INTELLIGENT COMPUTING METHOD FOR PROTEIN SECONDARY STRUCTURE PREDICTION

A Thesis
Submitted in partial fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY (PhD)

in
COMPUTER APPLICATION

By
Mrs. Sarneet Kaur

Registration Number: 41300093

Under the Supervision of
Dr. Ashok Sharma
Associate Professor
School of Computer Science and Engineering

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
2021

DECLARATION

The work embodied in the thesis entitled "Intelligent Computing Method for Protein Secondary Structure Prediction" submitted to the Faculty of Technology and Sciences, Lovely Professional University, Punjab, India, for the award of the degree of Doctor of Philosophy in the subject of Computer Application has been carried out by me.

The thesis is entirely based on my original piece of work and has not been submitted fully or partially elsewhere for the award of any degree. All the ideas and references have been duly acknowledged.

Sarneet Kaur

(Research Scholar)

Reg. No: 41300093

Certificate

It is to certify that Ms. Sarneet Kaur, Ph.D. Scholar, Department of Computer Application has carried out research work entitled **“Intelligent Computing Method for Protein Secondary Structure Prediction**”, for the award of the degree of Doctor of Philosophy in Computer Application, Faculty of Technology and Sciences, Lovely Professional University, Punjab, India.

Further certify that:

i. The thesis embodies the original work of the candidate and is not copied from any other sources.

ii. The candidate has worked under my supervision for the period required under statues.

iii. The candidate has put in the required attendance in the department during the period of research.

iv. The candidate has fulfilled all the requirements of the UGC- 2018 regulations.

Dr. Ashok Sharma
Supervisor

Acknowledgments

*To the Almighty **“Akalpurakh”,** the unique creator who made all things possible,*

*To the “**Akalpurakh”** greatest gift: love,*

Moral satisfaction of fruitful conclusions of a challenging task is never complete without the acknowledgment of the efforts of all those who helped in accomplishing that task. First of all, I thank Almighty **“Akalpurakh”** for bestowing upon me the courage to face the complexities of life, for standing by me through thick and thin and for guiding me in the right direction whenever I felt torn between purpose and doubt.

It is my sincere pleasure to express my deepest admiration and reverence for my esteemed guide **Dr.Ashok Sharma** and Ex-Guide **Dr. Babita Pandey** whose expert guidance, uncompromising standards of accuracy, unceasing interest, consistent encouragement, deep understanding of the subject and critical reviews made this work a success. They were always accessible and willing to help, as a result of which my study became a rewarding journey.

My sincere thanks to Department faculty members, **Prof. Lovi Raj Gupta**, **Dean LFTS**, **Sh. Ashwani Tewari Head of School**, for their support and guidance they provided me during my work. I acknowledge the help and cooperation received from Dr Rekha Head DRP, the office, library, IT section, and non-teaching staff of the Department of Computer Application, Phagwara, and Punjab.

The chain of my gratitude would be incomplete if I don’t acknowledge the contribution of my **Husband Paramvir Singh**, **Father in Law, Mother in Law, My Parents, My Sister and Brother in Law.** Their constant inspiration and guidance kept me focused and motivated. My husband deserves all my heartfelt appreciation and thanks for all his inspiring words of encouragement and for always lending me her helping hand which enabled me to complete this assignment successfully.

Sarneet Kaur **Date: 2/28/2021**

Abstract

Proteins are the fundamental molecules of all organisms which are having three-dimensional structures. Finding the protein structure from its amino acid sequence will help in understanding the relationship between structure and function. So, by a change in structure and synthesization of new proteins, some functions will be added or removed for getting desired functions.

While predicting protein structure, the main critical problem is identifying correct templates for similar structure sequences and how to refine the native closer template structure. Secondly, how to build a model for correcting topology from scratch for sequences without having correct templates.

This research work addresses many categories of data mining methods mainly classification and clustering-based techniques. The research work put an effort in enhancing various modern-day data mining approaches along with presenting new approaches in the same domain. The hybrid-based proposed approach helps in improving predictive power in terms of classification that is the most significant research work achievement. Due to growth in predictions and technology, it has been widely used in solving several real-world problems for instance; cancer detection, tennis match predictions, weather forecasting, and soil classification. Some developed powerful prediction models to help in getting the solution for these real-world problems. Further, traditional ML techniques are compared with these models for performing the quantitative analysis of classification performances.

In this thesis, hybrid model had been developed integrating feature selection and classification technique for improvement of prediction accuracy. Many ensemble approaches have been applied for the development of hybrid models. Also, before selecting the classifiers, combination of different classifiers had been evaluated with different techniques so that the model with highest accuracy can be developed. Since it has been observed that the clustering-aided approaches can enhance or helps in improving classification rate of predictions. So, in this thesis,

clustering approaches are used with combination of classification approaches to provide more accurate predictions and hence to increase its accuracy.

Abbreviations

3D	Three Dimensional
PSP	Protein Structure Prediction
AI	Artificial Intelligence
ID	One Dimensional
ML	Machine Learning
MS	Mass Spectrometry
CASP	Critical Assessment Of Protein Structure Prediction
FSH	Follicle-Stimulating Hormone
ACTH	Adrenocorticotropic Hormone
SVM	Support Vector Machine
NN	Neural Network
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
CNN	Convolutional Neural Network
BILSTM	Bidirectional Long-Short Term Memory
MLP	Multi-Layer Perceptron
RF	Random Forest
DT	Decision Trees
KNN	K-Nearest Neighbour
NB	Naive Bayes
CRF	Conditional Random Fields
DCNN	Deep Convolutional Neural Network
SMRNN	Segmented Memory Recurrent Neural Network
BSMRNN	Bidirectional Segmented-Memory Recurrent Neural Network
RESNET	Residual Neural Network
RBF	Radial Basis Function
MLP	Multi-Layer Perceptron
AUC	Accuracy
TP	True Positive
TN	True Negative

FN	False Negative
FP	False Positive
TPR	True Positive Rate
FPR	False Positive Rate
NPV	Negative Predictive Value
ROC	Receiver Operator Characteristic
GSA	Gravitational Search Algorithm
PSS	Protein Secondary Structure
FA	Firefly Optimization Algorithm
BILSTM	Bidirectional Long-Short Memory

TABLE OF CONTENT

Declaration	ii
Certificate	iii
Acknowledgments	iv
Abstract	v-vi
Abbreviations	vii-viii
List of Figures	xii-xvi
List of Tables	xvii-xix
Chapter- 1 Introduction	1-21
1.1 Overview	1-2
1.2 Introduction	3-9
1.3 Applications	9
1.4 Importance of Proteins	10
1.5 Biological Functions of Proteins	10-11
1.6 Different types of learning	11
1.8 Research Assumptions	15
1.9 Research gap, Hypothesis and Objectives of the research	15
1.10 Novelty	17
1.11Thesis Contribution	18
1.12 Organization of thesis	19-21
Chapter- 2 Literature Review	23-42

2.1 Overview	23-24
2.2 Year-wise Literature Survey	24-34
2.3 Protein Beta-turn Prediction	35
2.4 Comparison of Binding Patterns	35
2.5 Deep Learning	36
2.6 Convolutional Neural Network	36-37
2.7 ReLU, Batch Normalization, and Fully Connected Layer	37-38
2.8 Hybrid Methods	38-40
2.9 Base of hybridization	40-42
2.14 Summary	42
Chapter- 3 Protein secondary structure prediction using Multi-classifiers	46-55
3.1 Multi-classifiers comparison for protein secondary Structure prediction	46-51
3.2 Protein Secondary Structure prediction using Feed Forward artificial neural network and perceptron	52-55
Chapter-4 Proposed-RF Approach for enhancing protein structure prediction	56-76
4.1 Overview	56
4.2 Introduction	56-58
4.3 Proposed approach for protein structure prediction	58
4.4 Methodology	58
4.5 Proposed Algorithm	59-64
4.6 Benefit of hybridizing PSO and GSA	64
4.7 Performance Evaluation	64-65
4.8 Results and Discussions	65-72
4.9 Performance Evaluation for protein structural class D	73-75

4.10 Parameters used for Hybrid PSO-GSA Algorithm	75-76
4.11 Summary	76
Chapter-5 Proposed Hybrid CNN+BiLSTM approach	77-113
5.1 Overview	77
5.2 Introduction	77-78
5.3 Native Conformations of Proteins	78-80
5.4 Methodology	80-85
5.5 CNN Classifier	85-87
5.6 BILSTM	87
5.7 CNN+ BILSTM	88
5.8 Benefit of hybridization	88
5.9 Performance Evaluation	89
5.10 Results and Discussions	89-94
5.11 Comparison of Experiment I and II	95
5.12 Results of CNN+BILSTM on FC699 dataset	107
5.13 Summary	113
Chapter-6 Conclusion and Future Scope	114-120
6.1 Overview	114
6.2 Conclusion	115
6.3 Future scope of the study	119-120
References	121-151
Publications	152

List of Figures

1.1	Variants of data science	2
1.2	Shows an example of an amino acid molecule	4
1.3	Primary, secondary, tertiary and quaternary structures	5
1.4	Alpha helix	7
1.5	Beta sheet	7
1.6	Functions performed by proteins in human body	10
1.7	Types of Learning	11
1.8	Different Categories of Supervised Learning in Data Mining.	12
1.9	Different Categories of Unsupervised Learning in Data Mining	14
3.1	Structural Assignments of Various Datasets	46-47
3.2	ROC Curve for 3 Classes	47
3.3	Tree Grown Graph for a) Coil b) Sheet c) Helix	47-48
3.4	Functional Evaluation for Coil	48
3.5	Objective Function Model for coil	48
3.6	Functional Evaluation for sheet	49
3.7	Objective Function Model for sheet	49
3.8	Functional Evaluation for Helix	49
3.9	Objective Function Model for Helix	49
3.10	Accuracy Graph for Coil Class	50
3.11	Accuracy Graph for Sheet Class	50
3.12	Accuracy Graph for Helix Class	51
3.13	Molecular View of Protein Secondary Structure	51
4.1	Flowchart of the Proposed Technique	60
4.2	k –Means Clustering Algorithm	61
4.3	Variation of fitness function with no. of iterations for Class A	65
4.4	Performance of different parameters	66

4.5	Accuracy of different methods	66
4.6	Variation of fitness function with no. of iterations for Class-B	68
4.7	Performance of different parameters	68
4.8	Accuracy of different methods	69
4.9	Variation of fitness function with no. of iterations for Class-C	70
4.10	Performance of different parameters	71
4.11	Accuracy of different methods	71
4.12	Variation of fitness function with no. of iterations for Class-D	73
4.13	Performance of different parameters	73
4.14	Accuracy of different methods	74
4.15	Combined model to measure different parameters for all classes	75
5.1	Visual representation of protein folding into a compact conformation	80
5.2	Flowchart for present study	82
5.3	The architecture of Convolutional neural network (CNN) for fold classification	86
5.4	Detailed Schematic for CNN technique	87
5.5	Detailed Schematic for BiLSTM technique	87
5.6	Block-Diagram for CNN+BiLSTM technique	88

5.7(a)	Scoring Space and Winning path for sequence 1 and 2 for protein structural class A	90
5.7 (b)	Scoring Space and Winning path for sequence 1 and 2 for protein structural class B	90
5.7(c)	Scoring Space and Winning path for sequence 1 and 2 for protein structural class C	91
5.7(d)	Scoring Space and Winning path for sequence 1 and 2 for protein structural class D	91
5.8(a)	Fitness value outcome for protein structural class A	92
5.8(b)	Fitness value outcome for protein structural class B	92
5.8(c)	Fitness value outcome for protein structural class C	93
5.8(d)	Fitness value outcome for protein structural class D	93
5.9	Parameter value for proposed CNN+BiLSTM	94
5.10(a)	Accuracy comparison of both objectives for class A	95
5.10(b)	Accuracy comparison of both objectives for class B	96
5.10(c)	Accuracy comparison of both objectives for class C	96
5.10(d)	Accuracy comparison of both objectives for class D	97
5.11	Accuracy analysis for objective 1 and objective 2	97

5.12	Comparison of different parameters for various protein classes	100
5.13	Comparison of Accuracy with SVM and CNN models for various protein classes	101
5.14	Comparison of Sensitivity with SVM and CNN models for various protein classes	101
5.15	Comparison of Specificity with SVM and CNN models for various protein classes	102
5.16	Comparison of MCC with SVM and CNN models for various protein classes	102
5.17	Accuracy comparison of different methods	103
5.18	Comparison of Accuracy with SVM and CNN+BILSTM models for various protein classes	103
5.19	Comparison of Sensitivity with SVM and CNN+BILSTM models for various protein classes	104
5.20	Comparison of Specificity with SVM and CNN+BILSTM models for various protein classes	104
5.21	Comparison of MCC with SVM and CNN+BILSTM models for various protein classes	105
5.22	Overall Q3 accuracy of different methods	106
5.23	Comparison of different parameters for various protein classes	107
5.24	Comparison of different parameters for protein class A	107
5.25	Comparison of different parameters for protein class B	108
5.26	Comparison of different parameters for protein class C	108
5.27	Comparison of different parameters for protein class D	109
5.28	Accuracy Comparison of different methods on FC699 dataset	109
6.1	Performance of different parameters	116

6.2	Accuracy of different methods	116
6.3	Combined model to measure different parameters for all classes	117
6.4	Comparison of different performance parameters for various protein classes	118

List of Tables

1.1	Identified Research Gaps	16
1.2	Proposed Objectives	17
2.1	Year-wise literature survey	24-34
2.2	Comparative table of hybrid methods used in protein secondary structure prediction	38-40
2.3	Comparative table on accuracy, advantage, disadvantage and application of various approaches used in protein secondary structure prediction	41-42
3.1	Accuracy Table	49
3.2	Assigning Integer value to each sequence	53
3.3	Training data using 3 window sizes	53
3.4	Secondary sub-structure and their integer value	54
3.5	Normalized training data	54
3.6	Some feed forward neural network, Perceptron Neural Network parameters and their value	55
3.7	Performance comparison of two neural networks	55
4.1	Accuracy comparison of proposed RF method with other methods	67
4.2	Performance value of different parameters for proposed RF method	67
4.3	Accuracy comparison of proposed RF method with other methods	69
4.4	Performance value of different parameters for proposed RF method	70

4.5	Accuracy comparison of proposed RF method with other methods	72
4.6	Performance value of different parameters for proposed RF method	72
4.7	Accuracy comparison of proposed RF method with other methods	74
4.8	Performance value of different parameters for proposed RF method	74
4.9	Parameters of Hybrid PSO-GSA algorithm for clustering of protein structure	76
5.1	Hypermeters for Model Tuning	81
5.2	Comparative accuracy results of various classes	98
5.3	Comparison of sensitivity for SVM and CNN methods for various protein classes	99
5.4	Comparison of specificity for SVM and CNN methods for various protein classes	99
5.5	Comparison of MCC for SVM and CNN methods for various protein classes	100
5.6	Comparison of Accuracy for CNN+BILSTM method for various protein classes in two datasets	110
5.7	Comparison of Sensitivity for CNN+BILSTM method for various protein classes in two datasets	110
5.8	Comparison of Specificity for CNN+BILSTM method for various protein classes in two datasets	111

5.9	Comparison of MCC for CNN+BILSTM method for various protein classes in two datasets	111
5.10	Comparison of Q3 for CNN+BILSTM method for various protein classes in two datasets	112
5.11	Comparison of SOV for CNN+BILSTM method for various protein classes in two datasets	112

Chapter-1

Introduction

1.1 Overview

Proteins are the fundamental molecules of all organisms which are having three-dimensional structures. The three-dimensional or 3D protein structure determines the functional properties of the protein. There are various dissimilar biological functions in proteins that can act as enzymes of building blocks, as muscle fibers, or transport of oxygen-like transport function. Finding the protein structure from its amino acid sequence will help in understanding the relationship between structure and function. So, by the change in structure and synthetisation of new proteins, some functions may be added or removed for getting desired functions. For instance; by prediction of viral protein structures will allow researchers in proposing drugs for particular viruses.

Protein structure prediction (PSP) is a difficult problem that gains the attention of various researchers. Due to the significance of drug design, various researchers have started researching biotechnology and diagnostic methodologies. An important increasingly role is played by proteins secondary structure prediction for predicting tertiary structure and its function, different intelligence techniques have been used to solve this problem. Lots of techniques and tools have been proposed by various researchers for the prediction of protein structure but still not able to obtain good results in application [1.2].

Protein folding is the process by which a proteins can be folded into specific three dimensional structure. If it is not correctly folded then it will lead to many diseases. The main challenge in solving the protein folding problem is the prediction of folding routes and progression for each native structure. If the author will be able to solve the problem then they will come very close to the identification of protein structure. Then various challenges are outlined by David Searls in [1.3] for finding the protein structure:

> The physical root of protein structural constancy is not completely understood.
>
> Search space of the problem is excessively massive, due to the vast range of possible conformations of even relatively short polypeptides.

The primary sequence might not entirely state the tertiary structure.

There are plenty of techniques available for artificial intelligence, that efficiently provides good accuracy. Some empirical statistical methods are used by researchers before the introduction of AI techniques that are used to be very complex to use and have low accuracy. But now focus of researchers is moving to AI-based techniques for the prediction of protein structures that provide better accuracy. There are mainly two types of AI techniques that are used by researchers namely support vector machine and neural network. Both methods are very popular and powerful in AI field prediction tasks.

Generally, Data Science deals with the extraction of useful information from huge datasets by examining the data from different perspectives and converting the raw datasets into a knowledgebase that can be used to reduce losses, data maintenance costs, or both. For data extractions, we have many tools available depending upon the requirement for analysing data. In fact tools in this field have made the researcher's job very easy by allowing them to process datasets from different dimensions or angles, categorize the dataset, and providing consolidated patterns from the datasets leading towards a generalized knowledgebase [1.5].

We can also say that Knowledge extraction is a process of looking into correlations or relations among different fields in large datasets. Although data extraction is a relatively new term, most business-oriented organizations have used high-end systems to process a large volume of datasets and analyze market research reports for years. However, continuous innovations in IT have brought faster processors, disk storage devices, and statistical software which have dramatically increased the accuracy of analysis while driving down the cost. Figure 1.1 shows that Data Science is an interdisciplinary area and it is used in almost all areas of research.

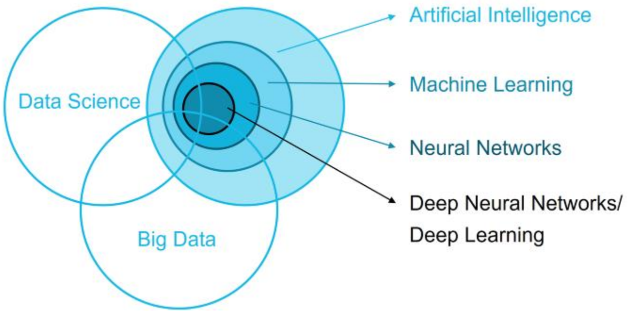


Figure 1.1 Variants of Data Science [1.43]

1.2 Introduction

In the early 19th century, the importance of proteins was recognized by chemists that include Swedish chemist Jons, Jacob Berzelius. The protein name came from proteos Greek word that means first place or primary. Proteins are consist of amino acids that are joined together to make long chains and mainly twenty amino acids are combined in different counts and sequences for the building block of the human body. This is the reason they are critically important for human lives. Proteins can also be used for making medicines and are useful in many other situations. For molecular design and biological medicine design knowing protein function is very important. In structural biology, there is a strong relationship between the functioning of proteins and their three-dimensional structure [1.6]. Because determining protein structure with experimental methods is costly, so 3-D structure prediction from the sequences of amino acid give an effective alternative and is one of the necessary motive in bioinformatics and theoretical chemistry [1.7]. Methods to predict the three-dimensional structure of proteins are divided into two main categories such as; template-based modelling and free modelling. The specified target proteins amino acid sequence detected similarity protein comes under template-based modelling. Then, three-dimensional structure prediction is computed using that protein as a template. If such a protein cannot be specified, a three-dimensional structure is predicted with free modelling. According to the thermodynamic hypothesis used by comparative and free modelling, proteins are folded to have minimum free energy in a physiological environment.

There are some one-dimensional (1D) structural characteristics such as solvent accessibility, profile matrix, secondary structure, and torsion angles that are used as features for predicting the 3D structure of the protein. Deducing these 1D characteristics with minimum error is necessary for the prediction of 3D structure. Till now, there are several machine learning algorithms developed for the prediction of 1D properties of protein and these ML algorithms dataset has critical importance in classifier performance and accuracy [1.8]. Having an abundance of features may lead to an increase in training time and a result in overfitting that reduces the accuracy of unseen data. Additionally, it can alter the training due to having more noisy features. On the other side, for having satisfactory training few features will not work well which is known as underfitting. So, proper and sufficient numbers of features need to be employed in Machine Learning models and for solving these problems feature selection and prediction methods like dimensionality reduction techniques need to be used [1.9]. The main difference between these two techniques is that in feature selection a subset of features are selected and used without any change, but in projection methods, the size of the dataset is reduced by using all features with the least information loss.

1.2.1 Protein Structure

Proteins form a major class of macromolecules found in every organism composed of consecutive attachments of amino acids by peptide bonds. There are twenty different amino acid types commonly found in nature.

Amino acid composition of proteins

The most common properties of all proteins are that they have long chains of alpha-amino or α-amino acids and their general structure is shown below figure. These acids are organic compounds that contain side-chain molecule (R), amino group (-NH_2), carbon atom (Ca), and carboxyl group (COOH). In amino acids there are α-carbon atoms in the molecules that carry carboxyl group (-COOH) and amino group (-NH_2) [1.9]. The amino acids are produced at the ribosomes and have different physical and chemical properties such as the electrostatic charge they carry, the hydrophobic states, acid dissociation constants (pKa), molecular size, and the functional group. These characteristics play an important role in determining the structure of proteins [1.10].

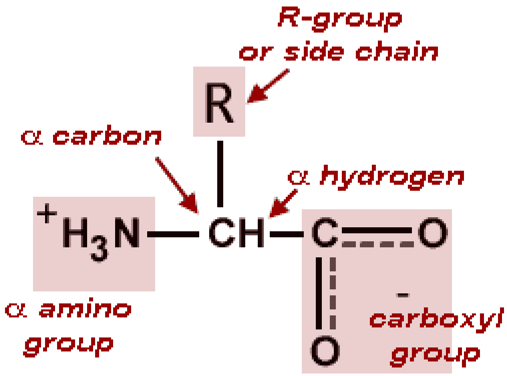


Figure 1.2 shows an example of an amino acid [1.42]

1.2.2 Protein structure Levels

There are mainly four levels of protein structure named primary, secondary, tertiary, and quaternary structure. The amino acid sequence is the primary structure, regular hydrogen bond patterns are the representation of the secondary structure and the 3D structure of a single amino acid chain is the tertiary structure and 3D protein structure is the quaternary structure, which might contain more than one amino acid chain [1.11].

Four different structural features are used to describe the structure of protein:

Amino acids sequence is denoted by the primary structure of proteins that make the chain of the polypeptide.

The polypeptide chain small sections form into regular shapes is described by secondary structures. The α-helix and β-strand or β-sheets are two main types of secondary structures where α-helix are like coiled spring and β-sheets are like pleat or concertina [1.12]

The overall shape of an individual molecule of protein makes the tertiary structure and polypeptide chain turns into a compact globular structure.

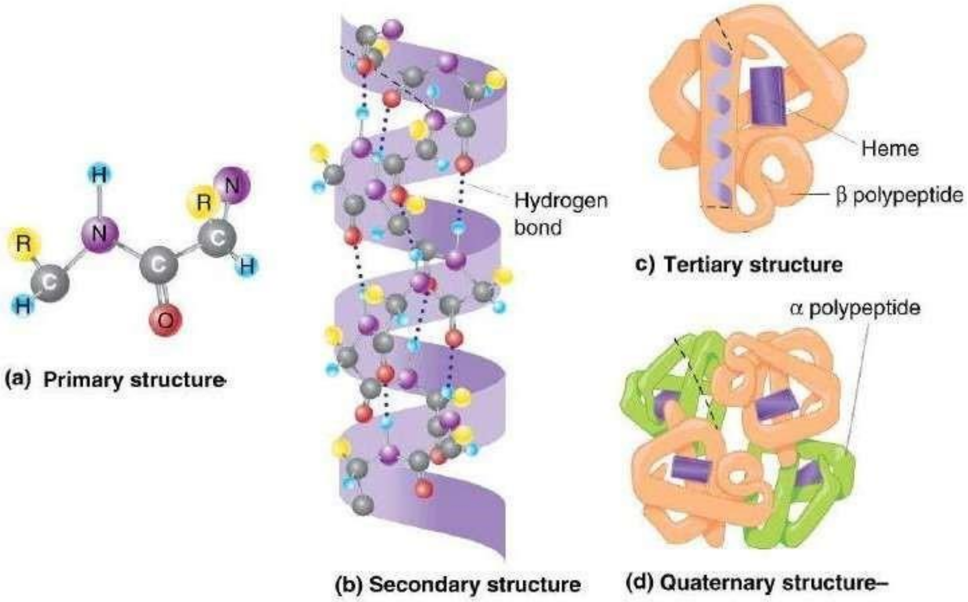


Figure 1.3 Primary, secondary, tertiary, and quaternary structures. Src: Figure after © 2010 PJ Russell, iGenetics 3rd ed.; all text material © 2014 by Steven M. Carr

When a protein subunit is formed by several molecules of protein then a quaternary structure has formed that acts as a single complex protein.

Four levels of protein structure are:

Primary structure

The primary structure is the amino acid sequence of a polypeptide chain. It stays together with peptide bonds that occur during protein synthesis. The primary structure of a protein is decided in vivo by the gene that encodes its amino acid content. The amino acid sequence serves as a signature for the protein dictating its structure and function. While this sequence can be determined by methods such as mass spectrometry (MS) or Edman degradation, typically it is identified by directly reading the sequence from the encoding gene [1.13].

Secondary structure

Secondary structure in proteins is formed by regular hydrogen bonds between neighbouring amino acids with similar dihedral angles. The pattern of a hydrogen bond is formed by two basic ideas namely bridge and rotation motif. Rotation motif is also known as an n-rotation motif in which hydrogen bond between amino acid at position i and i+ n and n take 3, 4, or 5 values. On other hand, there is a hydrogen bond between amino acids in the case of the bridge motif and these bonds are not closely related to each other in sequential order. Subsequent secondary structural elements are formed when the rotation and bridge motifs are successively brought to a certain layout. For example, the repeating 4-rotation motif forms the alpha helix, and the repeating bridge motif forms beta strands and beta sheets [1.14]. The three-dimensional structure of proteins can be thought of as the successive organization of secondary structural elements.

Helix

In this structure, the protein backbone adopts a helical structure (Figure 1.3). Alpha, $3_{10,}$ and pi are three types of the helix that can have numerous functional roles. These may consist of DNA-connected motifs and structures that pass through the membrane cell [1.15].

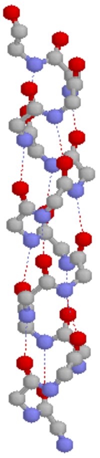


Figure 1.4 Alpha helix [1.44]

Beta Strands and Beta Sheets

Beta strands are the second most common regular units that stabilize the structure of proteins (Figure 1.3). A beta-strand consists of a polypeptide chain that has 3 to 10 amino acids. In beta-strands, the polypeptide typically has an extended conformation. Beta strands are aligned pairwise and consecutively in three-dimensional space interacting through hydrogen bonds. As a result of this interaction, beta-sheet units are formed, which contain at least two beta-strands. The interacting amino acid segments may be close to each other and linked by a short loop, or they may be separated by many other structures. Even though interacting beta strands are sequentially far from each other, they can come closer in the three-dimensional space as a result of the folding process. Protein aggregates and fibrils formed through the combination of beta strands play a role in the formation of various diseases like Alzheimer's [1.17].

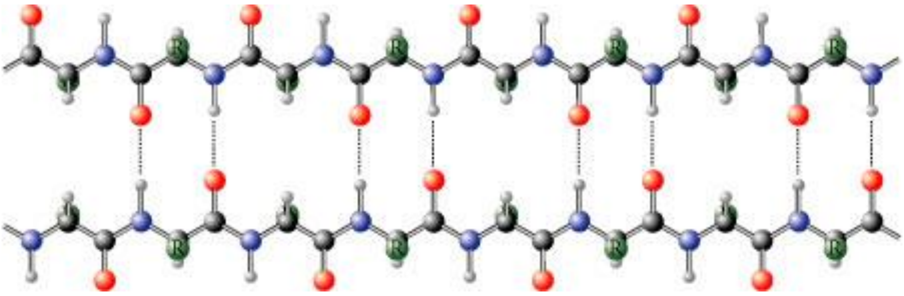


Figure 1.5 Beta sheet. Src: Chiral publishing company. Copyright 2013 Mark Bishop

Loop

Loops are structures usually located at the surface of the protein. They typically occur between helix and beta sheets with different lengths and configurations. Unlike the amino acids in the internal region of proteins, amino acids in loops are not exposed to spatial and environmental constraints. They also do not play an effective role in the regulation of secondary structural elements in the inner zone. That’s why there may be more mutations in the loops. Regions that have undergone this type of mutation in a series of alignments may indicate the loop structure. Loops are more inclined to contain cyclic charged and polarized amino acids that are mainly found in active regions [1.19]. Curl. Random and stitch coil are three types of loops.

Tertiary Structure

It is a 3D structure of a single protein molecule that is defined as the coordinates of atoms in 3D space. This compact structure is formed by folding sheets and strands that are guided by hydrophobic interactions. However, to stabilize the overall structure certain regions of a protein may be fixed with specific tertiary interactions [1.20].

Quaternary Structure

It is the agglomeration formed by some chain of polypeptides or proteins. The disulfide and non-covalent bonds stabilized the tertiary structure and most of the proteins do not have quaternary structures that make them function as a monomer. For instance, a hemoglobin protein is composed of four chains and carries oxygen in the blood [1.21].

Prediction

Its protein tertiary structure is made from only the sequence of its amino that is a very challenging protein but it becomes more tractable by using the definitions of simpler secondary structure. Almost all the secondary structure prediction methods are restricted to predict the random, sheet, or helix coil three predominant states. These methods are based on sheet or helix forming propensities of individual amino acids that are coupled with estimating rules of free energy formed secondary structure elements. The first most widely used approach for predicting protein secondary structure from the amino acid sequence was GOR and Chou–Fasman method [1.22]. These methods were able to achieve an accuracy of 60% in predicting helix, coil, or sheet states. This use of blind computed assessments shows that there was very low actual accuracy [1.23].

By exploitation of various alignment of sequence, there is significant increase has been seen accuracy that turns 89%. Some of these are the complete distribution of amino acids at a position with a vicinity of 7 residues on either side along with the evolution of throughput that gives a better picture of structural tendencies near to position [1.24] [1.25].

The Critical Assessment of protein Structure Prediction (CASP) experiments were used for evaluating secondary structure prediction methods along with continuous benchmarked for instance; the EVA benchmark. Based on these tests, the most accurate methods were described namely PROF, SAM, SABLE, and PORTER [1.26, 1.27, 1.28, and 1.29]. The main areas of improvement that started appearing in the β-strands prediction reside confidently in the β-strand prediction but the methods are used for overlooking some false negatives segments of β-strands. There are cases of having overall prediction accuracy of nearly 90% due to idiosyncrasies of standard method DSSP that assign the secondary structure, three classes, to PDB structures then marked the benchmarked predictions [1.30].

In the prediction of tertiary structure, a key element is the accurate prediction of secondary structure, and in this homology modelling is the simplest one. For instance; βαββαβ are six secondary structure elements is the signature of a ferredoxin fold [1.31]

1.3 Applications

Protein secondary structure prediction helps in drug design and also for predicting teriary structure prediction. Structure-based drug design is to predict the binding modes as well as affinities of various ccompounds.

In alignment with multiple sequences, both nucleic and protein acid secondary structures are used. By insertion of secondary structure information, these alignments can be made more accurate along with simple sequence information. In RNA this information is become less useful due to base more highly conserved base-pairing compared to the sequence. Reserved relations between proteins with primary unalienable primary structures can sometimes found using secondary structures. From the study, it has been found that the use of α-helices gives a more stable or robust outcome to mutations and designable as compared to βstrands in natural proteins [1.32] thus all α proteins designing function is easier compared to designing the proteins with both strands and helices [1.33].

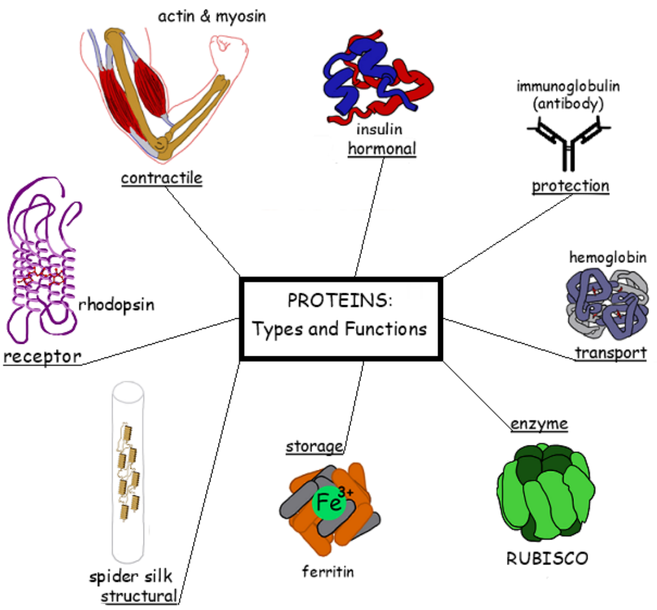


Figure 1.6 Functions performed by proteins in the human body. Src.: © A-Level Biology 2015-2021

1.4 Importance of proteins

Proteins are specific to species, which means there is a difference in species from one species to that from another species. These are specific to organs, for instance; within a single organism, muscle proteins are different from those of the liver and brain. The importance of proteins is directly linked with their amount in tissue or organism, on the other side, there is a very less amount of hormones and enzymes, which are the most important type of proteins. The importance of proteins is directly linked principally with their function. Most of the work in the cell is done by proteins and performs different jobs in the human being body as shown in the figure.

Protein is responsible for various functionalities in human beings that vary according to the bond of protein amino acids. Amino acids are the raw elements of protein.

1.5 Biological Functions of Proteins

Proteins are also known as a polymer that is macro-molecules and made up of subunits of amino acids namely monomers. These amino acids are attached for forming long linear chains known as polypeptides that are further folded into a specific shape in 3D. Most of the time these folded polypeptide chains function by themselves. Other times they combine it with additional chains of the polypeptide for forming the final structure of the protein. For example, four chains of polypeptide made a blood protein hemoglobin and each of these chains contains a heme molecule that is in a ring structure with an iron atom in the center.

Methods of ML are widely used in computational, systems biology, and bioinformatics. The prediction of protein structure is a complex problem that is often attacked and decomposed into four levels namely 1D prediction of structural features and amino acids primary sequence then the 2D prediction of amino acids spatial relationships, 3D prediction of protein tertiary structure, and 4D prediction of complex multiprotein quaternary structure.

1.6 Different Types of Learning in Data Mining

Assorted sets of supervised and unsupervised ML approaches are applied for years for tackling these problems and significantly prove to be better than a traditional prediction of protein structure.

Data Mining is usually portioned into two categories, Supervised and Unsupervised Learning while a few of the literature also discusses Semi-Supervised Learning. But basically, Semi-Supervised learning is a combination of both Supervised and Semi-Supervised Learning. Fig 1.7 depicts different types of learning in Machine learning and Data Mining.

