Sequence SARS-CoV-2-human PPI database Negative, dataset	LVQ for feature subset selection	CM	(SVM), (NB), (RF), (KNN) with multilayer perceptron	predict Protein -Protein Interaction between human protein and viruses discover anti-COVID drug various length of Protein human sequences	Time complexity Biological knowledge
[18]	RNA Sequence MSKCC dataset (https:// www.mskcc. org/) TCGAPRAD dataset (https://portal.gdc .cancer.gov/)	Segmentation	PCA	(RBP)LSTM	identify immune-related Retinol-binding proteins (RBP) are a family of proteins with diverse functions predict prognosis and therapy response in prostate cancer	Time complexity

[34]	*Extract protein related to Disease https://archive.ics.uci.edu/ml/datasets/Codon+usage# https://www.kaggle.com/code/peterwu19881230/ml-alzheimer/data*	*Dynamic Compression – three demission (DC-3D)*	▪ *Accuracy* ▪ *Loss* ▪ *CM*	*ADI DSN-WOA (DOLSTM)*	▪ *Identify all type of protein* ▪ *Accurate feature extraction* ▪ *Less computation* ▪ *Less parameter used* ▪ *optimized parameter used in Nerocomputing selected* ▪ *new biological rule* ▪ *for extracting only active Alzheimer gens*	-

2.3 Evaluations Challenge

The Time implementation is one of the main problem of any intelligent data analysis system in biological diagnosis therefore all researcher not work on different platform but run these work on CPU only , therefore there problem is time complexity. we implement our work on different platform to keep out the issues of time complexity .We used three different platform to solve this problem includes (GPU, TPU) that based on multicore.

[20] introduce ADEPT, a novel domain-independent sequence alignment method for GPU designs that supports the alignment of sequences from both genomes and proteins. Our suggested method employs GPU-specific optimizations that don't depend on the sequence's inherent properties. By using the Smith-Waterman algorithm and contrasting it with related CPU strategies as well as the quickest GPU methods for each area, we show the viability of this approach. The driver for ADEPT makes it possible for it to scale across numerous GPUs and makes easy integration into software pipelines that make use of massively parallel computing systems possible. We have demonstrated that for protein-based and DNA-based datasets, the ADEPT-based Smith-Waterman algorithm exhibits top performance of 360 GCUPS and 497 GCUPs.

[21] extracting protein from DNA using different method, A key epigenetic modification mechanism in the control of gene transcription is DNA methylation, which adds methyl groups to DNA molecules. They used CPU for time implementation and spent (121) hours for extracting useful knowledge from DNA. Our method implemented in different platform (CPU, GPU, TPU).

[26] predict protein structure, They use computational modelling to forecast protein structure in the future and can locate every atom in a protein molecule spatially based solely on its amino acid sequence. Using CPU the implementation of this wok implement by 2 month. Our method implemented in different platform (CPU, GPU, TPU).

TABLE (3) Description of different evaluation measures related with DNA applications.

Author(s)	Data set/ Database	Method	Platform	Time
[20]	RNA sequences	DNA methylation	CPU	121 hours
[21]	RNA sequences	Protein structures - prediction	CPU	2 month
[26]	RNA sequences	sequence alignment strategy	GPU	360 minutes

[32]	***Extract protein related to Disease https://archive.ics.uci.edu/ml/datasets/Codon+usage# https://www.kaggle.com/code/peterwu19881230/ml-alzheimer/data***	***(Deterministic selection network – Whale optimization algorithm using Dynamic Optimal Long short Term memory) (DSN-WOA -DOLSTM)***	▪ ***CPU*** ▪ ***GPU*** ▪ ***TPU***	***1.139 minute*** ***0.852 minute*** ***0.652 minute***

3. MORE-SPEED MODEL

This paper aims to address the challenging problem of accurately extracting specific proteins that cause deadly diseases. The evolution of these proteins in different diseases makes extraction tools inaccurate and unreliable. To assist specialists in obtaining accurate diagnostic data, this paper proposes a prediction model built using intelligent data analysis, including statistical, mathematical, and optimization approaches. The model also incorporates deep learning layers to optimize prediction results. Compared to conventional approaches, deep learning contributes significantly to discovering and selecting the best structure to construct the prediction model. The model consists of five main stages: pre-processing by Data Compression in three dimensions (DC-3D), building a deterministic structure network using the natural-inspired optimization algorithm called Whale Optimization Algorithm (DSN-WOA), constructing a new Biological Dynamic layer (BMLSTM) that matches protein names according to a biological matching table, determining protein activity using a Biological Rule (Bi-Rule), and finally, implementing the model in different runs. The main benefit of this approach is that it provides a reliable and accurate tool for specialists to diagnose specific diseases, even in times of uncertainty and doubt. time platform. Figure (2), (3) and Algorithm (1) illustrate the main prediction model descriptions.

