International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 01, Issue 02, November – December (2023)

RESEARCH ARTICLE

IOT-ENABLED PROTEIN STRUCTURE CLASSIFICATION VIA CSA-PSO BASED CD4.5 CLASSIFIER

T. Maris Murugan[1, *] and A Jeyam[2]

[1] Erode Sengunthar Engineering College, Perundurai, Erode, Thuduppathi, Tamil Nadu 638057 India.
[2] Department of Electrical and Electronics Engineering, Erode Sengunthar Engineering College, Perudurai, Erode, 638057 India

[*]*Corresponding e-mail: marismurugan428@outlook.com*

Abstract – Data mining is a technique for obtaining useful information from vast amounts of information. Big data refers to large amounts of complicated information that is processed, particularly in relation to biological processes. The investigation of protein structures has recently received a lot of attention from structural biologists. The majority of recent research projects have tried to improve protein structure identification in huge data. Feature selection-based protein structure identification in large data analysis, on the other hand, takes a long time. A hybrid crow search algorithm and particle swarm optimization (CSA-PSO) based CD4.5 (CP-CD) approach has been developed to increase Protein Structure Identification accuracy with less amount of time. First samples from the patients are given to IOT-enabled microscope and the details will be stored in big data and then the process will be divided into two steps. At first, feature selection is done using CSA-PSO algorithm, and the classification is done using CD4.5 classifier. This aids in identifying the protein structure and accurately diagnosing the condition, as well as lowering the false positive rate.

Keywords *– Protein structure classification, feature selection, Big data analysis, CD4.5 classifier, IOT-enabled microscope.*

1. INTRODUCTION

Proteins are important in biological activities and are collected of amino acids connected together by peptide bonds. The three-dimensional structure of atoms in a protein molecule is known as protein structure. Protein structures are analyzed using nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography. Protein sequence is a method of determining a protein's amino acid sequence or structure. The sequence and three-dimensional (3D) structure of a protein influence its function. At an unprecedented rate, large-scale genome sequencing efforts are supplying researchers with millions of protein sequences from numerous species. [1]. Enzymatic catalysis, transferring ions and chemicals from one organ to other, nutrition, the contractile system of muscles, tendons, cartilage, antibodies, and modulating cellular and physical processes are all roles performed by proteins [2]. Structure of protein shown in Figure 1.

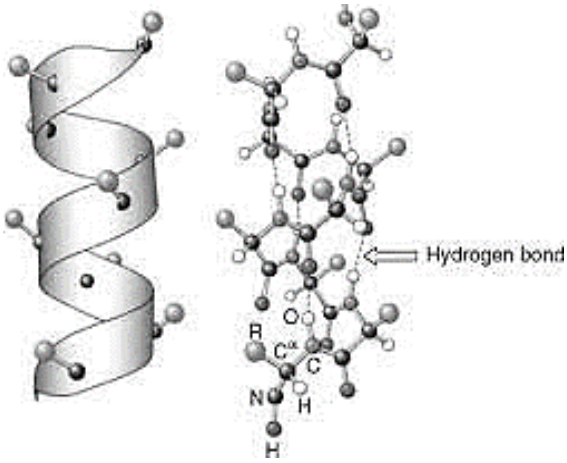

Figure 1. Structure of protein

The primary structure in a polypeptide chain is the arrangement of amino acids. The regular, repeating spatial groupings of nearby amino acid residues in a polypeptide chain are referred to as secondary structure. The amide hydrogens and carbonyl oxygens are tightly bonded together to form the peptide backbone. Helixes and structures are the most common types of secondary structures [2]. In medical diagnostics, identifying the protein structure is essential in determining the disorder. Feature selection is the process of choosing the most important inputs to process and analyze, or reducing the total amount of inputs. Increasing the predictive model's performance while reducing the modeling cost is the main goal of feature selection [3].

The Internet of Things, or IoT, is a network of individually addressable physical items that can interact and communicate with each other through the Internet. These objects have varied degrees of processing, sensing, and actuation capabilities. Therefore, the main objective of the IoT is to allow objects to connect with people and other objects at any time and from any place by using any network, technique, or service [4]. The simplest method to assess an algorithm is to check at all subsets of potential functions and see what decreases the proportion of errors. It is a systematic

exploration for space that benefits everyone in the calculation except the smallest feature set. Wrappers, filters, and embedding techniques are three different types of feature selection algorithms with different test evaluations. The term feature selection is often used in data mining to limit inputs to a manageable size for analysis and processing, with an emphasis on discovering relevant content without compromising the classification algorithm's accuracy [5].

Bioinformatics and theoretical chemistry work together in medical applications to analyse protein structure big data. The majority of recent research projects have tried to improve protein structure recognition in large datasets. However, in big data analysis, feature selection-based protein structure identification does not save time. As a result, feature selection is required to establish the structure of the protein. And for category label, the key characteristic has a powerful meaning and significance. The duplicated functions, on the other hand, not only affect the algorithm's classification performance but also contribute to the processing expenses. With the feature selection process, which selects the best feature subset of the original feature domain, it is critical to reduce unwanted and redundant features.

The abovementioned defect causes lots of new problems, including a failure to choose relevant features, lower classification accuracy and longer identification times, a large false positive rate, and so on. To deal with such problems, hybrid CSA-PSO algorithm based CD4.5 classification (CP-CD) method has been presented. In this method patient's samples will be acquired and then given to an IOT-enabled microscope. The findings will be kept as large data, and the procedure will be divided into two steps. In the first stage, a hybridised crow search and particle swarm optimization technique is used to identify features. The classification process is then accomplished in the second stage.

The other sections of the document are arranged under the respective topics. Section II provides a description of the literature review. Section III provides a description of the suggested method. Section IV provides a description of the findings and discussion. Section V provides a description of the conclusion.

2. LITERATURE REVIEW

Protein sequence classification is a crucial process for identifying the disease in humans. There is lot of research on this protein structure classification, among these a few are discussed here.

In 2021, Sequeira, et al, [6] proposed a method called ProPythia, a generic and modular Python program that permits users to quickly apply ML and DL algorithms to a variety of protein sequence prediction and classification challenges. It makes it easier to implement, compare, and validate the main errands in ML or DL pipelines, such as modules for reading and altering series, calculating protein features, preprocessing datasets, and dimensionality drop, feature selection gathering and diverse scrutiny, and training and optimizing ML/DL models and using them to create forecasts. They also compare the presentation of the various forms in four distinct protein categorization problems.