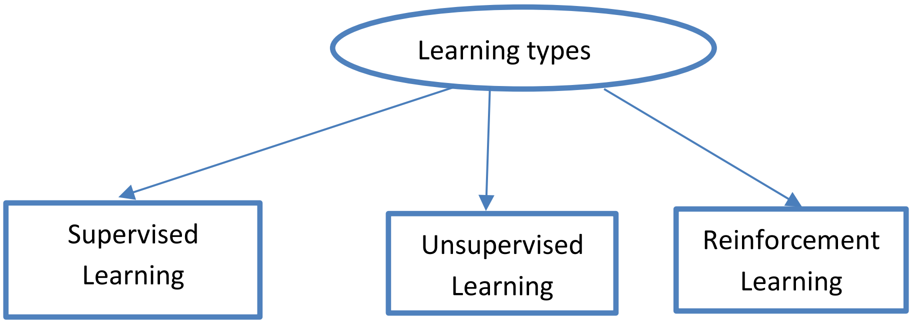


Figure 1.7 Types of Learning

1.6.1 Supervised Learning

Supervised Learning is a method of discovering a new pattern or relationship among input attributes (Independent Variable) and a target attribute (dependent variable). In this type of learning the supervision in the learning comes from the labelled examples in the training data set which supervise the learning of the classification model which later on are used for predicting the outcomes of unlabelled tuples.

The findings from the datasets are represented in a referred structure as a model that describes the information hidden in the dataset and uses the model's outcome for predicting the target attributes values that are input variables output. There are numerous applications in which supervised methods are used such as manufacturing, health, marketing, education, and finance. Different types of Supervised Learning categories are depicted in fig 1.8.

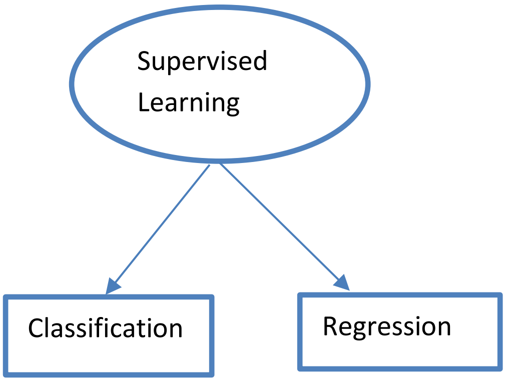


Figure 1.8 Different Categories of Supervised Learning in Data Mining

1.6.1.1 Classification

Classification is a process of supervised learning where the model is usually trained with training sets containing the class label, to predict the outcome of the test dataset. Classification is a widely applicable and demanding technique that has applications in most of the real-world issues in society related to predictions, outcomes of some activity, etc. For prediction whether the day will be sunny, rainy, or cloudy the weather forecasting department might make use of classification. Also, the doctors might evaluate the health conditions of a patient, which can be mapped and generalized with the use of classification to predict medical outcomes.

Classification performs the assignment of observations from the dataset into discrete categories instead of estimating the continuous quantities. Although most of the problems are binary several important queries can also be modelled in terms of multi-binary classification. Mostly used classification algorithms are KNN, Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF).

Firstly, some pre-processing tasks classification procedure is applied then pre-processed dataset is divided by methods into testing and training dataset two different part. These datasets need to be independent of each other for avoiding biases. The classifier is the other name given to classification and it contains two different steps namely development of a classification model that indicates a well-defined set of classes is the first type. This is also known as the training phase in which classification approaches construct the model by learning from available data set of training conveyed by their related attributes of class labels. This is the reason it is the form of supervised learning technique and after that, the classification model becomes suitable for predicting that is known as the testing phase. This step is used for finding the accuracy of the derived model by using a test dataset for checking the accuracy and classification approach used in the predictive data mining task.

On larges repositories of information, a classification procedure is applied for building identification models of diverse data classes. This analysis helps in getting profound insights into an improved understanding of large-scale datasets. The model achieved from it is based on a training analysis dataset and it can be used for various procedures like simple if-else rules, decision trees, mathematical formulas, and artificial neural networks. The classification technique-related software applications analyzed a large amount of dataset and developed meaningful patterns and classifications in the scientific researcher, industrial and commercial purposes dataset. There are some criteria given below that is used for getting the classification model performance:

Accuracy: Accuracy of classification model tell about its capacity of predicting the label class of previous or new unknown data.

Speed: This is linked with the required costs of computation for developing and using a given classifier or classification model.

Robustness: This is linked with the classification model's capability of making accurate predictions in the data presented that is noisy or has missing values.

Scalability: This tells the capacity of building a proficient classification model for a large amount of data.

Interpretability: This tells about the comprehensive and vision level offered by a given model of classification.

1.6.1.2 Regression

Regression analysis is a procedure of assessing the inter-variable relations in ML and statistics [1.40]. It consists of various approaches for examining and demonstrating various variables in which the main motive is to analyze the association between dependent and independent variables. Mainly regression analysis recognizes the independent variables that are closely linked with the dependent variable and discovers the forms of relationships linked with it.

1.6.2 Unsupervised Learning

Unsupervised learning is a method that enables an algorithm to learn from test data that has not been labelled, classified, or categorized. This is a necessary synonym for association and clustering and it is an unsupervised learning process as input examples are not labelled classes. Naturally, clustering is used for discovering classes within the data. Varied things can be classified by machines such as behavioural patterns, purchasing habits of a customer, hacker attacks, etc. The main concept behind unsupervised learning is to first train the machines to voluminous and varied data and then let the machine learn and extract information from data, depicted in fig 1.10.

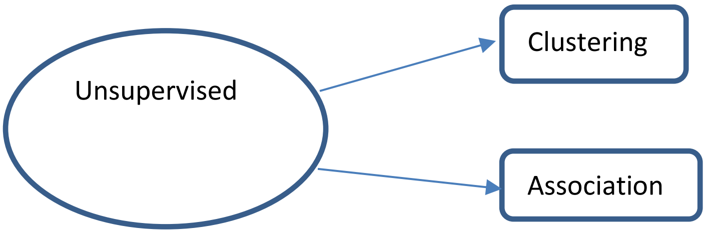


Figure 1.9 Different Categories of Unsupervised Learning in Data Mining

1.6.2.1 Clustering

Like clustering and classification examined the data objects without referring to a known label class and class labels are not present in the training dataset due to being unknown initially [1.10]. Research scholars apply clustering analysis for creating such labels that make it an unsupervised learning technique and is a classification pre-processing step. According to the maximizing rule, the intraclass and inter-class similarity is maximized and minimized respectively.

1.6.3 Semi-Supervised Learning

In this type of learning, models learn by making use of labelled as well as unlabelled training examples with labelled examples are used to train class models while unlabelled sets are used to improve the boundaries between classes. While working on the two-class problem the examples which belong to one particular class is considered a positive example while the other belonging to the other class is taken as negative examples.

Semi-supervised learning is done to find out and understand how learning behaviour can be changed by combining labelled and unlabelled data and design algorithms which takes take advantage of such an arrangement. Semi-supervised learning is of the incredible enthusiasm for AI and information mining because it can utilize readily accessible unlabelled information to improve managed learning assignments when labelled data are inadequate or expensive. Semi-supervised adapting likewise indicates potential as a quantitative apparatus to comprehend human class realizing, where the majority of the info is self-obviously unlabelled. The achievement of Semi-supervised learning depends basically on some basic presumptions.

1.6.4 Reinforcement learning

In Reinforcement learning decisions are taken sequentially by interacting with the environment. In artificial intelligence its work better as human interaction is favored.

1.7 Research Assumptions

A large number of sequences are added to the database on regular basis. But still, they are either not predicted correctly or very tough tasks to predict.

Accurate prediction of secondary structure and its substructure requires proper pre-processing before its classification is done.

Assumptions of our research is large number of sequences are added to the database regurally.So the need to develop a accurate prediction model. Secondly, classification requires proper pre-processing.

1.8 Research gap, Hypothesis and objectivies

A combination of various models and hybrid techniques as used by various researchers along with their observations and findings have been critically studied and described in the preceding sections. After carrying out the literature survey it has been observed that researchers have made extensive use of the techniques in protein structure prediction and there have been very fruitful results. However, there are many areas /techniques which are yet to be explored and require further investigation. The few Gaps identified by the investigators are listed in Table 1.1 given below:

Table 1.1 Identified Research Gaps

RG-01	There is too much focus on data entry rather than data analysis for decision-making.
RG-02	The data is being rarely used for planning purposes and optimal decision making.
RG-03	In previous work, no hybrid optimization technique has been implemented related to the prediction of the protein structures.
RG-04	Small datasets have been used in data mining for building a predictor model, for optimal decision making.
RG-05	Different approaches in Hybrid methods using data have not been used to test and analyze the efficiency of the predictor models.
RG-06	To improve the performance, analytical tools like data mining and pattern analysis need to be used for informed decision making

In the field of predicting protein structure, the main critical problem is the identification of correct templates for similar structure sequences and how to refine the native closer template structure. Secondly, how to build a model for correcting topology from scratch for sequences without having correct templates. The atomic simulations-based molecular structure has been focused by early efforts on the refinement of template structure that help in refining the low-resolution methods by using classical methods. Till now, there is no detailed physicochemical description of folding protein defined by the structural or evolutionary distance between the target and solved proteins in the Protein data bank or PDB library. For close templates protein, a template framework copying can be used to construct the full-length models. A current study was done in the same, which shows that if the best possible structures of the template are used in PDB then traditional model algorithms can be used to build high-quality models for single-domain proteins. The hypothesis of this research work was to:

Improve outcome by introducing hybrid PSO-GSA optimization with K-mean clustering then used selected cluster for training random forest classifier.

Get improved results in terms of various parameters by introducing hybridization of an algorithm for optimization.

Analysis of different dataset 25PDB dataset and FC699 dataset for Protein secondary structure (PSS).

A hybrid model using PSO and Firefly optimization is utilized for feature selection to improve accuracy results. It helps in selecting the best attribute of the secondary structure of the protein and improving the system accuracy and reducing the time complexity.

Utilize CNN+BILSTM as a classifier that will improve results

Research objectives as given below in Table 1.2. The main objective of this work is **t**o apply intelligent techniques for the prediction of protein secondary structure.

Table 1.2 Proposed Objectives

R0-1	To deploy intelligent techniques for sequence clustering.
R0-2	To deploy intelligent techniques for sequence alignment.
R0-3	To deploy intelligent techniques for the prediction of substructures in secondary structures.

1.9 Novelty

A new hybrid-based prediction model had been developed which uses the capability of hybridization of clustering, feature selection, feature extraction, and classification. In our proposed methodology different optimization algorithms had been used which are used for the first time in our research topic.

In the existing work of protein structural analysis [1.45], the outcome of all-α and all-β classes are not enough for explaining the effectiveness of the existing algorithm due to the same dataset used for both training and validation set. For a double layer of the SVM model, a 90% of 25PDB dataset was selected randomly for the training set and distinguishing all-α class from other three classes and rest need to be used for the validation set. They have repeated the experiment 100 times and got averaged accuracy of 100 times. The reduction in accuracy has been found due to poor extensibility of the model that means the accuracy achieved after validation of the training set from the validation set is not good. Another reason is that after dividing classes into other class by the model, the belonging of some proteins to the all-α class that will be re-classified into all-α class. This step has not been found in the double-layer SVM model that creates a need of reducing the accuracy. In existing work, the classification algorithm used for distinguishing between $\alpha+\beta$ and α/β classes is only dependent on the above two double-layer SVM models that make it lower than the accuracy of other classes. Although, lot of work has been done on the improvement still there is need to improve the existing problem by introducing new classification methods and feature selection optimization.

1.10 Thesis Contribution

In this thesis, different structural classes of protein are classified to make an understanding of different problems like folding and protein structure prediction, etc. Clustering is performed using the K-mean algorithm. In the current work a random forest (RF) classifier is proposed which is compared with conventional classifiers like SVM, Ada boost, RF, etc. in terms of accuracy for all four classes of protein. The accuracy of the proposed RF classifier is much higher than other classifiers. Also, the values of performance parameters like accuracy, recall, precision, and specificity are measured for different classes of protein. A Hybrid PSO-GSA algorithm was analyzed and its different parameters are analyzed for the classification of protein structure. As suggested in the literature, the proposed hybrid PSO-GSA algorithm has proved to achieve better results as compared to single algorithms.

The accuracy though significant yet is further improved by integrating more precise classifiers and by performing more intense pre-processing. Considering these key points as the motivation, another prediction model with altogether different architecture has been proposed. Sub-cellular localization of protein structure is attempted by numerous researchers by using several techniques of deep learning and machine learning. In the present study deep learning technique of CNN is utilized as a classifier which is compared with SVM concerning the accuracy, specificity, sensitivity, and MCC values for all four classes of protein. The accuracy of the CNN classifier is much higher than SVM classifier. Clustering is performed using the K-mean algorithm. A Hybrid PSO-Firefly algorithm is used for feature extraction of various classes of protein. 25 PDB dataset is used to analyze the protein structure in terms of various performance parameters. Also, scoring spaces and fitness values are evaluated for different classes of protein.

1.11 Organization of Thesis

This research work addresses numerous categories of data mining methods mainly classification and clustering-based techniques. The research work put an effort in enhancing various modern-day data mining approaches along with the presentation of new approaches in the same domain. The hybrid-based proposed approach helps in improving predictive power in terms of classification that is the most significant research work achievement. Due to growth in predictions and technology, it has been widely used in solving several real-world problems for instance; cancer detection, tennis match predictions, weather forecasting, and soil classification. Some of the developed powerful prediction models help in getting the solution for these real-world problems.

Further, a comparison of these models is performed with traditional ML techniques for performing the quantitative analysis of classification performances. For various research domain proposed model prove to be useful as compared to other traditional approaches. The complete research work is compiled in various chapters mainly seven are there and a chapter-wise description is also given below in this chapter.

Chapter-1: This chapter presents an insight into protein structure prediction and the role of technology and other analytical tools in this problem domain, about decision making. This chapter also presents an overview of the Techniques and Tools like classification, clustering, and Association for data analysis and also discusses different application areas along with various steps involved in the extraction of useful information. In later sections of this chapter, the Application of Data Science in the areas of protein prediction problem and various models proposed by some researchers in different studies. A brief outline of the objective, significance, justification, and scope of the study has also been extensively discussed and elaborated in this chapter.

Chapter-2: This Chapter presents the detailed literature review with an extensive collection of literature in the form of articles and research papers published by several researchers in this area. The role and importance of protein structure prediction information systems have been discussed. Literature survey on applications of various classification techniques like Decision tree, Bayesian networks, Random Forest, Logistic Regression, K-Nearest Neighbour, Artificial

Neural networks, Support Vector Machine, in the protein structure prediction domain along with a critical review of several research papers related to the area of study, has been presented in this chapter. The chapter also describes in detail various models that have been implemented using different Techniques.

This chapter is derived from:

A paper presented on "Intelligence Computing Methods Deployed for Protein Structure Prediction: A review" in an International Conference on Future and challenges of computational and integrated sciences held on 7th & 8th November 2014 in Hmv college, Jalandhar.

Chapter-3:

This chapter focuses on different mechanisms developed for pre-processing of protein sequence and comparative study on four intelligent techniques.

This chapter is derived from:

A Research paper titled "Protein Secondary Structure Prediction using Feed Forward Artificial Neural Network and Perceptron" is published in Shannon 100- Third International Conference on Computing Sciences (ICCS). International Journal of Control Theory and Applications ISSN: 0974-5572 Volume 9, Number 45, 2016.

A Research paper titled "Multi-Classifiers Comparison for Protein Secondary Structure Prediction" is published in ICCCIS-2019 i.e. IEEE International Conference ICCCIS-2019 held from 18-19th Oct 2019 at Sharda University, Greater Noida, U.P., India.

Chapter-4: In previous chapters, different classifiers that were shortlisted based on their performance on different protein structure data set were used and their performance was evaluated on the test data set. Since the last few years, the researches in data mining performance has moved in another direction to address the question of whether a combination of the classifier with different feature selection methods can help in enhancing the performance of the model. In this chapter, a hybrid model integrating feature selection and classification techniques for improvement of prediction accuracy has been evolved and discussed. Many ensemble approaches have been applied for the development of hybrid models; however, there are two primary concerns in hybrid methodologies, the algorithm selection for building a hybrid and the approach by which the result of different algorithms are integrated. This chapter discusses the hybrid model which can overcome the existing issues. It also demonstrates that the model build using machine learning ensemble methods with a feature extraction technique to select the key features in the data results in higher performance. Also, this chapter evaluates the Combination of different classifiers with different feature selection techniques so that the model with the highest accuracy can be developed.

Chapter-5: This chapter demonstrates the use of the Hybrid model which has been developed with a combination of Supervised and Unsupervised Techniques. Since it has been observed that the clustering-aided approaches can enhance or helps in improving Classification Rate for Predictions, the clustering aided approaches in combination with classification approaches have been used to provide more accurate predictions. In the Last Section of the chapter, outcomes of the Hybrid Model have been presented with the facts and findings.

This Chapter is derived from:

A Research paper titled" Hybrid of PSO-GSA based Clustering Approach for Predicting Structural Class Prediction using Random Forest Method" is published in European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 10, Pages 17-32.

Chapter-6: Here in this chapter we have discussed the results and interpretation of the various results obtained during all the phases of the experimentation process. The chapter also presents the conclusion drawn in terms of the usage of Meta and base classifiers for hybrid modelling. The Last Section of the chapter concludes with various findings in the direction of using hybrid modelling by combining Supervised and Unsupervised Learning and the future scope for the prospective young researchers interested to pursue their research in this field.

This Chapter is derived from:

A Research paper titled "Protein Structural Classes Prediction Based on Convolutional Neural Network Classifier with Feature Selection of Hybrid PSO-FA Optimization Approach" is published in European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 10, Pages 252-265.

Chapter- 2

Literature Review

2.1 Overview

For the last 30 years, the prediction of protein secondary structure is considered a very active area of research. The main challenge is to solve the protein folding problem that is linked with the prediction of folding routes and progression for reaching towards the native structure. If the use this ability to solve the existing problems then the user will come closer to the identification of protein structure.

In [2.1], David Searls highlighted various issues that need to be achieved for finding the protein structure:

The physical root of protein structural constancy is not completely understood.

Search space of the problem is excessively massive, due to the huge range of possible conformations of even relatively short polypeptides.

The primary sequence might not entirely state the tertiary structure.

Little improvement in the accuracy has a significant impact on various related research problems and software tools. Several ML approaches include various neural networks that are used by existing predictors of protein secondary structure. From the past few years, it has been found that the benchmark datasets accuracy changed from 69.7% by Sander and Rost (Yaseen and Li, 2014) [2.8] server falls into the template-based method for secondary structure prediction. It applied three separate neural networks: first for prediction, second for structure filter, and third for refinement using PSSM and modified SS scores. Their innovation to use statistical context-based scores as well as the structural information, as encoded features, to train neural networks to achieve the improvement on secondary structure up to 82.7%. The current state-of-the-art Deep CNF-SS (Wang et al., 2016) [2.9] is the first method using deep convolutional neural fields for secondary structure prediction. The DeepCNF model contains two modules: 1) Top and label layer come under Conditional Random Fields (CRF) module and input to layer cover deep convolutional neural network (DCNN) module. Experimental results show that the DeepCNF can obtain about 84% Q3 accuracy. Currently, significant improvement has been achieved on Q8 accuracy using deep neural networks.

Deep CNN with the conditional random field was proposed by Wang et al., 2016 that able to achieve the Q8 accuracy of 68.3% and 82.3% on the benchmark CB513 dataset. A multi-scale convolutional and deep neural network followed by three stacked bidirectional recurrent layers was proposed by Li and Yu, 2016 along with 69.7% Q8 on the same test dataset [2.10]. Further deep CNN with the next step conditional technique was proposed by Busia and Jaitly, 2017 and obtained 71.4% of accuracy [2.11].

2.2 Year-wise literature survey:

Table 2.1 Year wise literature survey

Author	Technique	Database	Application/ Accuracy
N. Qian et al., 1988	Non-linear neural network mode	_	It is used, non-linear neural network models. On proteins, non-homologous an accuracy of 64.3% was achieved on a testing dataset with the resulting training dataset [2.13].
B. Rost et al., 1994	Neural network	_	It shows an improvement of a neural network system using evolutionary conservation information. Overall 71.4% of accuracy is achieved using the proposed method [2.14].
D T Jones, 1999	Two-stage neural network	CASP3	A two-stage position scoring matrices-based neural network was used by Jones in which PSI-BLAST was used to generate the matrices for the prediction of protein secondary structure [2.15].
Anderson et al., 2001	Feed-forward neural network	PDB	Different schemes had been presented for prediction, this scheme stabilizes the secondary structures by directly predicting the hydrogen bonds [2.16].
Hua et al.,2001	SVM	RS126 & CB513	The SVM method achieved a good performance of segment overlap accuracy SOV=76.2% [2.17].
Cedric et al., 2001	_	_	Explains existing techniques of multiple sequence alignment and describes the potential strengths

			and weaknesses of the most widely used multiple alignment packages [2.18].
Kim et al., 2002	Maximum Entropy Markov Model	_	MEMM model has 58% accuracy and also recommended more improvement [2.19].
Pollastri et al., 2002	Recurrent Neural Networks and Profiles	R126, EVA, and CASP4	78% prediction accuracy has been achieved on different test sets [2.20].
Zhu et al.,2002	Multi-Modal Neural Networks	_	66% accuracy had been achieved for the prediction [2.21].
Ceroni et al.,2003	Combination of Support Vector Machines and Bidirectional Recurrent Neural Networks	CB513, PDB	A combination of local classifier and a filtering BRNN give good performance [2.91].
Yang et al.,2003	Bayesian inference model & the artificial neural network based on the model	_	Give good results for predicting the secondary structure of proteins [2.22].
Nguyen et al.,2003	Multi-Class Support	RS126 &	Two-stage SVMs gives good accuracy i.e. 79.5% [2.23].

	Vector Machines	PSIPRED	
Hall et al.,2003	Comparison between Neural Networks & Decision Trees	_	A number of distinctiveness of protein secondary structure prediction problem is better exploited [2.24].
Liu et al.,2004	Different combination methods	CB513	Experiment shows graphical models are superior to window-based methods [2.27].
WANG et al.,2004	Comparison between neural network and support vector machine	CB513	Q3 accuracies of neural network and support vector machine are 74.2% and 76.6% [2.29].
Chen et al.,2004	Bidirectional Segmented-memory recurrent neural network (BSMRNN).	RS126	An architecture Bidirectional segmented-memory recurrent neural network indicates an improvement in the prediction accuracy [2.27].
Wang et al., 2004	Support Vector Machine Based on a New Coding Scheme	CB513	It achieved a *Q*3 accuracy of 78.44% [2.28].

Hu et al., 2004	SVM with a new encoding scheme & an advanced tertiary classifier	RS126	Accuracy is 78.8%. The final Q3 accuracy is higher than the result of SVMPS: which claims the highest accuracy so far [2.25].
Doong et al.,2004	Support Vector Machine & Clustering	RS126, CB396, CB513, CASP	Accuracy using the Non-clustered method is 81.43%. Accuracy using GA Clustering is 93.97% [2.26].
Nakayama et al.,2004	Multimodal neural network (MNN)	_	Multimodal neural network (MNN) improved to 66%[2.32].
Bidargaddi et al.,2005	Hybrid two level modular architecture with Bayesian segmentation & neural networks	CB513 & PDB	Highest accuracy values for single sequence prediction methods. Accuracy is 71% [2.33].
Zhang et al.,2005	Machine learning (UMD-OAO + Bayesian system)	PDB & CB513	Suffers from unbalanced data problem as neural network systems were trained using the one-against-all modelling scheme. The accuracy achieved is 75.8% [2.34].
Zheng, 2005	Combination of hidden Markov models &	_	The accuracy achieved is 70% [2.35].

	sliding window scores		
Subramaniam et al.,2005	Analysis of the Effects of Multiple Sequence Alignments	_	Prediction accuracy is improved with multiple sequence alignment as compare to single alignment [2.36].
He et al.,2006	SVM_DT	RS126	The accuracy achieved is 88.9% [2.37].
Kim,2006	Fuzzy *k*-Nearest Neighbour Method and Its Parallel Implementation	_	In this work, the fuzzy *k*-nearest neighbor method uses the evolutionary profile; a parallel algorithm for protein secondary structure prediction has been developed [2.38].
Samani et al.,2007	Hybrid GMM/SVM Architecture	_	Our hybrid model achieved a good performance of three-state overall per residue accuracy Q_3 = 77.6% which is comparable to the best techniques available [2.41].
Chen et al.,2007	Two-stage Support Vector Regression	-	The proposed method is characterized by lower prediction error & reduction of the computational time required to develop the prediction model [2.42].

Nguyen et al.,2007	Two-stage multi-class SVMs	RS126, CB396, PSIPRED, CASP4, EVA	Helps in minimizing the generalization error in the prediction. On dataset RS126 & CB396 the accuracy achieved is 78.01% & 76.3% whereas on dataset PSIPRED, CASP4, EVA the accuracy is 77.0% & 79.05% [2.43].
Reyaz-Ahmed et al.,2007	Genetic Neural Support Vector Machines	RS 126,DSSP	A tertiary classifier (GNSVM) is introduced which is much better than other techniques [2.44].
Reyaz-Ahmed et al.,2008	New SVM-Based Decision Fusion Method Using Multiple Granular Windows	_	Multiple windows are compared with single window and it has been found that single window is not good in every case. A tertiary classifier is created and compared; also it has been proved better than other [2.45].
Kakumani et al.,2008	A Two-Stage Neural Network	RS126	The first stage of the network is used to link the protein sequence i.e. the input to bin and the second stage make use of neural prediction model for predicting secondary structure [2.46].
WANG et al.,2008	BP neural network and quasi-newton algorithm	DSSP, PDB	The prediction accuracy of 73.68% has shown that the combination of quasi-Newton algorithm and BP network can give better results for predicting protein secondary structure predicting the secondary structure of the protein [2.47].

Liu et al.,2008	Two-stage predictor (1st Predictor based on SVM & Bayesian discrimination in the 2nd stage)	RS126	Increase accuracy of binary classifier of protein secondary structures [2.50].
Dzikovska et al.,2008	Neural Networks	_	For each secondary structure elements the first level of NN classifier are separated. Speed has been improved as compare to other models [2.51].
Mansour et al.,2009	Scatter search algorithm	_	The algorithm can produce 3D Structures with good suboptimal energy values [2.52].
Tang et al.,2009	Large Margin Methods	CB513 and RS126	A new method has been designed in which the problem is considered as labeling of sequence [2.53].
Lakizadeh et al., 2009	Neural networks	PDB	For improving protein secondary structure prediction using neural networks it has been shown that the contact number can be used as a good source of information [2.54].
Lin et al.,2009	Grey neural network	_	The grey model GM (1, 1) has a larger error. The error can be reduced by using a neural network. It is demonstrated in this study that the fusion of the grey model and neural network in predicting the unknown amino acid of protein can remarkably improve the accuracy [2.55].
Hoqueet al., 2009	Genetic Algorithm	_	It gives ab initio prediction as a search problem with a genetic algorithm [2.56].

Thalatam et al.,2010	Artificial Neural Network	_	The prediction of secondary structure concerning the neural network works well and the analysis can conclude that it is possible for predicting protein secondary structure using neural network [2.57].
Rao et al.,2010	Pattern Recognition Neural Network	_	This method is a pattern reorganization technique which is statistical based and it achieved 72.3% Q8 accuracy [2.58].
Yang et al., 2010	Improved SVM Method in the compound pyramid model	RS126, CB513	Accuracy on data set RS126 is 83.06%. SOV99 accuracy increased to 80.6%. Accuracy on data set CB513 is 80.49%. SOV99 accuracy increased to 79.84% [2.59].
Bouziane et al., 2011	Combine k-NNs, ANNs, and Multi-class SVMs (M-SVMs)	RS126, CB513	Results shows that the classifiers which are used singly produce less good results than combination of classifiers used [2.60].
Yang et al.,2011	Large margin nearest neighbor model	_	The use of PSSM profiles is not mainly designed for the prediction of protein secondary structure that unable to use NN method for getting suitable accuracy. This method enhanced the prediction accuracy [2.61].
Patil et al.,2012	GA-SVM	_	Enhanced classification accuracy. Accuracy is 88.09% [2.62].
Bordoloi, 2012	Multiple Artificial Neural Network Classifier	_	Here majority voting is used to find the final structure of protein from various neural networks [2.63].

Chetia et al., 2012	ANN	-	ANN along with some statistical and image processing techniques to formulate the protein prediction. They had experimentally shown that how complexity and time of prediction are reduced as compared to certain traditional techniques [2.64].
Yang et al.,2013	Margin nearest neighbour classification	_	Better prediction accuracy compared with the previous one [2.65].
Jian-wei et al.,2013	Multilayer Feed-forward Neural Networks	RS126, CB513	Higher accuracy is achieved; the time complexity to train the model is often high. Hence, we utilize the contrastive divergence algorithm to solve it [2.66].
Sønderbyet al.,2015	A bidirectional recurrent neural network with long short term memory cells	CB513	On 8-class problems, they report good accuracy of 0.674 [2.92].
Liu et al.,2016	2D convolutional neural network(CNN)	25PDB	The performance of the network is enhanced by CNN features [2.93].
Li et al.,2016	Deep network	CB513, CASP10	Combined local and global contextual features are used with CNN and bidirectional neural networks

		, CASP11	consisting of gated recurrent units. This method achieved 69.7%, 76.9% and 73.1% accuracy on CB513, CASP10 and CASP11 datasets on 8-Class [2.94].
Hattori et al., 2017	Deep Recurrent Neural Network with Bidirectional Long Short-Term Memory (DBLSTM)	CB513	DBLSTM gives 68% accuracy on CB513 dataset [2.95].
Liu et al.,2017	Deep convolutional neural networks	25PDB, CB513, CASP9, CASP10, CASP11, CASP12	This methodology achieves 80+ accuracy in each dataset [2.96].
Zhang et al.,2018	A convolutional neural network, residual network, and bidirectional recurrent	CB513, CASP10,CASP11,CASP12	Achieved good accuracy on all datasets for both 8-Class and 3-Class prediction [2.97].

	neural network		
Khalatbari et al., 2019	Fuzzy k-nearest neighbor, Support vector machine, Ensemble prediction machine	-	In this prediction mechanism, different tasks are performed and achieve good accuracy [2.98].
Zhu et al., 2019	Evaluation of servers	Four dataset were selected based on homologous sequence with 30%, 50%, 70% and 90%.	Analyses the methods, results of many servers [2.99].
Ge et al., 2019	Double layer SVM	25PDB	Accuracy had been increased for the 4 classes of SCOP for prediction [2.100].
Mehta et al., 2020	Random Forest	ASTRAL 1.73, 25PDB, FC699	The maximum increase in the accuracy is then compared with other methods [2.101].

2.3 Protein Beta-turn Prediction

Then position, diversity conservation scoring function, and secondary structure like incorporated features with SVM were used by Hu and Li (2008) that improve the MCC of up to 0.47 [2.68]. From PSIPRED (Jones 1999) a predicted secondary structure was used by Zheng and Kurgan (2008) [2.69] along with PROTEUS2 (Montgomerie et al., 2008), TRANSEEC (Montgomerie et al., 2006), and JNET (Cole et al., 2008) for improving the performance. Further PSSM and predicted dihedral angles are used by Kountouris and Hirst (2010) [2.70] for the prediction of secondary structures and achieving 0.49 MCC. Then, MCC with NetTurnP server was developed by Petersen et al. (2010) [2.71] in which they have used independent four models for prediction of beta-turn four positions. To get 0.51 MCC a BetaTPred3 server was developed by Singh et al. (2015) [2.72] in which they have used a random forest approach and this ML approach was able to get useful results in beta-turn prediction. But there is a need for further improvement mainly in the prediction of beta-turns nine types and most of these approaches rely on 4-10 sliding window of amino acid residues for capturing short interactions. Also, existing NN having one or two layers could be able to extract the large level of features from input datasets so that no deep NN will be applied for beta-turn prediction. Data representations can be learned by deep neural networks with multiple levels of abstraction that provide new opportunities to these existing research problems.

2.4 Comparison of Binding Patterns

These methods recognize the similarity of local patterns which may serve as binding sites. Since proteins function by binding to other molecules, the similarity of binding sites and interactions may suggest the similarity of biological function. Furthermore, it was shown that the convergent evolution of the protein binding sites and their patterns is not a rare phenomenon [2.79, 2.80]. Till now, there are two types of approaches for comparing the binding sites of the protein. The first type recognized the specific 3-D patterns of amino acids like catalytic triads. These are triplets of amino acids with conserved identities and spatial arrangements, which are not necessarily sequentially ordered along the polypeptide chain [2.81, 2.82]. Such conserved catalytic residues are typical for protein families, like serine proteases, ribonucleases, and lysozymes.

Some of these patterns, like the ’catalytic triad’ of serine proteases, are shared by proteins with different trypsin and subtilizing folds. Several methods were developed to compare the specific patterns exhibited by certain sets of residues. Most of these methods were successfully applied to predict the protein function [2.83]. On the other side, there are various examples of biological variant proteins that share the same patterns of building such as estradiol or ATP without having any spatial pattern of residues with the same identity [2.84]. For addressing these problems, second types of approaches are compared with surface regions that are comprised of binding sites. Some approaches are presented for the recognition of similar binding sites without any assumptions regardless of fold or identity of amino acids [2.85]. These approaches were also shown to contribute to the functional annotation of novel proteins [2.86]. However, these approaches perform a comparison of any two molecules of protein. A large number of features come under pairwise alignments that are not required for binding.

2.5 Deep Learning

Compared with traditional neural networks, deep neural networks usually have many layers. In deep neural networks, each layer of nodes can extract a distinct set of features from its previous layer's activation output. The deeper the network, the more complex and higher-level feature the nodes will recognize. The deeper layer can aggregate and recombine features from shallow previous layers. The deeper layer can learn/extract higher-level features from the shallow layer. This hierarchy of increasing complexity and feature abstraction makes a deep neural network can handle a very large dataset. Deep neural networks usually have stacked a neural network, which means the networks can have several layers. Each layer has many nodes and a node is a place where multiply the inputs with weights and added by biases. The determination of protein structures experimentally is a difficult task and the use of genetic information in it has proved to be progressive. By homologous sequences covariation analysis it becomes possible to add residues of amino acid that help in the prediction of structures of the protein. The use of this information for constructing the potential of mean force helps to accurately describe the protein shape. Further, a simple gradient descent algorithm can be used for optimizing the resulting potential for generating structures without using complex sampling procedures. The AlphaFold, resulting system can achieve high accuracy even in the case of few homologous sequences.

2.6 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are within the category of Neural Networks. CNN's are effectively used and have shown successful results in image recognition and classification (Razavian et al., 2014 [2.86]; Ciregan et al., 2012 [2.87]; Krizhevsky et al., 2012) [2.88]. CNN's have also been used in recognizing faces

(Parkhi et al., 2015) [2.90], objects and traffic signs (Stallkamp et al., 2011) [2.89]. CNN's are very important for many machine learning applications today. Several new deep CNNs have been proposed recently which are all improvement based on the LeNet. In CNN's, several important terminologies are used and defined as followed:

> Channel: is used to describe a certain component of an image. Usually, an image taken from a standard digital camera consists of red, green, and blue channels. Each channel contains pixel values ranging from 0 to 255.
>
> Grayscale: is used to describe a one-channel image. The value of each pixel of a grayscale image is ranging from 0 to 255, where 0 is black and 255 is white.
>
> Depth: is the number of filters used to perform the convolution operation. For example, Conv(3) means using three distinct filters to perform convolution on the input image.
>
> Stride is the number of pixels that filter matrix is slid over the input image. For example, when the stride is 2, two pixels will be skipped at a time when the convolution filter is slid around.
>
> Padding: Usually zeros are padded to the input matrix around the board so that the convolution filter can be applied to the border of the input matrix so that the output has the same length as the original input. This is often used in full convolution. The Convolution operation from CNNs can effectively extract features from input images. It can extract/learn the spatial relationship between pixels using small convolution kernels.