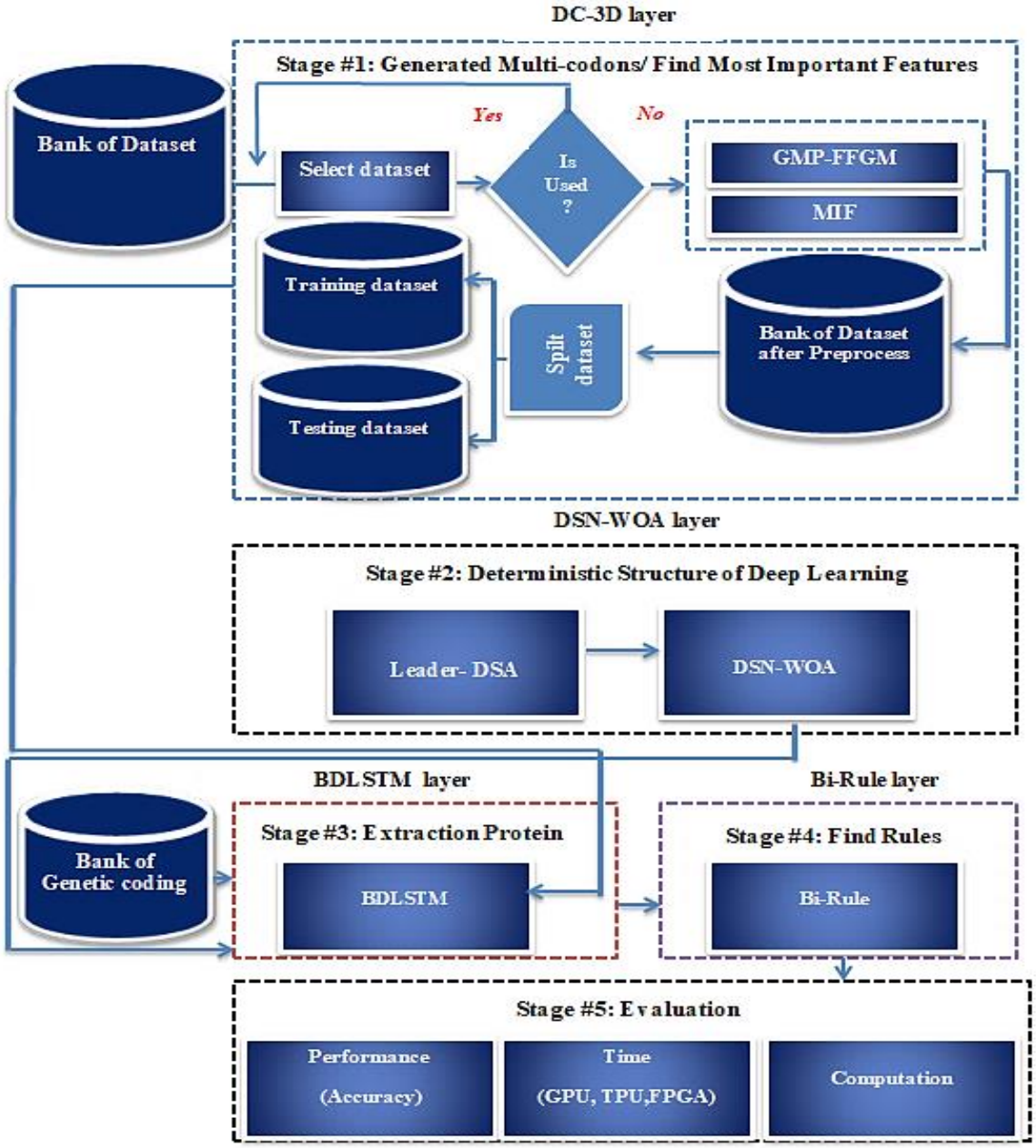

Figure 2 Block Diagram of More-SPEED Model

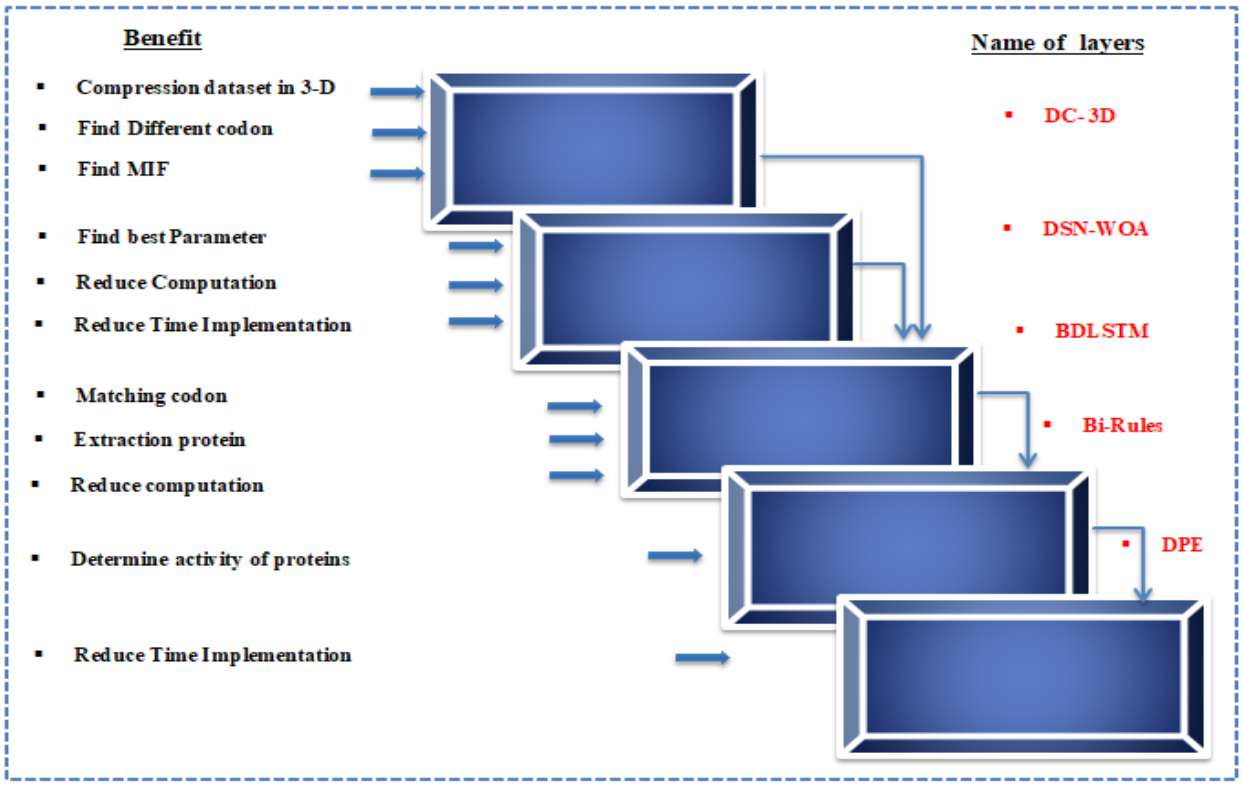

Figure (3) The Main Benefits of each Layer in More-SPEED and its benefit

- The first layer of this model compresses input data in three dimensions using the DC-3D algorithm and identifies different codons through the GMP-FFGM algorithm, dynamically adjusting the number of features based on the chosen dataset. The main benefit of this layer is the reduction of the dataset size.
- The second layer uses the Deterministic Selection Network- Whale Optimization Algorithm to extract the optimal structure of the network, including the optimal number of hidden layers, nodes in the hidden layers, four types of weight, type of objective function, and bias. The main benefit of this layer is the identification of the optimal network structure.
- The third layer uses the Biological Dynamic Long Short-Term Memory model to predict codons and proteins based on the optimal structure chosen from the previous layer and matches codons to specific names of proteins. The main benefit of this layer is accurate prediction.
- The fourth and final layer uses biological rules to determine the activity of the predicted protein, comparing it to real affected protein datasets. The main benefit of this layer is the validation of the predicted protein's activity.

