Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 6, 6970-6977

Machine Learning Tool Development And Use In Biological Information Decoding

Sheetalrani R Kawale[1] , kamalakar Ravindra Desai[2] , Parismita Sarma[3] , N. K. Darwante[4] , C M Velu[5] , Pundru Chandra Shaker Reddy[6]

[1]*Assistant Professor, Department of Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura, Karnataka, India.*
[2]*Professor, Department of Electronics and Telecommunication, Bharati vidyapeeths college of Engineering Kolhapur, Maharastra, India.*
[3]*Assistant Professor, Department of Information Technology, Gauhati University,Guwahati, Assam, India.*
[4]*Associate Professor, Department of Electronics and Telecommunication, Sanjivani College of Engineering, Kopargaon, Affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India.*
[5]*Professor, Department of CSE, Savertha School of Engineering, Saveetha University, SIMATS, Chennai, Tamilnadu, India.*
[6]*Associate Professor, School of Computing and Information Technology, REVA University, Bangalore, India,*

Abstract

DNA, RNA, and proteins are the main molecules of life, and the varied roles that proteins play determine the phenotypes of living organisms. Since proteins are polymers made up of amino acid molecules, it is crucial to understand their many roles and features in order to comprehend life at the molecular level. Complete protein sequences for many species have been obtained thanks to recent developments in high throughput deep sequencing methods. Experimental approaches to functionally annotating proteins are time-consuming, labor-intensive, and expensive. As a result, only a fraction of the total sequenced proteins have been annotated experimentally. Instead of using experiments to determine how proteins should be categorised, we may utilise machine learning techniques to train computer models using annotated proteins and then use those models to classify freshly sequenced proteins into their respective categories. Significant biological knowledge and computing ability are necessary for using machine learning. Machine learning algorithms, on the other hand, are meant to construct models without any human intervention. However, this is true only for numerical training data sets, since the vast majority of biological data are textual or otherwise qualitative in nature. Specific algorithms are needed to transform biological data into machine readable forms. Therefore, experimentalists rely on computer professionals to create models using machine learning for their data. Due to the need for assistance from computer professionals, the time it takes to generate hypotheses and uncover new information has increased.

Keywords: Machine Learning, RNA, DNA, Hypothesis, Medical, Framework.

1. Introduction

To put it another way, knowledge is the result of processing and analysing information. Recently, however, there has been an unparalleled exponential growth in the pace at which new data is created. Improvements in data creation and storage techniques have been crucial to this shift. It is hard to manually grasp, analyse the data to extract information, or formulate hypotheses due to the sheer number and dimensions of the data being generated[1]. A possible solution to this issue is to use AI (AI). There has been a lot of study in this area, and one of the offshoots, Machine Learning (ML), was created to help with classification and categorization by discovering previously undetected patterns in the data. Proteins and RNA (Ribonucleic acid) are the functional components of life that are translated from the

information contained in the genome. The ability to hypothesise about biological processes relies on a firm grasp of the function and interrelationship of these units[2]. Determining these functions experimentally is a time-consuming and expensive process. The fast sequencing of DNA, RNA, and proteins has been made possible by the development of high throughput technologies. Unfortunately, this trend has led to a significantly quicker growth in the number of uncharacterized sequences than in the number of characterised ones. Similarities with experimentally identified gene products and related correct biological activities might be discovered using the information buried in the unclassified sequences. Machine learning has been used to execute this work since human interpretation of such non-numeric, vast, and multi-dimensional data is nearly impossible. The gene products with the most research have been utilised to train ML algorithms[3]. These trained classifiers were then used to describe and categorise data that was previously unknown. Early on, the ability to use ML techniques on a broad scale was hampered by the availability of computing resources. But things are different now that computing power has expanded exponentially to back the more complicated ML algorithms, allowing for the analysis of ever-increasing volumes of data, as predicted by Moore's law. As a result, the data-driven model of research has emerged as a viable alternative to the conventional approach of hypotheses-driven research by virtue of its emphasis on the construction of models for data mining, hypothesis generation, and knowledge extension. This work continues its exploration of ML and its biological applications by focusing on the challenge of protein classification[4]. It is impractical and error-prone to undertake the repeated operations required for high-throughput data processing, hypothesis creation, and knowledge discovery by hand. However, computers excel at doing routine jobs accurately, making it crucial to create algorithms capable of automating pattern detection for massive datasets. Applications of ML, such as image classification, diagnosis, natural language processing, email spam filtering, etc., have made significant strides. Based on the nature of the training data, ML algorithms may be classified into two broad categories: I unsupervised learning and (ii) supervised learning. The purpose of unsupervised learning is to identify previously undetected groupings or patterns in data that has not been labelled. There is a higher degree of similarity between items within a group than between items in different groups[8]. Unsupervised learning, for instance, might be used to identify sub-phenotypes within a disease cohort[9]. Common unsupervised learning approaches include K-means, hierarchical clustering, and expectation-maximization.