2.7 ReLU, Batch Normalization, and Fully Connected Layer

ReLU stands for a rectified linear unit that has been used with convolution operation (Nair and Hinton, 2010). This unit is the most commonly used activation function in deep learning models and in case of receiving negative input, the function returns 0 and returns a value for any positive value. All negative values in the feature map get replaced to zero by the ReLU activation function. The motive behind applying ReLU activation after the convolution operation is for introducing non-linearity. The two main purposes of activation functions are to help a model account for the effects of interaction.

In different applications, tan-h or sigmoid activation functions are also used. The fully connected layer is usually used in the output layer with Softmax as the activation function. A fully connected layer means every neuron in the previous layer is fully connected with every neuron in the next layer. The output of CNN's is high-level features extracted from input data. Those features can then be fed into a fully connected layer for performing the classification task. Since the Softmax is used as an activation function, the output probabilities from the fully connected layer sum to 1.

2.8 Hybrid Methods

Till now, several techniques have been applied in protein secondary structure prediction such as PSO, GA, K-nearest neighbour, and many more. Same work has also been done using hybridization of approaches such as Deep Recurrent Neural Network with Bidirectional Long Short-Term Memory, Convolutional neural network, residual network, and bidirectional recurrent neural network, Fuzzy k-nearest neighbor, Support vector machine, Ensemble prediction machine, Hybrid GMM/SVM Architecture and many more in a different field that gives improved results but still lots of modification need to be done that's why a new hybrid approach has been proposed in our thesis work.

Table 2.2 Comparative table of hybrid methods used in protein secondary structure prediction

Author name	Approach used	Dataset	Outcome
Sønderbyet al.,2015	A bidirectional recurrent neural network with long short term memory cells	CB513	On 8-class problems, they report good accuracy of 0.674 [2.92].
Liu et al.,2016	2D	25PDB	The performance of the network is enhanced by CNN features [2.93].

	convolutional neural network(CNN)		
Li et al.,2016	Deep network	CB513, CASP10, CASP11	Combined local and global contextual features are used with CNN and bidirectional neural networks consisting of gated recurrent units. This method achieved 69.7%, 76.9% and 73.1% accuracy on CB513, CASP10 and CASP11 datasets on 8-Class [2.94].
Hattori et al., 2017	Deep Recurrent Neural Network with Bidirectional Long Short-Term Memory (DBLSTM)	CB513	DBLSTM gives 68% accuracy on CB513 dataset [2.95].
Liu et al.,2017	Deep convolutional neural networks	25PDB, CB513, CASP9, CASP10, CASP11, CASP12	This methodology achieves 80+ accuracy in each dataset [2.96].
Zhang et al.,2018	A convolutional neural	CB513, CASP10,CASP11,CASP12	Achieved good accuracy on all datasets for both 8-Class and 3-Class prediction [2.97].

	network, residual network, and bidirectional recurrent neural network		
Khalatbari et al., 2019	Fuzzy k-nearest neighbor, Support vector machine, Ensemble prediction machine	-	In this prediction mechanism, different tasks are performed and achieve good accuracy [2.98].
Zhu et al., 2019	Evaluation of servers	Four dataset were selected based on homologous sequence with 30%, 50%, 70% and 90%.	Analyses the methods, results of many servers [2.99].

2.9 Base of hybridization

In existing works, a double-layer SVM model-based step-by-step classification algorithm has been constructed for the prediction of the secondary structure of proteins. In this, the accuracy of two classes *α+β* and *α/β,* which was lower in the past is increased. For evaluation 25PDB and FC699 datasets are used, out of which 25PDB datasets contain 40% homologous sequences. There are still some limitations in the existing work that needs to be improved. For the same, there is a need of introducing all the information into the model for getting better classification outcomes then take the precision secondary structural sequences from PDB files experimental results. The existing work outcome creates a need of introducing hybrid approaches for feature selection and a new approach for classification.

In the existing papers, various works has been done by various researchers for prediction of protein structure. Researchers has employed SVM, PSO, GA, K-Nearest neighbour, fuzzy logic, deep learning, firefly algorithm, gravitational search algorithm, convolutional neural network, and many more algorithms either individually or in combination. The outcome of all these approaches shows that an improved outcome is achieved if two or more algorithms are combined in the work rather than using it individually. This inspired us to use a hybrid approach for protein secondary structure prediction.

Table 2.3: Comparative table of various approaches used in protein secondary structure prediction

Author	Technique	Database	Application
Kennedy et al.1995	Particle swarm optimization	_	Optimization of non-linear functions using particle swarm methodology had been introduced [2.102]
Rashedi et al.2009	Gravitational search algorithm	_	In solving non-linear functions, this algorithm gives high performance [2.103].
Fister et al. 2013	Firefly algorithm	_	This shows that through the hybridization of firefly with other algorithms any problem can be solved [2.104].
Liu et al.,2016	2D convolutional neural network(CNN)	25PDB	The performance of the network is enhanced by CNN features [2.93].

Ge et al., 2019	Double layer SVM	25PDB	Accuracy had been increased for the 4 classes of SCOP for prediction [2.100].
Mehta e al., 2020	Random Forest	ASTRAL 1.73, 25PDB, FC699	The maximum increase in the accuracy is then compared with other methods [2.101].

2.10 Summary

This chapter covers the introductory part of various methods and strategies that are used popularly. The literature review provides plentiful chances to the specialists to construct new and progressed proficient models like decision tree, neural network, Naive Bayes, Support Vector Machine, and many more classifiers. This main work is taken through reviewing this model in this chapter. Decision trees algorithm work faster and very easy to interpret and can be used to generate data-driven rules from a larger data set. Missing values are managed by neural networks and efficient categorical values are obtained. Naive Bayes insists that their numeric data should be normally distributed and provide predictive modelling on a small dataset. Various research work done using the Hybrid predictive data mining technique for knowledge discovery has also been reviewed extensively. It has been found that a Hybrid approach can help in developing a model with better performance as compared to the single learning base classifier and these techniques are specifically used in our model-building task.

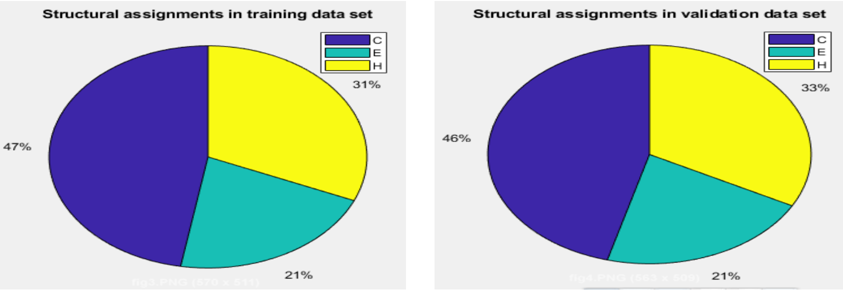
Chapter-3

Protein secondary structure prediction using Multi-classifiers

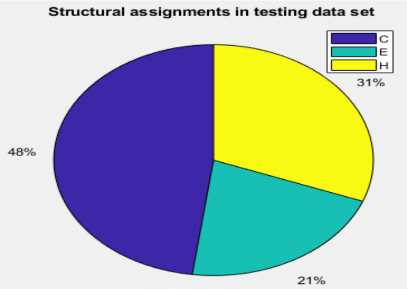
3.1 Multi-Classifiers Comparison for Protein Secondary Structure Prediction

In this chapter, four different classifiers had been used on the same dataset. The result obtained from the classifiers are then compared.Through prediction of structural properties such as secondary structure, solvent accessibility, dihedral angles, and contact maps, protein secondary structure can be predicted. Several algorithms are developed and classifiers are used for protein secondary structure prediction. Out of the classifiers used Support vector machines and neural networks has given good results compare to others. For subcellular location, receptor and many other biological purposes of protein structure, Adaboost as well as support vector machine are used [3.1] [3.2]. On RS126 dataset for predicting secondary structure of protein, genetic programming and neural network are used that gives good results [3.6]. With 3D structural information, protein structure is predicted using random forest feature extractor [3.3]. Clustering of the pockets are done by support vector machine [3.4]. SVM train accurate theoretical model by extracting efficient features from the dataset of protein sequence [3.5] [3.9]. Neural network base methods gives accuracy of 60% [3.7] , also it depends upon the type of protein used for analysis.

Results are shown in following figures starting from Figure 3.1 which are based on 3-dimensional structural dataset i.e. PDB. Figure 3.1 gives the structural assignments for different classes for training, testing and validation. Figure 3.2 ROC graph for different classes.



(a) (b)



(c)

Figure. 3.1: (a), (b), (c) Structural Assignment of Training, Validation and Testing Dataset

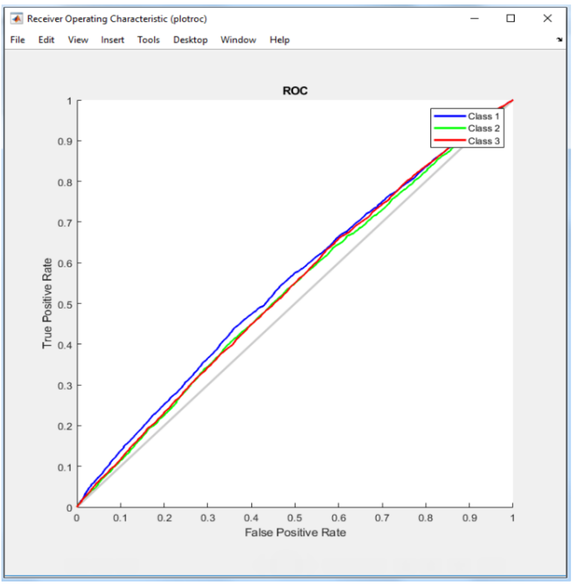
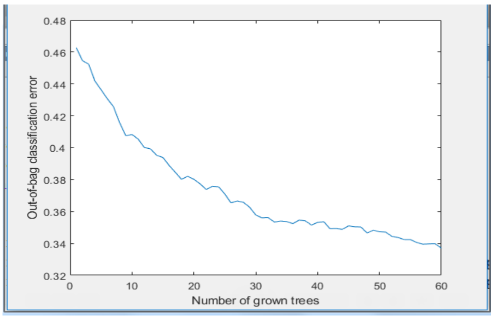
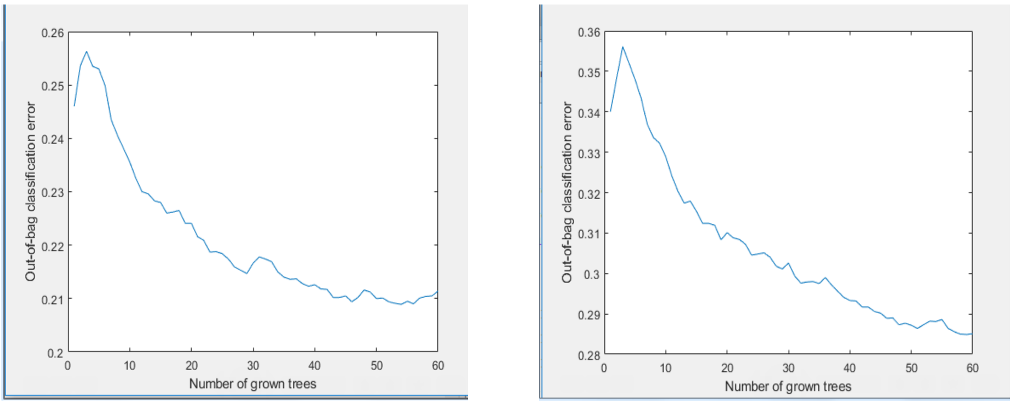


Figure. 3.2: Different Classes ROC curves



(a)



(b) (c)

Figure. 3.3: (a) , (b), (c) Tree grown graph for coil, sheet and helix

Figure. 3.4 depicts evaluation graph for function of the coil class and Figure. 3.5 depicts function model. Likewise, Figure. 3.6 depicts evaluation graph for function of the sheet class and Figure. 3.7 depicts function model. Similarly, Figure. 3.8 depicts evaluation graph for function of the helix class and Figure. 3.9 depicts function model .

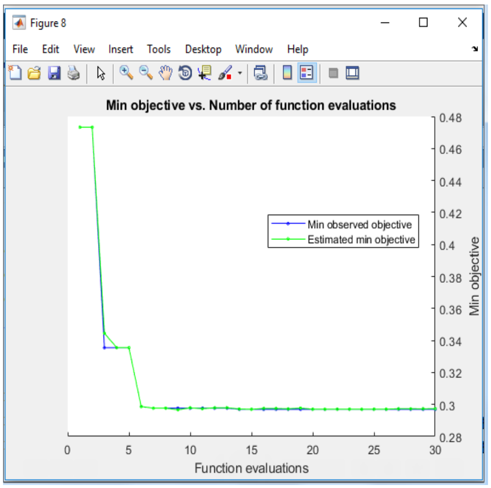


Figure. 3.4 Evaluation of the function for coil

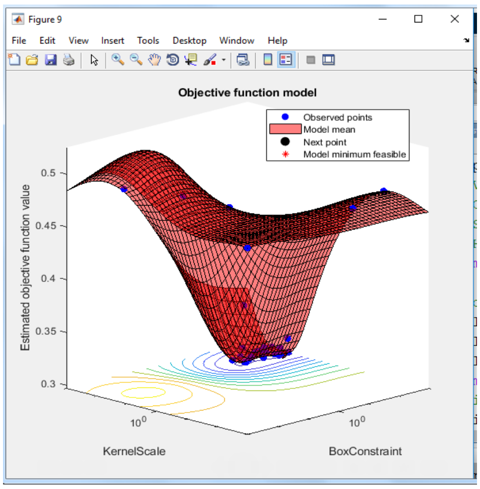
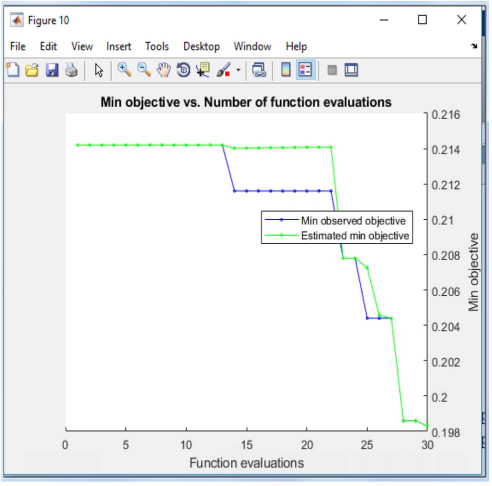


Figure. 3.5: Coil Class Objective Function Model

/



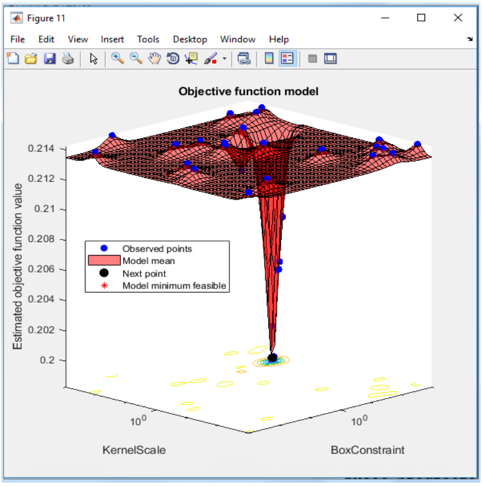
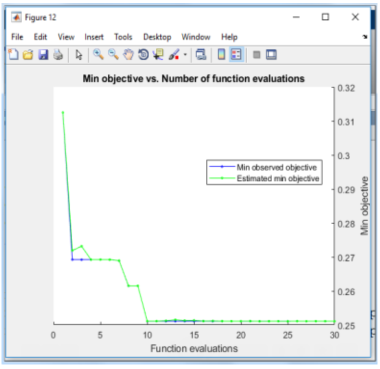


Figure. 3.6: Evaluation of the function for Sheet

Figure. 3.7: Objective Function Model of Sheet

Table 3.1. Accuracy table

Methods/ Classes	Coil	Sheet	Helix
ANN	68.68	79.47	71.19
Random Forest	67.73	79.36	70.95
AdaBoost	66.8	79.03	69.23
SVM	69.23	81.1	74.2



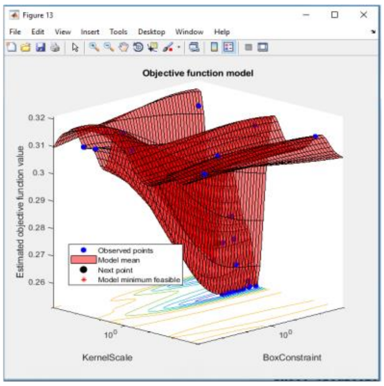


Figure. 3.8: Evaluation of the function for Helix

Figure. 3.9: Objective Function Model of Helix

Table. 3.1 shows different classifiers like ANN, Random Forest, AdaBoost, and SVM and their accuracy. Figure 3.10, Figure 3.11, and Figure 3.12 show the accuracy graphs for different class labels coil, sheet and helix respectively. At last, molecular view of the protein secondary structure is shown in Figure 3.13. Conclusion has been made from these graphs, results that Support vector machine gives good result comparatively.

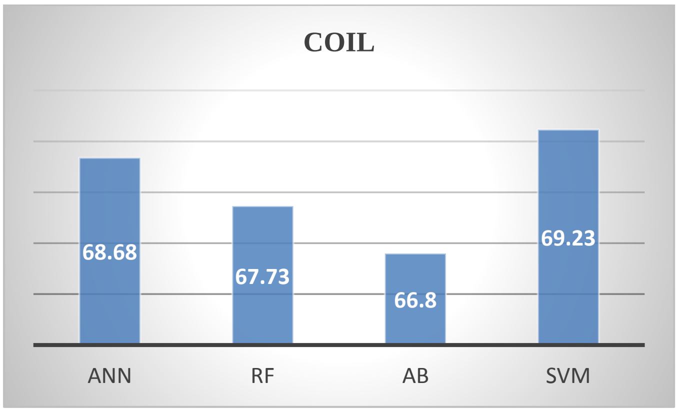


Figure. 3.10: Accuracy graph for the coil class

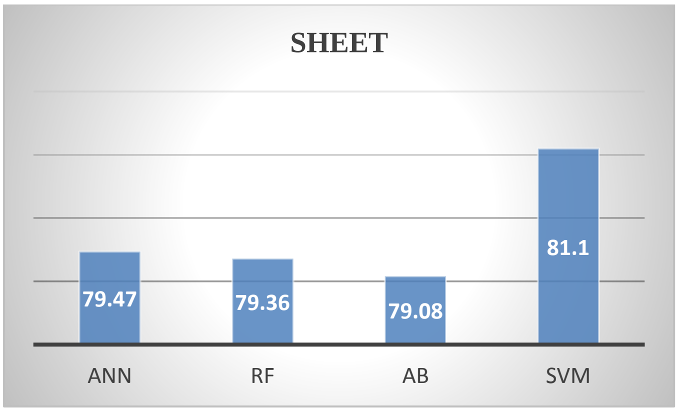


Figure. 3.11: Accuracy graph for the sheet class

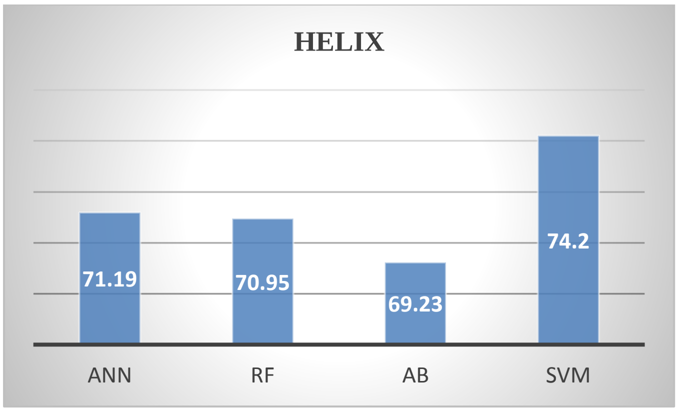
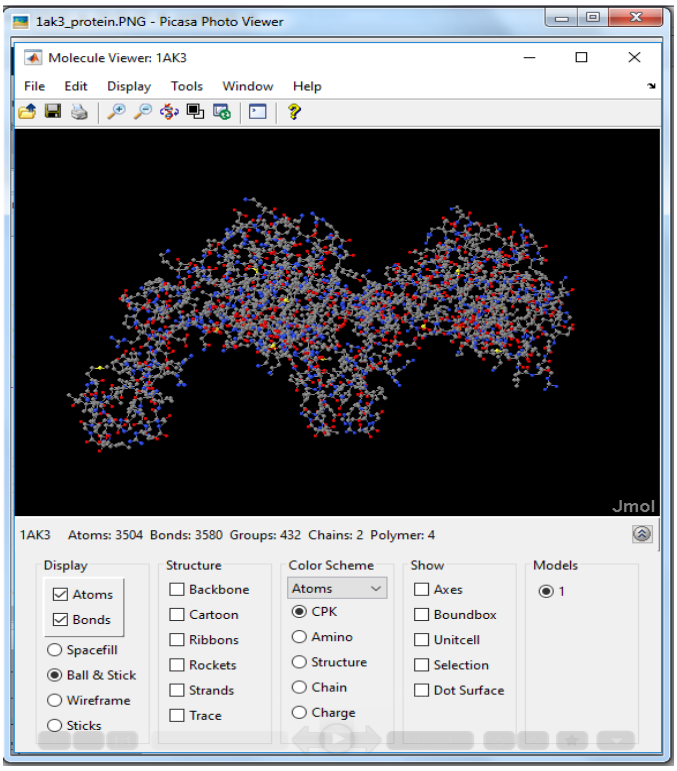


Figure. 3.12: Accuracy graph for the helix class



1AK3 Atoms: 3504 Bonds: 3580 Groups: 432 Chains: 2 Polymer: 4

Display
☑ Atoms
☑ Bonds
○ Spacefill
● Ball & Stick
○ Wireframe
○ Sticks

Structure
□ Backbone
□ Cartoon
□ Ribbons
□ Rockets
□ Strands
□ Trace

Color Scheme
Atoms
● CPK
○ Amino
○ Structure
○ Chain
○ Charge

Show
□ Axes
□ Boundbox
□ Unitcell
□ Selection
□ Dot Surface

Models
● 1

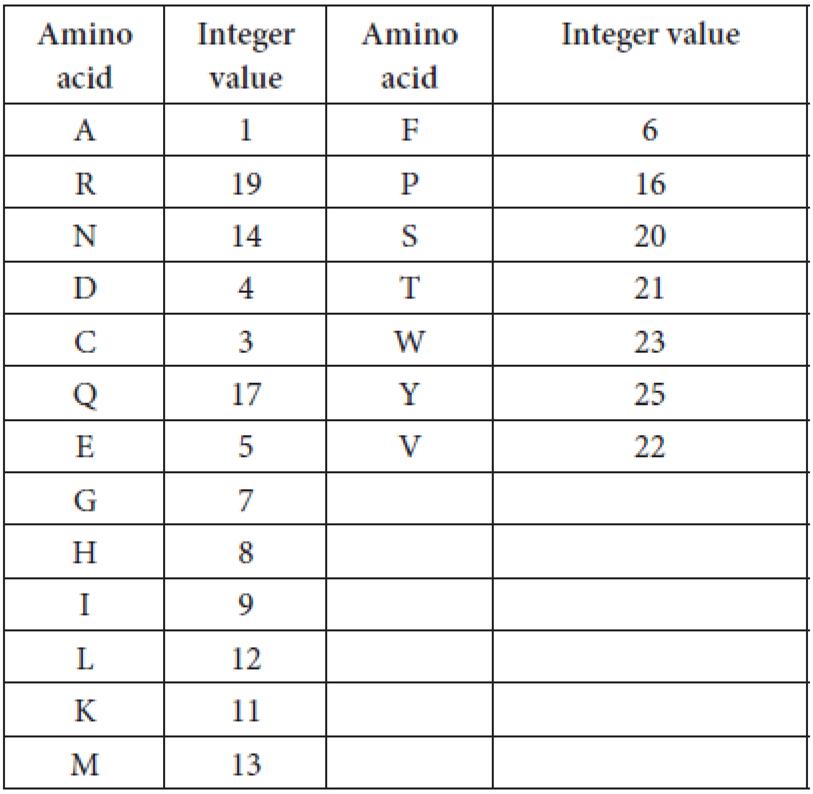
Figure. 3.13: Protein Secondary Structure Molecular View

Predicting secondary structure of protein plays vital role for predicting protein function and its tertiary structure. Results from the prediction gives good result with support vector machine as compare to other classifiers on different classes.

3.2 Implementing Feed Forward Artificial Neural Network and Perceptron for predicting protein secondary structure using newly proposed encoding method

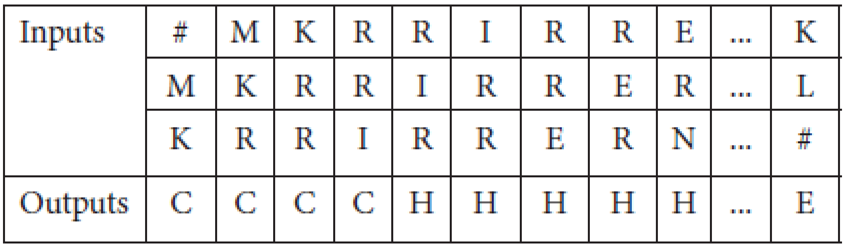
For designing the drug we should know the cause of the disease and its structure. Different protein structures are their, out of which tertiary structure are used for predicting the protein function. Predicting the tertiary structure is little tougher, so once we predict the secondary structure of protein it will be easy to predict its tertiary structure. Protein secondary structure can be predicted from different features of amino acids or directly from the primary sequence of protein. Many researchers [3.16–3.27] applied neural network for predicting secondary structure of protein. Main problem is with the large input data hence memory requirement is high and so its time of processing [3.28]. So there is need for encoding scheme that gives small input size. Here first a new method is defined to encode amino acid sequences, this has been implemented using MATLAB 7.10.0, nn tool are used for implementation by using neural network as a classifier as shown in Figure 3.14. An example of protein sequence has been shown in Table 3.3 with window size 3. Here integer encoding scheme has been designed in which each amino acid has been assigned an integer value from 1-26 as shown in Table 3.2. The symbol # has been assigned an integer value 0. In Table 3.3, each row of sequences has its ouput from different structure classes.

Table 3.2 Assigning Integer value to each sequence



Amino acid	Integer value	Amino acid	Integer value
A	1	F	6
R	19	P	16
N	14	S	20
D	4	T	21
C	3	W	23
Q	17	Y	25
E	5	V	22
G	7		
H	8		
I	9		
L	12		
K	11		
M	13		

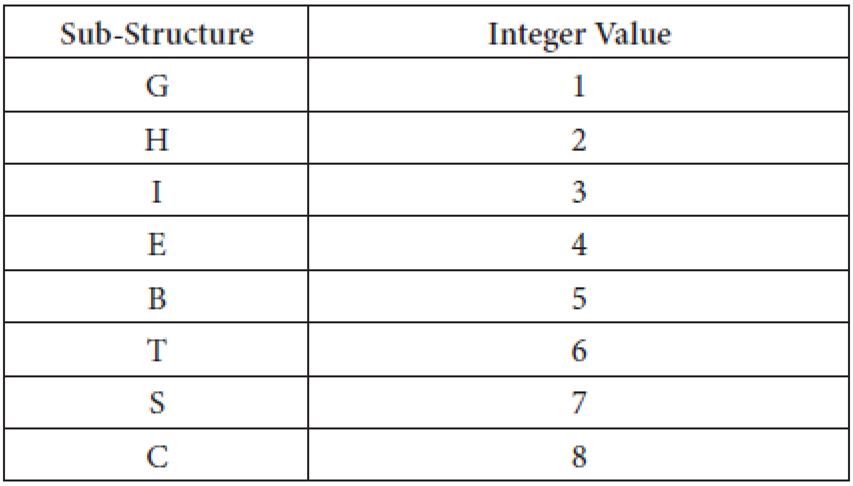
Table 3.3 Training data using 3 window sizes



Inputs	#	M	K	R	R	I	R	R	E	...	K
	M	K	R	R	I	R	R	E	R	...	L
	K	R	R	I	R	R	E	R	N	...	#
Outputs	C	C	C	C	H	H	H	H	H	...	E

DSSP define 8 sub-structures of the secondary structure Turns - T, Alpha Helix- H, 3/10 Helix-G, Coil-C, pi Helix-I, Bridge-B, beta ladder and beta sheet- E and Bend-S all are assigned an integer value as shown in Table 3.4.

Table 3.4 Secondary sub-structure and their integer value



Sub-Structure	Integer Value
G	1
H	2
I	3
E	4
B	5
T	6
S	7
C	8

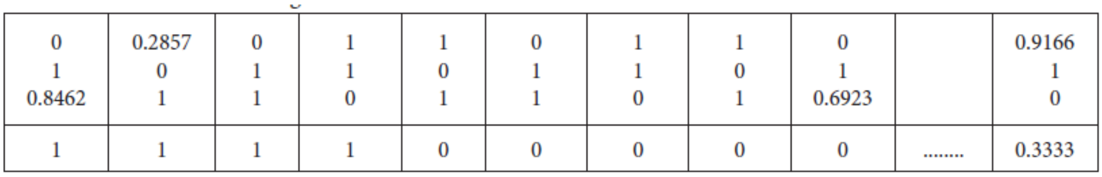
Artificial Neural Network is trained with input and output set which is normalized by the formula given in Equation given below:

$$Normalized\ (e_1) = e_1\ - E_{min}/E_{max}\ - E_{min}$$

The normalized set is shown in Table 3.5.

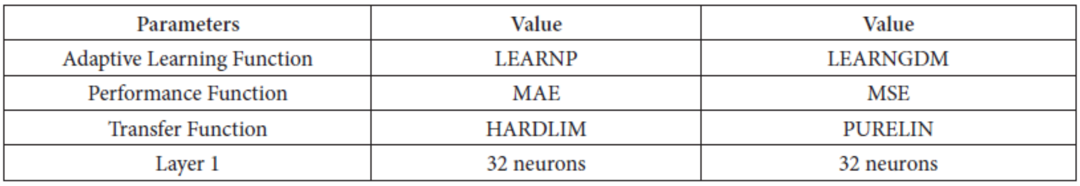
An neural nework consist of an input and output layer with window size of 3 is used. Here the dataset has been divided in to 60% and 40% for training and testing. Parameters of different classifier used is shown in Table 3.6 and the accuracy is given in Table 3.7.

Table 3.5: Normalized training data



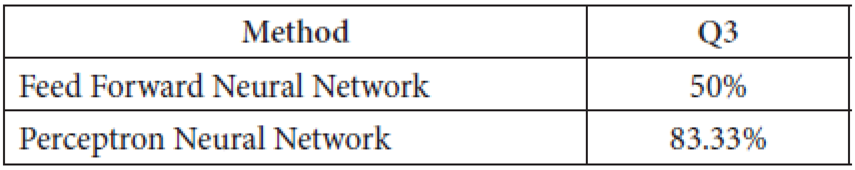
0 1 0.8462	0.2857 0 1	0 1 1	1 1 0	1 0 1	0 1 1	1 1 0	1 0 1	0 1 0.6923		0.9166 1 0
1	1	1	1	0	0	0	0	0	....	0.3333

Table 3.6: Some feed-forward neural network, Perceptron Neural Network parameters and their value



Parameters	Value	Value
Adaptive Learning Function	LEARNP	LEARNGDM
Performance Function	MAE	MSE
Transfer Function	HARDLIM	PURELIN
Layer 1	32 neurons	32 neurons

Table 3.7 Performance comparison of two neural networks



Method	Q3
Feed Forward Neural Network	50%
Perceptron Neural Network	83.33%

To encode the amino acid three blocks of input layer are used in the network belonging to window position. Using Q3 (alpha-helix, beta-strands, and irregular turns/ loops) formula accuracy of the netwok is calculated as shown in Equation below:

Q3 = No. of correctly classified secondary structures/ Total no. of amino acid residues *100%.

After implementing the feed-forward neural network we got Q3 (alpha-helix, beta-strands, and irregular turns/ loops) accuracy, calculated as follows:

Q3 = 12/24*100 = 50%.

After implementing Perceptron Neural Network we got Q3 (alpha-helix, beta-strands, and irregular turns/ loops) accuracy, calculated as follows:

Q3 = 20/24*100 = 83.33%.

In the method proposed, first data sequence is extracted from the database, input data is encoded then normalized and then finally trained using the network. Lastly accuracy has been calculated. Using Matlab, two different networks perceptron neural network and feed- forward neural network has been trained. Out of the two networks perceptron neural network gives good accuracy comparatively. In future, we will try to make use of Physio-chemical properties and other features of protein structure.So in this chapter we compared different classifiers on the same dataset and for pre-processing of the data new encoding scheme had been proposed.

Chapter-4

Proposed-RF Approach for enhancing protein structure prediction

4.1 Overview

The main aim is to get better accuracy by combining different classifiers. A lot of surveys had been done by the researchers on various possibilities in which they can combine classifiers. The result they got is compared with the best base-level classifiers.

The model developed using a hybrid technique is always expected to give better performance as compared to a model developed using an individual technique. Hence it is essential to identify a hybrid technique that would have better predictive accuracy than the existing techniques. Also, different data mining techniques are available for extracting the features present in the data. We have tried to identify novel features which can improve the performance of the universal classifiers.

Hybridization of particle swarm optimization and gravitational search algorithm has been proposed, due to which performance has been improved. Apart from these two algorithms, k-means algorithm has also been used. For classification purpose, Random forest has been used and the results are compared with the different classifiers used for the prediction of protein secondary structure on the same dataset.

4.2 Introduction

Proteins consists of a chain of amino acids (AA), which can be organized as secondary structures of three main types: helices (termed as α structure), the strands (termed as β structure), and the coils. Levitt and Chothia firstly defined and structured the protein classes [4.1]. Based on the pioneering work, the authors distinguished four different structure classes as defined by SCOP as (1) an all-α class, where only a small quantity of strands is included in proteins; (2) an all-β class in which only a small quantity of helices is included in proteins; (3) α/β class in which both helices, as well as the strands, are included and the strands are mainly parallel with each other; (4) an α + β class, where both helices, as well as the strands, are included and the strands are mainly anti-parallel with each other [4.2]. This structural class knowledge of proteins helps understand a wider problem called protein structure prediction (PSP). The knowledge of these structural classes is useful for predicting the accuracy of the secondary structure and reducing the possible conformations of search space for the tertiary arrangement [4.3, 4.4].

A lot of important biological functions are determined by the protein's spatial structure [4.5]. Presently, two processes are utilized for determining the protein structure which is Nuclear Magnetic Resonance (NMR) as well as X-Ray Crystallography (XRC). However, a lot of time and money is consumed in both processes. Accordingly, there is an enormous gap between the volume of decoded and cataloged sequences of protein structures [4.6]. The method that can predict the structure of a protein using different computational techniques is called Protein Structure Prediction (represented as PSP). PSP is a major concern/problem in analyzing the spatial structure of the protein [4.7].

For solving this PSP problem in numeral computational techniques are suggested by literature using several problem concepts classified as threading modelling, homology modelling, and ab-initio modelling, etc. The ab-initio modelling aims in predicting the protein's native conformation using its main sequence and physicochemical properties of amino acids (AA) like the hydrophilic or hydrophobic interaction [4.8]. The ab-initio modelling PSP problem is approaching with the use of off-lattice and on-lattice modelling. The on-lattice model limits the structure of a protein in a lattice. The Hydrophobic-Hydrophilic (HH) is an approach using an on-lattice type of modelling which is proposed by Dill (in 1985). This is possibly a lesser complex model ab-initio (on-lattice type) [4.9]. Despite its simplicity, this HH model is also verified by Berger & Leighton (in 1998). Hence, HH model is circulated to other abstractions of PSP in which a superior degree of freedom is presented. A distinguished disadvantage in the on-lattice model is that there is not enough detail in protein representations, so it is difficult to reproduce a genuine protein structure [4.10].

The reason for native structure predicting of small proteins using the ab-initio model is that these are inexpensive conformation evaluation, and it presents enormous and multimodal space for search. So the need of the hour is to design and develop simplified models for protein like HH model for reducing the time complexity and degree of freedom [4.11]. The main objectives of such models are testing, development, and contrasting several methods. A simplified three-dimensional model like AB off-lattice can be used for demonstrating two-phase optimization efficiency by utilizing Differential Evolution (DE) algorithms [4.12].