In 2020, Kalaiselvi, and Thangamani, [7] have proposed a Weighted Pearson Correlation based Improved Random Forest Classification (WPC-IRFC) Technique. The WPC-IRFC method was created with the goal of improving protein structure prediction accuracy while saving time. WPC-IRFC Technique achieves 7% FPR in an experimental assessment utilising 50-500 amino acid characteristics from VariBench DS, whereas previous techniques achieve 22 percent, 20 percent, 17 percent, and 14 percent. As a result, the FPR of the WPC-IRFC method is smaller than other approaches.

In 2020, Ge et al., [8] proposed a step-by-step classification approach on the basis of double-layer SVM model to calculate the proteins' secondary structure. This approach is evaluated using a frequently used dataset, the 25PDB dataset, which has a sequence similarity of less than 40%. Despite the fact that these two models' accuracy is somewhat reduced, the correctness of the α+β and α/β classes is upto 85.09 percent and 78.64 percent, respectively, and the correctness of the α+β class is greater than existing techniques. The findings reveal that this technique performs well, and the correctness of α+β class proteins is greatly enhanced by assuring the correctness of the other three structural classes of proteins.

In 2017, Shu, and Yong, [9] describes a method for classifying protein secondary structures on the basis of protein "signal-plotting" and digital signal processing using the Fourier methodology. It has been shown that a larger variety of protein secondary structures may be categorized using these indices, which are the hydrophobicity rate and the dominant frequency. Finally, it is hoped the discovery will usher in a whole new era of protein secondary structure analysis, as well as DNA and protein sequence analysis. The findings indicate that these newly proposed indices can classify a greater variety of protein secondary structures.

In 2017, Najibi, et al., [10] developed a nonparametric approach for estimating numerous bivariate density functions for a group of populations with protein backbone angles. The suggested approach would be more effective than previous methods. The adaptive basis expansion coefficients for the fitted densities give a low-dimensional depiction of the densities which may be used for conception, grouping, and identification. The proposed method takes a novel and innovative approach to two important and challenging problems in protein structure research: structure-based protein classification and angular-sampling-based protein loop structure prediction.

In 2019, Ahmad, and Hayat, [11] proposed a revolutionary high-throughput computational methodology for accurately identifying subGolgi proteins. The publicly accessible benchmark dataset is very unbalanced, with trans-Golgi sequences accounting for 72 percent of the entire dataset. The high-rank features are chosen using a maximum vote technique, which reduces the feature space by 85 percent. The results show that using a KNN classifier in conjunction with a hybrid feature space yielded good results. It has a jackknife cross-validation accuracy of 98 percent,

individual data accuracy of 94 percent, and a 10-fold cross-validation accuracy of 96 percent.

In 2020 Ahmad, et al., [12] have suggested a method that employs numerical descriptors based on sequences and evolution, primary protein sequences are constructed in this work. While evolutionary characteristics are gathered utilising a bigram scoring matrix customized to positions, sequential information is extracted employing K-space amino acid pair (KSAAP) and dipeptide composition. SVM with ideal features had a correctness of 97.54% for the training dataset and 93.71% for the independent dataset, respectively. Their suggested model was shown to outperform and provide the best results among the current computational models.

In 2020, Mirceva, et al., [13] presented a method for categorising protein shapes in this work. The results revealed that filtering 20 or 30 of the most significant attributes only slightly reduces performance in general. Only the C4.5 classifier is exempt from this, but this is due to the nature of this classifier, which picks the features with the best information gain throughout the model induction phase. The earlier results concerning the minor drop in accuracy by decreasing to 20 and 30 features are crucial since it suggests that the time required for training and testing the models might be cut in half with feature selection while still maintaining high accuracy.

In 2019, Mirceva, et al., [14] suggest a method for categorising protein structures in this work. They create models by combining several categorization algorithms. The proposed technique is thoroughly examined, as well as the advantages of using feature selection. The results demonstrate that feature selection produces superior outcomes in virtually all circumstances than when no feature selection is used. The investigation's overall conclusion was that most of the ways in the analysis perform better than the protein voxel-based descriptor, even though it beats several of the strategies.

In 2020, Ghosh, et al., [15] suggested a ML–based approach for classifying secondary structure of proteins into four categories: all-α, all-β, α+β, and/. On the four standard datasets 640, 1189, 25pdb, and fc699, the overall accuracies achieved using the proposed model are 86.89 percent, 92.93 percent, 91.38 percent, and 94.87 percent, respectively. In this comparison, the suggested model outperforms certain state-of-the-art approaches.

3. PROPOSED METHOD

Protein structure identification is crucial for disease detection in big data analysis. Several data mining techniques, such as gene structure, DNA sequences, and protein sequences, have been developed in the disease diagnostic area. To eliminate the problem of mystery cases and prediction analysis during illness diagnosis, the suggested approach, protein sequences identification, is used. The suggested method efficiently identifies protein structures for brain tumour diagnosis.

Protein structure is the three-dimensional configuration of atoms within an amino acid chain molecule. Peptide bonds are formed by the condensation of amino acids to create protein structures. The terminus of a peptide or protein sequence with a free carboxyl group is called the carboxy-terminus, or C-terminus. The termini of a sequence with a free -amino group are denoted by the terms amino-terminus and N-terminus. Proteins are composed of twenty different compounds called amino acids. The citric acid cycle, Glycolysis, and the pentose phosphate pathway all offer intermediaries that are used to make amino acids. The 20 amino acids are made up of both essential and non-essential amino acids. Nine amino acids are essential, whereas the remaining nine are non-essential. The genetic code determines the amino acid sequence in a protein as well as its function. 20 different types of Amino acids shown in Table 1.

Table 1. 20 different types of Amino acids

Glycine	Gly	G	Tyrosine	Try	Y
Alanine	Ala	A	Methionine	Mer	M
Serine	Ser	S	Tryptophan	Trp	T
Threonine	Thr	T	Asparagine	Asn	A
cysteine	Cys	C	Glutamine	Gln	G
Valine	Val	V	Histidine	His	H
Isoleucine	Ile	I	Aspartic Acid	Asp	A
leucine	Leu	L	Glutamic Acid	Glu	G
Proline	Pro	P	Lysine	Lys	L
Phenylalanine	Phe	P	Arginine	Arg	A

Protein materials are composed up of a precise order of amino acids. The amino acid sequence is indicated on these strings. As a result, the erection of a protein explains the specific classification in which amino acids are connected together by peptide bonds to create a protein.

Fig 2 represents the flow of proposed methodology. In bioinformatics, protein structure identification is a critical step. Many factors contained in the training data set may increase the risk of correctly identifying the protein structure in real-world applications. As a result, for protein structure big data analysis to diagnose brain tumour illness and reduce the risk in protein structure identification, feature selection and classification are necessary. Attribute selection from a big dataset is also known as feature selection. The protein structure is then identified using the classification technique.