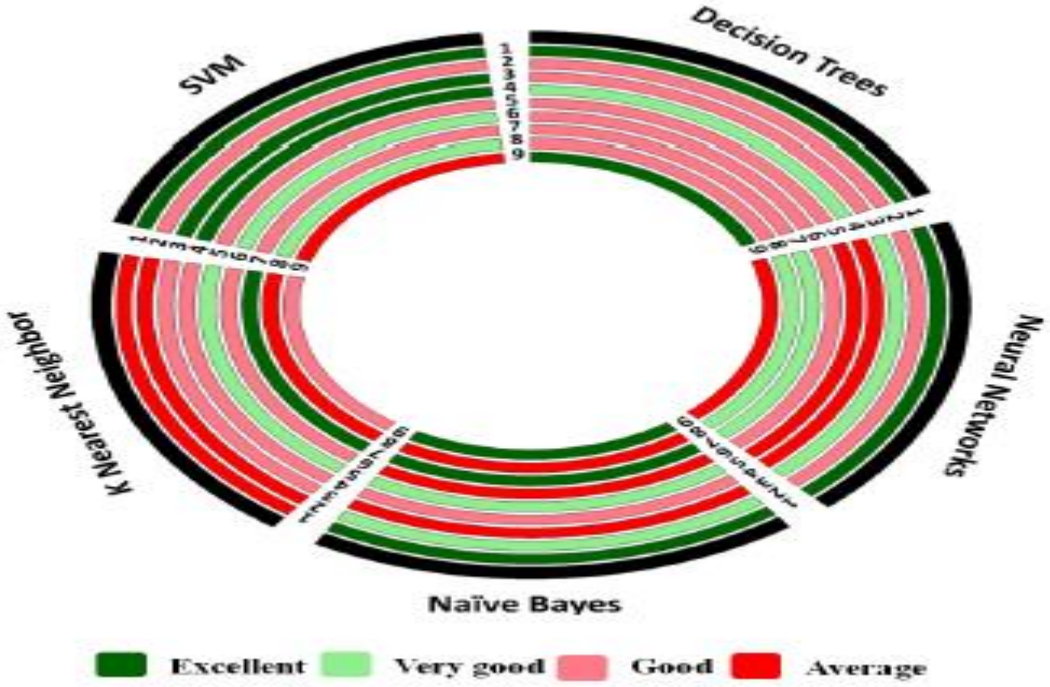

Figure.1. Comparisons of different machine learning algorithms

To classify unlabeled data, supervised machine learning approaches first optimise a decision function over a set of labelled data. Some examples of these algorithms include decision trees, neural networks, support vector machines, and random forests. You may compare and contrast the features of many well-known classifiers with the help of Figure.1. SVM is found to have higher accuracy and tolerance of duplicated characteristics than the other algorithms[5]. In contrast to the other methods, however, SVM models lack transparency. Traditional uses of ML algorithms focus on numerical data, however this might be a barrier when trying to analyse non-numerical data formats including text, photos, audio, and video. Feature vectors, which are numerical vectors of a defined length that include information retrieved from the non-numeric data, must be created from these data types[6]. Measurements of quantitative data such as glucose level, body mass index (BMI), heart rate, and blood pressure make it easy for ML algorithms to identify patients. It is not the same, however, to interpret an X-ray or a microscope picture. Consequently, it is necessary to extract quantitative aspects from such information. An expert may create a Feature Calculation Algorithm (FCA) to extract numerical features from various data formats with the use of past knowledge.