Protein-protein interactions can also be performed by using a Bayesian framework in a superior approach based on an unsupervised technique of learning, in which the models of network studies are presented in a given form of the protein [4.13]. Direct mapping match is undergone utilizing hyperparameters in PPI modelling form. For molecular search, parameterized BLOSUM metrics are used for sending back the alignment models of existing proteins. A simulation model is performed by considering the data value interconnections for identifying the model efficiency [4.14].

In this Chapter, various experiments that have been conducted to develop a hybrid model integrating feature selection and different classification techniques for the improvement of prediction accuracy have been explained.

4.3 Proposed approach for protein structure prediction

Protein secondary structures (PSS) described as primary folded structures, are produced inside polypeptide because of interactions among backbone atoms. PSS classification is vital for diverse biological functions which include: recognition of protein fold, prediction of tertiary structure, DNA-binding prediction, and conformation search area reduction. In the current article, a model based on machine learning for PSS classification is proposed. Here both sequence-based, as well as structure-based features, are considered. Firstly, pre-processing on protein data is applied, then a clustering technique i.e. hybrid model of PSO and GSA optimization with the collaboration of *K*-Mean clustering is proposed. Selected clusters are used for training of classifier random forest, and evaluation of performance parameters.

4.4 Methodology

1) Read data from excel and apply pre-processing on data for refining dataset. we used the FC699 dataset for Protein secondary structure (PSS).
2) After that apply k-mean clustering on data to make the initial cluster and find out centroid point that centroid points take input for optimization algorithm or take initial population for generate by k-mean clustering.
3) Initialize PSO and GSA parameters like C1, C2, and G0 and number of population, maximum iterations;
4) Generate the best solution of clustering with the help of a hybrid of PSO and GSA optimization.
5) Initialize random forest for classification and evaluation of performance parameters.

4.5 Proposed Algorithm

In this section, we have explained the proposed algorithm used for clustering and classification of various proteins. The flowchart of the proposed algorithm is shown in Figure 5.1. Next, the detailed description of each component of the flowchart is explained below.

4.5.1 K-Means algorithm

Macqueen in 1967 developed a simple clustering algorithm termed a k-means algorithm. This algorithm is based on an uncomplicated and unsupervised partitioned cluster algorithm in which the data is clustered based on a given *k*-value of data. An iteration technique is utilized for producing independent data produced into a variety of clusters with their data properties similar to each other. There are two separate segments in this algorithm. The first segment provides a methodology for random selection of *k* center by any user, whereas the second segment recalculates the average value of the different clusters formed previously. In the first segment, numerous metrics for distance calculation (like Euclidean distance) are considered for taking the individual object into the closest center. Thus each identified object in all the clusters is considered an early grouping of these objects is done in the same way to finish the first segment. In the next phase, the average value of previously shaped clusters is recalculated. This process of iteration is continued until the criterion function is allotted the highest value. Iteration is stopped when this value reaches to a minimum.

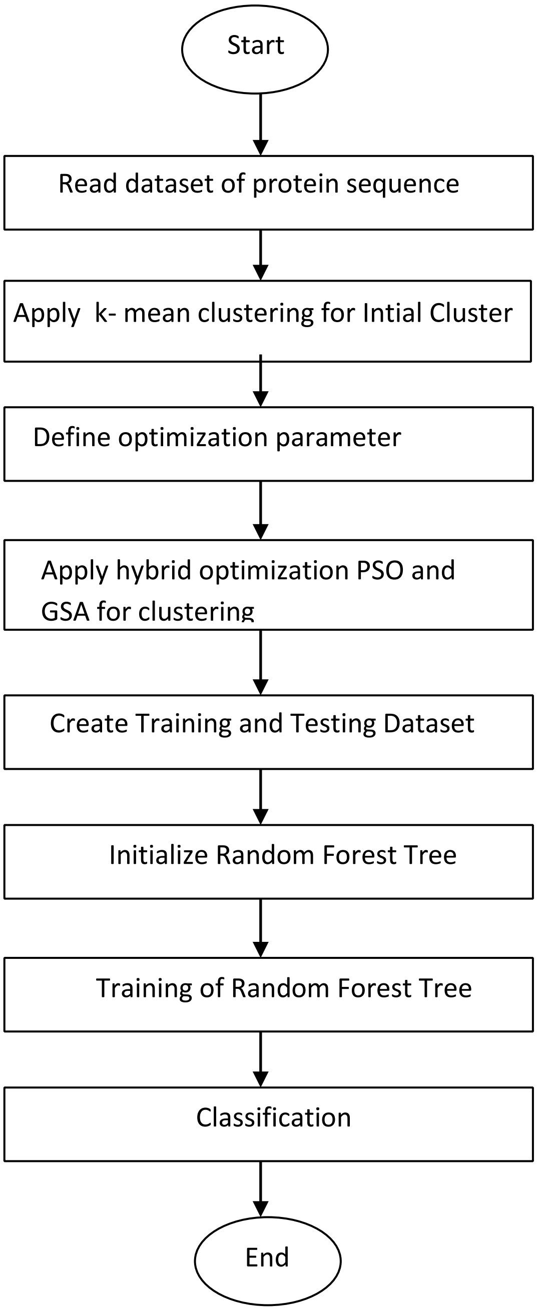

Start
Read dataset of protein sequence
Apply k- mean clustering for Intial Cluster
Define optimization parameter
Apply hybrid optimization PSO and GSA for clustering
Create Training and Testing Dataset
Initialize Random Forest Tree
Training of Random Forest Tree
Classification
End

Figure 4.1 Flowchart of the Proposed Technique

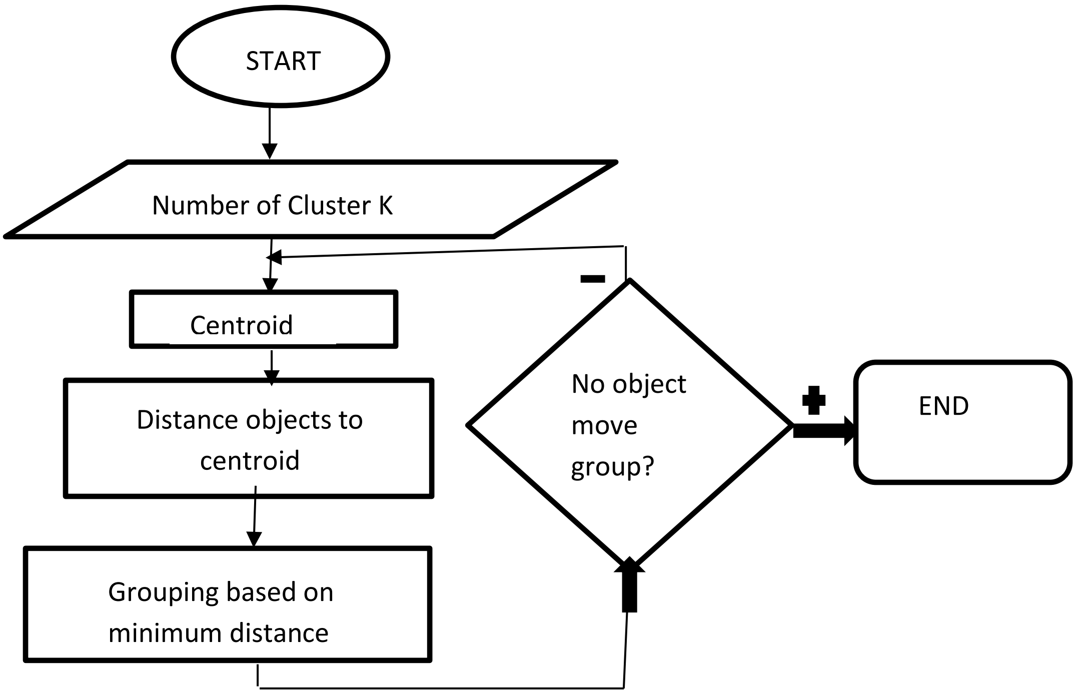

START
Number of Cluster K
Centroid
Distance objects to centroid
Grouping based on minimum distance
No object move group?
END
−
+

Figure 4.2: k –Means Clustering Algorithm

Various calculations needed for clustering by *k*-means algorithms are specified as:

$$d(n,z) = \min_{1<i<k} d(n, z_i) \qquad (1)$$

$$d(N,Z) = \frac{\sum d(n_i, Z)^2}{L} \qquad (2)$$

Here N = {*n*1,.… *nL*} is a set for k-centers, also $Z = \{Z_1,\ldots Z_k\}$ is a mean square distance which is computed between the cluster center and given data point. To complete the operation, a complexity analysis is performed.

Generally for the process of grouping, different steps performed by the k-means algorithm are indicated in Fig. 4.2. In the starting phase, similar average data are grouped assuming an initial value of neighboring average data. Afterward, the initial data is calculated by the average value of a cluster of individual data. Then, an initial data for the individual is again assumed for the identified group of neighboring average data. Lastly, the classification process is checked for the next data and then to the next until data is not changing and the same data value resembles the previous one. After this process, the clustering is stopped and the result is produced. If there is any failure in the checking process, the process is again repeated until the same data value is achieved.

4.5.2 Particle Swarm Optimization (PSO)

PSO algorithm is provoked by the organized movement of bird flocks and fishes [4.23]. The PSO is composed of a swarm of elements that interact with one another in a constant search area. A prospective solution of any problem can be represented with the position of each element and representation is done like an *n*-dimensional space vector. The particles in PSO “fly” throughout the n-dimensional search area and socio-cognitive affinity decides the possible change in their positions to imitate the success accomplished by further particles. The life experience of every particle in the swarm is different from other particles and the quality of each particle is evaluated by its own experiences. As an individual in a social gathering, each particle knows the behavior of its neighbors. The information of the cognitive factor is also termed as individual learning whereas information of social factor is termed as cultural transmission. Hence every individual’s decision is made by accounting for both cognitive as well as social factors, which lead the swarm population to an evolving behavior [4.24].

4.5.3 Gravitational Search Algorithm (GSA)

GSA is formed based on the gravitational law and the concept of interaction of masses [4.25]. The Newton theory of physics is utilized by the GSA algorithm and its search instruments are the mass collectors. GSA contains an isolated organism of different masses. Based on gravitational force, each mass in the organism can notice the condition of the other mass. So by using the gravitational force, the information can be transferred between diverse masses. In GSA, an agent is an object whose performance is calculated with its mass. All such objects interact with one another by the gravitational force which causes the combined movement of these objects towards the heavier mass object. The heavy masse objects form a superior solution to this problem. Agent’s position is the solution to the given problem which is used for determining its mass [4.26].

The algorithm depicting the hybridization of PSOGSA is as indicated by Algorithm 4.1.

Algorithm 4.1: Hybrid of PSOGSA Optimization Algorithm for Clustering

Input: **n** number of population, **t** maximum iteration, **d** number of clusters, **C1** and **C2** are constants, **w** inertia weight, and **G0** gravitational constant.

Notations: $m_i$is active gravitational mass, m_j is passive gravitational mass, R_{ij} Euclidean distance between two agents i and j, ε is a small constant, $rand$ **is** random vectors in [0, 1], d_n is a dataset of protein structure and $gbest$ is global best fitness.

Output: idx number of centre index and **C** number of centre

```
Initialization n number of $\boldsymbol{x_i}$ positions generate by the K-Means algorithm
      Initialization of velocity, force, G, and acceleration
      for i=1 to t
      Update G by equation
      G= $G0 * e^{-23\frac{i}{t}}$
      Calculate fitness function of each population by equation
          $fitness = min(\sum_{i=1}^{n}\sqrt{(x_i - d_n)^2})$
      for j=1 to n
      Update force by equation
       $f_d$=$G * \frac{m_i - m_j}{R_{ij}+ε}(x_i - x_j)$
      $force = \sum_{i=1}^{n} rand * f_d$
      Update acceleration by equation
      $acceleration = \frac{force}{m_{ij}}$
      Update $x_i$ positions by equation

      $velocity = w + velocity + c1 * rand * acceleration + c2 * rand * (gbest - x_i)$
      $x_i = velocity + x_i$
```

```
end for
end for
```

4.6 Benefit of Hybridizing PSO and GSA

- Hybridization of Particle Swarm Optimization (PSO) and Gravitational Search Algorithm (GSA),for integrating the skill of escapade in PSO and probing in GSA, to combine the power of both [4.27].
- Combining the advantage of both PSO and GSA, velocity equations are updated. Also the position equations are updated using a mobility factor, which increases speed and accuracy [4.28].

4.7 Performance Evaluation

K-mean clustering algorithm can be used for predicting the structures of four classes of protein α (A), β (B), α/β (C), and α + β (D). The performance can be analyzed feature-wise and class-wise in terms of different parameters like true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). TP illustrates the correctly marked positive samples of protein. TN illustrates the correctly marked negative samples of protein. FP illustrates the incorrectly marked positive samples of protein. FN illustrates the incorrectly marked negative samples of protein. FP is also termed as a type-I error. FN is also termed as a type-II error. For understanding the PSP problem, both these errors are taken into consideration.

Various calculations of these parameters (TP, TN, FP, and FN) can be considered for evaluating the performance of various parameters such as Accuracy, Precision, Specificity, and Recall (Sensitivity) as provided in Equations (3-6). Accuracy is one of the most frequently used parameters for indicating the performance of the appropriately classified samples out of total samples. The precision determines the preciseness in a model for correctly classifying the correct positive samples out of total positive samples. Recall determines the correct positive samples out of all available correct positive samples (TP+FN). Specificity determines actual negative samples out of total negative samples.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (3)$$

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (6)$$

4.8 Results & Discussions

The main objective of this clustering process is a grouping of similar objects in the same group (cluster). A set of measurements or attributes is used to define each object. For determining any similar objects, the similarity is measured between them. Numerous similarity measures are provided in the literature. In the current article, Euclidean distance is used for calculating the similarity between different objects. Euclidean distance is provided by the following equation:

$$distance(o_i, o_j) = \left(\sum_{p=1}^{m} |o_{ip} - o_{jp}|^{\frac{1}{2}}\right)^2 \qquad (7)$$

Here, m represents total no. of attributes, o_{ip} represents the attribute number's value, p for the object 'i' (o_i). For solving the problem of data clustering, the standard algorithms (PSO and GSA) are adapted for reaching the centroid of clusters.

4.8.1 Performance Evaluation for Protein Structural Class A

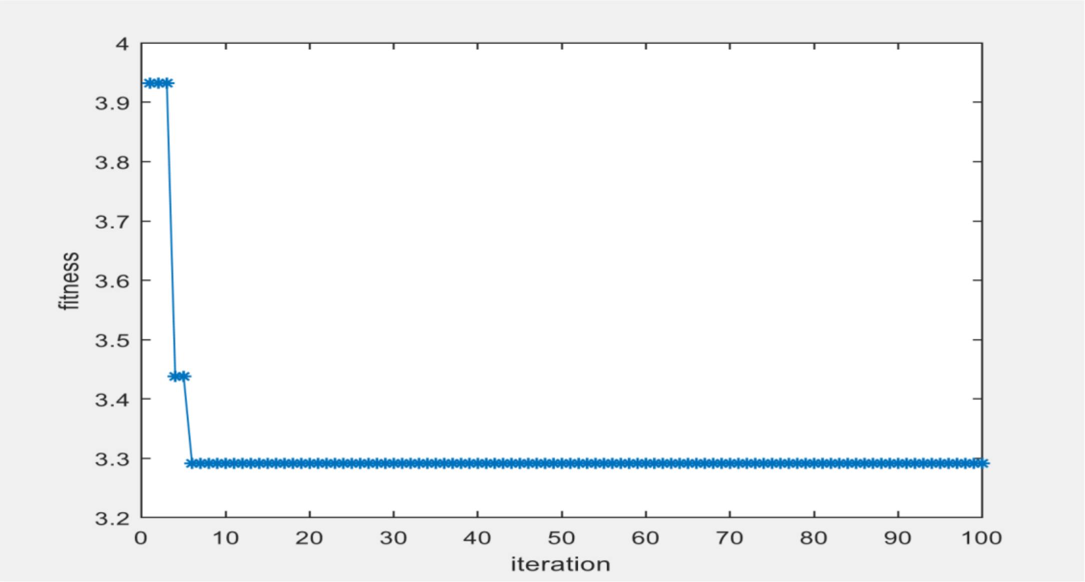


Figure 4.3: Variation of fitness function with no. of iterations for Class A

Figure 4.3 provides the variation of the fitness function as per the number of iterations for Protein Structural Class A. Figure 4 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in %) accomplished by the proposed Random Forest (RF) classifier with FC699 represented test data. Figure

4.5 (Protein Structural Class A) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with another classifier like SVM, Ada boost, RF, etc. As provided by figure 4.5, the accuracy of the proposed RF classifier is much higher than other classifiers.

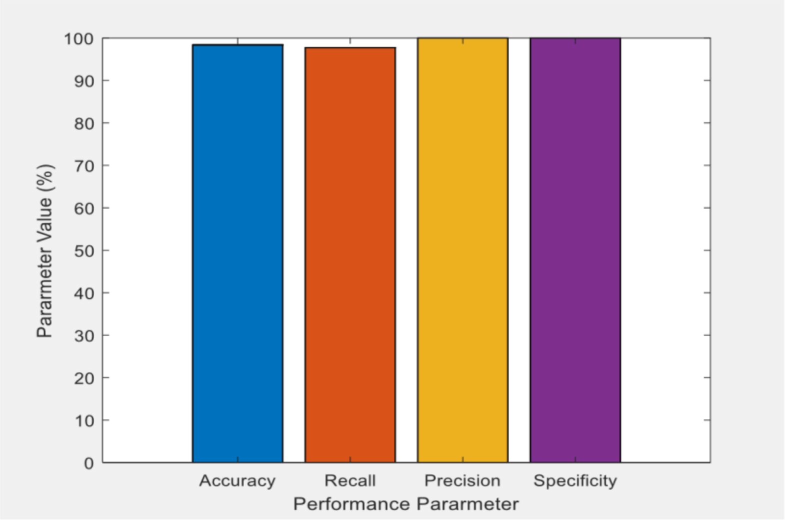


Figure 4. 4 Performance of different parameters

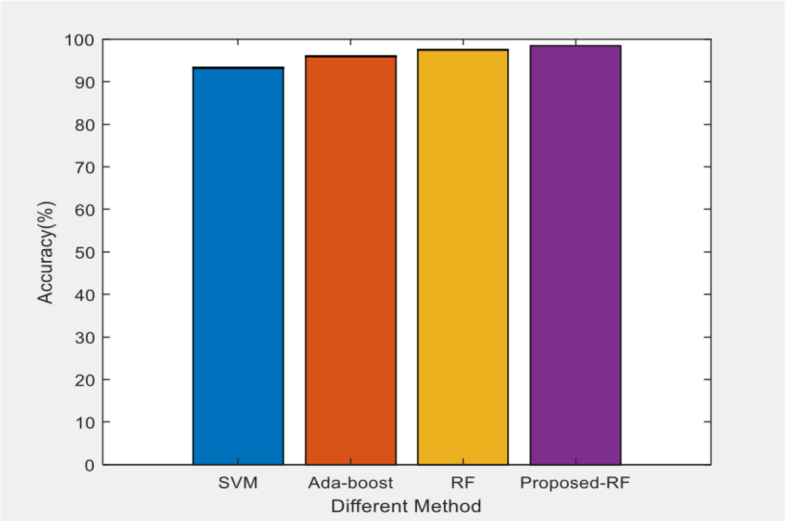


Figure 4.5 Accuracy of different methods

Table 4.1: Accuracy comparison of proposed RF method for Class A with other methods

Sr.No.	Technique	Accuracy (%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.36

Table 4.1 compares the accuracy values of the proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class A using FC699 Data sets. Table 4.2 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 4.2 Performance value of different parameters for proposed RF method of Class A

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.36
2	Recall	97.72
3	Precision	100
4	Specificity	100

4.7.2 Performance Evaluation for Protein Structural Class B

Figure 4.6 provides the variation of the fitness function as per the number of iterations for Protein Structural Class B.

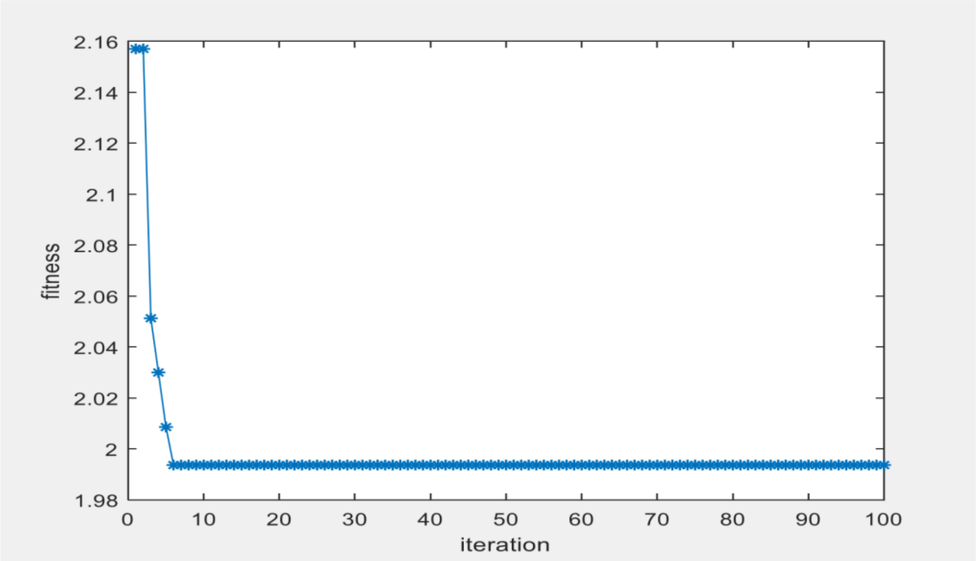


Figure 4.6 Variation of fitness function with no. of iterations for Class-B

Figure 4.7 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in %) accomplished by the proposed Random Forest (RF) classifier with FC699 represented test data. Figure 4.8 (Protein Structural Class

B) Highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with another classifier like SVM, Ada boost, RF, etc. As provided by Figure 4.8, the accuracy of the proposed RF classifier is much higher than other classifiers.

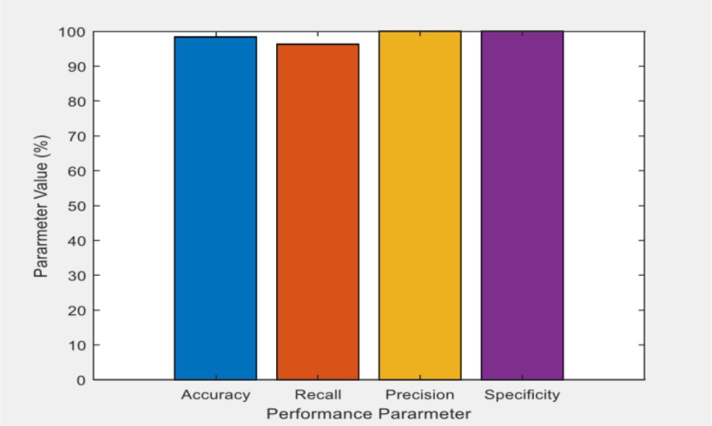


Figure 4.7 Performance of different parameters

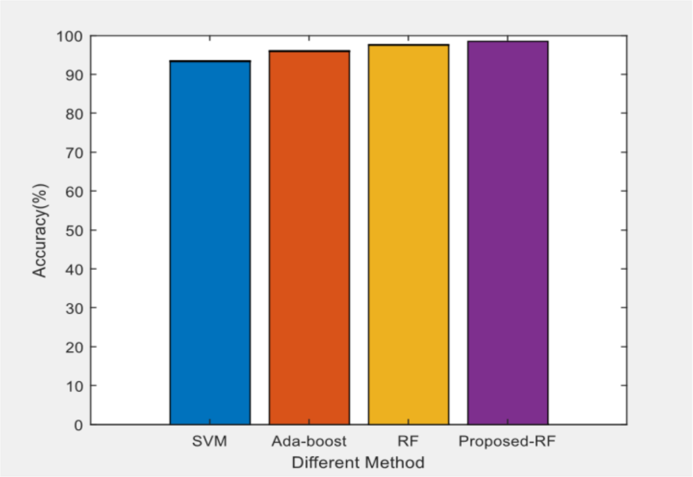


Figure 4.8 Accuracy of different methods

Table 4. 3: Accuracy comparison of proposed RF method for Class B with other methods

Sr.No.	Technique	Accuracy (%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.47

Table 4.3 compares the accuracy values of the proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class B using FC699 Data sets. Table 4.4 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 4.4: Performance value of different parameters for proposed RF method of Class B

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.47
2	Recall	96.36
3	Precision	100
4	Specificity	100

4.7.3 Performance Evaluation for Protein Structural Class C

Figure 4.9 provides the variation of the fitness function as per the number of iterations for Protein Structural Class C.

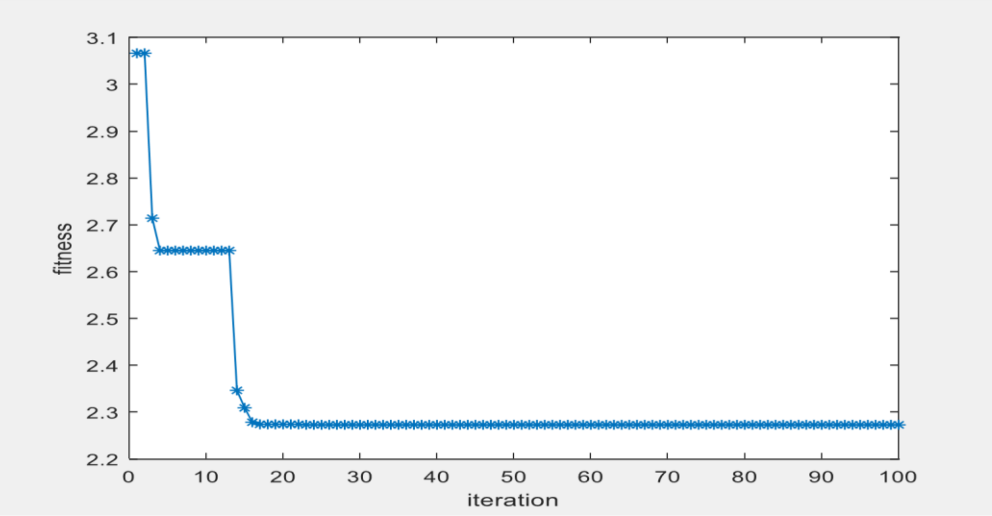


Figure 4.9 Variation of fitness function with no. of iterations for Class-C

Figure 4.10 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in %) accomplished by the proposed Random Forest (RF) classifier with FC699 represented test data. Figure 4.11 (Protein Structural Class C) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with another classifier like SVM, Ada boost, RF, etc.

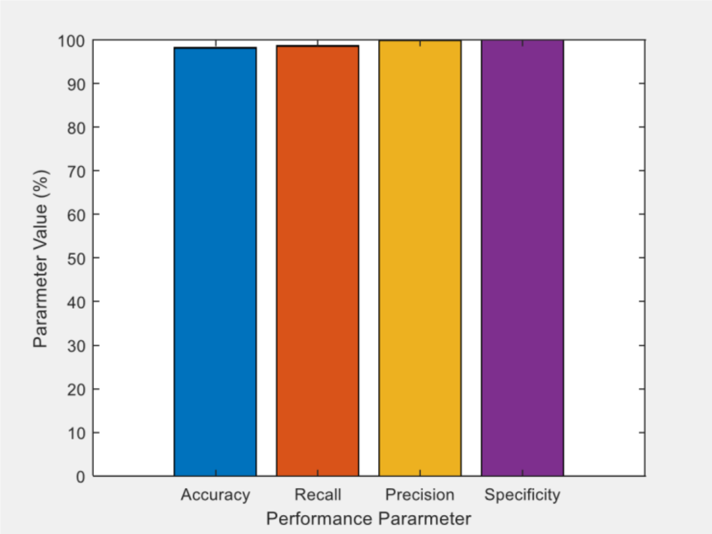


Figure 4.10: Performance of different parameters

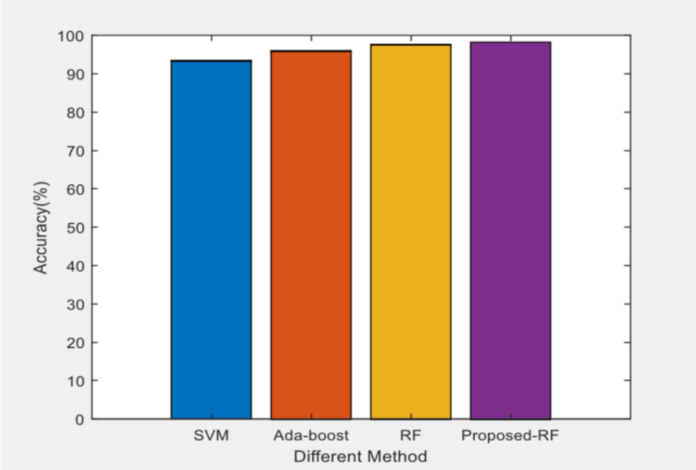


Figure 4.11 Accuracy of different methods

As provided by figure 4.10, the accuracy of the proposed RF classifier is much higher than other classifiers.

Table 4.5: Accuracy comparison of proposed RF method for Class C with other methods

Sr.No.	Technique	Accuracy (%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.22

Table 4.5 compares the accuracy values of the proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class C using FC699 Data sets. Table 4.6 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

Table 4.6: Performance value of different parameters for proposed RF method of Class C

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.22
2	Recall	98.70
3	Precision	100
4	Specificity	100

4.9 Performance Evaluation for Protein Structural Class D

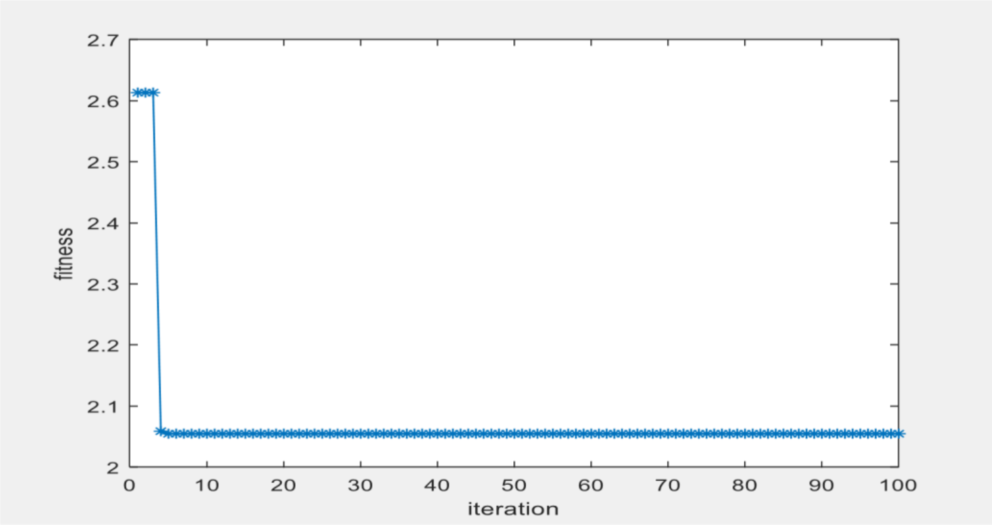


Figure 4.12: Variation of fitness function with no. of iterations for Class-D

Figure 4.12 provides the variation of the fitness function as per the number of iterations for Protein Structural Class D. Figure 4.13 highlights the performance of different parameters (accuracy, recall, precision, and specificity) values (in %) accomplished by the proposed Random Forest (RF) classifier with FC699 represented test data. Figure 4.14 (Protein Structural Class D) highlights the comparison of accuracy values by the proposed RF classifier with the FC699 dataset with another classifier like SVM, Ada boost, RF, etc. As provided by Figure 4.13, the accuracy of the proposed RF classifier is much higher than other classifiers.

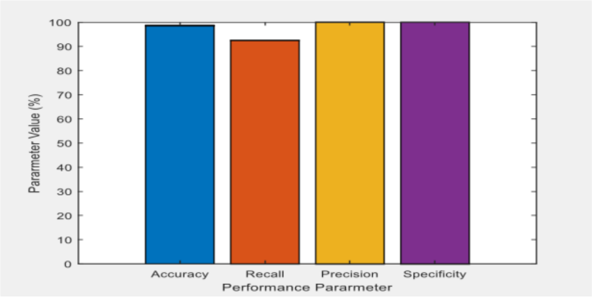


Figure 4.13 Performance of different parameters

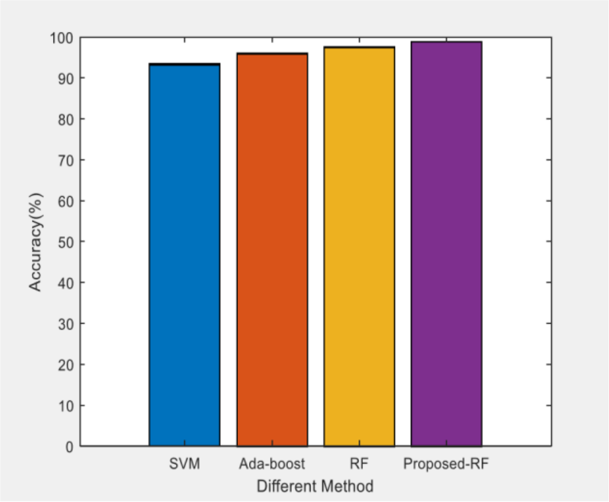


Figure 4.14 Accuracy of different methods

Table 4.7: Accuracy comparison of proposed RF method for Class D with other methods

Sr.No.	Technique	Accuracy (%)
1	SVM	93.36
2	Ada-boost	96
3	RF	97.56
4	Proposed-RF	98.70

Table 4.8: Performance value of different parameters for proposed RF method of Class D

Sr.No.	Performance (%)	Proposed RF
1	Accuracy	98.70
2	Recall	94.73
3	Precision	100
4	Specificity	100

Table 4.7 compares the accuracy values of the proposed RF method with different methods like SVM, Ada-boost, and RF for prediction of protein structure for class D using FC699 Data sets. Table 4.8 provides the performance parameters values (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier.

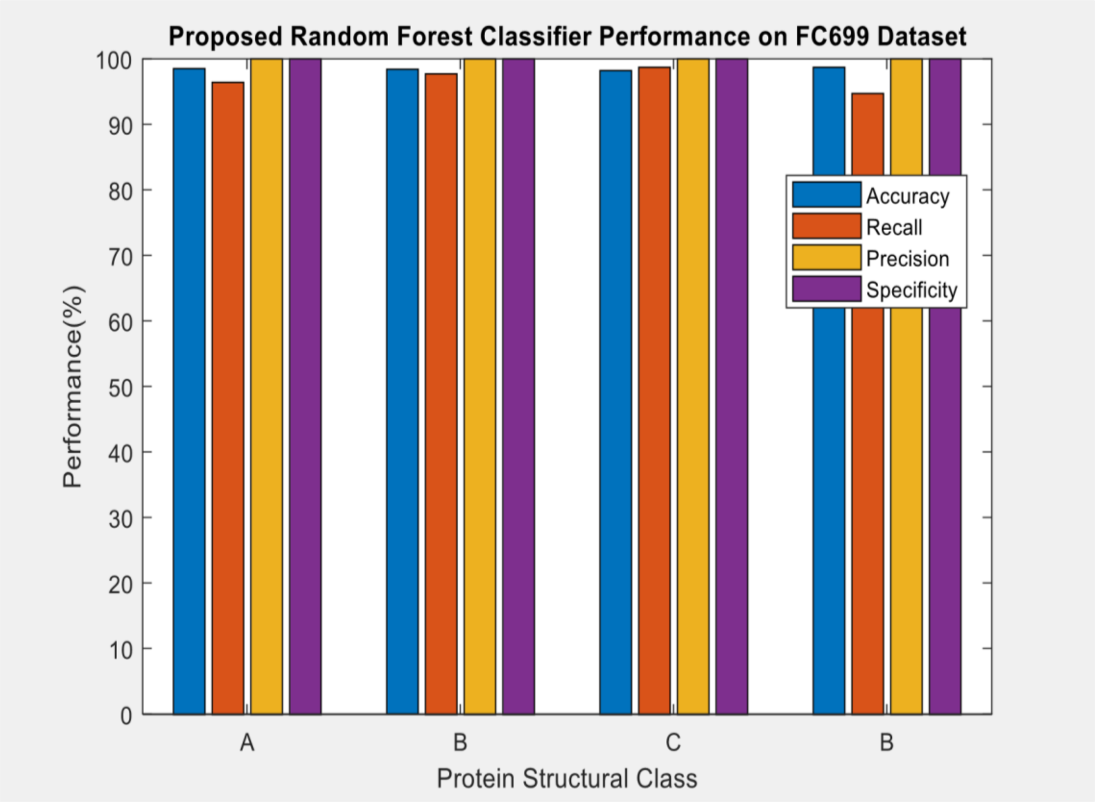


Figure 4.15: A combined model to measure different parameters for all classes

Figure 4.15 provides the performance values of different parameters (accuracy, recall, precision, and specificity) accomplished by the proposed RF classifier for the FC699 data set for different structural classes of Protein.