In the proposed technique, the samples of the patients are collected and tested using an IOT enabled microscope and the details will be automatically send to the big data cloud and also it informs the hospital so that the information can be accessed remotely. The proposed CP-CD technique consists of two processing steps: feature selection and classification, which allow for quick protein structure identification. For feature selection in the initial stage, a hybrid CSA-PSO approach is applied. The CD4.5 classifier is used to classify the selected features in the second stage. This aids in improving the efficiency of bioinformatics data processing while also saving time.

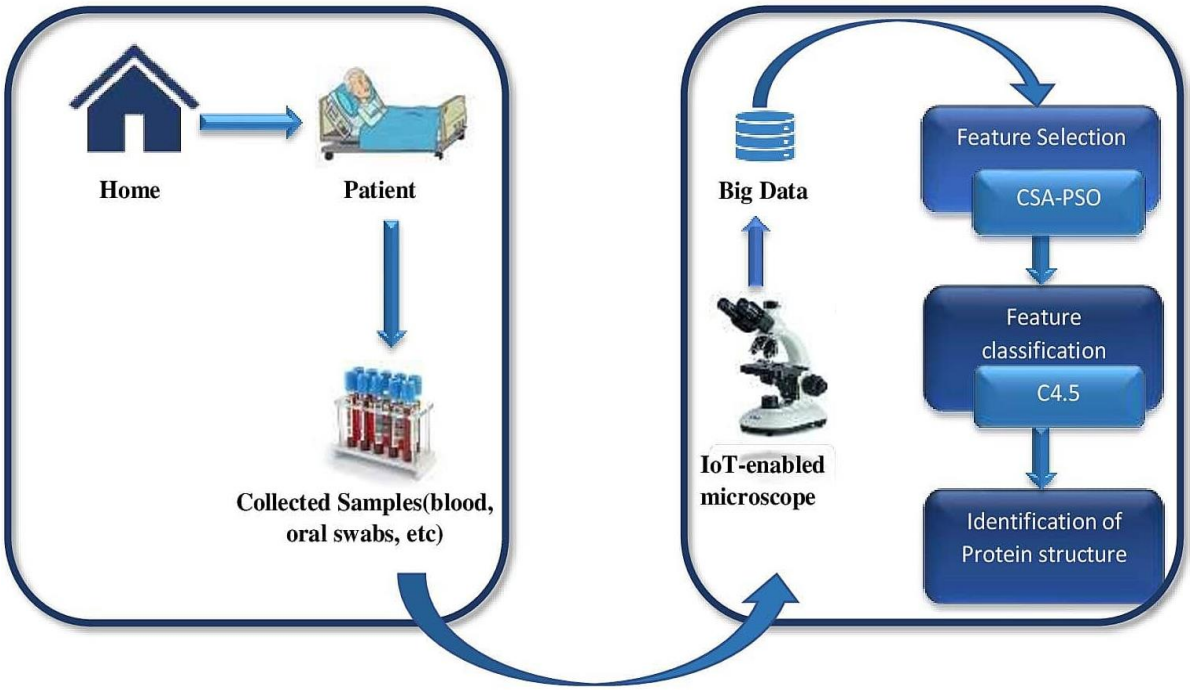

Figure 2. Schematic representation for proposed framework

3.1. Feature selection using CSA-PSO algorithm

The hybrid Crow Search Algorithm and Particle Swarm Optimization (CSA-PSO) algorithm. This algorithm hybrids the properties of crow search algorithm and particle swarm optimization algorithm which will give better results for feature selection from the large dataset.

a) Crow search algorithm

Crows were used as inspiration for the crow search algorithm because of their habit of keeping food in a secret area and recovering it after several months. Crows, like other social animals, may participate in thievery at some point by carefully studying other Crows' food concealing locations and then robbering their food. When a crow doubt that other is following him, he flees to a position faraway from where the food is hidden to deceive the thief. The CSA technique is linked with chaotic series in this method and it is represented as,

$$s_x^{(r+1)} = \begin{cases} s_x^{(r)} + J_y * Ht_y^{(r)} * \left(n_y^r - s_x^{(r)}\right) & J_i \geq BQ_t^r \\ choose\ a\ random\ position & otherwise \end{cases} \quad (1)$$

b) Particle Swarm Optimization algorithm

The attribute choices on the basis of social characteristics related with bird flocking to resolve optimization issues attract a lot of academic attention in Particle Swarm Optimization (PSO). PSO, which is a type of swarm intelligence optimization, has been shown to be lesser computationally expensive and to settle more quickly. Every solution in PSO may be seen as a swarm of particles, each with its own velocity and position.

c) Opposition Based Learning:

The OBL approach searches in both directions in the search space. One of these two pathways contains the initial answer, while the other indicates the opposite direction. The opposite location in M-Dimensional space with s (s_1……s_m) and $s_1 \in [\delta, \gamma]$, x=1,2,3….M is calculated in the below equation

$$s_x^{\overline{opp}} = \delta_x + \gamma_x - s_x^r \quad (2)$$

d) CSA-PSO algorithm:

The concepts of binary CSA and binary PSO algorithms have been mixed, resulting in a technique called CSA-PSO that benefits from their inclusion. For example, in CSA-PSO approach, only aiming particular crows with better foods improves the execution of randomly following each crow. The Opposition Based Learning approach is then used to create the crows' opposite positions, which are subsequently utilized to upgrade the post in the PSO. This is achieved so that both methods can examine the exploration space in turn, without being impacted by the results of the other.

Fig 3 represents the feature selection process using this CSA-PCO method. The first step is preprocessing, that is to get the details we needed from the big dataset are to preprocessed for further process. In CP-CD technique, CSA-PSO is the hybrid technique of crow search and particle swarm optimization method.