2. Literature Survey

Large volumes of biological data, often diverse and noisy, are generated by high throughput methods. Hidden patterns in complicated data may be inferred using machine learning (ML), and the unknown can be characterised using classifiers. Wet-lab tests, which need a lot of time and money to do, may be avoided by using ML to verify the data's functional annotation. Because they are based on mathematics, ML approaches like Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) can only process numerical information[7]. Biological sequences, which are not quantitative data, cannot be analysed. To quantify biological data, feature calculation algorithms (FCA) are used. These qualities should have some bearing on the biological phenomena in question if you want to use ML to analyse it. Domain specialists and computational biologists must work together to choose appropriate criteria for feature calculation. After sequence data is transformed into feature vectors, ML techniques standardise and choose features that are most informative for the purpose of developing a classifier. Test data is then utilised to assess the classifier's performance before it is put to use for prediction[8].

Many characteristics may be generated from the instances in a typical FCA application. If there is a large discrepancy between the sample size and the total number of features, it is hard to ensure that all potential cases were included in the training data. This is what is known in the jargon as the "curse of dimensionality." To get around this issue, we employ dimensionality reduction algorithms like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to build feature vectors with a less number of features than the original feature vector. Selecting suitable characteristics for the input of the ML algorithm is an alternative to generating new features that may provide the same or even better results. The inclusion of noisy features may substantially impact the learning performance of the ML algorithm yielding poor models[9]. To filter out redundant characteristics, Feature Selection Methods (FSM) have been developed based on “best first”, “greedy search” and Genetic Algorithms. Some of these algorithms are separately built from the ML algorithms while others are incorporated into the underlying model building process[10]. Annotating DNA, RNA, and proteins, as well as aiding in medical diagnostics, are just a few of the many areas where ML has been put to use. This thesis focuses on using ML to solve protein categorization challenges. Since proteins are a cell's functional entities, deciphering the information encoded in their sequence might provide light on how cells work. Proteins are a series of amino acids that are translated from genes in the genome. Most of these proteins become functional when they attain a certain 3D shape. Sequential and structural information are both utilised in the feature vector computations for ML applications[11]. The sequence information for proteins is readily accessible from sequence databases developed utilising high throughput sequencing technologies. Due to the technological challenges associated with crystallisation and

X-ray diffraction for the identification of protein structures, the requisite structural information for comparison is not readily accessible. Comparatively few proteins have been crystallised, despite the large number of proteins that have been sequenced (Figure 1.2). So the usage of sequence information more prevalent than structural information for the use of ML. Experimentally defined proteins were utilised to train new classifiers to recognise and characterise freshly sequenced proteins[12]. As a general rule, the number of proteins in a positive sample (those that have been identified) is substantially lower than the number of proteins in a negative sample (those that have not been identified). In other words, some groups of data are under-represented with regard to others. This may bias the final classifier towards highly represented classes. Such unbalanced training sets are inevitable in many classification applications and hence the negative dataset should be built carefully. Furthermore, when training data is created from diverse animals, it is feasible that homologous proteins might introduce redundancy to the data set. To construct a non-redundant data set, homologous proteins were identified by sequence similarity and deleted from the training data set by applying tools like BLASTClust, CD-HIT, etc. The protein sequences are transformed into feature vectors once the training data has been prepared and FCA have been theorised that are unique to the classification issue. Some of these characteristics take evolutionary data into consideration, while others take into account the protein's sequence pattern and physicochemical qualities. To construct evolution-based features that rely on non-redundant databases outside of the training set, position specific scoring matrices (PSSMs) were widely employed. Compositionally, transitionally, and distributively, the features estimated only from the training set differ in order to quantify latent patterns in amino acid sequences or physicochemical attributes using signal processing operations as Fourier transform or wavelet analysis[13]. Previous classifiers have made use of these traits to correctly predict protein folds, enzyme subfamilies, protein structural and functional classes, protein–protein interactions, sub-cellular locations, etc. To prevent overfitting and ensure that key aspects of biological processes are properly identified, feature selection approaches have been utilised in certain research.