4.10 Parameters used for Hybrid PSO-GSA Algorithm

Table 4.9 provides the parameters of Hybrid PSO-GSA algorithms for clustering of protein structure classes. Here, C1, C2 are acceleration coefficients, W is inertia weight (PSO), G_0 is used for controlling the search accuracy for GSA algorithm.

Table 4.9: Parameters of Hybrid PSO-GSA algorithm for clustering of protein structure

Sr.No.	Parameter	PSO-GSA
1	Population size	50
2	Iteration	100
3	C1	2
4	C2	2
5	W	0.72
6	G0	1

4.11 Summary

In this chapter, different structural classes of protein are classified to make an understanding of different problems like folding and protein structure prediction, etc. Clustering is performed using the K-mean algorithm. In the current work a random forest (RF) classifier is proposed which is compared with conventional classifiers like SVM, Ada boost, RF, etc. in terms of accuracy for all four classes of protein. The accuracy of the proposed RF classifier is much higher than other classifiers. Also, the values of performance parameters like accuracy, recall, precision, and specificity are measured for different classes of protein. A Hybrid PSO-GSA algorithm was analyzed and its different parameters are analyzed for the classification of protein structure. As suggested in the literature, the proposed hybrid PSO-GSA algorithm has proved to achieve better results as compared to single algorithms.

The accuracy though significant yet can be further improved by integrating more precise classifiers and by performing more intense pre-processing. Considering these key points as the motivation, another prediction model with altogether different architecture has been proposed in Chapter 5.

Chapter-5

Proposed Hybrid CNN+BiLSTM approach

5.1 Overview

Literature survey demonstrated that the clustering-aided approach can contribute to improving the classification rate for predicting structure. Classification of the protein is based on four different classes A (All A (All α), B (All β), C (α+β), and D (α/β). Particle swarm optimization (PSO) and Firefly algorithm (FFA) are used pre- processing of the data. Whereas Convolution neural network, are used for classification individually as well as it is combined with Bidirectional long short term memory for predicting different classes of the protein structure. The performance of CNN classifier, CNN+BILSTM classifier is compared with SVM classifier. For all the performance parameters the classifier used gives very good values for Class A, B and somewhat lower for class C, D although better values has been gained as compare to other classifiers. For feature selection, a hybrid PSOFFA algorithm has been used. Scoring spaces and fitness values are used for the evaluation of all the classes of protein secondary structure.

Two models had been developed one with CNN classifier only and another with a combination of CNN+BILSTM classifier.

5.2 Introduction

Sub-cellular localization of protein structure is attempted by numerous researchers by using several techniques of deep learning and machine learning. Most of these researchers have classified the handcrafted image features of protein using CNN techniques like Resnet, Inception, and Vgg16. A lot of time is taken by these networks for training and considerable memory is used to store such networks. In most of the previous attempts Sub-cellular localization of the protein, chiefly machine learning approaches are utilized.

Five different categories (Hoechst, Giantin, NOP4, LAMP2, and Tubulin) of protein are classified by Boland et al. [5.1] using several features like Haralick texture and Zernike moments provided to the neural network (88%) and classification tree (66%). Boland and Murphy [5.2] localized the protein structure in ten distinct cell organelles utilizing HeLa dataset and utilizing different features like Haralick grain, SLF (Subcellular location feature), and Zernike moments fed to the neural networks (83%). Multi-resolution (MR) decomposition is carried out by Chebira et al. [5.3] which is followed by the processes of feature extraction and then the classification of images is done for every MR space. Accuracy of 95% was achieved with a 2-D Hela dataset utilizing the NN classifier for extracting 26 different features. Hung and Murphy [5.4] performed a comparison of five different techniques of ML and observed that the Ada boost technique of optimization of neural networks for 2D Hela images provides 88.2% accuracy. Nanni and Lumini [5.5] obtained 85% accuracy with the Hela dataset utilizing the SVM technique on Invariant digital patterns. Although the problem of protein localization was tackled successfully using machine learning methods for extracting discriminate features from different images. Litjens et al. [5.6] emphasized that the tendency of relating CNN for classifying microscopic descriptions is growing over the years. Different steps of feature extraction are reduced using deep learning and the system is allowed to learn features of the image by itself. Kraus et al. [5.7] utilized eleven layers CNN model (DeepLoc) to classify budding yeast cell images of proteome into fifteen different categories and achieved 84% accuracy. Parnamaa and Leopold [5.8] trained neural network (Deep Yeast) for classifying fluorescent protein with sub-cellular localization and achieved 91% accuracy. Liimatainen et al. [5.9] trained a Fully Convolutional Network (FCN) to detect protein in thirteen different cell organelles for Human Protein. Xiao et al. [5.10] utilized transfer learning to classify deep yeast protein images for depicting ten different classes. Eleven layers of Vgg and Resnet were trained and an accuracy of 87% and 88% were obtained for these two datasets respectively. Pre-trained networks such as InceptionV3, ResNet50, and InceptionResnetV2 were applied by Kensert et al. [5.11] for classifying mechanism of action datasets with 95-97% accuracy. Thus organelle proteome was efficiently classified by CNN. Training of CNN can be performed using fine-tuned or scratch as per database size. Human protein can be easily classified into major cell compartments. There are limited cell organelles and classified for obtaining single-cell images. Machine learning and feature extraction techniques are used to obtain excellent results. Results can be further improved using image resizing and cropping. But for protein structure learning, only a few CNN models are used to date.

5.3 Native Conformations of Proteins

Proteins are considered molecular instruments that are used to express genetic information. Protein is created by the human body using data received from human Deoxyribonucleic Acid

(DNA) that is composed of a linear chain of deoxyribonucleotides. DNA codes are used for producing a protein with a respective linear chain of amino acids. This resulted in the folding of the protein into a meticulous 3D shape which is called native conformation. The 3D structure of protein also called conformation is accountable directly for its operation [5.12]. Proteins are created from naturally-occurring similar sets of twenty amino acids. From these different combinations, a cell can be used to produce proteins with remarkably different activities and properties [5.13]. Based on their side chain properties there are five main classes of amino acids: (1) hydrophobic (water-hating or non-polar); (2) hydrophilic (water-liking or polar); (3) aromatic; (4) negatively charged; (5) and positively charged [5.13]. Protein's shape is specified by a sequence of its amino acid. To obtain an accurate protein fold an important role is played by the cellular environment. The shape of a protein can be determined by the hydrophobic force of clusters. Alberts et al. [5.14] described that hydrophobic molecules of protein are liable to be enforced together in a liquid environment to minimize the effect of hydrogen-bonded networks on different molecules of water. So, non-polar side sequences in proteins tend to bunch in the inside of the molecule, whereas the polar groups are likely to be arranged outside of the molecule. Therefore, hydrogen bonds can be formed with the combination of water and polar molecules of protein. Figure 5.1 illustrates how protein is folded into its compact conformation. It is noted that hydrophobic core regions are established in the inside of protein whereas hydrophilic amino acid is wrapping the interior hydrophobic acid. Currently, there are two main methods for predicting the neighboring 3D structures for protein, i.e. using an X-Ray Crystallographer (XRC) and a Nuclear Magnetic Resonance (NMR). Although, XRC is a costly technique concerning time and economy, yet during the crystallization process of protein, the problem can occur, and there is a possibility that the final conformation obtained may not be the native one. NMR is the most recent method to predict the protein structure and is not restricted to the number of molecules to be crystallized.

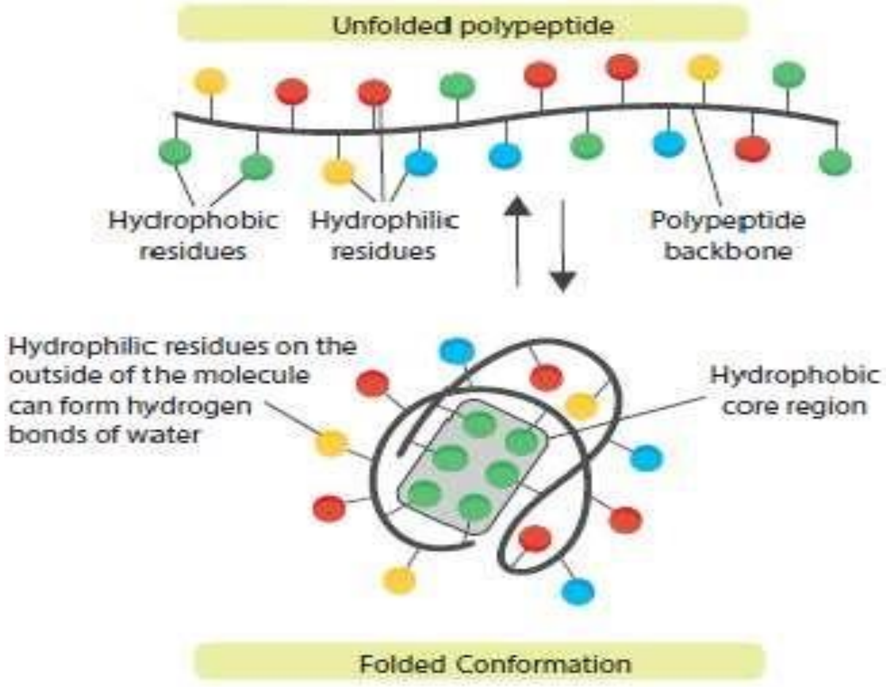


Figure 5.1: Visual representation of protein folding into a compact conformation Src.: © 2021 Quizlet Inc.

However, as with the case of XRC, NMR also presents a small amount of uncertainty in predicting the 3D protein structure. Furthermore, significant human efforts, as well as vastly equipped laboratories, are required in both processes. In the current scenario, studies are involved in silico methods for predicting the native protein structure with an aim for reducing the gap between sequence and structure, the economic cost, and time efforts. The PSP is a problem to find the protein's native structure, with a known sequence of several amino acids [5.13, 5.15]. Computational methods for approaching the PSP are divided into three major categories: (1) comparative modelling or homology, (2) ab-initio, and (3) fold recognition or threading.

5.4 Methodology

In the first step, dataset information is extracted from excel of the 25PDB dataset. After this pre-processing of data is performed where the dataset is refined and then training and testing modules are separated.

In the second step, sequence alignment is applied to a secondary protein structure.

The next feature selection technique is applied to a secondary protein structure. A hybrid model using PSO and Firefly optimization is utilized for feature selection.

In the fourth step, the CNN layer is initialized for the training of the CNN model.

After the initialization of CNN, it is trained with different classes of proteins, and the secondary structure of a protein is predicted.

In the last step parameter performance of the projected CNN model is assessed.

Different training options for CNN models are highlighted in Table 5.1. Max Epochs, Learn Rate Drop Factor, Initial Learn Rate, Learn Rate Drop time, and Mini Batch Size is used to train the currently utilized CNN model.

Table 5.1: Hypermeters for Model Tuning

Sr. No.	Training Option	Parameter Value
1	Max Epochs	100
2	Learn Rate Drop cause	0.1
3	Learn Rate Drop Time	20
4	Initial Learn Rate	0.001
5.	Mini Batch Size	8

The flowchart of the present study is shown in Figure 5.2 and its various components explained above.

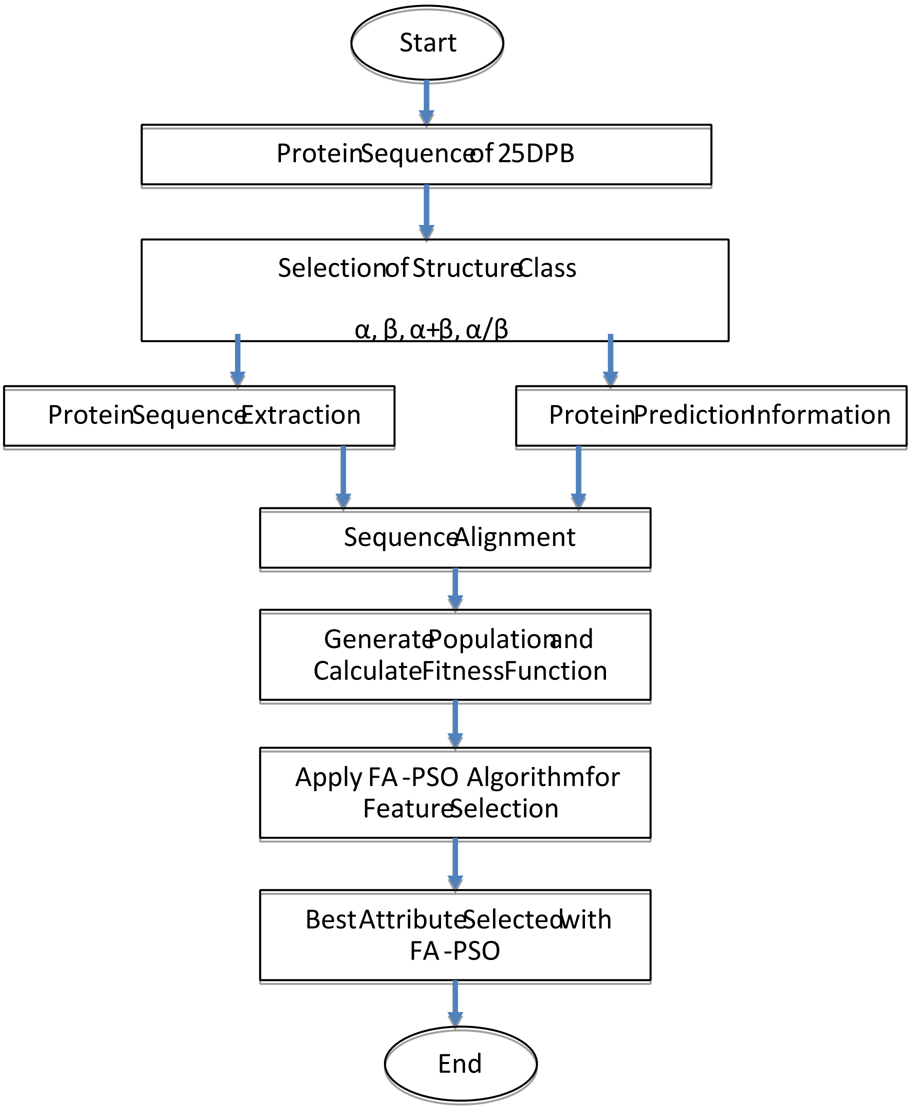

Start
ProteinSequenceof 25DPB
Selectionof StructureClass
α, β, α+β, α/β
ProteinSequenceExtraction
ProteinPredictionInformation
SequenceAlignment
GeneratePopulationand CalculateFitnessFunction
Apply FA-PSO Algorithmfor FeatureSelection
BestAttributeSelectedwith FA-PSO
End

Figure 5.2 Flowchart for the present study

5.4.1 Firefly algorithm (FFA)

A Firefly algorithm is proposed by X-S Yang is inspired by nature, a multimodal metaheuristic algorithm, and is based on the blinking performance of fireflies [5.16]. Unique tiny rhythmic flashes are produced by every species of fireflies and the process of producing flashes is called bioluminescence. Firefly algorithm is designed based on three ideal principles: (i) as all fireflies are having a unisex nature, so a firefly can be attracted toward another firefly despite their sex factor. (ii) Attractiveness is directly related to the luminance level of the fireflies, therefore the less bright firefly is always attracted by a bright firefly. (iii) An objective function is used to calculate the brightness level in a firefly [5.17]. Brightness and attractiveness are directly proportional to distance, so if the distance is increased then both these properties are decreased [5.18]. If any firefly does not find another firefly in its surrounding space, then its movement will be random in any direction. Flashing light is the main property in a firefly algorithm [5.19] that is accountable for attracting the neighbouring fireflies. Fireflies can charge and discharge their light at a regular interval, thus they are having an oscillatory behavior. Generally, fireflies stay mostly active for the period of the night times of the summer season [5.20]. When any firefly comes in contact with a neighbouring firefly, mutual coupling occurs between both the fireflies.

Any male firefly tries to attract the neighbouring female firefly through its signals [5.21]. In response to these signals by the male firefly, the female firefly discharges its flashing lights. Consequently, distinct illuminating patterns of male, as well as female fireflies, are produced to encode the information like sex and identity of the species [5.22]. Generally, a female firefly can be more attracted to any male firefly with brighter illuminating light. Blinking intensity is inversely proportional to the source distance of fireflies. In some unique cases, a female firefly is unable to differentiate between the weakest and strongest flash, which are generated by distant or neighbouring male fireflies respectively.

A firefly's brightness can be established by an objective function. Firefly attractiveness directly depends on light intensity perceived by neighbouring fireflies; variation in attractiveness (β) can be defined concerning distance (r) and is provided by relationships in equation 1 as:

$$\beta = \beta_0 e^{-\gamma r^2} \quad (1)$$

Here $\beta_0$is the attractiveness value at distance r = 0. A particular firefly's movement towards a brighter firefly 'j' can be determined by equation 2 as:

$x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r^2} (x_j^t - x_i^t) + \propto_t \in$ (2)

In equation 2, the second term is because of attraction. The last term is because of randomization and $\propto$ is the parameter of randomization, also $\in_i^t$ is a vector with all the numbers as random which is drawn using Gaussian distribution function at any time (t).

The case with β_0= 0 is considered a random walk. Additionally, the randomization parameter $\propto_t$ can be expanded to other distribution functions like Levy flight function.

Algorithm 5.1 Hybrid of PSOFA Optimization Algorithm for Feature Selection

Input: **n** number of population, **t** maximum iteration, **d** number of attributes, **C1** and **C2** are constants, **w** inertia weight, **γ** is light absorption coefficient, and **β0** is the attractiveness.

Notations: $gbest$ is global best fitness, $pbest$ is local best fitness, r is the distance between two fireflies

Output: $\boldsymbol{b_a}$ is the best attribute for training

```
Initialization n number of  $\boldsymbol{x_i}$ positions with several d
Initialization of velocity,
for i=1 to t
Calculate fitness function of each population by equation
            $Std = \sqrt{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$
    $fitness = min(\frac{1}{n}\sum_{i=1}^{n} Std)$
for j=1 to n
Update β by equation
  $β=β0e^{-\gamma r^2}$
 $r = (x_i - x_j)$
Update $x_i$ positions by equation
```

```
$velocity = w + velocity + c1 * rand * \beta * (pbest - x_i) + c2 * rand * (gbest - x_i)$
$x_i = velocity + x_i$

end for
end for
```

5.5 CNN Classifier

CNN architecture is used to map the protein chains into folds and is provided in figure 5.3. It consists of a total of fifteen layers including one input layer, ten convolution layers,

One pooling layer, one hidden layer, and one flattening layer which is shown in figure 5.4. Softmax function is utilized and applied to the output layer nodes for predicting the fold probability of proteins. Positional information of protein sequences is represented by L × 45 input numbers of a protein sequence having variable length L. CNN network accepts variable sequence protein features as input, which are changed into hidden features using ten hidden layers of CNN. Two windows of size 6 and size 10 are used. CNN can alternate between the pooling and convolution layers and the output can be available at fully connected layers which include nonlinear classifiers, like Softmax classifier, used to estimate the conditional probability for each class. Nonlinearity is introduced in CNNs by using rectified linear units (ReLU) which is an activation function with nonlinear transformations resulting in 10 x L hidden features.

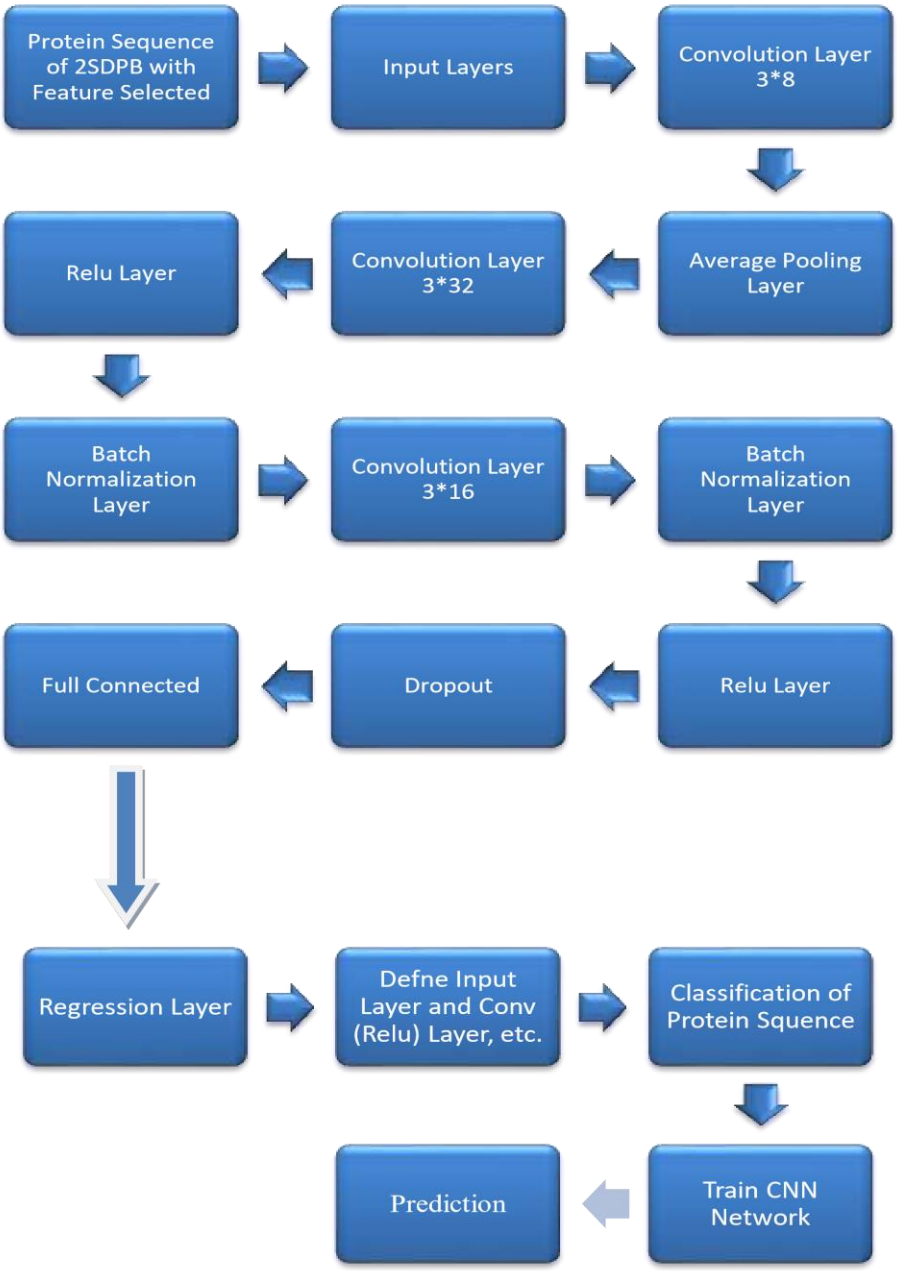


Figure 5.3 the architecture of Convolutional neural network (CNN) for classification

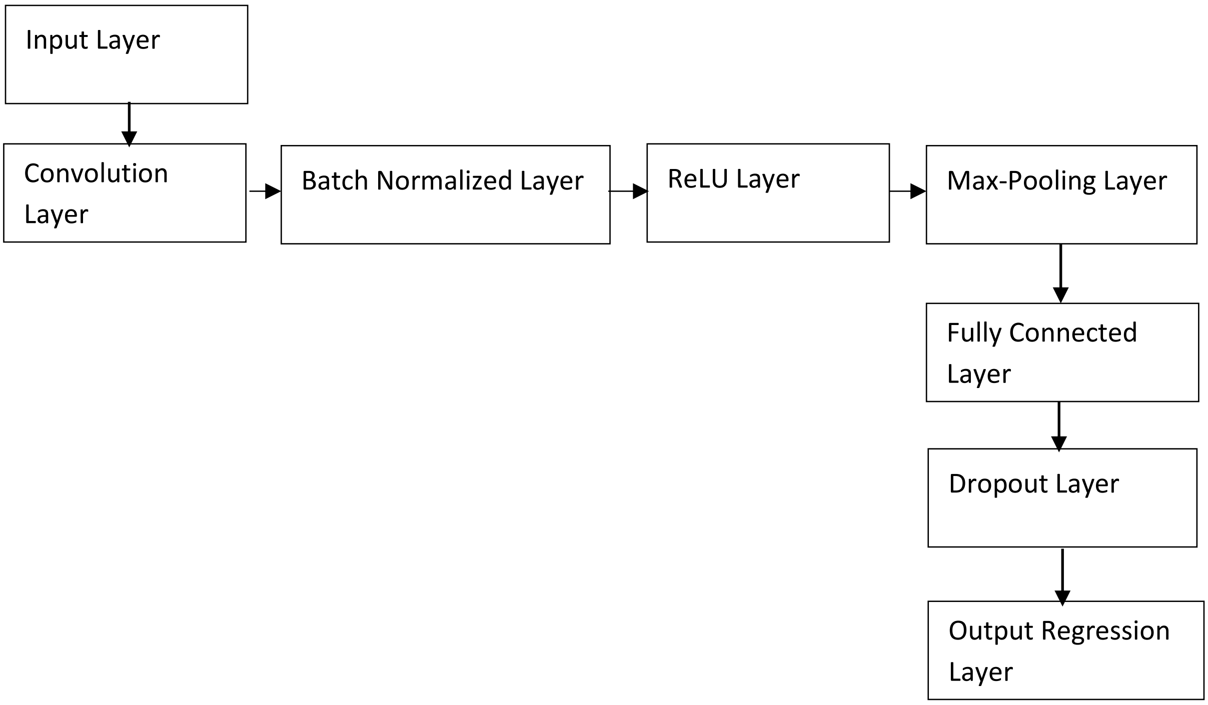


Figure 5.4. Detailed Schematic for CNN technique

5.6 BILSTM

Bilstm consists of two LSTM one taking input in the forward direction, another taking input in the backward direction. Bilstm is a sequence processing model, due to its processing capability it is very efficient for protein secondary structure prediction. Since sequence information is not lost and dependability can be maintained. It had been used by many authors in combination with another intelligent technique, Hattori et al., who used deep recurrent neural network with BILSTM for protein secondary structure prediction but they can give only 68% accuracy.

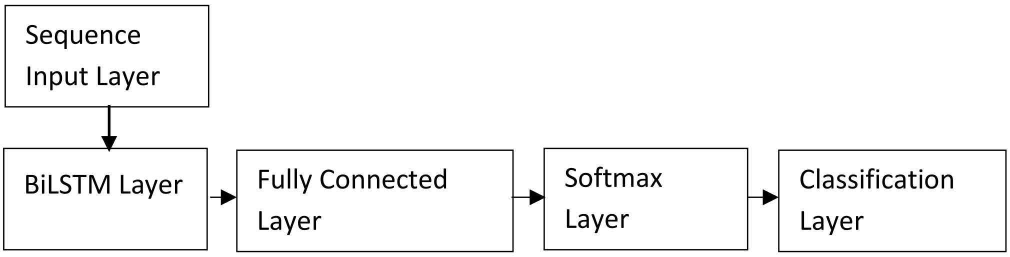


Figure 5.5: Detailed Schematic for BiLSTM technique

BILSTM automatically discover the features without any human direction. Where the learning problem is sequential BILSTM is used, each layer of BILSTM gives bidirentional long term dependencies between sequence data.

5.7 CNN+BILSTM

Pre-processing, feature selection, feature extraction are similar to the method we proposed for CNN. The difference here is the output of CNN classifier and BILSTM are combined to form a hybrid method as shown in figure 5.6. The number of layers used in CNN is 18 and the number of layers used in BILSTM is 5. The different layers used are shown in figure 5.5. The layers are the sequence input layer, BILSTM layer, fully connected layer, softmax layer, and a classification layer.

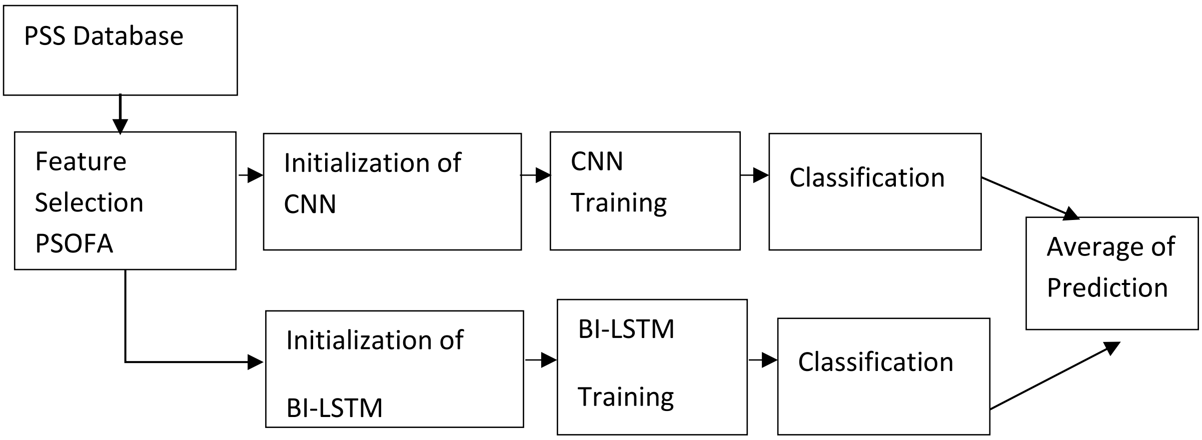


Figure 5.6: Block-Diagram for CNN+BiLSTM technique

5.8 Benefit of hybridization

- Hybridization of Firefly algorithm (FFA) and Particle Swarm Optimization (PSO), FFA algorithm gives optimal solution by the light intensity attraction which depends on random direction of search and move into local optima can’t reach global optima . With PSO operator and modified light intensity attraction step the local search is performed [5.23].

5.9 Performance Evaluation

For evaluating the quality of classification, four different parameters are used frequently, which include individual sensitivity (denoted as ‘Sens’), specificity

(Denoted as ‘Spec’), Matthew's correlation coefficient (denoted as ‘MCC’), and overall accuracy (denoted as OA) for each structural class over the entire dataset.

Equations (3-6) are used to represent these parameters:

$$Sens_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|C_j|} \qquad (3)$$

$$Spec_j = \frac{TN_j}{FP_j + TN_j} \qquad (4)$$

$$MCC_j = \frac{(TP_j * TN_j) - (FP_j * FN_j)}{\sqrt{(FP_j + TP_j)(TP_j + FN_j)(FP_j + TN_j)(TN_j + FN_j)}} \qquad (5)$$

$$OA = \frac{\sum_j TP_j}{\sum_j |C_j|} \qquad (6)$$

Where Cj is the structural class, TPj is true positives, TNj is true negatives, FPj is false positives, and FNj is false negatives.

5.10 Results & Discussions

This section contains the outcome obtained after executing the algorithms used in experiment II as mentioned in the above section of this chapter. The results are executed in MATLAB for scoring space and winning path for different values of sequences.

Figures 5.7 (a-d) provides the scoring spaces for protein structures of Class A, B, C, and D. Scoring spaces are heat maps used to display the best score for the entire fractional alignments of both sequences. The best score is represented by a pair of two subsequences i.e. Seq1 (s1:n1) and a Seq2 (s2:n2). Here n1 is the position of Seq1, n2 is a position of Seq2, s1 is a Seq1 position ranging between 1:n1, and s2 is a Seq2 position which is ranging between 1:n2. The best score for a pair of the definite subsequence is calculated by scoring the entire possible alignments for given subsequences by accumulating gap and match penalties.