Logistics map:

$$s_x^{r+1} = bs_x^r(1 - s_x^r) \quad b = 0.4\ and\ s_1 = 0.7$$

Exponential map: $s_x^{r+1} = s_x^r e^{2(1-s_x^r)}$ s_1=0.7

$$J_{r+1} = v + s_x^{r+1},$$

Where v is the energetic parameter that controls s_x^ractivity. When v steps up, s_x^r undergoes further bifurcations, eventually resulting in pandemonium. The current situation would change and the Crow would move to

the right answer if a predetermined random number was less than this threshold value.

$$U_{shape} = \left|\frac{2}{\pi}\ arctan(\frac{\pi}{2}s_x^r)\right| \quad (3)$$

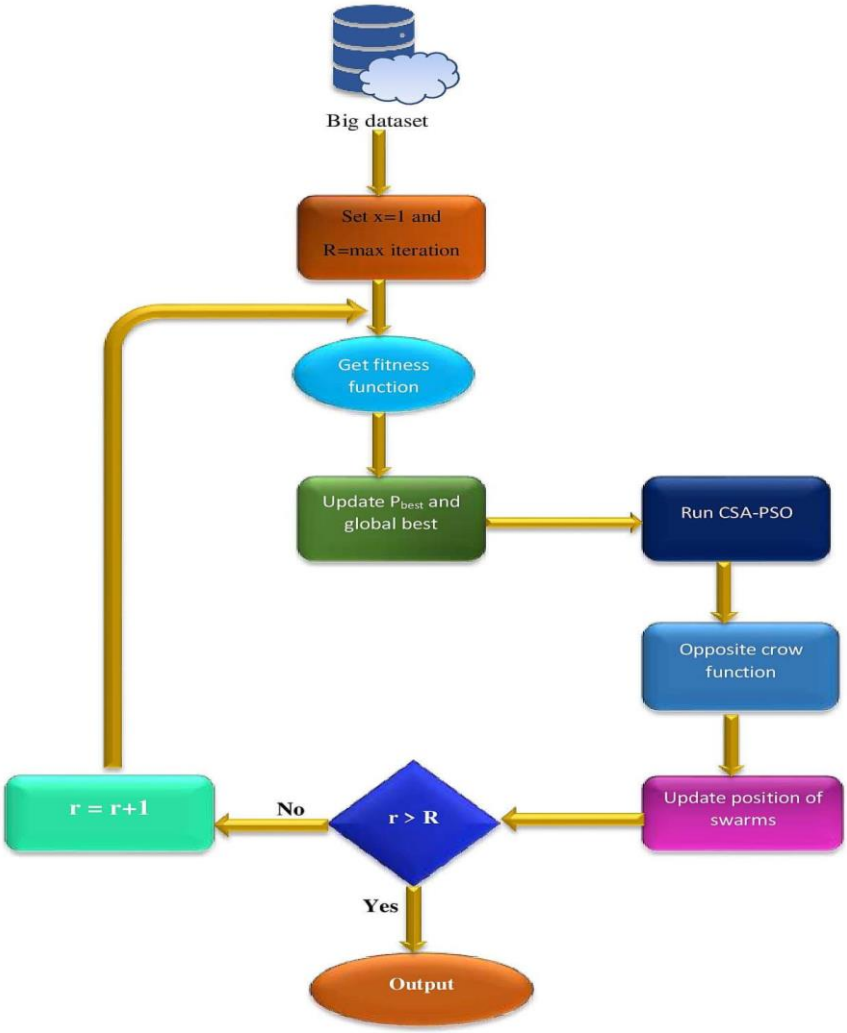

Figure 3. Flow diagram that represents feature selection using CSA-PCO

e) Fitness function

Algorithm 1: CSA-PSO based feature selection algorithm

Algorithm for CSA-PSO

1. Begin
2. Initialize x=0
3. Wgt= wgt_{max}-iteration $(\frac{wgt_{max}-wgt_{min}}{max-iteration})$
4. Evaluation of fitness value
5. P_{best}, g_{best} values are set
6. Run CSA-PSO with S_x^r population
7. Reversely alter the position that the CSA-PSO returned.
8. for (x=1; x≤ $D; x++$) do
9. if$(f(S_x^r) \geq crowmin) do$
10. $s_x^{\overline{opp}}=\delta_x + \gamma_x - s_x^r$
11. end if
12. end for
13. upgrade swarm position
14. for v=1 to QQ
15. $U_x^{r+1} = wgtu_s^r + J_1 n_{x1}(P_{best_x}^{\ r} - S_x^r) + J_2 n_{x2}(G_{best_x}^{\ r} - S_x^r)$
16. end
17. for v=1 to QQ
18. for y=1 to N
19. if (u(x,y) > U_{max})
20. u(x,y) = U_{max}
21. end
22. if(u(x,y)< -U_{max})
23. u(x,y)=-U_{max}
24. end
25. q=$\frac{1}{1+e^{-u(x,y)}}$
26. If (nand < q)
27. $S_{x,y}^{r+1} = 1$
28. else
29. $S_{x,y}^{r+1} = 0$
30. end
31. end
32. end
33. r=r+1
34. Produce best results

The equation defines the fitness function for finding results to attain a balance between the two objectives.

$$fitness = \delta\Delta_D(N) + \gamma\frac{|Z|}{|R|} \quad (4)$$

$\Delta_D(N)$ represents the error rate of classifier, $|Z|$ represents the subset's size which the method chooses and $|R|$ represents the absolute number of features in the existing dataset. δ is a parameter ∈ [0,1] associating to weight of error rate for classification. $\gamma = 1 - \delta$ represents the importance of decrease in feature.

The given algorithm is utilized to choose pertinent characteristics for categorization from a large dataset. To construct a protein structure, the properties that are most closely associated to the amino acid are chosen. This contributes to a higher true positive rate.

f) CD4.5 machine learning classifier

The classification is done using the c4.5 machine learning classifier after the relevant features from the huge dataset have been selected. The C4.5 technique is employed in data mining as a Decision Tree Classifier, which may be used to decide on the basis of a sample of data. Classification using CD4.5 classifier shown in Figure 4.

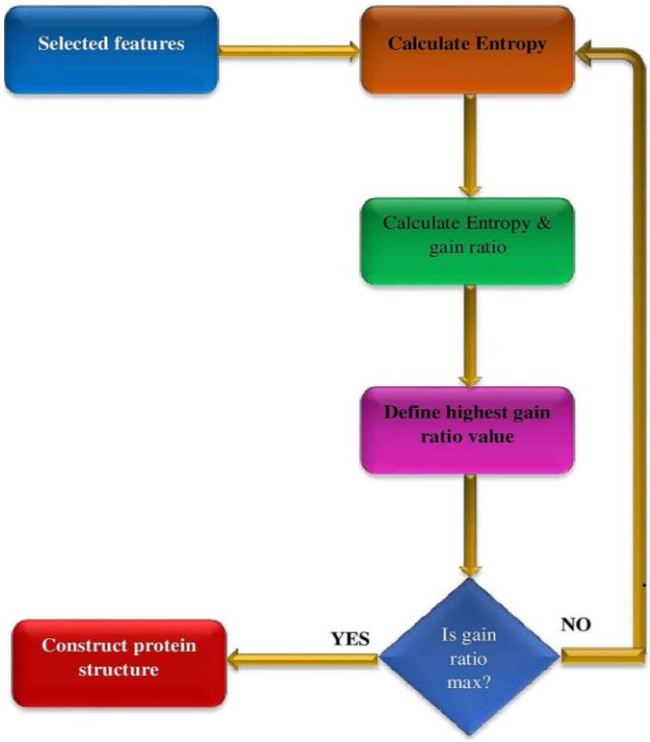

Figure 4. Classification using CD4.5 classifier

Ross Quinlan created the C4.5 algorithm, which is used to build decision trees. C4.5 extends the ID3 methodology. C4.5 is referred to as a statistical classifier since it generates decision trees that may be used to categorize data. Similar to ID3, C4.5 uses the concept of information entropy to build decision trees from a collection of training data.