In order to create new classifiers, you may have access to some of these frequently used FCA via specific software packages. After machine learning has been successfully used, online or desktop apps such as TargetMiner, ProPred, WoLF PSORT,Cell-PLoc, NRpred, etc. are created to make the technology openly available and simply usable by the research community. Updates to current models, as well as the creation of new models to account for newly available experimental data, are required. Several competitions, such as CASP (Critical Assessment of protein Structure Prediction), CAFA (Critical Assessment of protein Function Annotation), Breast Cancer Prognosis Challenge (BCC), etc., have been organised to bring together a large pool of human resources to work on pressing issues as quickly as possible. According to the aforementioned research, several different classifiers have been created to help with protein prediction and functional annotation. Our understanding of the underlying biological processes has been bolstered by these classifiers[13]. Due to experimental challenges, only a small fraction of the sequenced proteome has been described experimentally (UniProtKB/Swiss-Prot). These classifiers might be used to categorise the novel proteins[14]. Swiss-Prot, which contains humanly annotated proteins (experimental findings with scientific conclusion), and TrEMBL, which contains computationally annotated proteins, are two examples of the databases maintained by UniProt. Protein 3D structures may also be accessed via the Protein Data Bank (PDB). When compared to experimentally confirmed or crystallised proteins, the number of unclassified proteins in these databases has grown exponentially over time (Figure.2.). To properly annotate and categorise proteins, it is sometimes necessary to create new models in light of the fact that novel activities or classes of proteins are sometimes uncovered by experimental approaches.

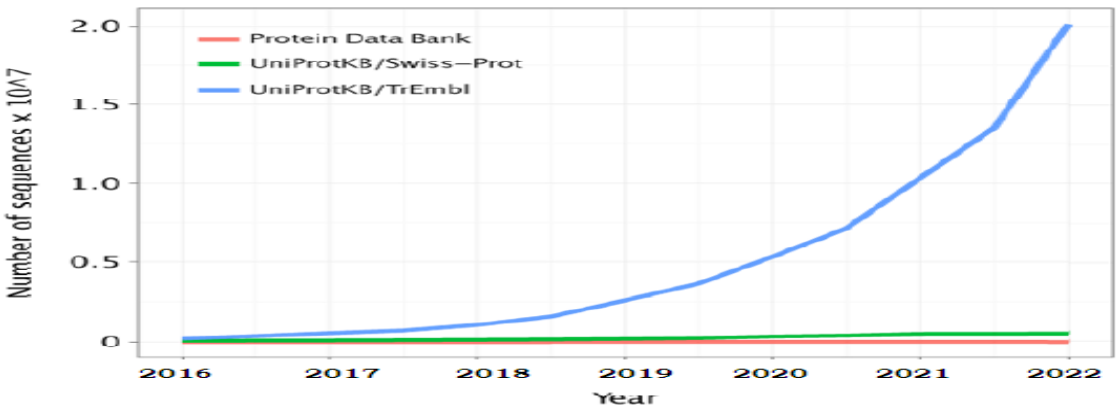

Figure .2. Growth of protein databases

When new entries are introduced to Swiss-Prot, the current models need to be updated on a regular basis. The current classifiers have limitations since they were designed to predict very particular protein activities and cannot be used to generate new classifiers from training data. The development of new models requires the iterative implementation of prerequisites such as feature extraction, feature analysis, ML model construction, and an application interface. A FCA is a time-consuming yet necessary stage in the building of an ML classifier, since it is responsible for translating biological objects into useful features vectors. After creating the feature vectors, using ML with common programmes like Weka, R, or Matlab is a breeze. The absence of such established algorithms is a technical barrier that limits the availability of ML approaches. FCA implementation, as indicated before, needs either biological expertise (background information) or programming abilities, both of which hinder the development of ML applications. Instead of starting from scratch when faced with a new biological issue, it is possible to leverage the FCA already established to create models for the new issue. A huge amount of features will be generated by using numerous FCA, which may increase noise and the likelihood of over-fitting, hence decreasing the performance of machine learning algorithms. The elimination of noisy features and the improvement of outcomes are both possible via the use of various feature selection approaches. The ability to accurately predict protein structure class and sub-mitochondrial sites has previously been shown using various combinations of feature selection and factor-based classification (FCA) methods[15]. Few programmes now exist that provide the FCA utilised in these classifiers; however, they lack an API that would have allowed for simple integration with new ML software. As a result, while solving a new categorization issue, most characteristics are ignored. From the preceding analysis, we may deduce that it could be feasible to create a "intelligent" system that can automatically assess both common and uncommon characteristics. After that, it may create a general-purpose classifier that applies sequence information to a variety of protein-classification issues. The study and identification of unique characteristics from various classes of proteins would be greatly aided by such a technology.