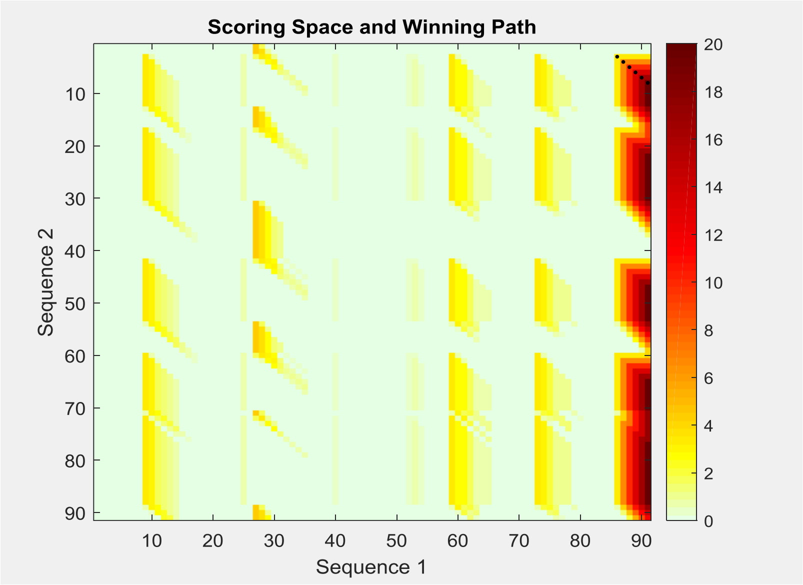


Figure 5.7 (a): Scoring Space and Winning path for sequence 1 and 2 for protein structural class A

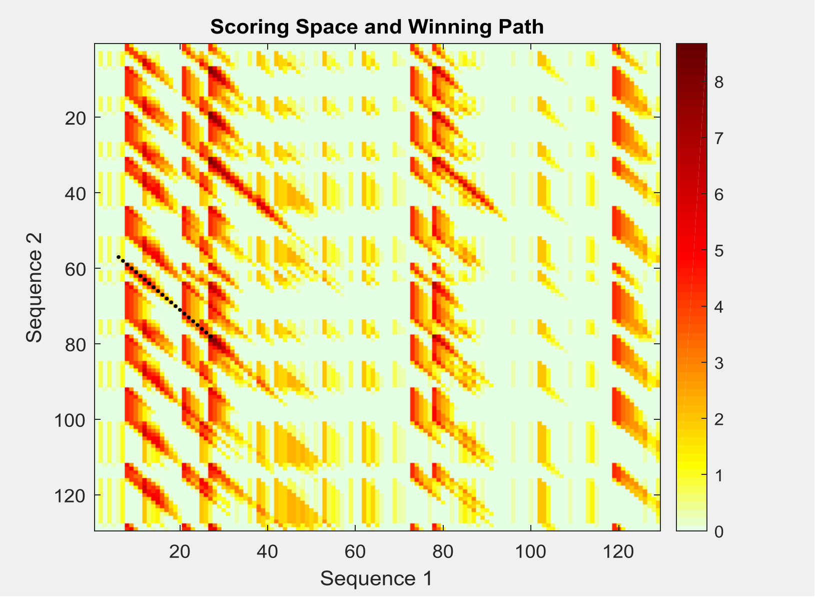


Figure 5.7 (b) Scoring Space and Winning path for sequence 1 and 2 for protein structural class B

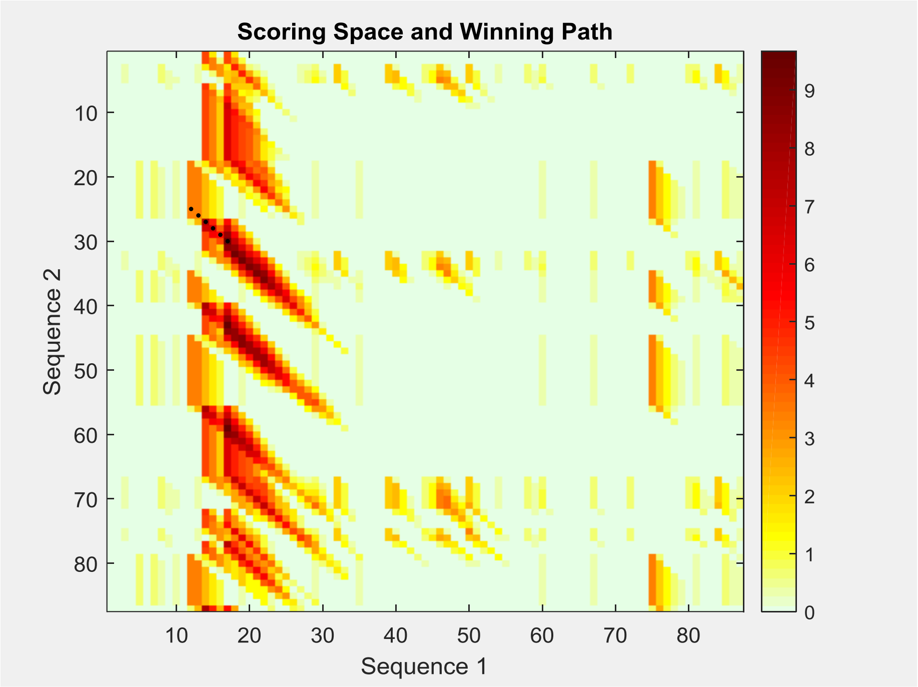


Figure 5.7 (c): Scoring Space and Winning path for sequence 1 and 2 for protein structural class C

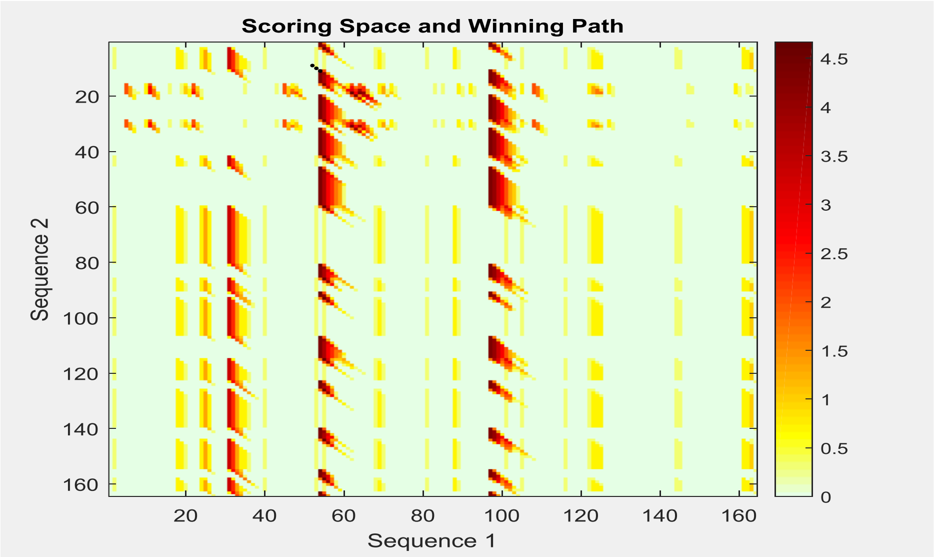


Figure 5.7 (d): Scoring Space and Winning path for sequence 1 and 2 for protein structural class D

Black dots in the scoring space represent the winning path. Positions pairing is illustrated as the best possible local alignment. Also, the color of the last point in the lower right portion of the winning path signifies the alignment score of best local for these two sequences.

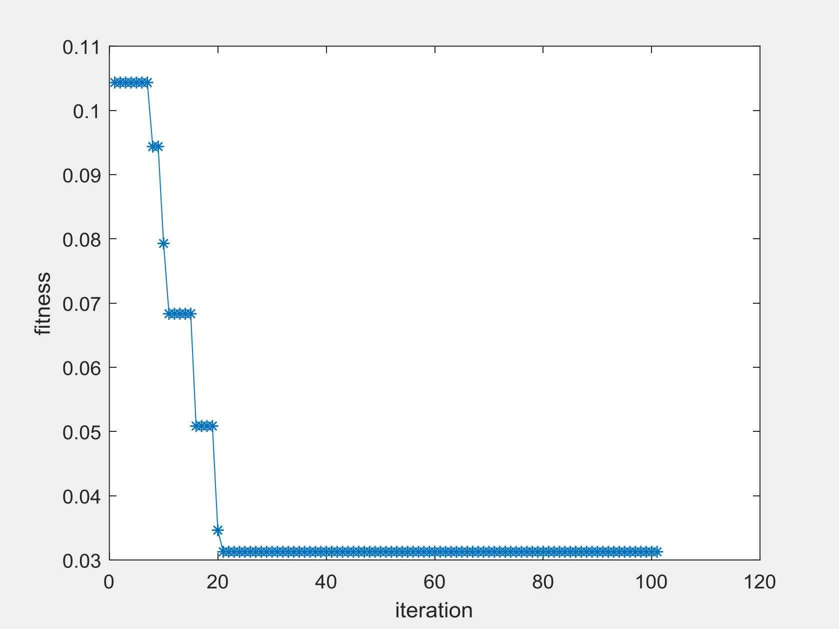


Figure 5.8 (a): Fitness value outcome for protein structural class A

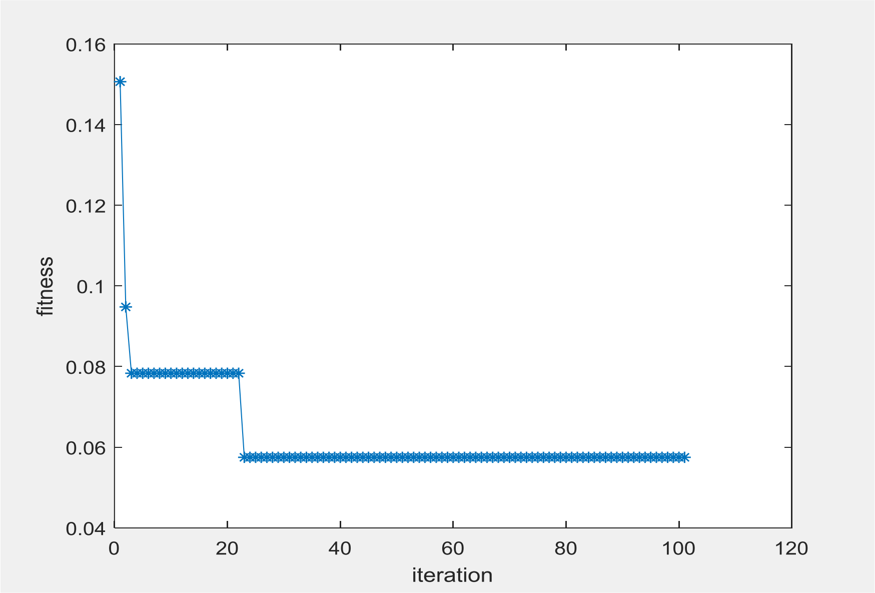


Figure 5.8 (b): Fitness value outcome for protein structural class B

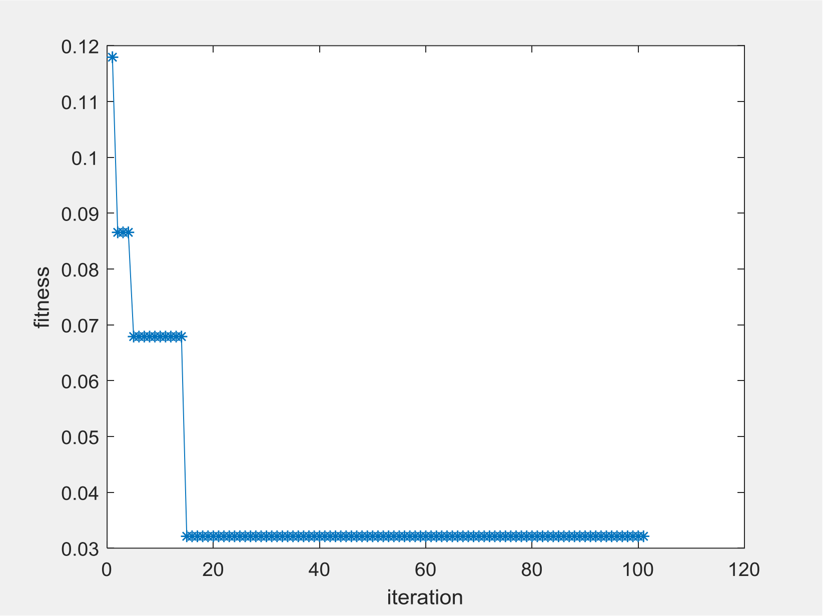


Figure 5.8 (c): Fitness value outcome for protein structural class C

Above figure 5.8 a,b,c and below figure d shows fitness value of selected feature from protein dataset w.r.t different iteration for protein structural class A,B,C and D. Convergence curve is a graphical representation of the evolution of the optimization as a function of the number of individuals evaluated.

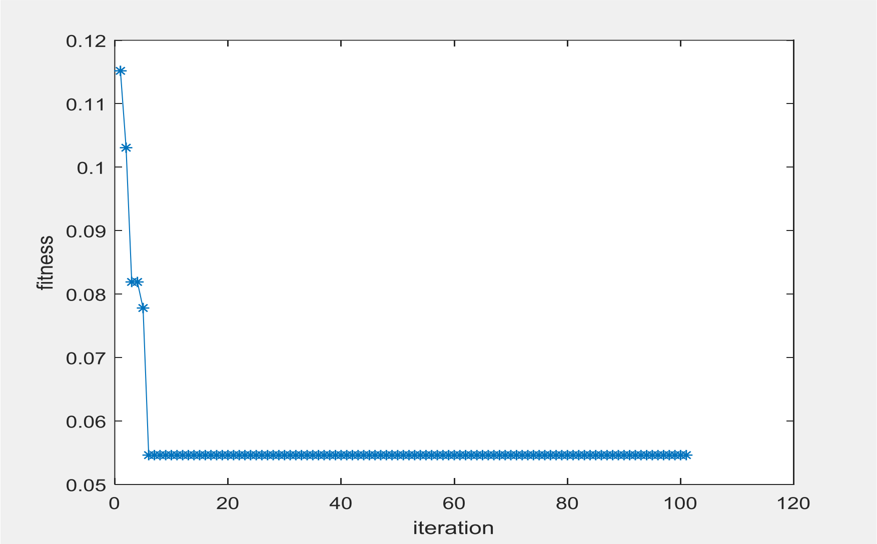


Figure 5.8 (d): for protein structural class D

The convergence history graph lets you know how the optimization problem is converging to the optimal solution. Convergence Curve of Firefly Algorithm shown in figure 5.8.

For evaluating the results of classification the four parameters which are used are sensitivity, specificity, Matthew's correlation coefficient, overall accuracy denoted by Sens, Spec, MCC and OA. These parameters are detailed as follows:

$$\mathrm{Sens}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j} = \frac{\mathrm{TP}_j}{|\mathrm{C}_j|}$$

$$\mathrm{Spec}_j = \frac{\mathrm{TN}_j}{\mathrm{FP}_j + \mathrm{TN}_j} = \frac{\mathrm{TN}_j}{\sum_{k\neq j} |\mathrm{C}_k|}$$

$$\mathrm{MCC}_j = \frac{\mathrm{TP}_j \times \mathrm{TN}_j - \mathrm{FP}_j \times \mathrm{FN}_j}{\sqrt{(\mathrm{FP}_j + \mathrm{TP}_j)(\mathrm{TP}_j + \mathrm{FN}_j)(\mathrm{TN}_j + \mathrm{FP}_j)(\mathrm{TN}_j + \mathrm{FN}_j)}}$$

$$\mathrm{OA} = \frac{\sum_j \mathrm{TP}_j}{\sum_j |\mathrm{C}_j|}$$

In the above formulas, TPj is the number of true positives, TNj is the number of true negativies, FPj is the number of false positives, FNj is the number of false negatives and Cj is the protein in structure class Cj.

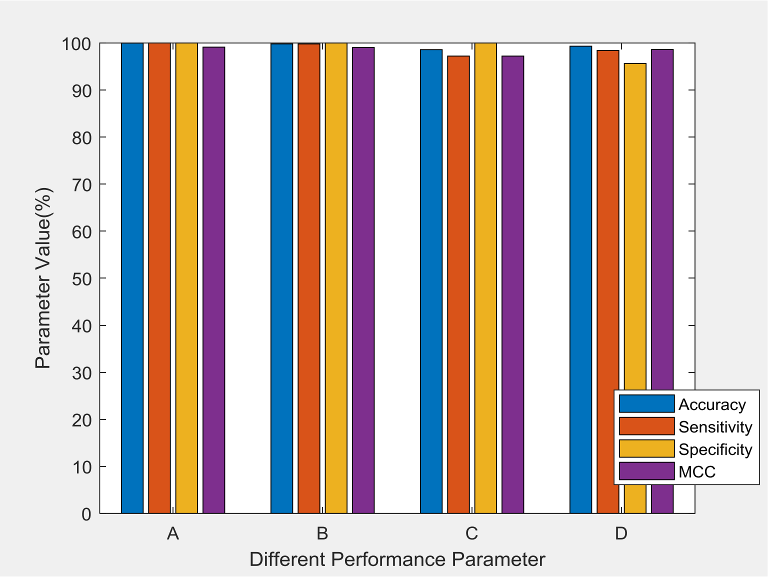


Figure 5.9: Parameter value for proposed CNN

Above figure 5.9 shown performance parameters are accuracy, sensitivity, specificity and MCC. The performance is calculated using the proposed CNN on 25PDB dataset for different protein structural classes A, B, C, D.

5.11 Comparison of Experiment I and II

Particle swarm optimization (PSO) and Firefly algorithm (FFA) are used pre- processing of the data. Whereas Convolution neural network, are used for classification individually as well as it is combined with Bidirectional long short term memory for predicting different classes of the protein structure. The performance of CNN classifier, CNN+BILSTM classifier is compared with SVM classifier. For all the performance parameters the classifier used gives very good values for Class A, B and somewhat lower for class C, D although better values has been gained as compare to other classifiers. For feature selection, a hybrid PSOFFA algorithm has been used. Scoring spaces and fitness values are used for the evaluation of all the classes of protein secondary structure. Convolution neural network classifier is compared with the performance of random forest classifier. All the values of CNN, CNN+BILSTM are much better as compared with RF classifier.

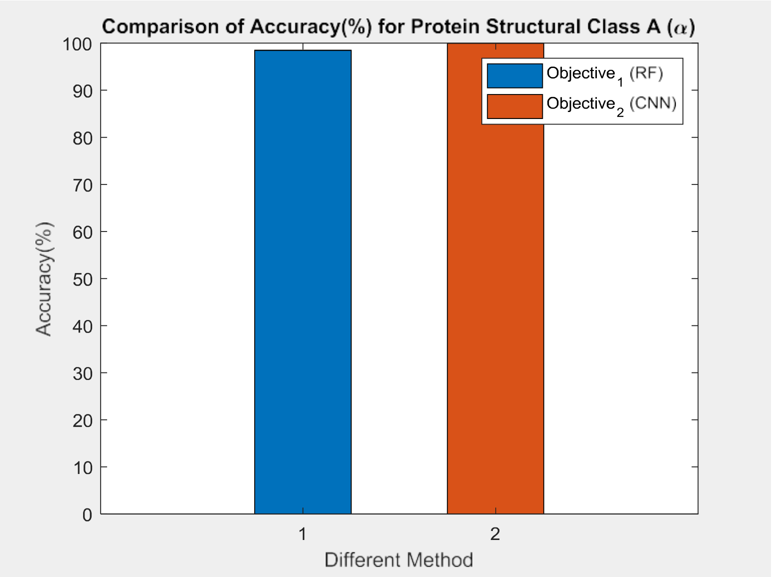


Figure 5.10 (a): Accuracy comparison of both objectives for class A

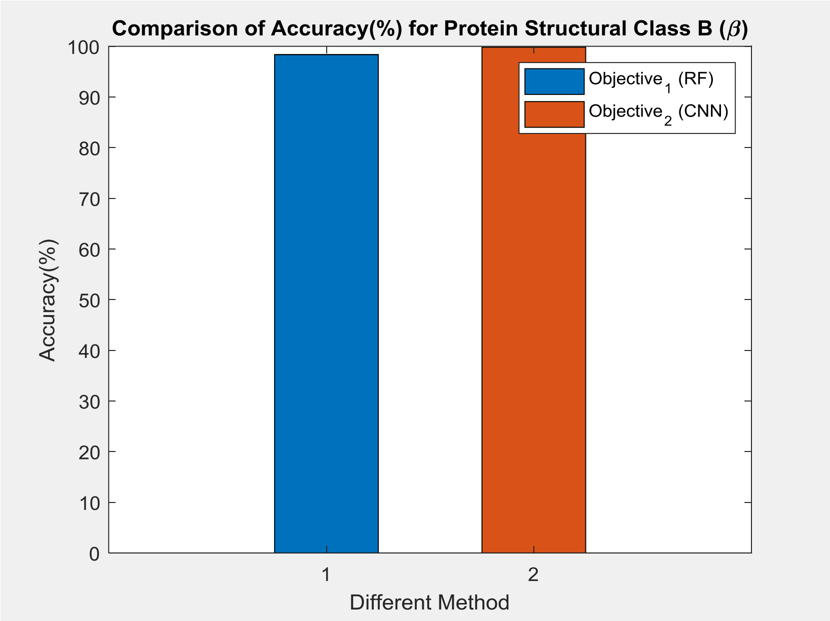


Figure 5.10 (b): Accuracy comparison of both objectives for class B

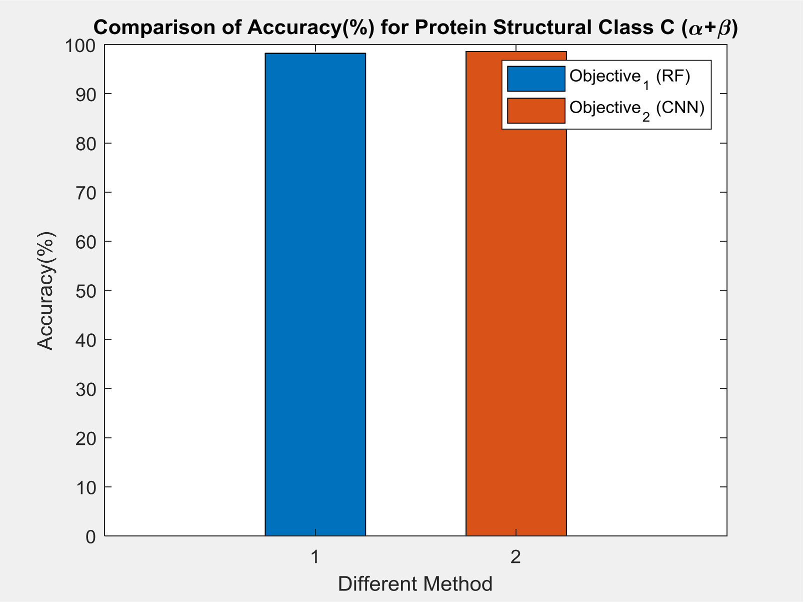


Figure 5.10 (c): Accuracy comparison of both objectives for class C

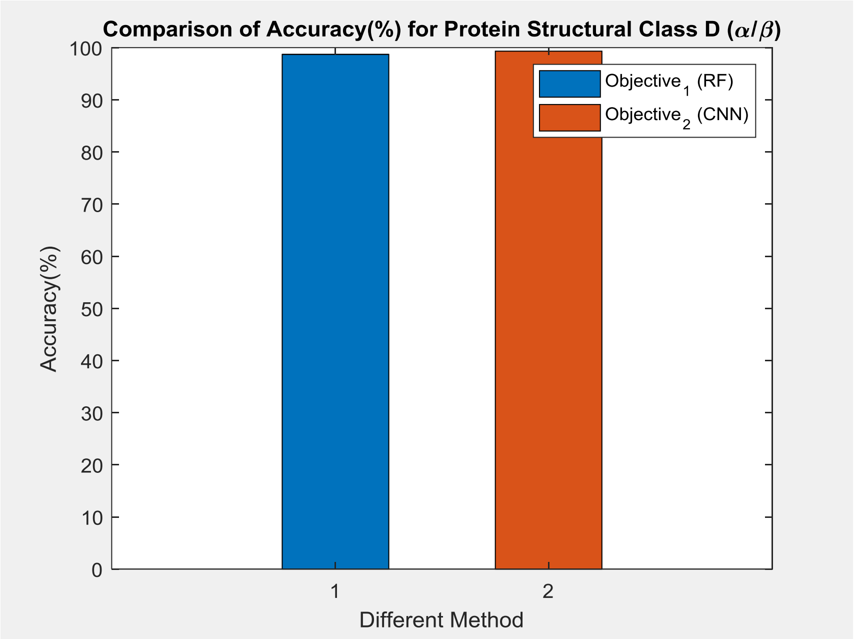


Figure 5.10 (d): Accuracy comparison of both objectives for class D

Above figure 5.10 (a), (b), (c), and (d) gives the comparative results obtained for both the objectives of proposed work in which RF and CNN classifier is used for obtaining the accuracy results.

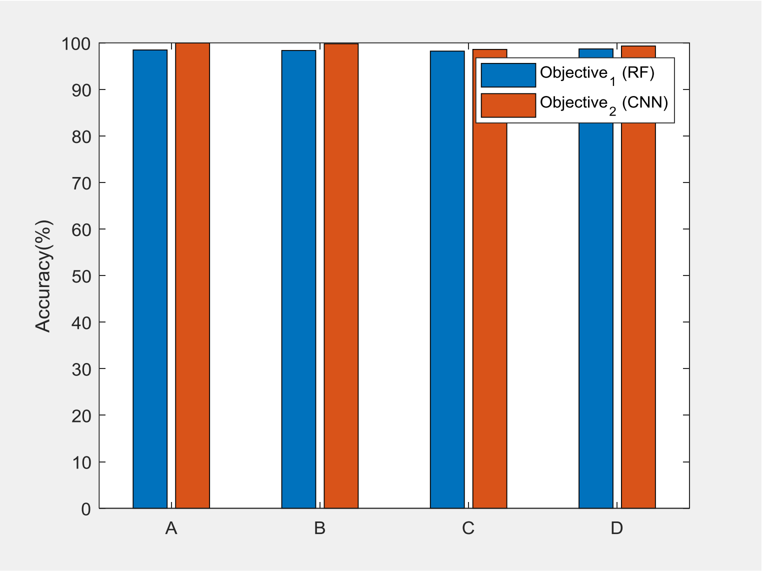


Figure 5.11 Accuracy analysis for objective 1 and objective 2

Sub-cellular localization of protein structure is attempted by numerous researchers by using several techniques of deep learning and machine learning. In the present study deep learning technique of CNN is utilized as a classifier which is compared with RF classifier concerning the accuracy, specificity, sensitivity, and MCC values for all four classes of protein. The accuracy of the CNN classifier is much higher than RF classifier.

Table 5.2: Comparative accuracy results of various classes

Techniques	Accuracy (%)			
	all-α	all-β	α+β	α/β
RF (Objective 1)	98.47	98.36	98.22	98.70
CNN (Objective 2)	99.96	99.96	99.96	99.96

Table 5.2 shows a comparison of our techniques in Objective I and Objective II in terms of accuracy values. The table suggested that the accuracies of all classes are better for the CNN classifier as compared to RF.

Table 5.3 provides a comparison of CNN and SVM techniques in terms of sensitivity values. The table suggested that the sensitivity of classes A and B is nearly 100%. The results indicated that the currently utilized CNN method provides greater sensitivity values for all classes of protein.

Table 5.3 Comparison of sensitivity for SVM and CNN methods for various protein classes

		Sensitivity (%)	
Sr. No.	Classes	SVM	CNN
1	All α (A)	99.77	100
2	All β (B)	99.77	99.81
3	α+β (C)	85.09	97.22
4	α/β (D)	78.64	98.41

Table 5.3 provides a comparison of CNN and SVM techniques in terms of specificity values. The table suggested that the specificity of class A and B, and C are 100% for the CNN technique. The results indicated that the currently utilized CNN method provides greater specificity values for all classes of protein.

Table 5.4 provides a comparison of CNN and SVM techniques in terms of specificity values. The table suggested that the specificity of class A and B, and C are 100% for the CNN technique. The results indicated that the currently utilized CNN method provides greater specificity values for all classes of protein.

Table 5.4 Comparison of specificity for SVM and CNN methods for various protein classes

Specificity (%)			
Sr. No.	Classes	SVM	CNN
1	All α (A)	99.51	100
2	All β (B)	99.42	100
3	α+β (C)	94.59	100
4	α/β (D)	95.45	95.65

Table 5.5 provides a comparison of CNN and SVM techniques in terms of MCC values. The table suggested that the MCC values for class A and Bare nearly 100% for the CNN technique. The results indicated that the currently utilized CNN method provides greater MCC values for all classes of protein.

Table 5.5 Comparison of MCC for SVM and CNN methods for various protein classes

Sr. No.	Classes	MCC%	
		SVM	CNN
1	All α (A)	98.93	99.12
2	All β (B)	98.77	99.05
3	α+β (C)	79.63	97.22
4	α/β (D)	75.10	98.62

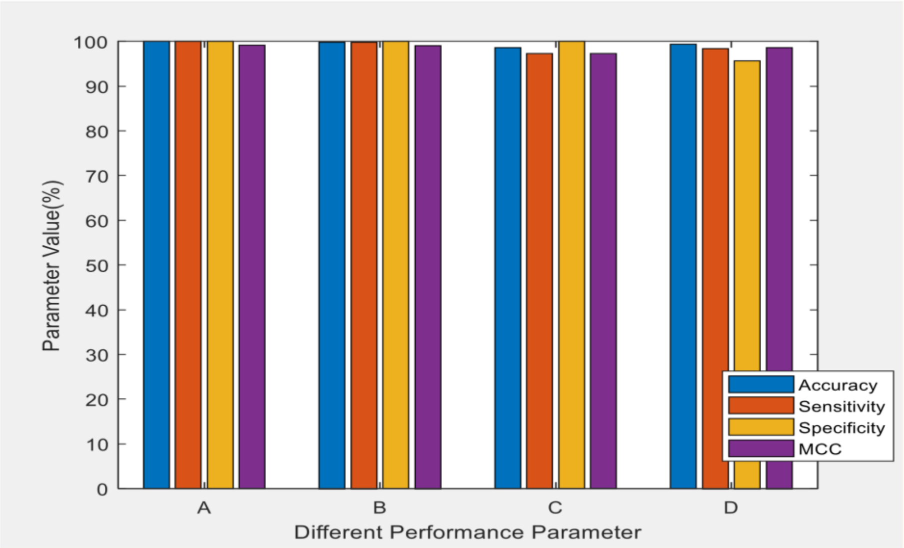


Figure 5.12 Comparison of different parameters for various protein classes

Figure 5.12 combines all the results of tables 5.2-5.5 and provides the values of different performance parameters like accuracy, sensitivity, specificity and MCC calculated using CNN technique on 25PDB dataset for different protein structural classes A, B, C, and D.

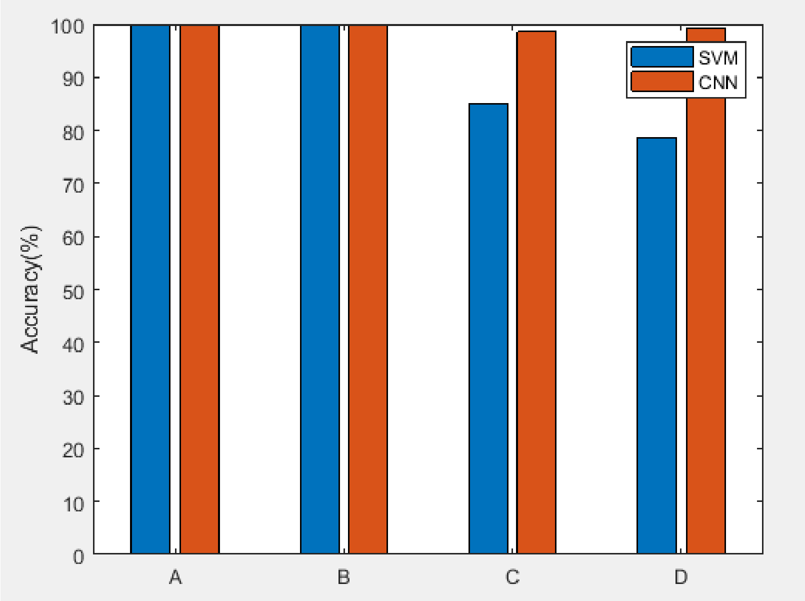


Figure 5.13 Comparison of Accuracy with SVM and CNN models for various protein

Figure 5.13 provides a comparison of accuracy values using SVM and CNN techniques with all the four classes. The results indicate currently utilized CNN technique gives good results for accuracy.

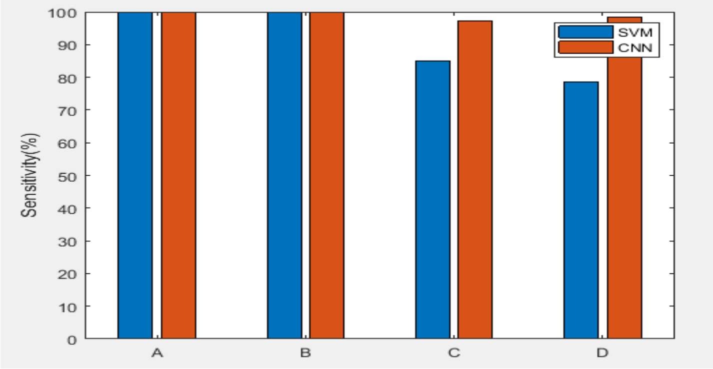


Figure 5.14 Comparison of Sensitivity with SVM and CNN models for various protein classes

Figure 5.14 provides a comparison of sensitivity values using SVM and CNN techniques with all the four classes. The results indicate currently utilized CNN technique gives good results for sensitivity.

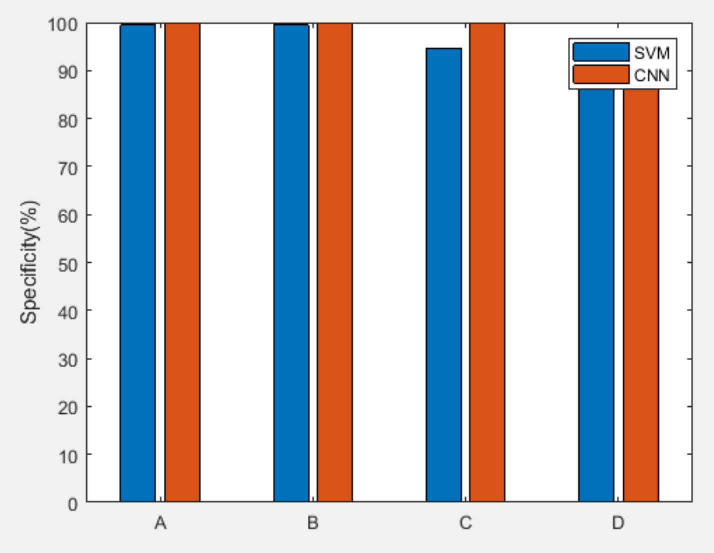


Figure 5.15 Comparison of Specificity with SVM and CNN models for various protein classes

Figure 5.15 provides a comparison of specificity values using SVM and CNN techniques with all the four classes. The results indicate currently utilized CNN technique gives good results for specificity.

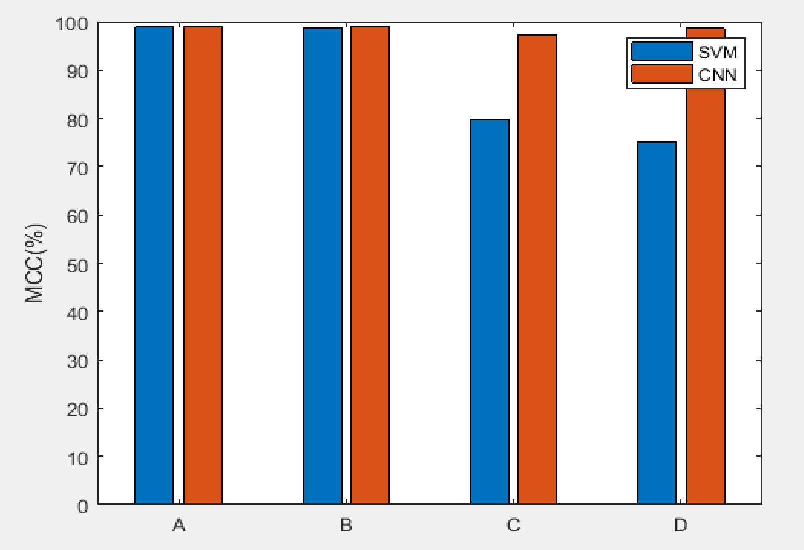


Figure 5.16 Comparison of MCC with SVM and CNN models for various protein classes

Figure 5.16 provides a comparison of MCC values using SVM and CNN techniques with all the four classes. The results indicate currently utilized CNN technique gives good results for MCC.

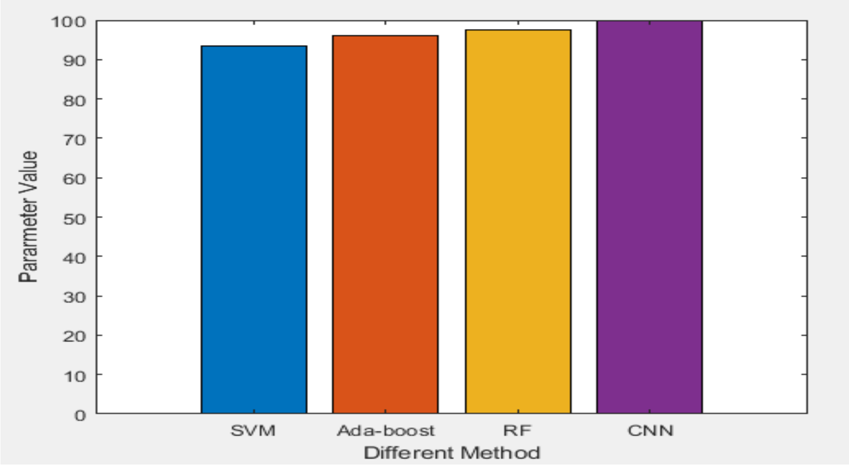


Figure 5.17 Accuracy Comparison of Different methods

Figure 5.17 provides a comparison of different methods used for classification. From the graph, the bar depicts that CNN got the highest accuracy.