g) Information Entropy

Information gain is the reduction in entropy produced by modifying a dataset, and it is commonly used in the training of decision trees. To measure information gain, the entropy of a dataset before and after a mutation is employed.

$$H(z) = -\sum_{x=1}^{m} p(y_j)p(y_j) \quad (5)$$

If a distinct classification for the resultant feature can be established for each of the feature values, the information gain is equivalent to the total entropy for that attribute. The relative entropies removed from the overall entropy are 0 in this scenario. The training dataset is a set D=d_1,$d_{2,...}$ of already classified samples. Each sample displaystyle d_i is made up of a m-dimensional vector display style ($y_{1,j}$,$y_{2,j}$,….$y_{m,j}$), where the y_i reflect the sample's attribute values or features, and the class in which y_j falls. C4.5 picks the properties of data which efficiently separates the sample set into subsets overloaded in one class or the other at every node. The normalized information gain is employed as a dividing criterion. The characteristic with the largest standardized information gain is selected for selections. The C4.5 method then iterate through the subdivided subsets. The CD4.5 decision tree is constructed as follows, using entropy and information gain calculations. Decision tree structure for CD 4.5 classifier shown in Figure 5.

Algorithm for CD4.5 classifier

Input: Training dataset D, features selected

Output: Classification of protein structure

Step:1 Examine the above-mentioned base cases.

Step:2 Find the standardized information gain ratio from dividing on a for each attribute d.

Step:3 Assume that d_{best} has the maximum normalized information gain.

Step:4 Make a decision node that splits based on the value of d_{best}.

Step:5 Recur on the subsets formed by separating on a best and append them as children to node.

End

The above diagram represents the decision tree structure for CD 4.5 classifier. A, B, C are the features. There are root nodes and leaf nodes in a typical decision tree. Features are used to represent the nodes. A subset of characteristics is represented by each node's decision. With a class label, the leaf node is also known as the tree's terminal node. As a selection, the feature with the largest information gain is picked. Every route from the root node to the leaf node in the decision tree creates a categorization rule. A decision tree is a type of recursive classification classifier. Every leaf node in the diagram represents a categorization judgement of characteristics to construct a protein structure. This reduces the number of false positives.

There are a few fundamental uses for this technique.

- All of the lists' collections belong to the same category. All that happens is that a leaf node telling the user to choose that class is added to the decision tree.
- None of the characteristics yield any information. In this case, C4.5 builds a decision node based on the anticipated rate.

- A class instance that had never been observed before occurred. C4.5 builds a decision node using the anticipated rate.

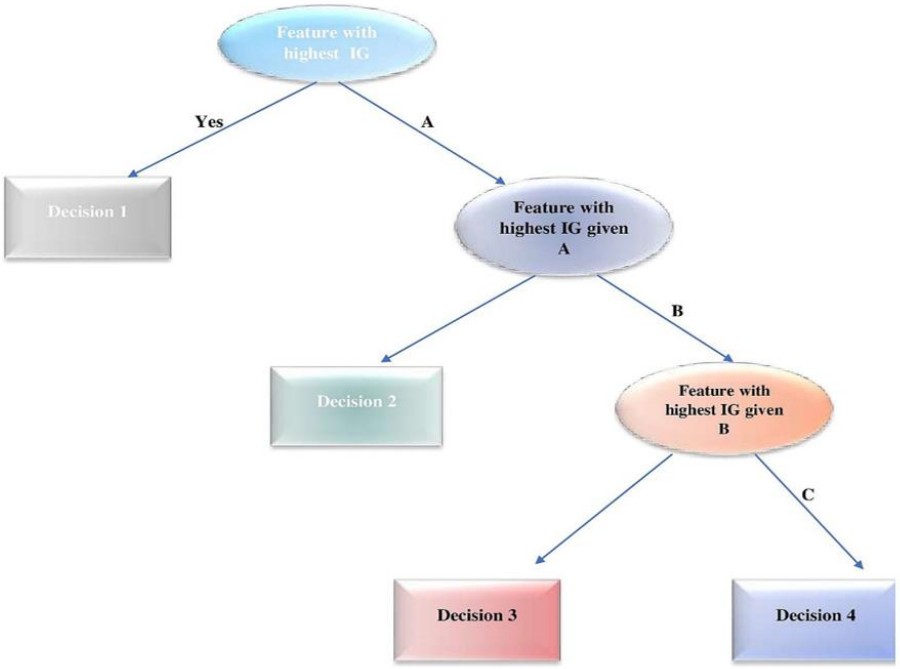

Figure 5. decision tree structure for CD 4.5 classifier

The algorithmic description of a CD4.5 decision tree method to pick the attributes of amino acid sequence to determine the protein structure is shown in algorithm 2. To categorise the best and greatest gain splitting features, entropy value and information gain are assessed for each picked feature from the huge dataset. After determining the best features, every node denotes a feature test, and every leaf node denotes a class label, the amino acid sequence may be identified. To build a protein structure, certain amino acid sequence characteristics are chosen. These speeds up the process of identifying the protein structure.

4. RESULTS AND DISCUSSION

For determining the protein structure, a CSA-PSO and CD4.5 classifier (CP-CD) approach is tested utilising the JAVA language and the Weka tool. The proposed CSA-PSO based CD 4.5 classification (CP-CD) technique is compared to 4 existing models: the Logistic Regression based Iterative Dichotomiser 3 classification (LR-ID3) technique, the Hydrogen-deuterium Exchange measured by Nuclear Magnetic Resonance (HDX-NMR) technique, the TAndem PrOtein detector (TAPO) method, and the Protein secondary structure prediction (PSSP) technique. Five datasets (i.e., Protein Data Bank (PDB), ProteinNet, PROSITE, sidechainNet, and pfam dataset) are utilized to effectively identify the protein structure to illustrate the benefit of the proposed CP-CD approach. It contains data on 3D protein shapes, nucleic acids, and complex assemblies, all of which are used to determine the importance of proteins in terms of health and illness.