3. Materials and methods

One classifier was made for nuclear receptors while the other was made for fungal adhesins and adhesin-like proteins. FaaPred and NR-2L, two free, publicly accessible programmes, were used to evaluate the results. New classifiers were developed and tested using the original training and test data from these applications. Using PFMpred's training data set and GeneDB's test data set (including information on 108 MPs and 125 non-MPs), we constructed a classifier to distinguish between the two types of PF MPs. In previous research, when the sample size was much less, it was inevitable that some duplicated sequences would be left in the highly homologous training set. As part of a thorough examination of the classifier, we eliminated sequences from both the training and testing sets that were identical to each other by a factor of 100 using the programme CD-HIT. There were 40 MPs and 125 non MPs in the last non redundant training set (nrPfM165). The classifiers were put through their paces using a non-redundant test set (nrPfM205) that included 90 MPs and 115 non-MPs.

Normalization was performed on the features before proceeding with feature selection and classifier construction. The classifiers were constructed using a support vector machine (SVM), and their efficacy was evaluated using 5-fold cross-validation. The radial basis kernel function was used to construct the classifiers. Grid search (similar to grid.py in LIBSVM) has been performed to determine the optimal values for the gamma and cost parameters. Two different classifiers, SVMF-score and SVMFCBF, corresponding to the F-score and FCBF feature ranking algorithms, have been developed for each classification task. Classifiers have been constructed using LIBSVM and Weka.

4. Results and Discussion

Developed Pro-Gyan, software that implements the integrated schema (Figure.3.). To facilitate the creation and dissemination of protein classifiers, Pro-Gyan was designed. It's a free, public-domain, Java-based desktop programme. It can run on any machine that has Java Runtime Environment (JRE) version 6 or above. Only the labelled training data in FASTA format is required, and the user may upload it with a few clicks of the mouse thanks to the intuitive GUI (GUI). The application and code for Pro-Gyan may be found at http://code.google.com/p/pro-gyan/.

Figure .3.The main window of Pro-Gyan helps the user to launch two different functionalities

Each FCA is a Java object, making it simple to update the FCA repository built on top of the Spring(http://www.springsource.org/) framework using Java reflection. Weka, a free and open-source machine learning algorithm framework, makes use of the SVM implementation supplied by LIBSVM. All of the libraries and programmes used to develop this application are freely available to the public. It's a standalone programme that biologists will find simple to use and it doesn't need any specific platform to function.

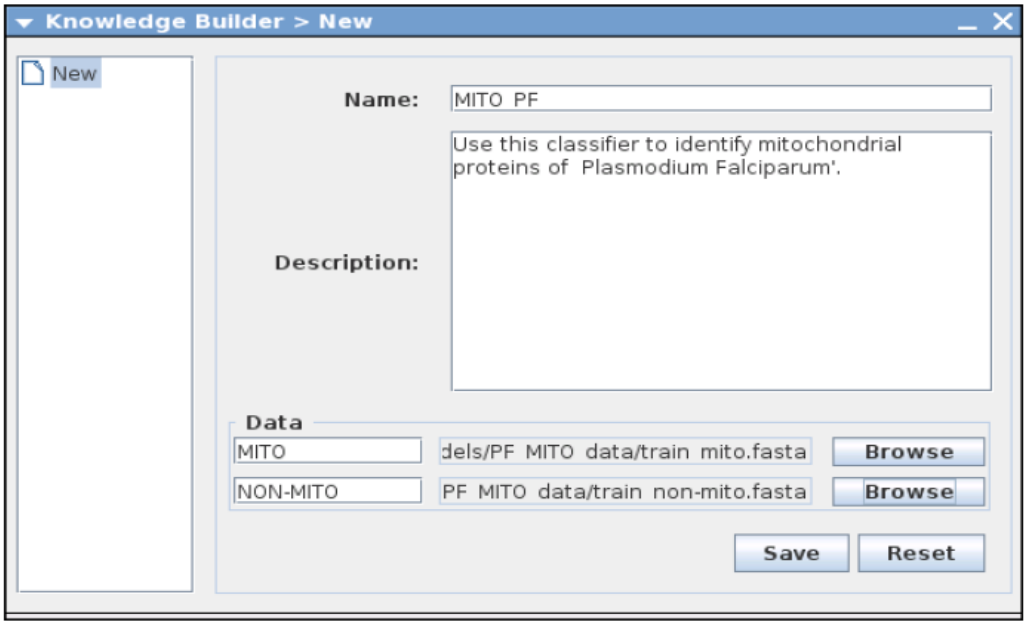

Figure .4. Input window to build a new classifier from two set of proteins in FASTA format.