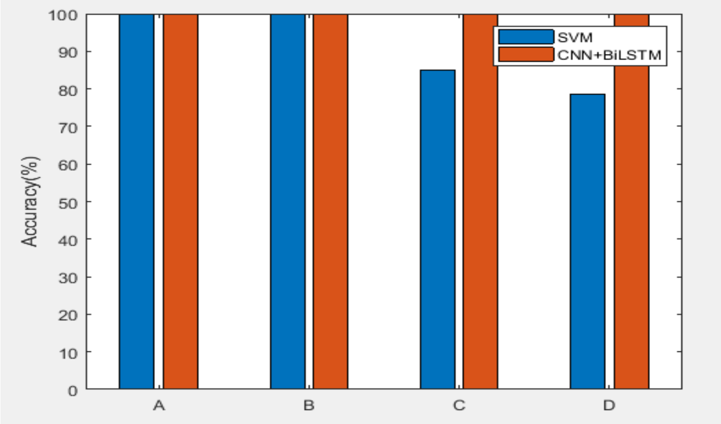


Fig.5.18 Comparison of Accuracy with SVM and CNN+BILSTM models for various protein classes

Figure 5.18 provides a comparison of accuracy values using SVM and CNN+BILSTM techniques for protein structure with all four classes. The results indicate that the currently utilized CNN+BILSTM technique gives good results for accuracy.

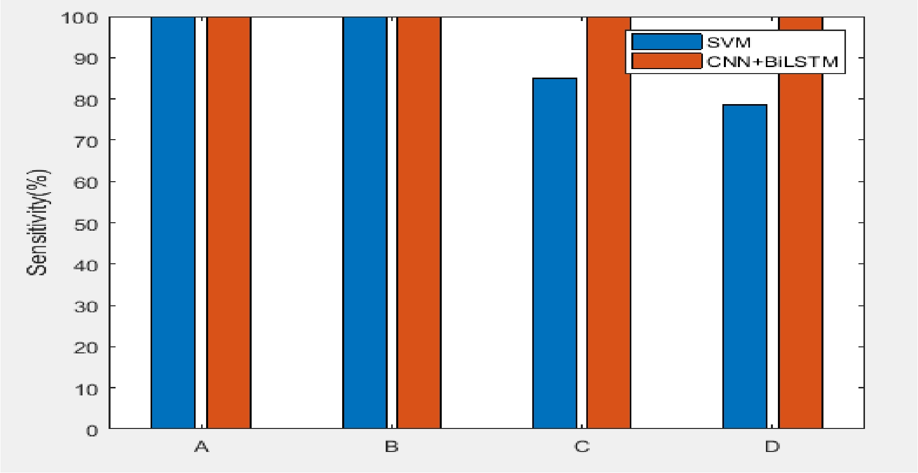


Figure 5.19 Comparison of Sensitivity with SVM and CNN+BILSTM models for various protein classes

Figure 5.19 provides better results in terms of sensitivity. provides a comparison of sensitivity values using SVM and CNN+BILSTM techniques with all four classes. The results indicate that the currently utilized CNN+BILSTM technique gives good results for sensitivity.

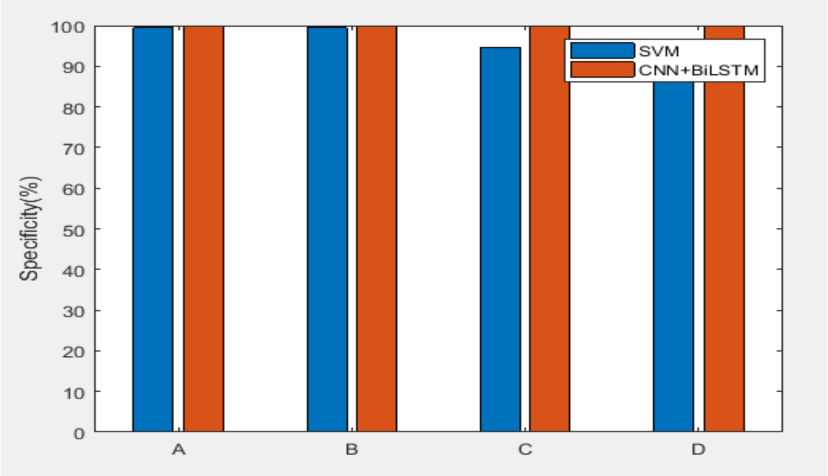


Figure 5.20 Comparison of Specificity with SVM and CNN+BILSTM models for various protein classes

Figure 5.20 provides a comparison of specificity values using SVM and CNN+BILSTM techniques with all four classes. The results indicate that the currently utilized CNN+BILSTM technique gives good results for specificity.

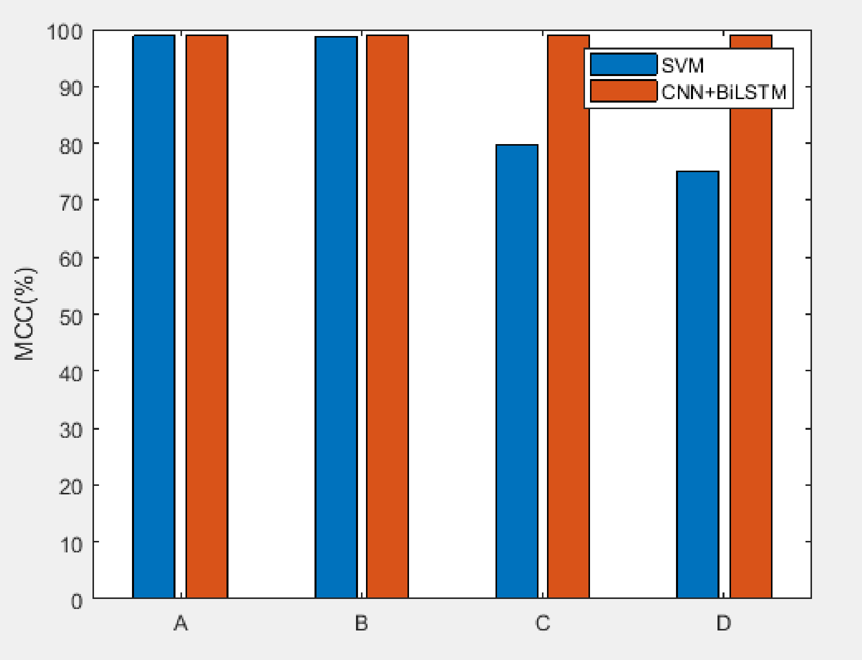


Figure 5.21 Comparison of MCC with SVM and CNN+BILSTM models for various protein classes

Figure 5.21 provides a comparison of MCC values using SVM and CNN+BILSTM techniques for with all four classes. The results indicate that the currently utilized CNN+BILSTM technique gives good results for MCC.

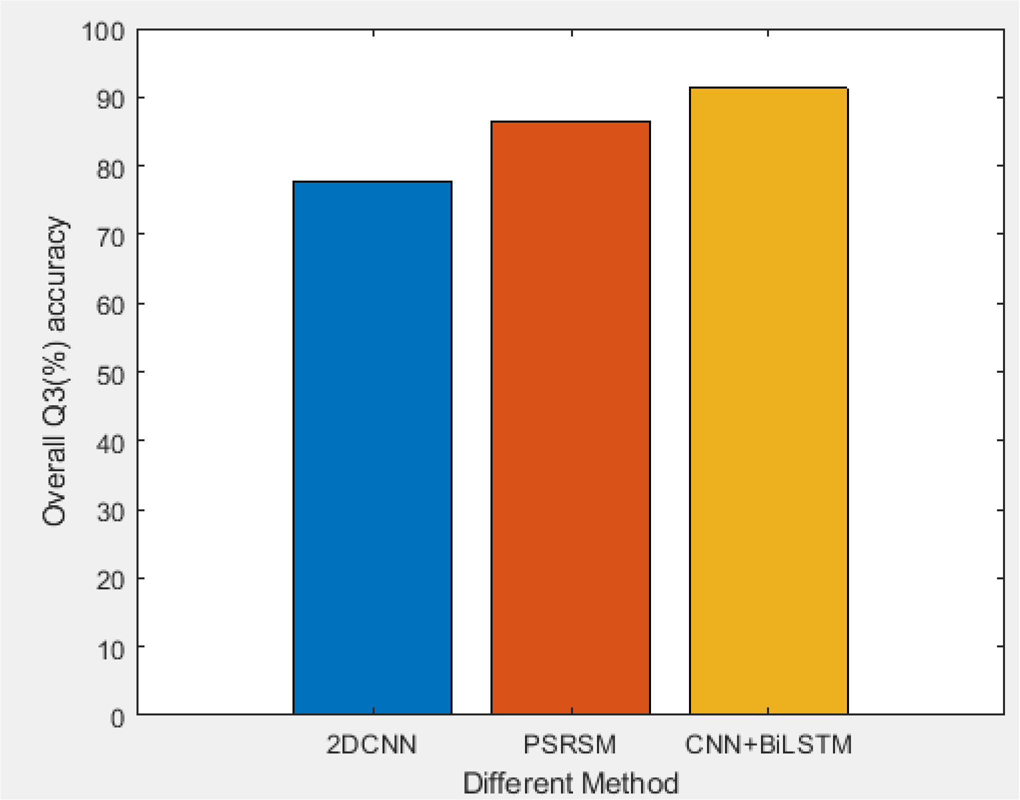


Figure 5.22 Overall Q3 accuracy of different methods

Figure 5.22 depicts the Q3 accuracy for different methods. From the bar, it had been concluded that for the method proposed reached highest accuracy.

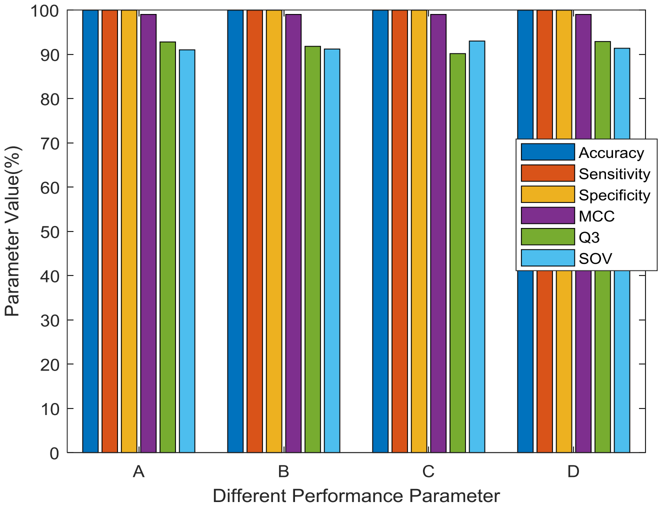


Figure 5.23 Comparison of different parameters for various protein classes

Figure 5.23 combines all the results of tables 5.17-5.20 and provides the values of different performance parameters like accuracy, sensitivity, specificity, MCC, Q3, and SOV calculated using CNN+Bilstm technique on 25PDB dataset for all the protein structural classes.

5.12 Results of CNN+BILSTM on FC699 dataset

The hybrid technique has also been tested on FC699 dataset. The results of the same has been shown in following figures.

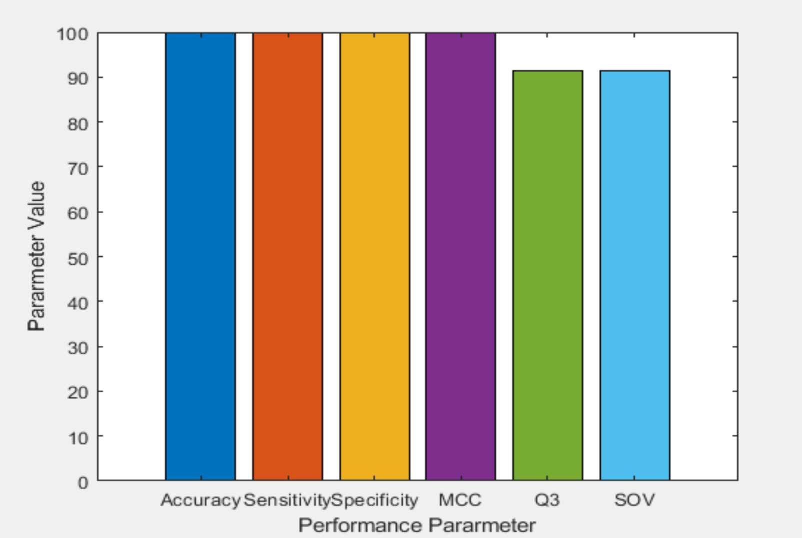


Figure 5.24 Different parameters for protein class A

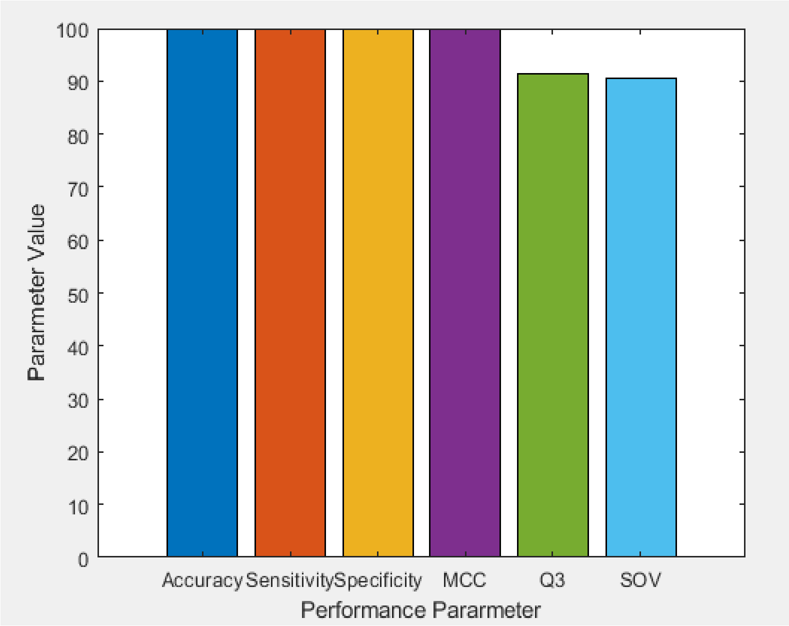


Figure 5.25 Different parameters for protein class B

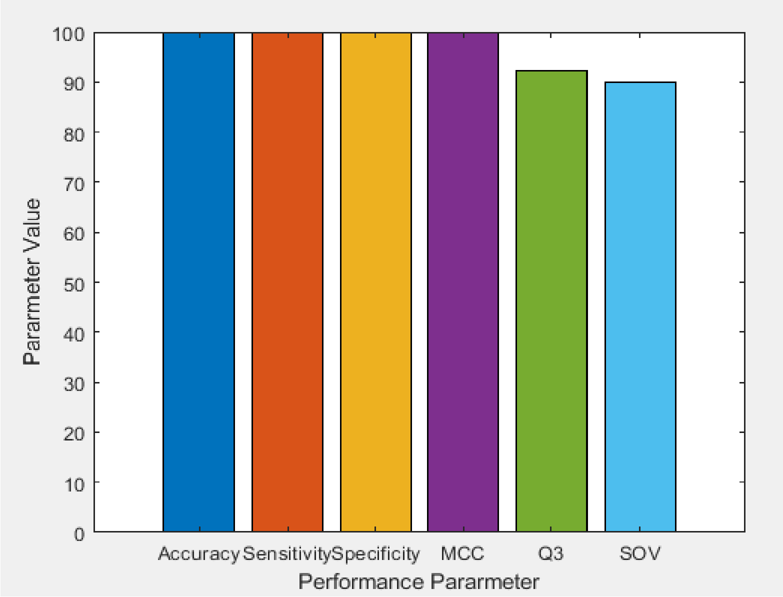


Figure 5.26 Different parameters for protein class C

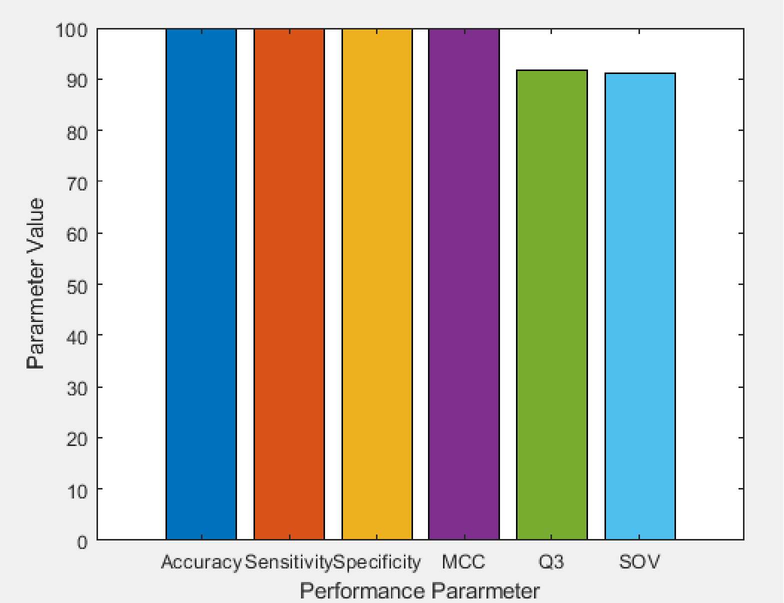


Figure 5.27 Different parameters for protein class D

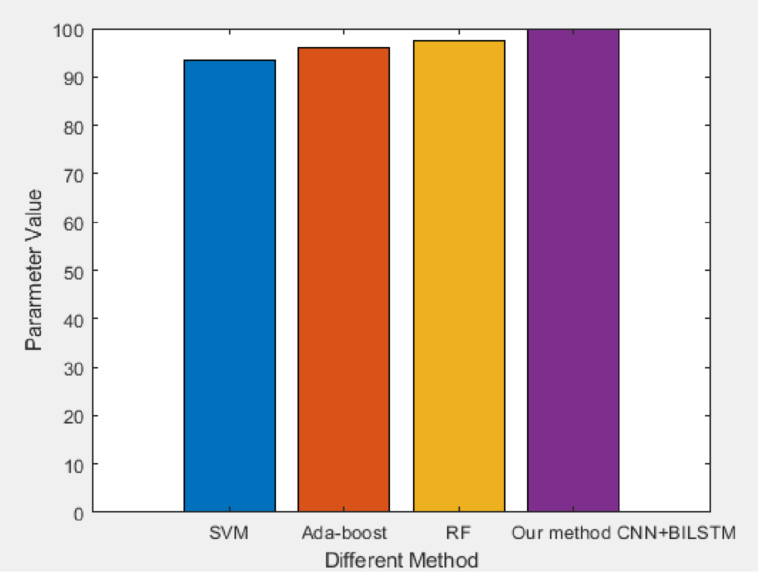


Figure 5.28 Accuracy comparison of different methods on FC699 dataset

Figures from 5.24 to 5.27, provides the values of different performance parameters like accuracy, sensitivity, specificity, MCC, Q3 and SOV, calculated using CNN+BILSTM technique on FC699 dataset for all the protein structural classes.

Table 5.6 Comparison of Accuracy for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	Accuracy%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	99.98	99.955
2	All β (B)	99.955	99.955
3	α+β (C)	99.957	99.955
4	α/β (D)	99.974	99.955

Table 5.7 Comparison of Sensitivity for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	Sensitivity%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	99.985	99.955
2	All β (B)	99.97	99.955
3	α+β (C)	99.97	99.955
4	α/β (D)	99.976	99.955

\

Table 5.8 Comparison of Specificity for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	Specificity%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	100	99.955
2	All β (B)	99.96	99.955
3	α+β (C)	99.974	99.955
4	α/β (D)	99.974	99.955

Table 5.9 Comparison of MCC for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	MCC%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	99.988	99
2	All β (B)	99.989	99
3	α+β (C)	99.988	99
4	α/β (D)	99.98	99

Table 5.10 Comparison of Q3 for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	Q3%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	92.527	90.813
2	All β (B)	91.622	90.576
3	α+β (C)	90.748	90.693
4	α/β (D)	91.197	90.652

Table 5.11 Comparison of SOV for CNN+BILSTM method for various protein classes in two datasets

Sr. No.	Classes	SOV%	
		CNN+BILSTM	
		FC699 dataset	25PDB dataset
1	All α (A)	90.393	92.435
2	All β (B)	90.458	90.43
3	α+β (C)	90.964	91.887
4	α/β (D)	90.391	92.756

Tables from 5.6 to 5.11 provides the values of different performance parameters like accuracy, sensitivity, specificity, MCC, Q3 and SOV, calculated using CNN+BILSTM technique on FC699 and 25PDB dataset for all the protein structural classes.

5.13 Summary

Sub-cellular localization of protein structure is attempted by numerous researchers by using several techniques of deep learning and machine learning. In the present study deep learning technique of CNN and CNN+BILSTM is utilized as a classifier which is compared with SVM concerning the accuracy, specificity, sensitivity, and MCC values for all four classes of protein. The accuracy of the CNN classifier and CNN+BILSTM is much higher than SVM classifier. Clustering is performed using the K-mean algorithm. A Hybrid PSO-Firefly algorithm is used for feature extraction of various classes of protein. 25PDB and FC699 dataset are analyzed based on various performance parameters. Also, scoring spaces and fitness values are evaluated for different classes of protein.

The proposed method CNN+BILSTM is compared with other methods and conclusion has been made that our method gives good accuracy.

Chapter-6

Conclusion and Future Scope

6.1 Overview

This chapter presents the findings and outcomes of the hybrid model along with the scope of improvement of the Hybrid model for the prediction of substructures in secondary structure. Literature Survey demonstrated that the clustering-aided approach can contribute to improving the Classification Rate for predicting patient outcomes. The objective of the thesis is to contribute to the prediction of the secondary structure of the protein. A thorough understanding of protein folding and structure recognition affects researchers in the fields of biology and chemistry for better drug design. In the area of designing new proteins, the specific function or the mechanism that determines the function of protein's knowledge is required. A large amount of protein and its complex structure forced the researcher to discover a computational algorithm for faster prediction. Protein are classified into four different classes, the classes are class A (All α), class B (All β), class C (α+β), and class D (α/β). Till today X-ray crystallography magnetic resonance method is the only method of 100% fold recognition of any given protein. But these methods are slow and cost for research. Looking at the large database researcher have opted for an alternate i.e., computational method. After careful investigation of results obtained using different techniques, we have come up with an optimized hybrid model with enhanced performance. The thesis has been concluded with careful reviews and discussion on the significance of our contributions, recommendations, and future scope of this research work.

To fasten the process of structure prediction nature-inspired algorithms are proposed in our thesis. Therefore, in each chapter, a significant amount of effort was devoted towards step-by-step prediction using machine learning through evolutionary algorithms. Chapter 1, deals with the fundamental of protein with their life cycle and importance. It deals with the different structures of protein and classifier. Chapter 2, deals with the literature survey, where all the recent work to date has been described.

Chapter 3, focuses on different mechanisms developed for pre-processing of protein sequence and Comparative study on four intelligent techniques. Chapter 4, proposes a new novel mechanism where we used the data from excel and apply pre-processing on data for refining the dataset. We used the FC699 dataset for Protein secondary structure (PSS) k-mean clustering on data to make an initial cluster and find out centroid point that centroid points take input for optimization algorithm or take initial population for generating by k-mean clustering. The best solution of clustering with the help of a hybrid of PSO and GSA optimization is generated to predict secondary structure using bi-clustering. The method is based on the idea that there can be many dissimilar primary structures of a protein that share similar secondary structures. Chapter 5 focuses on a novel approach for predicting different classes using a single Convolution neural network classifier (CNN) and a hybridization of CNN and Bidirectional long short term memory (BILSTM), also optimization algorithms are hybridized for feature selection. At last, Chapter 6 focuses on the conclusion and future scope in predicting the structure of the protein.

6.2 Conclusion

The conclusion of this study mainly consists of two aspects. Firstly deploy intelligent techniques for sequence clustering. Second, to deploy intelligent techniques for sequence alignment. And, finally deployment of intelligent techniques for prediction of substructures in secondary structure.

Different structural classes of protein are classified to make an understanding of different problems like folding and protein structure prediction etc. Clustering is performed using the K-mean algorithm. In the current work a random forest (RF) classifier is proposed which is compared with conventional classifiers like SVM, Ada boost, RF, etc. in terms of accuracy for all four classes of protein. The accuracy of the proposed RF classifier is much higher than other classifiers. Different performance parameters are measured for various classes of proteins. Varoius optimization algorithms are used for clustering and feature selection such as particle swarm optimization, gravitational search algorithm and K-mean clustering algorithms.The proposed classifier proved better than other classifiers in terms of accuracy and can help predict the protein structures. A Hybrid PSO-GSA algorithm was analyzed and its different parameters are analyzed for the classification of protein structure. As suggested in the literature, the proposed hybrid PSO-GSA algorithm has proved to achieve better results as compared to single algorithms. The accuracy though significant yet can be further improved by integrating more precise classifiers and by performing more intense pre-processing. Highlights of the performance of different parameters (accuracy, recall, precision, and specificity) values (in %) accomplished by proposed Random Forest (RF) classifier with FC699 represented test data. Figure 6.1 and 6.2 (Protein Structural Class A) highlights the comparison of accuracy values accomplished by the proposed RF classifier with FC699 test data with another classifier like SVM, Ada boost, RF, etc.Random forest classifier accuracy is good as compare to other classifiers.

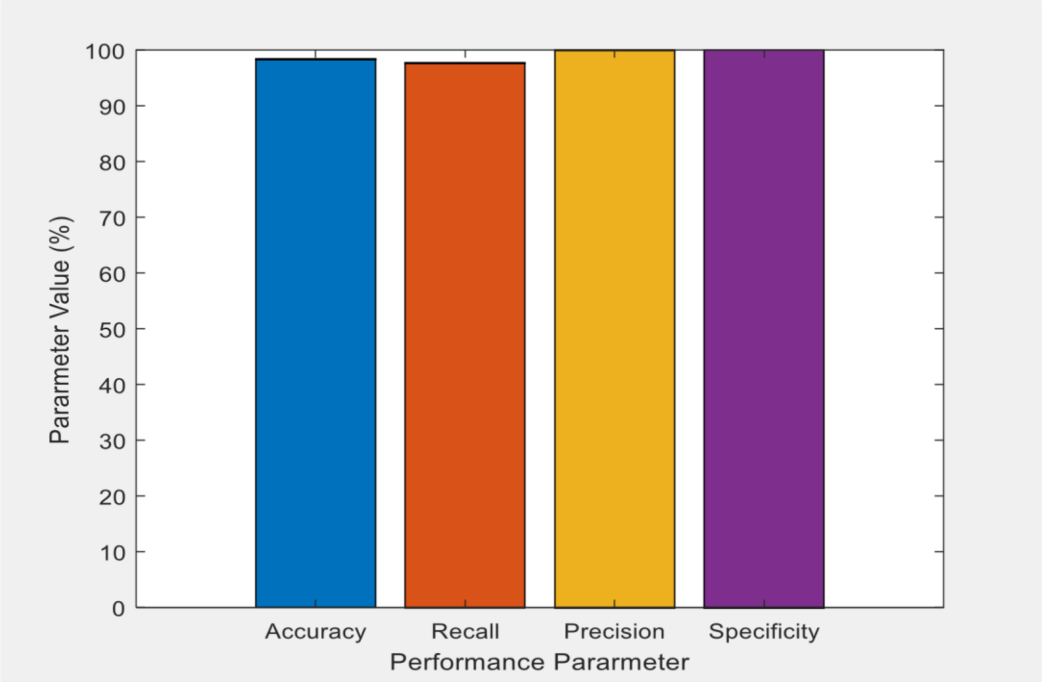


Figure 6.1 Performance of different parameters

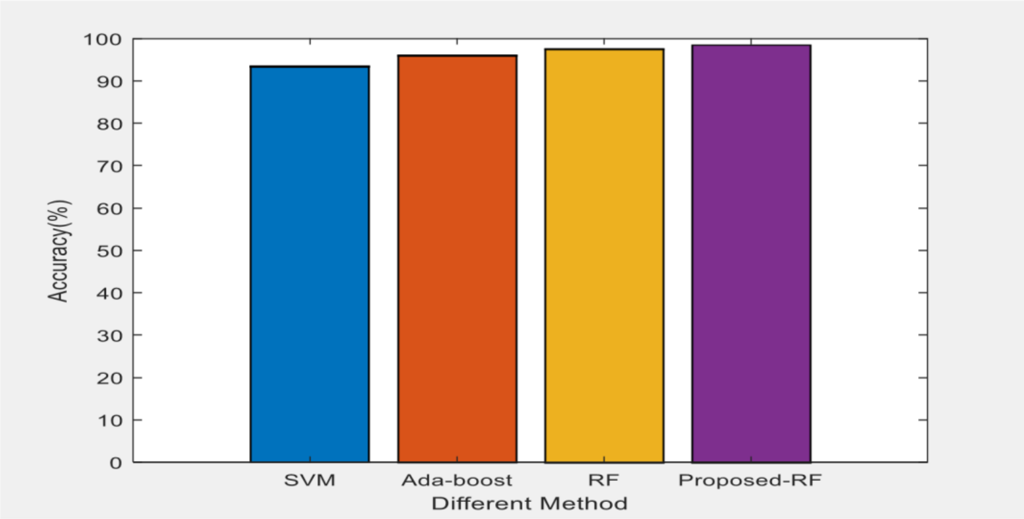


Figure 6.2 Accuracy of different methods

Figure 6.3 provides the performance values of different parameters accomplished by random forest classifier for the FC699 data set for different structural classes of Protein.

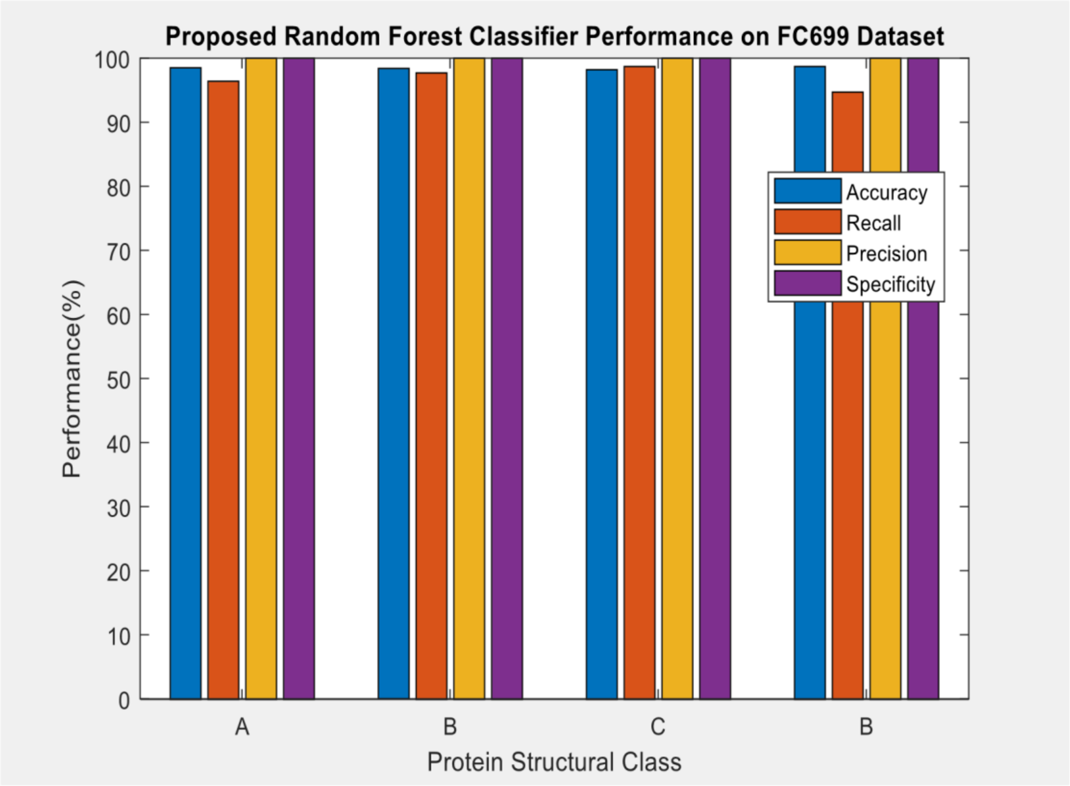


Figure 6.3 Combined model to measure different parameters for all classes

Sub-cellular localization of protein structure is attempted by numerous researchers by using several techniques of deep learning and machine learning. In the present study deep learning technique of CNN is utilized as a classifier which is compared with SVM concerning the accuracy, specificity, sensitivity, and MCC values for all four classes of protein. The accuracy of the CNN classifier is much higher than SVM classifier. Clustering is performed using the K-mean algorithm. A Hybrid PSO-Firefly algorithm is used for feature extraction of various classes of protein. 25PDB dataset is used to analyze the protein structure in terms of various performance parameters. Also,various classes of protein are evaluated by scoring spaces and fitness values. Particle swarm optimization (PSO) and Firefly algorithm (FFA) are used pre-processing of the data. Whereas Convolution neural network, are used for classification individually as well as it is combined with Bidirectional long short term memory for predicting different classes of the protein structure. The performance of CNN classifier, CNN+BILSTM classifier is compared with SVM classifier. For all the performance parameters the classifier used gives very good values for Class A, B and somewhat lower for class C, D although better values has been gained as compare to other classifiers. For feature selection, a hybrid PSOFFA algorithm has been used. Scoring spaces and fitness values are used for the evaluation of all the classes of protein secondary structure.

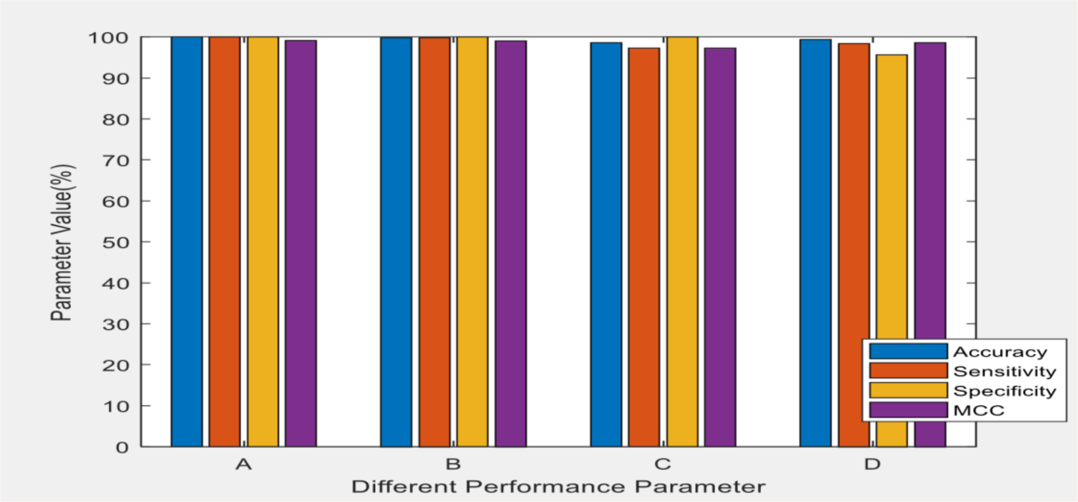


Figure 6.4 Comparison of different performance parameters for various protein classes

The following conclusion has been drawn from this work:-

This research draws a conclusion that the prediction of protein structure is the most significant problem, for optimal utilization of the resources.

It introduces an approach for the prediction of protein structure, clustering is performed using the K-mean algorithm. A Hybrid PSO-Firefly algorithm is used for feature selection of various classes of protein. Utilizing the FC699 dataset has been introduced which integrates the features of supervised and unsupervised learning.

The accuracy of the proposed RF classifier is much higher than other classifiers.

Different performance parameters are measured for various classes of proteins.

For feature selection, a hybrid PSOFFA algorithm has been used. Proposed classifier gives good accuracy in terms of comparing it with other classifiers.

The forward selection method emerges as the best Wrapper Feature Selection method among the three feature selection methods Used.

The final results showed that the Random Forest classification algorithm in combination with PSO and GSA emerges as the best classification algorithm for prediction.

The result with the cluster-aided approach greatly affects the performance of the classifier as compare to non-clustering predictions.

The study also verifies the result of previous studies that the Hybrid approach has better performance than the single learning classifiers and this is an approving outcome in the literature.

Convolution neural network, are used for classification individually as well as it is combined with Bidirectional long short term memory for predicting different classes of the protein structure. The performance of CNN classifier, CNN+BILSTM classifier is compared with SVM classifier. For all the performance parameters the classifier used gives very good values for Class A, B and somewhat lower for class C, D although better values has been gained as compare to other classifiers.

Hybridization of CNN-BILSTM had been used on 25PDB and FC699 dataset. On both datasets this method had achived good results.