Algorithm #1: Prediction Active Protein

Input: *Dataset related to DNA*

Output: *Prediction protein effect in disease*

Initialization: *codon=Triple item, TNC=64* *// TNC= Total Number of codons*

// Preprocessing Stage (DC-3D) *//Data Compression in three dimision*

```
For each datasets i                          //i =1,2,3
    For each sequence j                      // j= 1...,m
        Call GMP-FFGM                        //Generate Multi Protein
    End For
    For each sample j                        //j=1....,m
        For each feature k                   //j=1....,l
            Call MIF
        End For
    End For
End For
For each Pi                                  //Pi=Parent of indevidual l
    For each prameter
        Call DSN-WOA        // Determnistic Selection Network-Whale Optimazation
    End for
End for
// Build Dynamic Maching Long Short Term Memory
For each datasets i
    For each sequence j
        Call BDLSTM         // Dynamic Maching Long Short Term Memory
    End for
    Call Bi-Rule                             // Bilogical Rule
For each datasets i
    For each sequence j
        Call Performance musurment
    End for
End for
For each datasets i
    For each sequence j
        Call Time implementation (GPU, TPU and FPGA)
    End if
End if
End Main Algorithm
```

The proposed model that uses intelligence data analysis techniques to analyze different groups of data related to various diseases. The model has three stages: Pre-Process, Process, and Post-Process. In the Pre-Process stage, algorithms are used to clean the dataset of undesired data. In the Process stage, deep neural network layers are built with optimization algorithms to predict the name and activity of proteins in specific genetic codes. Finally, in the Post-Process stage, the proposed model's performance is evaluated using different hardware platforms.

4. RESULTS OF MORE-SPEED MODEL

The More-SPEED (More Selective Protein Effect in Each Disease) model include five main stage to find activity of protein these stages are: DC-3D layer, DSN-WOA layer, BDLSTM layer, Bi-Rule layer and DPE layer. In this case study will determine the activity of protein extracted from Biological DNA sequence dataset through apply More-SPEED Model, The main parameters of model shown in Table 4. In each layer

TABLE (4) **main parameter used in each More-SPEED Model layers**

#	Tools	Parameter
1	*FFGM-GMP*	*Min-ts, Max-ts, Buffer, gn(G,k+1)F(1),F(k+1), Buffer(codon)*
2	*DSN-WOA*	*r1,r2,ub,lb,F(OBJ#1),A,C,f(h),f(Nh),F(W),F(b)I,p(i)*
3	*BDLSTM*	*L,N,W[4],AF, b*
4	*Bi-Rule*	*Active p , Passive p*
5	*DPE*	*MSE, Accuracy, Recall, Precision, CPU, GPU, TPU*

The dataset of case study is (Mutated DNA) dataset published on the NCBI Gene Bank website (NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome). This type of dataset includes two FAST file (query DNA) and its (reference DNA), also contained Excel file format explain how query DNA mutated according to reference DNA position in FASTA file of reference DNA,

4.1 Results of GMP-FFGM

The More-SPEED model implemented (GMP-FFGM) to query and reference DNA FASTA file. The main benefit of this step is to reduce computation by using part of DNA instead of all DNA sequence in both (query and reference DNA) and also find different codon latter. The Tool used is GMP-FFGM (Generate multiple codons by Fast Frequent Graph Mining), The main Parameters used in GMP-FFGM are shown in Table 5.

TABLE (5) main parameter used in GMP-FFGM layer

#	Parameter	Description	value
1	*min-ts*	*Minimum DNA Length*	*9500*
2	*maxn-ts*	*Maximum DNA Length*	*950000*
3	*Buffer(seg)*	*set of non-frequent segment*	*31*
4	*gn(G,k+1)*	*Incedence matrix normalized for all segments*	*31*
5	*F(1)*	*set of one-frequent edge*	*16 edge*
6	*F(k+1)*	*set of two-frequent edge*	*64 edge*

The best case extracted by using GMP-FFGM from this sequence determined by enter all dataset to model because the length of all DNA is between (9500 , 95000) so the segment DNA sequence according to equation 3 of seg mentation that was describe in chapter three, The number of segments is thirty one segment means extract all poss ible sixty four codon. On the other hand the worst case when enter part of dataset to GMP-FGM step because the numb er of segmentation is determined

The GMP-FFGM started by convert DNA to mRNA For both Query and Reference DNA by replace every (T base into U base) . After converting DNA, we must extract only different triple for farther mapped each one of thes e triples (codons) into protein. GMP-FFGM work on graph dataset only means we must convert sequence of mRNA to graph by set connection edge to every two base adjacent to deal with mRNA as a graph. Figure (4) and Table (6) illustr ate that work.

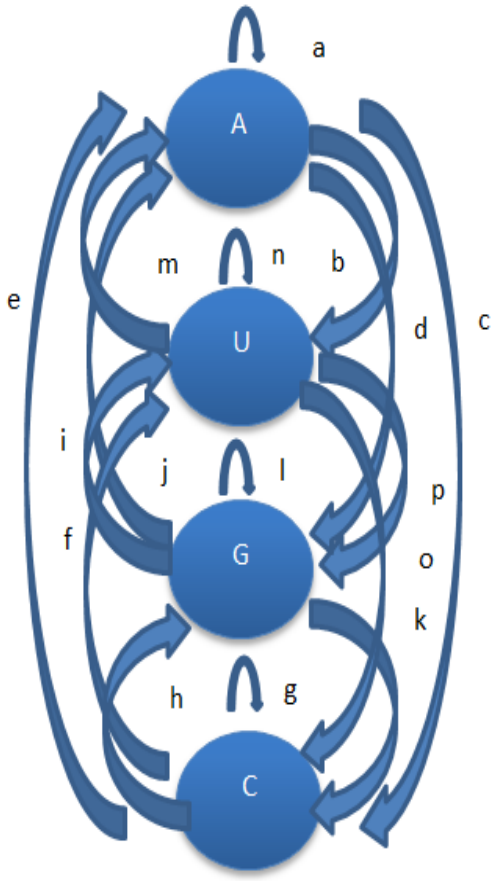

Figure (4) four node and sixteen edges connection

TABLE (6) Set of a Sixteen Edges

#	Node	Edge
1.	AA	A
2.	AU	B
3.	AC	C
4.	AG	D
5.	CA	E
6.	CU	F
7.	CC	G
8.	CG	H
9.	GA	I
10.	GU	J
11.	GC	K
12.	GG	L
13.	UA	M
14.	UU	N
15.	UC	O
16.	UG	P

Since we have four base the number of edges is 16 because (4^2) equal to sixteen different edge setting, we take every base with three different bases at each time. After set connection we segment mRNA into multiple segments to delete fr equent segments by equation of segmentation that depends on DNA length describe in chapter three, the thirty-one diff erent segments is extracted from all mRNA. every segment created draw in form of incidence matrix normalized to find the frequency of all edge in each segment. These matrices represent how each node in each graph connected with other as a number represent how many times in each graph that's two nodes connected. While (0) means no connections, aft er this step see for all matrix what is edge not connected to graph to added to it then compute one frequent edge for all g raph illustrate in table (7).