Using a graphical user interface, users may choose from a variety of labelled input (Figure.4) and classifier-generation choices. Pro-Gyan also gives prediction results, complete with performance metrics and a ROC curve, so that freshly constructed classifiers may be assessed using test data. The classifiers, together with their feature normalisation data, feature selections, and SVM files, may be shared with other researchers in pgc (Pro-Gyan classifier) format. In addition, the newly constructed classifier's performance may be estimated quickly with the help of multi-threaded Pro-Gyan. A single thread should be used to construct the final classifier.

5. Conclusions

FCA library development, FSM integration, and FCA library connection to an SVM library are all covered. This effort yielded Pro-Gyan, a user-friendly and platform-independent tool for constructing protein classifiers from training data provided in the form of amino acid sequences. New experimental and high-throughput investigations constantly update biological understanding. Thus, classifiers developed with biological data should be updated often with the advent of fresh data for more trustworthy predictions and those predictions should be tested again. Due to the iterative nature of the process, knowledge discovery may be sped up by using a quick automated cyclic technique. Although there has been considerable progress in experimental technology, the implementation of ML still requires the assistance of computer specialists and programmers. In addition, there are several valuable classifiers for intriguing issues that have been difficult to get. It aids in the prediction of fresh data properties and enables users to create readily distributable new classifiers without the need for programming or ML expertise. However, it cannot verify data redundancy, thus the training set must be meticulously constructed.

References

1. Nouretdinov, I., et al., Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. Neuroimage, 2011. **56**(2): p. 809-813.
2. Monti, S., et al., Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 2003. **52**(1-2): p. 91-118.
3. Saria, S., et al. Combining Structured and Free-text Data for Automatic Coding of Patient Outcomes. in AMIA Annual Symposium Proceedings. 2010. American Medical Informatics Association.
4. Saeys, Y., I. Inza, and P. Larra+Ýaga, A review of feature selection techniques in

bioinformatics. Bioinformatics, 2007. **23**(19): p. 2507-2517.
5. Shigeo, A. Modified backward feature selection by cross validation. in ESANN 2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2005, Proceedings. 2005.
6. Gao, M. and J. Skolnick, From Nonspecific DNAâ€—Protein Encounter Complexes to the Prediction of DNAâ€—Protein Interactions. PLoS Comput Biol, 2009. **5**(3): p. e1000341.
7. Sharma, R., et al., Recognition and analysis of protein-coding genes in severe acute respiratory syndrome associated coronavirus. Bioinformatics, 2004. **20**(7): p. 1074-1080.
8. Sujihelen L, Boddu R, Murugaveni S, Arnika M, Haldorai A, Reddy PC, Feng S, Qin J. Node Replication Attack Detection in Distributed Wireless Sensor Networks. Wireless Communications and Mobile Computing. 2022 May 31;2022.
9. Reddy, P.C., Sucharitha, Y.A.D.A.L.A. and Narayana, G.S., Development of Rainfall Forecasting Model Using Machine Learning With Singular Spectrum Analysis. IIUM Journal of Engineering, 23(1), pp.172-186.
10. Reddy, P.C. and Sureshbabu, A., 2019. An adaptive model for forecasting seasonal rainfall using predictive analytics. International Journal of Intelligent Engineering and Systems, 12(5), pp.22-32.
11. Shaker Reddy, P.C. and Sureshbabu, A., 2020. An enhanced multiple linear regression model for seasonal rainfall prediction. International Journal of Sensors Wireless Communications and Control, 10(4), pp.473-483.
12. Balamurugan, D., Aravinth, S.S., Reddy, P., Rupani, A. and Manikandan, A., 2022. Multiview Objects Recognition Using Deep Learning-Based Wrap-CNN with Voting Scheme. Neural Processing Letters, pp.1-27.
13. Verma, R., G. Varshney, and G.P.S. Raghava, Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids, 2010. **39**(1): p. 101-110.
14. Li, W. and A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006. **22**(13): p. 1658-1659.
15. Gribskov, M., A.D. McLachlan, and D. Eisenberg, Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences, 1987. **84**(13): p. 4355-4358.