The purpose of the PDB database is to classify and describe protein structures while also giving biological data. SEQRES (i.e. Residues in the Sequence) entries in the PDB database include the sequences of the three peptide chains A, B, and C. It is used in a variety of fields such as molecular biology, structural biology, and computational biology.

ProteinNet is a standardised data collection for protein structure machine learning ProteinNet relies on the biennial CASP evaluations, which include making blind predictions of freshly resolved but publicly accessible protein structures, to provide test sets that push the boundaries of computational methods.

PROSITE was the world's first secondary database. Most protein families have certain highly conserved motifs that may be decoded to determine diverse biological activities. When a new sequence is found, we may quickly determine the protein family by utilising a database tool like this. PROSITE is essential in this regard. A regular expression is used to encode motifs in PROSITE (called patterns).

Sidechain Net is an extension of ProteinNet1 dataset for protein structure prediction. In particular, Sidechain Net replaces the protein backbone with measurements for protein angles and coordinates that characterize the whole, all-atom protein structure (backbone and sidechain, excluding hydrogens).

The Pfam database's major goal is to give a precise and detailed identification of protein sequences. The goal of building the database is to increase genome annotation efficiency.

The true positive rate, protein structure identification accuracy, false positive rate, and protein structure

identification time, recall, precision, and F-measure are measure against existing approaches to calculate the performance of CP-CD technology.

a) True Positive rate

It is expressed as a percentage of the amount of correctly picked features when measure against the total number of characteristics in the big data for protein structure identification (%).

$$True\ Positive\ Rate(TPR) = \frac{Number\ of\ features\ selected\ correctly}{Total\ number\ of\ features} \times 100 \quad (6)$$

The procedure is believed to be more efficient if the true positive rate is much higher. Table 2 shows the experimental findings for true positive rate based on the number of characteristics.

Table 2. Comparison of true positive rate

Dataset	True Positive Rate				
	PSSP	HDX-NMR	TAPO	LR-ID3C	CP-CD
PDB	72	78	80	85	90
ProteinNet	77	80	83	86	92
PROSITE	82	84	87	88	93
SidechainNet	84	85	88	92	94
pfam	85	88	92	94	97

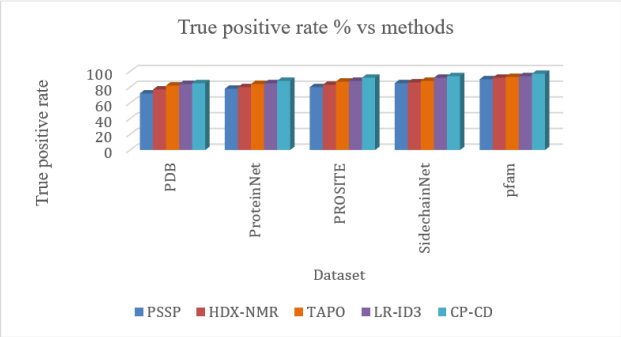

Figure 6. Comparison of true positive rate of CP-CD technique with other existing methods

Figure 6 represents that the suggested CP-CD methodology achieves a true positive rate of 97 percent for 500 protein characteristics, whereas current techniques such as LR-ID3C, TAPO, HDX-NMR and PSSP provide true positive rates of 94 percent, 92 percent, 88 percent, and 85 percent, respectively, as shown in Table 2.

b) Protein structure identification accuracy

The relation of number of features with least entropy and highest information gain employed to create the decision tree to the number of features in the database is described as protein structure identification accuracy in the CP-CD approach. The formula for determining the accuracy of protein structure identification is as follows:

$$PSIA = \frac{Number\ of\ features - Number\ of\ maximum\ Information\ gain}{Number\ of\ features} \times 100 \quad (7)$$

Where PSIA represents the Protein Structure Identification Accuracy.

Table 3. Comparison of protein structure identification accuracy

PSIA					
Dataset	PDB	proteinNet	PROSITE	SidechainNet	pfam
PSSP	83	84	86	88	90
HDX-NMR	88	89	90	92	93
TAPO	92	93	94	94	95
LR-ID3C	93	95	96	97	98
CP-CD	95	96	97	98	99

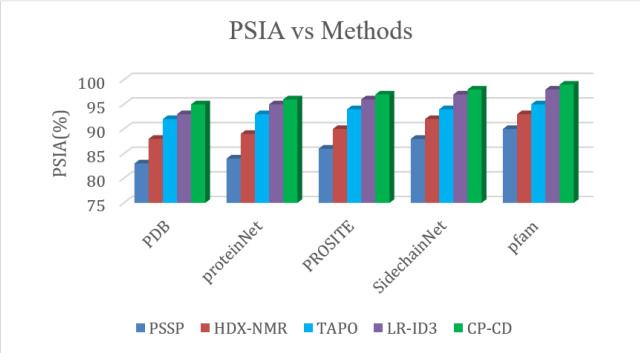

Figure 7. Comparison of PSIA for CP-CD method with existing techniques

Figure 7 shows the results of five protein datasets' protein structure identification accuracy. Table 3 shows that the proposed CP-CD technique produces higher protein structure identification accuracy results of 99% for 500 number of features (PDB), whereas existing methods such as LR-ID3C, TAPO, HDX-NMR, and PSSP produce protein structure identification accuracy results of 98 percent, 95 percent, 93 percent, and 90 percent, respectively.

c) False Positive Rate:

Table 4. Comparison of False positive rate

False Positive Rate					
Datasets	PDB	ProteinNet	PROSITE	SidechainNet	pfam
PSSP	24	22	20	19	15
HDX-NMR	23	20	19	17	13
TAPO	22	19	18	16	12
LR-ID3C	18	16	15	14	11
CP-CD	14	13	11	10	9

By dividing the number of features by the improperly detected features, one may calculate the false positive rate. It has a percentage (%) as its expression.

$$False\ Positive\ Rate = \frac{incorrectly\ identified\ feaures}{Number\ of\ features} \times 100 \quad (8)$$

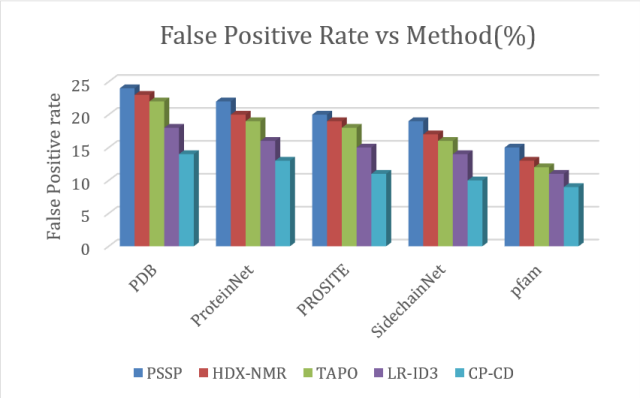

Figure 8. Comparison of False positive rate for CP-CD technique with other existing techniques

The false positive rate of the proposed CP-CD methodology and four existing techniques were compared in Figure 8. The suggested CP-CD methodology achieves a minimal false positive rate of 9% for 500 samples (PDB) as shown in the figure, but other techniques such as LR-ID3C, TAPO, HDX-NMR, and PSSP create false positive rates of 11%, 12 percent, 13 percent, and 15percent, respectively, as shown in Table 4.

d) Protein structure Identification Time

The time it takes to find a protein structure with the most information gain characteristics from a dataset is called protein structure identification time. The following is how time is calculated:

$$Protein\ structure\ identification\ Time = Number\ of\ features \times time \quad (9)$$

Figure 9 shows the time it took five algorithms to identify protein structures on five protein datasets with varying numbers of characteristics.