TABLE (7) compute one frequent edge

#	One Edge	1FEDC
1.	'U', 'n', 'U'	3159
2.	'A', 'a', 'A'	2872
3.	'U', 'p', 'G'	2542
4.	'U', 'm', 'A'	2401
5.	'A', 'b', 'U'	2326
6.	'C', 'f', 'U'	2100
7.	'C', 'e', 'A'	2080
8.	'A', 'c', 'C'	2027
9.	'G', 'j', 'U'	1988
10.	'A', 'd', 'G'	1762
11.	'G', 'i', 'A'	1634
12.	'U', 'o', 'C'	1471
13.	'G', 'k', 'C'	1220

14	'G', 'l', 'G'	1134
15	'C', 'g', 'C'	927
16.	'C', 'h', 'G'	538

Table (7) includes three column first column represent number of edge connected sixteen as explain in table (6), colu mn two represents every two base in query DNA and their adjacent set edge according to set forward of set backward c olumn three represents the one frequent edge counter is the frequency of each one edge connected every successive bas e opposite. after this step must compute two frequent edges illustrate in table (8).

TABLE (8) compute two frequent edge

Two Edge	2FEDC	Two Edge	2FEDC
'U', 'n', 'U', 'n', 'U'	486	'G', 'j', 'U', 'm', 'A'	255
'A', 'a', 'A', 'a', 'A'	471	'G', 'k', 'C', 'f', 'U'	241
'A', 'b', 'U', 'n', 'U'	429	'A', 'b', 'U', 'm', 'A'	239
'U', 'n', 'U', 'm', 'A'	419	'C', 'e', 'A', 'b', 'U'	234
'U', 'p', 'G', 'j', 'U'	419	'G', 'i', 'A', 'b', 'U'	233
'U', 'n', 'U', 'p', 'G'	396	'C', 'e', 'A', 'd', 'G'	232
'A', 'c', 'C', 'e', 'A'	394	'C', 'e', 'A', 'c', 'C'	229
'U', 'm', 'A', 'a', 'A'	365	'G', 'l', 'G', 'j', 'U'	227
'A', 'a', 'A', 'b', 'U'	362	'U', 'm', 'A', 'd', 'G'	211
'C', 'f', 'U', 'n', 'U'	358	'C', 'g', 'C', 'e', 'A'	180
'A', 'b', 'U', 'p', 'G'	351	'G', 'k', 'C', 'e', 'A'	179
'A', 'c', 'C', 'f', 'U'	347	'A', 'c', 'C', 'g', 'C'	174
'C', 'e', 'A', 'a', 'A'	346	'G', 'i', 'A', 'c', 'C'	173
'G', 'j', 'U', 'n', 'U'	328	'A', 'd', 'G', 'k', 'C'	172
'U', 'm', 'A', 'b', 'U'	321	'A', 'd', 'G', 'l', 'G'	167
'U', 'm', 'A', 'c', 'C'	314	'C', 'f', 'U', 'o', 'C'	166
'U', 'p', 'G', 'i', 'A'	297	'C', 'g', 'C', 'f', 'U'	165
'A', 'a', 'A', 'c', 'C'	299	'G', 'i', 'A', 'd', 'G'	160
'U', 'p', 'G', 'i', 'A'	297	'A', 'b', 'U', 'o', 'C'	157
'A', 'a', 'A', 'c', 'C'	290	'G', 'l', 'G', 'i', 'A'	143
'U', 'o', 'C', 'e', 'A'	286	'G', 'j', 'U', 'o', 'C'	142
'G', 'i', 'A', 'a', 'A'	286	'G', 'l', 'G', 'k', 'C'	129
'A', 'd', 'G', 'i', 'A'	282	'U', 'o', 'C', 'g', 'C'	114
'U', 'o', 'C', 'f', 'U'	281	'A', 'c', 'C', 'h', 'G'	106
'A', 'a', 'A', 'd', 'G'	280	'G', 'k', 'C', 'g', 'C'	105
'C', 'f', 'U', 'm', 'A'	277	'C', 'h', 'G', 'j', 'U'	86
'U', 'p', 'G', 'l', 'G'	273	'G', 'l', 'G', 'l', 'G'	75
'G', 'j', 'U', 'p', 'G'	273	'C', 'g', 'C', 'g', 'C'	71
'U', 'p', 'G', 'k', 'C'	268	'U', 'o', 'C', 'h', 'G'	68
'C', 'f', 'U', 'p', 'G'	265	'C', 'h', 'G', 'k', 'C'	66
'A', 'd', 'G', 'j', 'U'	258	'C', 'h', 'G', 'i', 'A'	60
'U', 'n', 'U', 'o', 'C'	257	'G', 'k', 'C', 'h', 'G'	60

Table (8) includes four columns first and third columns represents every three base in query DNA and between every t wo is edge setting ,column two and four represents the frequency of each two edge connected every successive base. We stop calculated next frequent because codon mean three base connections no more so, take from each triple combin e one triple related to mutation then removing edges from mRNA and return to DNA. The length of DNA extracted is ***1 98 base*** for reduce the computation instead of working on ***all DNA*** sequence we select the different codon to reduce co

mputation by GMP-FFGM rather than working on all sequence together, GMP-FFGM also work on reference DNA of l ength (29903) to get new reduced reference DNA opposite query DNA . both table (9) and table (10) illustrate the ***resul ts of GMP-FFGM.***

TABLE (9) Results of GMP-FFGM layer on query DNA

ATT	TAA	ACG	GGT	TTT	TAT	TAC	CCT
TTC	CCC	CAG	ACA	AAA	ACC	CAA	ACT
TCA	AAT	TCT	TTG	GTA	AGA	ATC	CTG
GTT	AAC	CGA	CTT	TTA	TGT	TGG	GCT
GGC	GCA	ATG	AGT	TGC	CAC	CTC	CGC
ATA	GTC	CGT	GAC	GGA	GAG	GCC	CGG
CCG	GAT	CTA	AGG	TCC	TGA	AAG	GAA
CCA	AGC	GTG	TCG	TAG	GCG	CAT	GGG

TABLE (10) Results of GMP-FFGM layer on reference DNA

ATT	TAA	AAG	GGT	TTT	TAT	TAC	CCT
TTC	CCC	CAG	ACA	AAA	ACC	CAA	ACT
TCG	GAT	TCT	TTG	GTA	AGA	ATC	CTG
GTT	AAC	CGA	CTT	TTA	AAT	TGT	TGG
GCT	TCA	GGC	GCA	ATG	AGT	TGC	CAC
CTC	CGC	ATA	GTC	CGT	GAC	GGA	ACG
GAG	AGG	CGG	TCC	AGC	CCG	CTA	TGA
GCC	GAA	CCA	GTG	TAG	GCG	CAT	GGG

4.2 Results of DSN-WOA

The More-SPEED model implemented next layer called (DSN-WOA) to DNA dataset. The Benefit of this step is to reduce computation by using deterministic selection algorithm using whale optimization algorithm to find best stru cture deep Neurocomputing network including best (objective function, weight, hidden layer, node in hidden layer). Th e tools used is DSN-WOA, The main Parameters used in DSN-WOA are ***:(r1,r2,ub,lb,F(OBJ#1), A,C, f(h), f(Nh), f(w) , f(b),I, p (i))***. Table (11) show the main parameters used in this stage.

TABLE (11) main parameter used in DSN-WOA layer

#	*Parameter*	*Description*	*Value*
1	*r1, r2*	*Random numbers*	*range [0, 1]*
2	*ub*	*Upper Boundary*	*10*
3	*lb*	*lower Boundary*	*-10*
4	*F(OBJ#1)*	*Fitness Faction (OBJ#1)*	*(hyperbolic, polynomial)*
5	*A*	*Random numbers*	*range [0, 1]*
6	*C*	*Random numbers*	*range [0, 1]*
7	*f(h)*	*Number of hidden layers*	*(2,256)*
8	*f (Nh)*	*Number of neurons in each layer*	*(1,256)*
9	*F(w)*	*Wight of network*	*(0,1)*
10	*F(b)*	*Bios of network*	*(0,1)*
11	*i*	*Whale iteration*	*50*
12	*P(i)*	*Whale population*	*100*

The best structure by DSN-WOA for selecting structure that have low computation including minimum num ber of hidden layers, minimum number of nodes in each hidden layer. Furthermore, leader of DSN-WOA not selected r andomly but in Deterministic Selection Algorithm (DSA) called leader-DSA, main steps of apply it:

Firstly, create Random Population contain 100 an individuals, , All population are entered to deterministic s election algorithm (DSA) for finding leader to an optimization algorithm (WOA) by leader-DSA by splits into five grou p and each group of length five structures are sorted ascending then select middle from each middle structure as leader to an optimization algorithm latter. In our work the leader-DSA choose structure that contain 29 hidden layer and each l ayer have nodes [109, 133, 18, 131, 12, 161, 189, 112, 66, 122, 32, 38, 12, 166, 14, 75, 29, 250, 106, 21, 114, 130, 135, 44, 152, 211, 113, 211, 4] to enter to whale algorithm to select best individual have min computation and entered it to next iteration as leader to continue for searching optimal structure. Table (12) is a sample of DSN-WOA.

TABLE (12) Sample of DSN-WOA

#	#Leader		# Best Individual	
#	**#Layer**	**#Node**	**#Layer**	**#Node**
1	29	***[109, 133, 18, 131, 12, 161, 189, 112, 66, 122, 32, 38, 12, 166, 14, 75, 29, 250, 106, 21, 114, 130, 135, 44, 152, 211, 113, 211, 4]***	27	***[120,34,12,34,65,44,3,2,64,22, 45,87,66,45,36,44,33,1,21,31, 14,1,12,43,65,77,8,45]***
2	27	***[120,34,12,34,65,44,3,2,64,22,45,87,66,45,36,44, 33,1,21,31,14,1,12,43,65,77,8,45]***	24	***[36, 233, 33, 89, 82, 50, 95, 215, 22, 219, 41, 157, 74, 230, 59, 40, 86, 247, 76, 48, 108, 175, 18, 178]***
3	24	***[36, 233, 33, 89, 82, 50, 95, 215, 22, 219, 41, 157, 74, 230, 59, 40, 86, 247, 76, 48, 108, 175, 18, 178]***	23	***[242, 94, 210, 66, 46, 52, 158, 43, 89, 90, 207, 248, 175, 62, 89, 101, 156, 103, 128, 77, 252, 7, 201]***
4	23	***[242, 94, 210, 66, 46, 52, 158, 43, 89, 90, 207, 248, 175, 62, 89, 101, 156, 103, 128, 77, 252, 7, 201]***	18	***[112, 109, 208, 52, 109, 116, 252, 12, 183, 94, 25, 54, 113, 242, 92, 39, 221, 138]***
.	.	.	.	.
.	.	.	.	.
30	5	***[173, 71, 159, 187, 80]***	***3***	***[20,13,28]***
.	.	.	.	.
.	.	.	.	.
50	***3***	***[20,13,28]***	***3***	***[20,13,28]***

4.3 Results of BDLSTM

Biological Dynamic Long Short Term Memory (BDLSTM) used the best structure results from DSN-WOA layer and the dataset results from GMP-FFGM layer, The main parameters used in BDLSTM are :(L, N, W[4], AF,B) as sho wn in Table (13).

TABLE (13) Parameter used in BDLSTM layer

#	*Parameter*	*Description*	*Value*
1	*L*	*Best Layer from DSN-WOA*	*3*
2	*N*	*Best nodes from DSN-WOA*	*[20, 13,28]*
3	***Wc***	*Best cell weight from DSN-WOA*	*0.232*
4	*Wi*	*Best input weight from DSN-WOA*	*0.244*
5	*Wf*	*Best forget weight from DSN-WOA*	*0.665*
6	*Wo*	*Best output weight from DSN-WOA*	*0.302*
7	*Af*	*Best Activation function from DSN-WOA*	*sigmoid*
8	*b*	*Best bios from DSN-WOA*	*1*
9	*GC*	*Genetic code of protein –codon mapping*	*22 protein*

The results of BDLSTM comparing with Genetic Coding protein table to extract name of protein efficiency; through enter all codons extracted from GMP-FFGM to BDLSTM, Then BDLSTM used to extracting name of protein by comp aring it with Genetic coding of protein-codon mapping (GC). Table (14) shown the results of BDLSTM.