Table 5. Comparison of protein structure identification time

Protein structure identification time					
Datasets	**PDB**	**ProteinNet**	**PROSITE**	**SidechainNet**	**pfam**
PSSP	25	27	29	31	33
HDX-NMR	23	24	26	28	29
TAPO	20	21	23	26	28
LR-ID3C	16	18	20	24	26
CP-CD	11	12	14	16	20

The proposed CP-CD technique produces a least protein structure identification time of 11 ms for 500 features (PDB) as shown in the figure, whereas other existing methods such as LR-ID3C, TAPO, HDX-NMR, and PSSP produce protein structure identification times of 16 ms, 20 ms, 23 ms, and 25 ms as shown in Table 5.

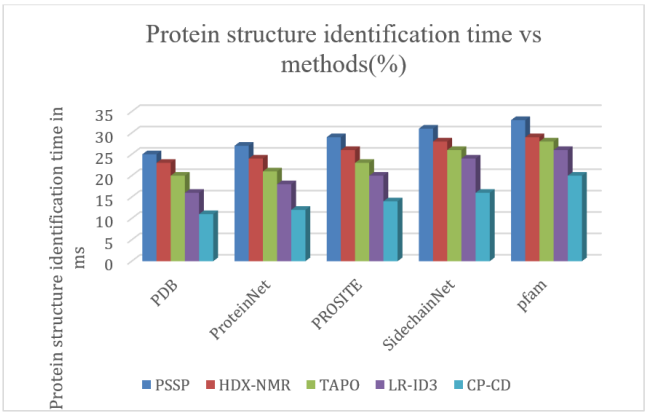

Figure 9. Comparison of Protein structure identification time for CP-CD technique with other existing methods

e) Precision

Divide the true positive rate by the total of the collection's true positive and false positive rates to get the precision. The exact mathematical formula is as follows:

$$precision(p) = \frac{True\ positive}{True\ positive + False\ positive} \quad (10)$$

Table 6. Comparison of precision

Precision					
Datasets	**PDB**	**ProteinNet**	**PROSITE**	**SidechainNet**	**pfam**
PSSP	87	86	82	79	76
HDX-NMR	88	87	85	84	78
TAPO	89	87	84	83	80
LR-ID3C	92	90	89	85	87
CP-CD	94	91	90	89	90

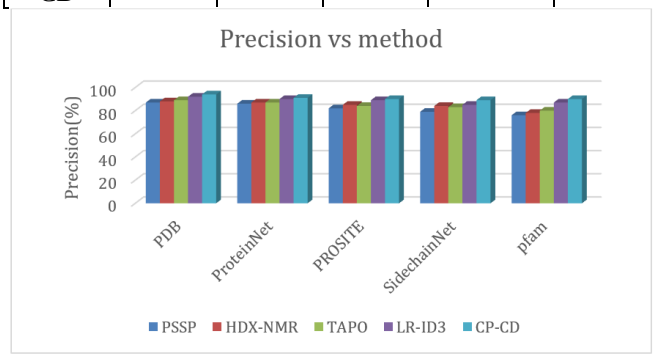

Figure 10. Comparison of Precision for CP-CD method with other existing methods

Figure 10 displays the accuracy findings of five different approaches. The proposed CP-CD method acquires higher precision results of 90% for 500 features (amino acid sequence from PDB), whereas other methods such as LR-ID3C, TAPO, HDX-NMR, and PSSP provide precision

results of 87 percent, 80 percent, 78 percent, and 76 percent, respectively, as shown in Table 6.

f) Recall

The rate of number of relevant characteristics to the the absolute number of features that really correspond to the relevant features is how recall is calculated. The recall value is calculated in the following way:

$$Recall(R) = \frac{True\ positive}{True\ positive + False\ positive} \quad (11)$$

Table 7. Comparison of Recall

Recall					
Datasets	**PDB**	**ProteinNet**	**PROSITE**	**Sidechain Net**	**pfam**
PSSP	78	79	83	84	86
HDX-NMR	79	82	84	85	87
TAPO	82	83	84	86	89
LR-ID3C	84	85	86	89	90
CP-CD	85	86	87	90	92

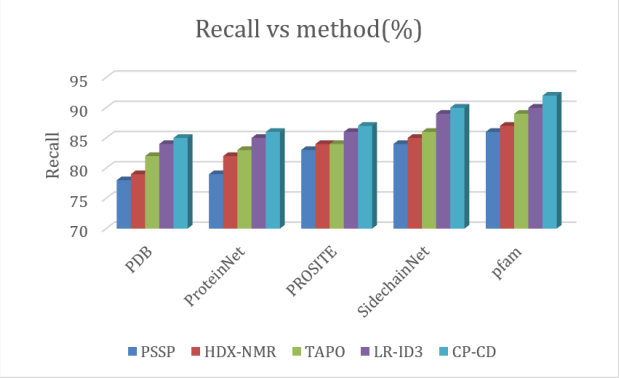

Figure 11. Comparison of Recall for CP-CD method with other existing methods

According to the Figure 11, the CP-CD method acquires larger recall results of 92 percent for 500 number of features of amino acid sequence from the PDB, whereas other methods such as LR-ID3C, TAPO, HDX-NMR, and PSSP produce recall results of 90 percent, 89 percent, 87 percent, and 86 percent values, which are mentioned in Table 7.

g) F-measure

The F-measure is a single positive class test measure. It's the weighted mean of a test's precision and recall. It is written down as follows:

$$F - measure = \frac{P \times R}{P + R} \quad (12)$$

Figure 12 shows that the proposed CP-CD method achieves higher F-measure results of 92 percent for 500 features (PDB), whereas other existing methods such as LR-ID3C, TAPO, HDX-NMR, and PSSP produce F-measure results of 90 percent, 88 percent, 87 percent, and 86 percent, respectively, as shown in Table 8.