TABLE (14) Results of BDLSTM

Codon	Protein	Codon	Protein	Codon	Protein	Codon	Protein
ATT	lle	**ACG**	Thr	**GGT**	Gly	**TAA**	Stop
TTC	Phe	**CAG**	Gln	**ACA**	Thr	**CCC**	Pro
TCA	Ser	**TCT**	Ser	**TTG**	Leu	**AAT**	Asn
GTT	Gly	**CGA**	Arg	**CTT**	Leu	**AAC**	Asn
GGC	Gly	**ATG**	Met	**AGT**	Ser	**GCA**	Ala
ATA	lle	**CGT**	Arg	**GAC**	Asp	**GTC**	Val
CCG	Pro	**CTA**	Leu	**AGG**	Arg	**GAT**	Stop
CCA	Pro	**GTG**	Val	**TCG**	Ser	**AGC**	Ser
TTT	Phe	**TAT**	Tyr	**TAC**	Tyr	**CCT**	Pro
AAA	Lys	**ACC**	Thr	**CAA**	Gln	**ACT**	Thr
GTA	Val	**AGA**	Arg	**ATC**	lle	**CTG**	Lue
TTA	Leu	**TGT**	Cys	**TGG**	Trp	**GCT**	Ala
TGC	Cys	**CAC**	His	**CTC**	Lue	**CGC**	Arg
GGA	Gly	**GAG**	Glu	**GCC**	Ala	**CGG**	Arg
TCC	Ser	**TGA**	Stop	**AAG**	Lys	**GAA**	Glu
TAG	Stop	**GCG**	Ala	**CAT**	His	**GGG**	Gly

4.4 Results of Bi-Rule

Bi-Rule layer used to find the type of protein extracted from BDLSTM layer; where the activity of proteins is eith er Active or Passive this achieve by comparing the protein with reference dataset., The main Parameters used in Bi-Rul e are ***(Active P, Passive P)*** as shown in Table (15)

TABLE (15) Parameter used in Bi-Rule layer

#	*Parameter*	*Description*	*Value*
1	*Active P*	*Active Protein*	***43****(codons)*
2	*Passive P*	*Passive Protein*	***21****(codons)*

We used the format (***IF conditions Then Action***) to find the activity of protein through matching every prot ein (64 codon) with reference DNA dataset, if codon is mutated in some point set Active Protein else set Passive Protei n.

4.5 Results of DPE

DPE layer used to evaluate model through both sides software and hardware. The main parameters used in DPE are ***:(CM, MSE, CPU, GPU, TPU)***. While, the results of this layer shown in Table (16).

TABLE (16) Parameter used in DPE layer

#	*Parameter*	*Description*	*Value*		
1	***CM***	*Confusion Matrix*	96%, 94%,96%		
2	***MSE***	*Mean Square Error*	*0.04*		
3	***platform***	*CPU, GPU, TPU*	***0.335***	***0.331***	***0.032***

5. Conclusions

The More-SPEED model that was proposed in the paper aims to predict protein activity in high accuracy and reduce time implementation. The following major conclusions were drawn from the development and application of the More-SPEED model:

A. The dataset utilized for protein activity prediction was intricate, involving the handling of diverse data types, including pure DNA sequences, Working with DNA sequences as a data type proved to be considerably challenging due to the complexity of the extracted codons during the pre-processing stage.

B. The DC-3D layer of the More-SPEED model performs preprocessing on various types of datasets to prepare them for the subsequent stage. This layer incorporates with algorithms, namely GMP-FFGM; The primary advantage of this layer is to identify and extract all unique codons from long and medium DNA sequences, leading to a reduction in dataset computation. By leveraging the GMP-FFGM algorithm, the More-SPEED model achieves improved performance and efficiency in processing DNA sequence data.

The following observations were made regarding GMP-FFGM:

- When working with long DNA sequence data types (containing more than 95,000 bases), GMP-FFGM produces superior results. This is due to the higher frequency of codons saved in the buffer, which reduces the implementation and computation time.
- For medium DNA sequence data types (ranging between 9,500 and 95,000 bases), GMP-FFGM yields good results. The codons' frequency saved in the buffer contributes to a reduction in implementation and computation time, while ensuring that all sixty-four different codons are extracted.
- However, when dealing with short DNA sequence data types (less than 9,500 bases), GMP-FFGM produces inferior results since not all sixty-four different codons are extracted.

C. The DSN-WOA is a structure designed to address a common challenge encountered in Neurocomputing. The primary benefit of the DSN-WOA structure lies in its ability to optimize the parameters of the BDLSTM (Bidirectional Long Short-Term Memory) model. These parameters include the objective function, optimal hidden layers, and the optimal number of nodes in each hidden layer. By employing the optimization algorithm, the DSN-WOA enables efficient parameter selection, leading to reduced processing time during implementation. The key advantages of the DSN-WOA are as follows:

- The DSN-WOA eliminates the need for manual parameter selection in the neural network by avoiding the trial-and-error approach. This significantly reduces the time and effort involved in fine-tuning the network.
- The DSN-WOA algorithm leverages the optimization algorithm called WOA (Whale Optimization Algorithm). This algorithm relies on principles derived from the behavior of whale agents, enabling it to dynamically select the optimal parameters for the network.

D. The BDLSTM layer of the More-SPEED model plays a crucial role in predicting proteins using DNA sequences, The primary advantage of the BDLSTM layer is its ability to match codons and predict the names of proteins. By leveraging the power of the BDLSTM algorithm, the More-SPEED model reduces computational complexity while accurately predicting proteins. This enables efficient analysis and understanding of biological datasets, contributing to advancements in protein research. The following observations were made regarding the BDLSTM algorithm:

- The BDLSTM algorithm demonstrated effectiveness in extracting and matching codons from DNA sequences. By comparing them with the Biological Genetic code, the algorithm successfully extracted the names of proteins and achieved high accuracy in prediction.