Table 8. comparison of F-measure

F-measure					
Datasets	**PDB**	**Protein Net**	**PROSITE**	**SidechainNet**	**pfam**
PSSP	76	78	79	83	86
HDX-NMR	78	82	84	85	87
TAPO	79	83	85	87	88
LR-ID3C	82	84	86	89	90
CP-CD	85	87	88	90	92

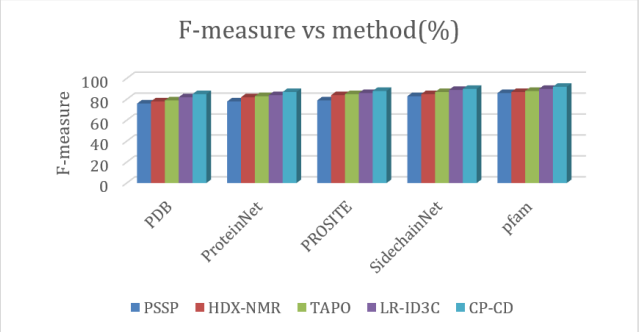

Figure 12. Comparison of F-measure for CP-CD method with other existing methods

5. CONCLUSION

In biotechnology, determining the function of proteins is a critical step. Because of the time necessary to execute the categorization tasks, the excessive characteristics render computer systems unproductive. As a result, determining protein structure based on amino acid sequence requires careful attention. Therefore, a method called CSA-PSO based CD4.5 Classification technique is used. Here, the samples of the patients are collected and tested through an IOT enabled microsope and the details will be stored in the big data cloud.Then it have two steps: Initially, feature selection will takes place through a hybrid crow search algorithm and particle swarm optimization algorithm(CSA-PSO). It aids in increasing the chances of specific outcomes. This aids in improving the true positive rate in a short amount of time. Later, classification is done by using CD4.5 classifier, which is an extension of ID3 classifier. For making a choice to identify the structure, characteristics with the least entropy are picked. This improves the accuracy of protein structure recognition and lowers the percentage of false positives. Different protein datasets are used to evaluate the CP-CD approach in the experiment. When compared to the other protein dataset samples, PDB datasets provide better performance results. When comparing to other techniques, the suggested CP-CD methodology considerably enhances true positive rate, protein structure identification accuracy, precision, recall, F-measure with minimal protein structure identification time, and false positive rate. Further studies will be taken to use the offered approaches to deal with issues in protein structure identification and a high-dimensional perspective of structure in the future.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] N. Eswar, D. Eramian, B. Webb, M.Y. Shen, and A. Sali, "Protein structure modeling with Modeller", *In Structural proteomics*, pp. 145-159, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[2] M.M. Gromiha, "Protein bioinformatics: from sequence to function", academic press. 2010. [CrossRef] [Google Scholar] [Publisher Link]

[3] K. Kira, and L.A. Rendell, "A practical approach to feature selection", *In Machine learning proceedings* 1992, pp. 249-256, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[4] A.R.H. Hussein, "Internet of things (IOT): Research challenges and future applications", *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp.77-82, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] J. Masood, M. Shahzad, Z.A. Khan, V. Akre, A. Rajan, S. Ahmed, and F. Masood, "Effective Classification Algorithms and Feature Selection for Bio-Medical Data using IoT," *In 2020 Seventh International Conference on Information Technology Trends (ITT)*, IEEE pp. 42-47, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] A.M. Sequeira, D. Lousa, and M. Rocha, "ProPythia: A Python package for protein classification based on machine and deep learning", *Neurocomputing*. 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] B. Kalaiselvi, and M. Thangamani, "An efficient Pearson correlation based improved random forest classification for protein structure prediction techniques", *Measurement*, Vol. 162, pp.107885, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Y. Ge, S. Zhao, and X. Zhao, "A step-by-step classification algorithm of protein secondary structures based on double-layer SVM model", *Genomics*, vol. 112, no. 2, pp. 1941-1946, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] J.J. Shu, and K.Y. Yong, "Fourier-based classification of protein secondary structures", *Biochemical and biophysical research communications*, vol. 485, no. 4, pp.731-735, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] S.M. Najibi, M. Maadooliat, L. Zhou, J.Z. Huang, and X. Gao, "Protein structure classification and loop modeling using multiple Ramachandran distributions", *Computational and structural biotechnology journal*, vol. 15, pp.243-254, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] J. Ahmad, and M. Hayat, "MFSC: Multi-voting-based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components", *Journal of Theoretical Biology*, vol. 463, pp. 99-109, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] A. Ahmad, S. Akbar, M. Hayat, F. Ali, and M. Sohail, "Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs-based descriptors incorporating with ensemble feature selection", *Biocybernetics and Biomedical Engineering*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] G. Mirceva, A. Naumoski, and A. Kulakov, "Classification of Protein Structures by Making Fuzzy-Rough Feature Selection", *In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1-42020. [CrossRef] [Google Scholar] [Publisher Link]

[14] G. Mirceva, I. Ivanoska, A. Naumoski, and A. Kulakov, "Feature Selection for Improved Classification of Protein Structures", *In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1013-1018, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[15] K.K. Ghosh, S. Ghosh, S. Sen, R. Sarkar, and U. Maulik, "A two-stage approach towards protein secondary structure classification", *Medical & Biological Engineering & Computing*, vol. 58, pp. 1723-1737, 2020. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

T. Maris Murugan Associate Professor and Head, Department of Electronics and Instrumentation Engineering, at Erode Sengunthar Engineering College, Erode, Tamilnadu, India. He was awarded B.E. degree in Electronics and Instrumentation Engineering from Government College of Technology, Coimbatore and M.E. degree in Embedded Systems from Karpagam University Coimbatore, Tamilnadu, India. He was awarded Ph.D. degree from Anna University, Chennai, Tamilnadu. He has 14 years in Industry and as well as 15 years in academic experience with expertise in the field of Process Control, Industrial Automation, Machine Learning and Internet of Things.

A. Jeyam received his Master Degree (M.E) in Computer Science and Engineering GCT-Coimbatore during 2012-2014. He has very good knowledge in Computer related subjects like Java, Python, Database systems, Foundations of Computer Systems, Website Designing, Software Engineering, Database Management Systems, Object Oriented Programming, Image Processing and Data Mining in university and college level, good knowledge in Computer software, Hardware and Networking. He has done many projects in Image Processing. He has worked as a Lecturer and Network Administrator for more than 05 years in Research Point India.

Arrived: 29.11.2023

Accepted: 28.12.2023