E. The Bi-Rule layer of the More-SPEED model plays a key role in predicting the activity of proteins using DNA sequences. The main advantage of the Bi-Rule layer is its ability to determine the activity of proteins, particularly in the context of different diseases. By utilizing the Bi-Rule algorithm, the More-SPEED model facilitates the identification and understanding of protein activity, contributing to advancements in disease research and potential therapeutic interventions. The following observations were made regarding the Bi-Rule algorithm:

- The Bi-Rule algorithm demonstrated effectiveness in presenting the activity of proteins when implemented on DNA sequences. It achieved this in a relatively shorter time compared to other approaches. This indicates that the algorithm efficiently analyzes the DNA sequence data to determine protein activity.

F. Finally; The More-SPEED model was developed to accurately predict protein activity while minimizing the risk of fatalities. Through its evaluation, several key findings were observed:

- Dataset Complexity: The dataset used in the More-SPEED model consisted of diverse data types, including DNA sequences, codons, and Alzheimer's disease-related proteins. Dealing with segmented codons yielded the best results, while processing long DNA chains posed challenges.
- DSN-WOA Structure: The DSN-WOA structure was implemented to overcome issues commonly encountered in Neurocomputing. It successfully eliminated the need for manual parameter selection, reducing the time and effort involved in optimizing the network's performance.

- BDLSTM and Biological-Rule Structures: The BDLSTM structure, combined with the Biological-Rule, enhanced the accuracy of protein activity prediction while reducing computational requirements. The BDLSTM algorithm effectively matched codons and predicted protein names. The Biological-Rule leveraged reference datasets to improve prediction accuracy.
- Importance of Reference Dataset: The Bi-Rule structure, utilizing a reference dataset, proved effective in predicting protein activity. However, manually setting rules without a reference dataset led to poor performance. This highlights the significance of leveraging relevant data for accurate predictions.
- Evaluation Metrics: The More-SPEED model was evaluated based on various metrics, including time, accuracy, and performance. Different hardware platforms, such as GPUs, TPUs, and FPGAs, were utilized to reduce computation time and improve overall efficiency.

As result, the More-SPEED model demonstrated its effectiveness in predicting protein activity accurately and with reduced computational demands. The incorporation of optimized structures and the utilization of reference datasets contributed to improved performance and enhanced understanding in the field of protein research.

The analysis of protein prediction activity led to several recommendations in the dissertation that could improve the accuracy and efficiency of the process. These recommendations included:

A. For working in preprocessing step: it is possible to use other pre-processing techniques statistically like Principal Component Analysis (PCA) for reducing the dimension of biological data or using person correlation (PC) that depend on regression model to remove the unwonted feature from most biological data , these two methods illustrated in two-equation bellow:

$$T(k) = x.w \quad \text{.....(1)}$$

$$rc = \frac{\sum(xi-\bar{xi})(yi-\bar{yi})}{\sqrt{\sum(xi-\bar{xi})^2\sum(yi-\bar{yi})^2}} \quad \text{.....(2)}$$

Where, Rc=Correlations coffins ; Xi=values of x-variable in a sample ; $\bar{xi}$=mean of x-variable in a sample; yi=values of y-variable in a sample ; $\bar{yi}$=mean of y-variable in a sample

B. For working in building a new structure neural network: for second part of our work include building an optimal structure Nerocomputing that build depend on optimization techniques, it is possible to build a structure using other techniques that depend on the concept of trees on data mining techniques , such as Genetic programing (GA) for finding the optimal structure for Nerocomputing network.

C. For Employing Deep Learning: it is possible to build a hybrid Nerocomputing techniques or using data mining techniques like Mars to minimize error rate (loss) and increase (accuracy) in protein prediction strategies.

D. For Employing Rule generation: It is possible to use fp-Growth algorithm that depend on growing rule by tree principles to or build a mathematical model description to establish rule generation for determining protein activity

E. For implementing evolution measurements: Verifying the accuracy of the system that predict protein by using evaluation metrics such as The molecular weight that depends on Moller mass of the protein. So that biologists can know the importance of protein before and after the prediction process.

$$M = \frac{\sum NiMi^{1+a}}{\sum NiMi} \cdots\cdots (3)$$

where N and Mi. is weight of single atom.

F. Finally, it is possible to verifying the time implementation of the system on a cloud system to eliminate the needing of huge memory requirements while working on personal computer on physical hardware.

Funding

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

None

REFERENCES

[1] S. Al-Janabi, A. Patel, H. Fatlawi, K. Kalajdzic, and I. Al Shourbaji, "Empirical rapid and accurate prediction model for data mining tasks in cloud computing environments," in *2014 International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, 2014, pp. 1-8. doi: 10.1109/ICTCK.2014.7033495.

[2] S. H. Ali, "A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining," in *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Sousse, 2012, pp. 951-961. doi: 10.1109/SETIT.2012.6482042.

[3] I.H. Sarker, A.S.M. Kayes, S. Badsha et al., "Cybersecurity data science: an overview from a machine learning perspective," *J. Big Data*, vol. 7, no. 1, p. 41, 2020. Available: https://doi.org/10.1186/s40537-020-00318-5.

[4] S. Chakraborty, A. Saha, and A. Neelavar Ananthram, "Comparison of DNA extraction methods for non-marine molluscs: Is the modified CTAB DNA extraction method more efficient than DNA extraction kits?," *3 Biotech*, vol. 10, no. 2, p. 69, 2020. Available: https://doi.org/10.1007/s13205-020-2051-7.

[5] J. Gong, J. Wang, X. Zong, Z. Ma, and D. Xu, "Prediction of protein stability changes upon single-point variants using 3D structure profile," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1237-1244, 2021. Available: https://doi.org/10.1016/j.csbj.2021.03.003.

[6] A. Syberfeldt and F. Vuoluterä, "Image Processing based on Deep Neural Networks for Detecting Quality Problems in Paper Bag Production," *Procedia CIRP*, vol. 93, pp. 1224-1229, 2020. Available: https://doi.org/10.1016/j.procir.2020.04.158.

[7] S. H. Ali, "Miner for OACCR: Case of medical data analysis in knowledge discovery," in *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Sousse, 2012, pp. 962-975. doi: 10.1109/SETIT.2012.6482043.

[8] S. Al-Janabi and M. A. Mahdi, "Evaluation prediction techniques to achieve optimal biomedical analysis," *Int. J. Grid and Utility Computing*, vol. 10, no. 5, pp. 512-527, 2019.

[9] H. Aouani and Y. Ben Ayed, "Speech Emotion Recognition with deep learning," *Procedia Comput. Sci.*, vol. 176, pp. 251-260, 2020. Available: https://doi.org/10.1016/j.procs.2020.08.027.

[10] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Comput. Sci.*, vol. 208, pp. 460-470, 2022. Available: https://doi.org/10.1016/j.procs.2022.10.064.

[11] A. Haleem, M. Javaid, M. A. Qadri, R. P. Singh, and R. Suman, "Artificial intelligence (AI) applications for marketing: A literature-based study," *Int. J. Intell. Networks*, vol. 3, pp. 119-132, 2022. Available: https://doi.org/10.1016/j.ijin.2022.08.005.

[12] I.H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," *SN Comput. Sci.*, vol. 3, no. 2, p. 158, 2022. Available: https://doi.org/10.1007/s42979-022-01043-x.

[13] H. Tao, X. Niu, L. Fu et al., "DeepRS: A Library of Recommendation Algorithms Based on Deep Learning," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, pp. 45-56, 2022. Available: https://doi.org/10.1007/s44196-022-00102-8.

[14] S. S., V. C., and H. S., "Nature-inspired meta-heuristic algorithms for optimization problems," *Computing*, vol. 104, pp. 251-269, 2022. Available: https://doi.org/10.1007/s00607-021-00955-5.

[15] M. A. Mahdi and S. Al-Janabi, "A Novel Software to Improve Healthcare Based on Predictive Analytics and Mobile Services for Cloud Data Centers," in *Big Data and Networks Technologies*, Y. Farhaoui, Ed. Cham: Springer, 2020, vol. 81. doi: 10.1007/978-3-030-23672-4_23.

[16] S. Al-Janabi and A. F. Alkaim, "A Comparative Analysis of DNA Protein Synthesis for Solving Optimization Problems: A Novel Nature-Inspired Algorithm," in *Innovations in Bio-Inspired Computing and Applications*, A. Abraham et al., Eds., Cham: Springer, 2021, vol. 1372. doi: 10.1007/978-3-030-73603-3_1.

[17] I.H. Sarker, "Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects," *Ann. Data Sci.*, 2022. Available: https://doi.org/10.1007/s40745-022-00444-2.

[18] M.A. Abdou, "Literature review: efficient deep neural networks techniques for medical image analysis," *Neural Comput. & Applic.*, vol. 34, no. 17, pp. 5791-5812, 2022. Available: https://doi.org/10.1007/s00521-022-06960-9.

[19] S. Al-Janabi and A.F. Alkaim, "A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation," *Soft Comput.*, vol. 24, pp. 555-569, 2020.

[20] I.H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," *SN COMPUT. SCI.*, vol. 3, no. 2, pp. 158-169, 2022. doi: 10.1007/s42979-022-01043-x.

[21] I. Ahmed and G. Jeon, "Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses," *Interdiscip Sci.*, vol. 13, no. 3, pp. 1-16, Aug. 2021. doi: 10.1007/s12539-021-00465-0.

[22] D. Narmadha and A. Pravin, "An intelligent computer-aided approach for target protein prediction in infectious diseases," *Soft Computing*, vol. 24, no. 18, pp. 13585-13598, Oct. 2020. doi: 10.1007/s00500-020-04481-5.

[23] S. Al-Janabi, A. Alkaim, E. Al-Janabi, et al., "Intelligent forecaster of concentrations (PM2.5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP)," *Neural Comput & Applic*, vol. 33, pp. 14199-14229, 2021. doi: 10.1007/s00521-021-06067-7.

[24] S. Al-Janabi, A. Alkaim, "A novel optimization algorithm (Lion-AYAD) to find optimal DNA protein synthesis," *Egyptian Informatics Journal*, vol. 23, no. 2, pp. 271-290, 2022. doi: 10.1016/j.eij.2022.01.004.

[25] X. Zhao, Y. Zhang, and X. Du, "DFpin: Deep learning–based protein-binding site prediction with feature-based non-redundancy from RNA level," *Computers in Biology and Medicine*, vol. 142, article 105216, 2022. doi: 10.1016/j.compbiomed.2022.105216.

[26] M. Wang et al., "SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting," *Neural Computing and Applications*, vol. 32, no. 6, pp. 1945-1954, Mar. 2020. doi: 10.1007/s00521-020-04792-z.

[27] F. Li et al., "Computational analysis and prediction of PE_PGRS proteins using machine learning," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 3822-3833, Jan. 2022. doi: 10.1016/j.csbj.2022.01.019.

[28] A. Khan et al., "Detecting N6-methyladenosine sites from RNA transcriptomes using random forests," *Journal of Computational Science*, vol. 4, article 101238, 2020. doi: 10.1016/j.jocss.2020.101238.

[29] M. Wang et al., "Malsite-Deep: Prediction of protein malonylation sites through deep learning and multi-information fusion based on NearMiss-2 strategy," *Knowledge-Based Systems*, vol. 240, article 108191, Mar. 2022. doi: 10.1016/j.knosys.2022.108191.

[30] A. Rajangam, S. Jacob, and R. Rajavel, "Protein Sequence Based Anomaly Detection for Neuro-Degenerative Disorders Through Deep Learning Techniques," in *Proceedings of ICBDCC18*, 2019, pp. 562-569. doi: 10.1007/978-981-13-1882-5_48.

[31] G. S. Mohammed, S. Al-Janabi, "An innovative synthesis of optimization techniques (FDIRE-GSK) for generation electrical renewable energy from natural resources," *Results in Engineering*, vol. 16, 2022. doi: 10.1016/j.rineng.2022.100637.

[32] Z. A. Kadhuim, S. Al-Janabi, "Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis," *Results in Engineering*, vol. 17, 2023. doi: 10.1016/j.rineng.2022.100847.

[33] C. Zhang and J. Han, "Data Mining and Knowledge Discovery," in *Urban Informatics*, W. Shi, M.F. Goodchild, M. Batty, M.P. Kwan, and A. Zhang, Eds., Springer, Singapore, 2021, vol. 142, pp. 497-509. doi: 10.1007/978-981-15-8983-6_42.

[34] Z.A. Kadhuim and S. Al-Janabi, "Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis," *Results in Engineering*, vol. 17, article 100847, 2023. doi: 10.1016/j.rineng.2022.100847.