6.3 Future Scope of the Study

Data mining has a large pool of classification and analysis techniques. These techniques can be applied to all kinds of datasets to extract hidden patterns for decision-making. The present study is focused on protein structures. As a future scope, Future work in the field will be for prediction tertiary and quaternary protein structures using the best algorithms. Moreover, this algorithm is tested with a benchmark protein dataset but still, there is a large amount of protein available that needs to be tested with this algorithm. Future work is also laid in refining the parameter and using a tri-clustering algorithm with better classification. The use of different evolutionary algorithms and fusion strategies can also bring revolution in prediction strategy.

The main objective of this research was to develop a prediction model that can predict with a high accuracy rate. For achieving this objective combination of Decision Tree, Random Forest, Logistic Regression, SVM, a genetic algorithm with different feature selection methods, along with clustering has been used. This resulted in a Hybrid model that gave a better performance as compared to models based on a single learning classifier.

To make these systems more practical, future work could include the following

In the future, such models can be designed with different advanced data mining techniques such as Association mining, constraint learning, structured prediction, and many others. Also, the data used in the present work is collected from the secondary source with some missing values. For further improvement, data may be collected from other data sources to explore the beauty of rich patterns available in the data.

In future research, other Feature Selection algorithms such as PSO, F-score, ant colony, LDA, and rough sets can be used and different Classification and Clustering algorithms can be used and the comparison can be made with the results of the Model. Also applying more clusters or ideally finding the optimized number can be the subject for further study.

A comprehensive parameter setting can be performed to search and find better results.

The innovative technique can be further evaluated through application in another field with different databases.

Finally, there is a need to develop software based on the proposed method, to achieve potential benefits from the proposed methods.

References

[1.1] Zhang X, "A Hybrid Algorithm for Determining Protein Structure", IEEE Expert, vol. 94, pp. 66-74, 1994.

[1.2] J. Lee, S. Wu, Y. Zang, "Ab Initio Protein Structure Prediction, From protein to structure to function in Bioinformatics", Springer, pp. 3-25, 2009.

[1.3] S.L. Salzberg, D.B. Searls, S. Kasif, "Grand challenges in computational biology", In Computational Methods in Molecular Biology, (eds.), Elsevier, 1998.

[1.4] Qin, Zhao, et al. "Artificial intelligence method to design and fold alpha-helical structural proteins from the primary amino acid sequence", Extreme Mechanics Letters 36, 2020.

[1.5] Goncalves, Ariadne Barbosa, et al. "Feature extraction and machine learning for the classification of Brazilian Savannah pollen grains", PloS one 11.6, 2016.

[1.6] Alberts, Bruce, et al. "Analyzing protein structure and function", Molecular Biology of the Cell. 4th edition. Garland Science, 2002.

[1.7] Biro, J. C. "Amino acid size, charge, hydropathy indices and matrices for protein structure analysis", Theoretical Biology and Medical Modelling 3.1, 2006.

[1.8] Senior, Andrew W.,et al. "Improved protein structure prediction using potentials from deep learning", Nature 577.7792, pp. 706-710, 2020.

[1.9] Gorissen, Stefan HM, et al. "Protein content and amino acid composition of commercially available plant-based protein isolates", Amino acids, pp. 1685- 1695, 2018.

[1.10] S. Hua, Z. Sun, “A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach”, J. Mol. Biol., vol. 308, pp. 397-407, 2001.

[1.11] Alberts, Bruce, et al. "The shape and structure of proteins", Molecular Biology of the Cell. 4th edition, Garland Science, 2002.

[1.12] Eisenberg, David. "The discovery of the α-helix and β-sheet, the principal structural features of proteins." Proceedings of the National Academy of Sciences, pp. 11207-11210, 2003.

[1.13] Harrow, Jennifer, et al. "Identifying protein-coding genes in genomic sequences", Genome Biology, pp. 201, 2009.

[1.14] Watkins, Andrew M., Paramjit S. Arora, "Anatomy of β-strands at protein–protein interfaces", ACS Chemical Biology, pp. 1747-1754, 2014.

[1.15] Y. Kim, “Application of Maximum Entropy Markov Models on the Protein Secondary Structure Predictions”, San Diego La Jolla, CA 92093.

[1.16] G. Pollastri, D. Przybylski, B. Rost, P. Baldi, “Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and

Profiles", PROTEINS: Structure, Function, and Genetics, Wiley- Liss, Inc., vol. 47, pp 228–235, 2002.

[1.17] H. Zhu, L. Yoshihara, K. Yamamom, "Prediction of protein structure by multi-modal feed-forward neural network", Proceedings of the 2002 International Joint Conference on, Vol. 1, pp. 280 – 285, 2002.

[1.18] A. Ceroni, P. Frasconi, A. Passerini, A. Vullo, "A Combination of Support Vector Machines and Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction", Springer-Verlag Berlin Heidelberg, pp. 142–153, 2003.

[1.19] G.-H. Yang, C.-G. Zhou, C.-Q. Hu, Z.-Z. Yu, "An algorithm based on improved bayesian inference network model for prediction protein secondary structure", Proceedings of the Second IEEE Conference on Machine Learning and Cybernetics, Xi'an, 2-5 November, 2003.

[1.20] Rehman, Ibraheem, and Salome Botelho. "Biochemistry, Tertiary Structure, Protein", 2017.

[1.21] Yu, Xiaojing, Chuan Wang, Yixue Li, "Classification of protein quaternary structure by functional domain composition", BMC Bioinformatics, pp. 1-6, 2006.

[1.22] S.H. Doong; C.Y. Yeh, "Secondary Structure Prediction Using SVM and Clustering", Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 2004.

[1.23] Y. Liu, J. Carbonell, J. Klein-Seetharaman, V. Gopalakrishnan, "Comparison of probabilistic combination methods for protein secondary structure prediction", Bioinformatics, vol. 20, pp. 3099–3107, 2004.

[1.24] Chen, N.S.Chaudhari, "Capturing Long-Term Dependencies for Protein Secondary Structure Prediction", 2004.

[1.25] L.-H.Wang, J. Liu, Y. Li, H. Zhou, " Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme", Genome Informatics, vol. 15(2), pp. 181–190, 2004.

[1.26] S.Akkaladevi, A.K.Katangur, S.Belkasim, Y.Pan, "Protein Secondary Structure Prediction using Neural Network and Simulated Annealing Algorithm", Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA September 1-5, 2004.

[1.27] M.N. Nguyen, J.C. Rajapakse, "Two-stage multi-class support vector machines to protein secondary structure prediction", Pacific Symposium on Biocomputing, vol. 10, pp. 346-357, 2005.

[1.28] N. P. Bidargaddi, M. Chetty, J. Kamruzzaman, "An Architecture Combining Bayesian segmentation and Neural Network Ensembles for Protein Secondary Structure Prediction", IEEE, 2005.

[1.29] B. Zhang, Z. Chen, Y. L. Murphey, "Protein Secondary Structure Prediction Using Machine Learning", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.

[1.30] J. He, H.-J. Hu, R. Harrison, P. C. Tai, and Y. Pan, “Rule Generation for Protein Secondary Structure Prediction with Support Vector Machines and Decision Tree”, IEEE transactions on nanobioscience, vol. 5, no. 1, March 2006.

[1.31] K. Chen, M. Kurgan, L. Kurgan, “Improved Prediction of Relative Solvent Accessibility Using Two-stage Support Vector Regression”, IEEE, 2007.

[1.32] M.N. Nguyen, J.C. Rajapakse, “Prediction of protein secondary structure with two-stage multi-class SVMs”, 2007.

[1.33] A. Reyaz-Ahmed, Y.-Q. Zhang, “Protein Secondary Structure Prediction Using Genetic Neural Support Vector Machines”, IEEE, 2007.

[1.34] R. Unger., “The Genetic Algorithm Approach to Protein Structure Prediction, Applications of Evolutionary Computation in Chemistry Structure and Bonding”, Vol. 110, pp 153-175, 2004.

[1.35] J.S.Bhalla, A. Aggarwal, “Prediction of Protein Structure using Parallel Genetic Algorithm”, International Journal of Computer Applications, vol. 81, 2013.

[1.36] J. Yang, “Protein Secondary Structure Prediction based on Neural Network Models and Support Vector Machines”, CS229 Final Project, Dec 2008.

[1.37] K.S. Guimaraes, J.C.B. Melo, G. D. C. Cavalcanti, “Combining few neural networks for effective secondary structure prediction”, BIBE’03, 2003.

[1.38] H.Zhu, I.Yoshihara, K.Yamamori, M. Yasunaga, “A multimodal neural network with single state predictions for protein secondary structure”, Artificial Life and Robotics, vol. 8, Issue 2, pp 168-173, 2004.

[1.39] K.W; S.C. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features”, Biopolymers, vol. 22 (12), pp. 2577–2637, 1983.

[1.40] L.O. Hall, X. Liu, K.W. Bowyer, Robert Banfield, “Why are Neural Networks Sometimes Much More Accurate than Decision Trees: An Analysis on a Bio- Informatics Problem”, IEEE, pp. 2851-2856, 2003.

[1.41] H.-J. Hu, Y. Pan, R. Harrison, P.C. Tai, “Improved Protein Secondary Structure Prediction Using Support Vector Machine with a New Encoding Scheme and an Advanced Tertiary Classifier”, IEEE Transactions On Nano-bioscience vol. 3 no. 4, December 2004.

[1.42] University of Arizona. “The Biology Project Department of Biochemistry and Molecular Biophysics”, 2003.

[1.43] Karen Lin. “Role of Data Science in Artificial Intelligence”, 2019.

[1.44] C.Geourjon & G. Deleage, “Protein Engineering”, 7, 157-164, 1994.

[1.45] Yongzhen Ge, Shuo Zhao, Xiqiang Zhao, “A step-by-step classification algorithm of protein secondary structures based on double-layer SVM model”, Genomics, 2019.

[2.1] S.L. Salzberg, D.B. Searls and S. Kasif, “Grand challenges in computational biology”, In Computational Methods in Molecular Biology, (eds.), Elsevier, 1998.

[2.2] Rost, Burkhard, Chris Sander, and Reinhard Schneider, "Redefining the goals of protein secondary structure prediction." Journal of molecular biology, pp. 13-26, 1994.

[2.3] Hall, Robert E., and Charles I. Jones, "Why do some countries produce so much more output per worker than others?", The quarterly journal of economics 114.1, pp. 83- 116, 1999.

[2.4] Dor, Ofer, and Yaoqi Zhou, "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training", Proteins: Structure, Function, and Bioinformatics, pp. 838-845, 2007.

[2.5] Cheng, Jill, et al. "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution", science 308.5725, pp. 1149-1154, 2005.

[2.6] Heffernan, Rhys, et al. "Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning", Scientific reports, pp. 1-11, 2015.

[2.7] Mirabello, Claudio, Gianluca Pollastri, "Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility", Bioinformatics, pp. 2056-2058, 2013.

[2.8] Yaseen, Ashraf, and Yaohang Li. "Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features", BMC bioinformatics, 2014.

[2.9] Wang, Niya, et al. "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues", Scientific reports, pp. 1- 12, 2016.

[2.10] Li, Yan-peng, et al. "Effect of exogenous phytohormones treatment on glycyrrhizic acid accumulation and preliminary exploration of the chemical control network based on glycyrrhizic acid in root of Glycyrrhiza uralensis", Revista Brasileira de Farmacognosia, pp. 490-496, 2016.

[2.11] Uddin, Mostofa Rafid, et al. "SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction", bioRxiv, 2019

[2.12] Heffernan, Shane M., et al. "COL5A1 gene variants previously associated with reduced soft tissue injury risk are associated with elite athlete status in rugby", BMC genomics, pp. 29-37, 2017.

[2.13] N. Qian and T. J. Sejnowski, “Predicting the Secondary Structure of Globular Proteins Using Neural Network Models”, Journal of Mol. Biol., vol. 202, pp. 865-884, 1988.

[2.14] B. Rost, C. Sander, and R.Schneider, “Evolution and Neural Networks - Protein Secondary Structure Prediction Above 71 % Accuracy”, Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences, 1994.

[2.15] David T. Jones, “Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices”, J. Mol. Biol., vol. 292, pp. 195-202, 1999.

[2.16] Claus A. Andersen, Henrik Bohr, Soren Brunak, “Protein secondary structure: category assignment and predictability”, FEBS Letters, pp. 6-10, 2001.

[2.17] S. Hua and Z. Sun, “A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach”, J. Mol. Biol., vol. 308, pp. 397-407, 2001.

[2.18] C. Notredame, “Recent progresses in multiple sequence alignment: a survey”, Ashley Publications Ltd ISSN 1462-2416, 2001.

[2.19] Y. Kim, “Application of Maximum Entropy Markov Models on the Protein Secondary Structure Predictions”, San Diego La Jolla, CA 92093.

[2.20] G. Pollastri, D. Przybylskim, B. Rost, P. Baldi, “Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles”, PROTEINS: Structure, Function, and Genetics, Wiley- Liss, Inc., vol. 47, pp 228–235, 2002.

[2.21] H. Zhu, I. Yoshihara, K. Yamamori, “Prediction of protein secondary structure by multi-modal neural networks”, Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02, 2002.

[2.22] G.-H. Yang, C.-G. Zhou, C.-Q. Hu, Z.-Z. Yu, “An algorithm based on improved bayesian inference network model for prediction protein secondary structure”, Proceedings of the Second IEEE Conference on Machine Learning and Cybernetics, Xi'an, 2-5 November 2003.

[2.23] M.N. Nguyen, J.C. Rajapakse, “Multi-Class Support Vector Machines for Protein Secondary Structure Prediction”, Genome Informatics, vol. 14, pp 218–227 2003.

[2.24] L.O. Hall, X. Liu, K.W. Bowyer, Robert Banfield, “Why are Neural Networks Sometimes Much More Accurate than Decision Trees: An Analysis on a Bio-Informatics Problem”, IEEE, pp. 2851-2856, 2003.

[2.25] H.-J. Hu, Y. Pan, R. Harrison, P.C. Tai, “Improved Protein Secondary Structure Prediction Using Support Vector Machine with a New Encoding Scheme and an Advanced Tertiary Classifier”, IEEE Transactions On Nanobioscience vol. 3 no. 4, December 2004.

[2.26] S.H. Doong, C.Y. Yeh, “Secondary Structure Prediction Using SVM and Clustering”, Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS’04), 2004.

[2.27] Y. Liu, J. Carbonell, J. Klein-Seetharaman, V. Gopalakrishnan, “Comparison of probabilistic combination methods for protein secondary structure prediction”, Bioinformatics, vol. 20 no. 17, pp. 3099–3107, 2004.

[2.28] J.Chen, N.S.Chaudhari, “Capturing Long-Term Dependencies for Protein Secondary Structure Prediction”, 2004.

[2.29] L.-H.Wang, J. Liu, Y. Li, H. Zhou, “Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme”, Genome Informatics, vol. 15(2): pp. 181–190, 2004.

[2.30] L-H Wang, J. Liu, H-B Zhou, “A comparison of two machine learning methods for protein secondary structure prediction”, Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004.

[2.31] M.N. Nguyen, J.C. Rajapakse, “Two-stage multi-class support vector machines to protein secondary structure prediction”, Pacific Symposium on Biocomputing, vol. 10, pp. 346-357, 2005.

[2.32] K. Nakayama, A. Hirano, K-I Fukumura. On Generalization of Multilayer Neural Network Applied to Predicting Protein Secondary Structure. IEEE International Joint Conference on Neural Networks, 2004, Proceedings, pp.1209 – 1213, 2004

[2.33] N. P. Bidargaddi, M. Chetty and J. Kamruzzaman, “An Architecture Combining Bayesian segmentation and Neural Network Ensembles for Protein Secondary Structure Prediction”, IEEE, 2005.

[2.34] B. Zhang, Z. Chen and Y. L. Murphey, “Protein Secondary Structure Prediction Using Machine Learning”, Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.

[2.35] Z.WM, “Protein secondary structure prediction by combining hidden Markov models and sliding window scores”, Int J Bioinform Res Appl. , vol. 1(4), pp. 420-428, 2003.

[2.36] G. J. Pappas Jr., S. Subramaniam. “Analysis of the Effects of Multiple Sequence Alignments in Protein Secondary Structure Prediction”, Advances in Bioinformatics and Computational Biology, vol. 3594, pp 128-140, 2005.

[2.38] Z.Aydin, Y.Altunbasak, and M.Borodovsky, “Protein secondary structure prediction for a single-sequence using hidden semi-Markov models”, BMC Bioinformatics, vol. 7, pp. 178-182, 2006.

[2.39] J. Chen and N. S. Chaudhari, “Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction”, Soft Comput., vol. 10, pp. 315–324, 2006.

[2.40] E.B. Samani; M.M. Homayounpour; H.Gu, “A Novel Hybrid GMM/SVM Architecture for Protein Secondary Structure Prediction”, WILF '07 Proceedings of the 7th international workshop on Fuzzy Logic and Applications: Applications of Fuzzy Sets Theory, pp. 491-496, 2007.

[2.41] A. Reyaz-Ahmed and Y.-Q. Zhang, “Protein Secondary Structure Prediction Using Genetic Neural Support Vector Machines”, IEEE, 2007.

[2.42] C.JA; B.GJ., “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction”, Proteins, vol. 34, pp. 508-519, 1999.

[2.43] M.N. Nguyen, J.C. Rajapakse, “Prediction of protein secondary structure with two-stage multi-class SVMs”, 2007.

[2.44] R. Kakumani, V. Devabhaktuni, M.O. Ahmad, “A Two-Stage Neural Network Based Technique for Protein Secondary Structure Prediction”, 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August pp. 20-24, 2008.

[2.45] A. Reyaz-Ahmed; Y. Zhang., “A New SVM-Based Decision Fusion Method Using Multiple Granular Windows for Protein Secondary Structure Prediction”, Rough Sets and Knowledge Technology Lecture Notes in Computer Science, vol. 5009, pp. 324-331, 2008.

[2.46] R. Kakumani, V. Devabhaktuni, M.O. Ahmad, “A Two-Stage Neural Network Based Technique for Protein Secondary Structure Prediction”, 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008.

[2.47] J.Wang; J.-P.Li, “Protein Secondary Structure Prediction based on BP Neural Network and Quasi-Newton Algorithm”, International Conference on Apperceiving Computing and Intelligence Analysis, (ICACIA 2008), pp.128 – 131, 2008.

[2.48] W. Y. Liu, S.X. Wang, B.W. Wang and J.X. Yu, “Protein Secondary Structure Prediction Using SVM with Bayesian Method”, IEEE (2008).

[2.49] L. Regad, F. Guyon, J. Maupetit, P. Tufféry, A.C. Camproux, “A Hidden Markov Model applied to the protein 3D structure analysis”, Computational Statistics and Data Analysis, vol. 52, pp. 3198 – 3207, 2008.

[2.50] W. Y. Liu, S.X. Wang, B.W. Wang and J.X. Yu, “Protein Secondary Structure Prediction Using SVM with Bayesian Method”, IEEE, 2008.

[2.51] V. Dzikovska, M. Oreskovic, S. Kalajdziski, K. Trivodaliev, D. Davcev, “Protein Secondary Structure Prediction Method based on Neural Networks”, IEEE, 2008.

[2.52] N. Mansour; C. Kehyayan; H. Khachfe, “Scatter Search algorithm for Protein Structure Prediction”, International Journal of Bioinformatics Research and Applications, vol. 5 Issue 5, pp. 501-515, 2009.

[2.53] B. Tang, X. Wang, X. Wang, “Protein Secondary Structure Prediction using Large Margin Methods”, Eight IEEE/ACIS International Conference on Computer and Information Science, 2009.

[2.54] A. Lakizadeh, S.-A. Marashi, “Addition of contact number information can improve Protein secondary structure prediction by neural Networks”, EXCLI Journal, vol. 8, pp.66-73, 2008.

[2.55] W.-Z. Lin, X. Xiao, “Using grey neural network to predict protein primary structure”, IEEE, 2009.

[2.56] Md. T.Hoque; M.Chetty; A. Sattar, “Genetic Algorithm in Ab Initio Protein Structure Prediction Using Low Resolution Model: A Review”, Biomedical Data and Applications Studies in Computational Intelligence , vol. 224, pp. 317-342, 2009.

[2.57] M.V.Thalatam, P. V. Rao, K. Varma, N. Murty, A. Apparao, “Prediction of Protein Secondary Structure using Artificial Neural Network”, International Journal on Computer Science and Engineering(IJCSE), vol. 02, No. 05, pp.1615-1621, 2010.

[2.58] P.V. N. Rao, T. U. Devi, D. Kaladhar, G.R. Sridhar, A.A.Rao, “Protein Secondary Structure Prediction using Pattern Recognition Neural Network,” International Journal of Engineering Science and Technology, vol. 2(6), pp.1752-1757, 2010.

[2.59] B. Yang; W. Qu; Y. Zhai; H. Sui, "Protein Secondary Structure Prediction Based on Improved SVM Method in Compound Pyramid Model", IEEE, 2010.

[2.60] H.Bouziane; B. Messabih and A. Chouarfia, "Profiles and Majority Voting-Based Ensemble Method for Protein Secondary Structure Prediction", Evolutionary Bioinformatics, vol. 7, pp. 171–189, 2011.

[2.61] Yang, Kuanquan Wang, Wangmeng Zuo, "Prediction of protein secondary structure using large margin nearest neighbor classification", 3rd International Conference on Advanced Computer Control (ICACC), 2011.

[2.62] N. Patil, Durga Toshniwal, Kumkum Garg, "Effective framework for protein structure prediction", International Journal of Functional Informatics and Personalised Medicine (IJFIPM), vol. 4, No. 1, 2012.

[2.63] H. Bordoloi, K.K. Sarma, "Protein Structure Prediction Using Multiple Artificial Neural Network Classifier", Soft Computing Techniques in Vision Science Studies in Computational Intelligence, vol. 395, pp 137-146, 2012.

[2.64] S. Chetia, K.K.Sarma, "Protein structure prediction using certain dimension reduction techniques and ANN", IRNet Transactions on Electrical and Electronics Engineering (ITEEE) ISSN 2319 – 2577, vol-1, 2012.

[2.65] K.Wang, "Prediction of protein secondary structure using large margin nearest neighbor classification", International Journal of Bioinformatics Research and Applications 01/2013; vol. 9(2), pp. 207-219, 2012.

[2.66] L. Jian-wei, C. Guang-hui, L. Hai-en, L. Yaun Prediction of protein secondary structure using multilayer feed-forward neural networks", Control and Decision Conference (CCDC), 25th Chinese, pp.1346 – 1351, 2013.

[2.67] Fuchs, Angelika, Andreas Kirschner, and Dmitrij Frishman, "Predicting residue and helix contacts in membrane proteins", Structural Bioinformatics of Membrane Proteins. Springer, Vienna, pp. 187-203, 2010. [2.68] Hu, Huiqing, et al. "Functional analyses of the CIF1–CIF2 complex in trypanosomes identify the structural motifs required for cytokinesis", Journal of cell science, vol. 24, pp. 4108-4119, 2013.

[2.69] Ce Zheng, Lukasz Kurgan, "Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments", BMC bioinformatics, vol. 9, pp. 430-435, 2008.

[2.70] Kountouris, Petros, and Jonathan D. Hirst. "Predicting β-turns and their types using predicted backbone dihedral angles and secondary structures", BMC bioinformatics, vol. 11.1, pp. 407, 2010.

[2.71] Lindberg, Sara M., et al., "New trends in gender and mathematics performance: a meta-analysis", Psychological bulletin, vol. 136.6, pp. 1123-1128, 2018.

[2.72] Singh, Shantanu, et al., "Morphological profiles of RNAi-induced gene knockdown are highly reproducible but dominated by seed effects", PloS one 10.7, 2015

[2.73] Pham, Can G., et al. "Oxygen JNKies: phosphatases overdose on ROS", Developmental cell 8.4, pp. 452-454, 2005.

[2.74] Fang, Rong, et al. "Music therapy is a potential intervention for cognition of Alzheimer’s Disease: a mini-review", Translational neurodegeneration, vol. 6.1, 2017.

[2.75] Fang, Bohao, et al., "Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent", Molecular phylogenetics and evolution, vol. 142, 2020.

[2.78] Lindström, Martin, "Commentary on Wang et al., “Differing patterns of short-term transitions of nondaily smokers for different indicators of socioeconomic status (SES)", Addiction 112.5, pp. 873-874, 2017.

[2.79] W.Z. Lin, X. Xiao, “Using grey neural network to predict protein primary structure”, IEEE, 2009.

[2.80] M.V.Thalatam, P. V. Rao, K. Varma, N. Murty, A. Apparao, “Prediction of Protein Secondary Structure using Artificial Neural Network”, International Journal on Computer Science and Engineering(IJCSE), vol. 02, No. 05, pp.1615-1621, 2010.

[2.81] A. Lakizadeh, S.-A. Marashi, “Addition of contact number information can improve Protein secondary structure prediction by neural Networks”, EXCLI Journal, vol. 8, pp.66-73, 2009.

[2.82] P.V. N. Rao, T. U. Devi, D. Kaladhar, G.R. Sridhar, A. A.Rao., “Protein Secondary Structure Prediction using Pattern Recognition Neural Network”, International Journal of Engineering Science and Technology, vol. 2(6), pp.1752-1757, 2010.

[2.83] M.V.Thalatam, P. V. Rao, K. Varma, N. Murty, A., “Apparao. Prediction of Protein Secondary Structure using Artificial Neural Network,” International Journal on Computer Science and Engineering(IJCSE), vol. 02, No. 05, pp.1615-1621, 2010.

[2.84] M.N. Nguyen and J.C. Rajapakse, “Two-stage multi-class support vector machines to protein secondary structure prediction”, Pacific Symposium on Biocomputing, vol. 10, pp. 346-357, 2005.

[2.85] B. Zhang; Z. Chen and Y. L. Murphey, “Protein Secondary Structure Prediction Using Machine Learning”, Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005.

[2.86] Sharif Razavian, Ali, et al., "CNN features off-the-shelf: an astounding baseline for recognition", Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014.

[2.87] Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber, "Multi-column deep neural networks for image classification", IEEE conference on computer vision and pattern recognition, IEEE, 2012.

[2.88] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, 2012.

[2.89] Stallkamp, Johannes, et al., "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition", Neural networks, vol. 32, pp. 323-332, 2012.

[2.90] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition", 2015.

[2.91] Alessio Ceroni, Paolo Frasconi, Andrea Passerini, Alessandro Vullo, " A Combination of Support Vector Machines and Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction", Advances in Artificial Intelligence, pp 142-153,2003.

[2.92] Søren Kaae Sønderby, Ole Winther, "Protein Secondary Structure Prediction with Long Short Term Memory Networks", 2015.

[2.93] Yihui Liu, Yehong Chen, Jinyong Cheng, "Feature extraction of protein secondary structure using 2D convolutional neural network", IEEE, 2016.

[2.94] Z.Li, Y.Yu. "Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks", IJCAI, 2016.

[2.95] Leandro Takeshi Hattori, Cesar Manuel Vargas Benıtez, Heitor Silverio Lopes. "A Deep Bidirectional Long Short-Term Memory Approach Applied to the Protein Secondary Structure Prediction Problem", IEEE, 2017.

[2.96] Y. Liu, J. Cheng, Y. Ma, Y. Chen, "Protein secondary structure prediction based on two dimensional deep convolutional neural networks", 2017 3rd IEEE International Conference on Computer and Communications (ICCC), IEEE, pp. 1995–1999, 2017.

[2.97] Buzhong Zhang, Jinyan Li and Qiang Lu, "Prediction of 8-state protein secondary structures by a novel deep learning architecture", BMC Bioinformatics, vol. 19:293, 2018.

[2.98] Leila Khalatbari, M.R. Kangavari, Saeid Hosseini, Hongzhi Yin, Ngai-Man Cheung, "P: A multi-component learning machine to predict protein secondary structure", Computers in Biology and Medicine, vol. 110, pp. 144–155, 2019.

[2.99] Shuping Zhu, Yihui Liu, “Protein Secondary Structure Online Server Predictive Evaluation”, J. Phys.: Conf. Ser. 1237 052005, 2019.

[2.100] Yongzhen Ge, Shuo Zhao, Xiqiang Zhao, “A step-by-step classification algorithm of protein secondary structures based on double-layer SVM model”, Genomics, 2019.

[2.101] Apurva Mehta, Mazumdar Himanshu, “Predicting structural class for protein sequences of 40% identity based on features of primary and secondary structure using Random Forest algorithm”, Computational Biology and Chemistry, vol. 84, pp. 107-164, 2020.

[2.102] J. Kennedy and R.C. Eberhart, “Particle swarm optimization”, In Proc. of the IEEE Int. Conf. on Neural Networks, pp 1942–1948, 1995.

[2.103] Esmat Rashedi, Hossein Nezamabadi-pour, and Saeid Saryazdi, “GSA: A gravitational search algorithm”, Information Sciences, vol. 179(13), pp. 2232–2248, 2009.

[2.104] Iztok Fister, Iztok Fister Jr., Xin-She Yang, Janez Brest. A comprehensive review of firefly algorithms. Swarm and Evolutionary Computation, vol. 13, pp. 34–46, 2013.

[3.1] Niu B, Cai Y.D, Lu W.C, Li G.Z and Chou K.C., “Predicting Protein Structural Class with Adaboost Learner”, Protein and peptide letters, vol. 13, No. (5), pp.489-492, 2006.

[3.2] Niu B, Jin Y.H, Feng K.Y, Lu W.C, Cai Y.D and Li G.Z, “Using Adaboost for the Prediction of Subcellular Location of Prokaryotic and Eukaryotic Proteins”, Springer, Molecular diversity, Vol. 12, No. (1), pp. 41, 2008.

[3.3] Yu D.J, Li Y, Hu J, Yang X, Yang J.Y and Shen H.B, “Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural Information and Random Forest Regression”, IEEE/ACM transactions on computational biology and bioinformatics, vol. 12, No. (3), pp. 611-621, 2015.

[3.4] Wong G.Y, Leung F.H and Ling S.H, “Predicting Protein-Ligand Binding Site using Support Vector Machine with Protein Properties”, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), Vol. 10, No. (6), pp. 1517-1529, 2013.

[3.5] Wang Y, Cheng J, Liu Y, and Chen Y, “Prediction of Protein Secondary Structure using Support Vector Machine with PSSM Profiles”, IEEE, In Information Technology, Networking, Electronic, and Automation Control Conference, IEEE, pp. 502-505, 2016.

[3.6] Zamani M and Kremer S.C, 2015, “A Multi-Stage Protein Secondary Structure Prediction System using Machine Learning and Information Theory”, IEEE, In Bioinformatics and Biomedicine (BIBM), IEEE International Conference on IEEE, pp. 1304-1309, 2015.

[3.7] Hasic H, Buza E and Akagic A, “A Hybrid Method for Prediction of Protein Secondary Structure Based on Multiple Artificial Neural Networks”, In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 40th International Convention on IEEE, pp. 1195-1200, 2017.

[3.8] Wei L, Liao M, Gao X, and Zou Q, “An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information”, IEEE transactions on nanobioscience, Vol. 14, No. (4), pp. 339-349, 2015.

[3.9] Dehzangi A, Paliwal K, Lyons J, Sharma A, and Sattar A, “A Segmentation-Based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition”, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), Vol. 11, No. (3), pp.510-519, 2014.

[3.10] He Z, Ma W, Zhang J and XuD, “A New Hidden Markov Model for Protein Quality Assessment using Compatibility Between Protein Sequence And Structure”, IEEE, Tsinghua science and technology, Vol. 19, No. (6), pp. 559-567, 2014.

[3.11] Zamani M and Kremer S.C, “Protein Secondary Structure Prediction Through A Novel Framework of Secondary Structure Transition Sites and New Encoding Schemes”, In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference on (pp. 1-7). IEEE, 2016.

[3.12] Xie S, Li Z, and Hu H, “Protein Secondary Structure Prediction Based on The Fuzzy Support Vector Machine with the Hyperplane Optimization”, Elsevier, Gene, vol. 642, pp. 74-83, 2018.

[3.13] Wang Y, Mao H, and Yi Z, “Protein Secondary Structure Prediction by using Deep Learning Method”, Elsevier, Knowledge-Based Systems, vol. 118, pp. 115-123, 2017.

[3.14] Cai Y.D, Liu X.J, Xu X.B, and Chou K.C, “Prediction of Protein Structural Classes by Support Vector Machines”, Elsevier, Computers & chemistry, Vol. 26, No. (3), pp. 293-296, 2002.

[3.15] Nguyen M.N, Zurada J.M, and Rajapakse J, “Toward Better Understanding of Protein Secondary Structure: Extracting Prediction Rules”, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 8, No. (3), pp. 858-864, 2011.

[3.16] Carlos G., “Artificial Neural Networks for Beginners”, 2003.

[3.17] Liu. J, Chi. G, Li. H, Liu. Y, Luo.X., “Prediction of Protein Secondary Structure Using Multilayer Feed-forward Neural Networks”, IEEE, pp. 1346 – 1351, 2013.

[3.18] T.W.B., “Feed-forward neural networks for secondary structure prediction”, Journal of Molecular Graphics, vol. 13, pp. 175-183,1995.

[3.19] Hanxi. Z, Ikuo. Y, Kunihito.Y., “Prediction of Protein Secondary Structure by Multi-Modal Neural Networks”, IEEE, vol. 1, pp. 280 – 285, 2002.

[3.20] Gianluca. P, Darisz. P, Burkhard. R, and Pierre.B., “Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles”, PROTEINS: Structure, Function, and Genetics, pp. 228–235, 2002.

[3.21] Kenji.N, Akihiro.H, Ken-ichi.F., “On Generalization of Multilayer Neural Network Applied to Predicting Protein Secondary Structure”, Neural Network, vol. 2, pp. 1209 – 1213, 2004.

[3.22] Jinmiao.C and Narendra S. C., “Capturing Long-Term Dependencies for Protein Secondary Structure Prediction”, Advances in Neural Network, vol. 3174, pp. 494-500, 2004.

[3.23] Rajasekhar. K, Vijay.D, M. A. “A Two-Stage Neural Network Based Technique for Protein Secondary Structure Prediction”, IEEE Eng Med Biol Soc., pp. 1355-1358, 2008.

[3.24] Vasilka.D, Mile.O, Slobodan.K, Kire.T, Danco.D. “Protein Secondary Structure Prediction Method based on Neural Networks”, IEEE, 2008.

[3.25] Amir. L, Sayed-Amir.M. “Addition of contact number information can improve Protein secondary structure prediction by neural Networks”, vol. 8, pp. 66-73, 2009.

[3.26] MN Vamsi. T, P Venkata.R, KVSRP.V, NVR.M, Allam.A. “Prediction of Protein Secondary Structure using Artificial Neural Network”, International Journal on Computer Science and Engineering, vol. 2, pp. 1615-1621, 2010.

[3.27] P.V.N.R, T. U, DSVGK.K, G.R.S, Allam Appa.R. “Protein Secondary Structure Prediction using Pattern Recognition Neural Network”, International Journal of Engineering Science and Technology, vol. 2, pp. 1752-1757, 2010.

[3.28] H. B and K. K. S. “Protein Structure Prediction Using Multiple Artificial Neural Network Classifiers”, Soft Computing Techniques in Vision Science, Studies in Computational Intelligence, vol. 395, pp. 137-146, 2012.

[3.29] S. M, B. H. C. “Input dimension reduction in neural network training-case study in transient stability assessment of large systems”, Intelligent Systems Applications to Power Systems, pp. 50-54, 1996.

[4.1] M. Levitt, C. Chothia, “Structural patterns in globular proteins”, Nature, vol. 261, pp. 552– 557, 1996.

[4.2] M. Gromiha, S. Selvaraj, “Protein secondary structure prediction in different structural classes”, Protein Eng., vol. 11, pp. 249–251, 1998.

[4.3] K.C. Chou, C.T. Zhang, “Prediction of protein structural classes”, Crit. Rev. Biochem. Mol. Biol., vol. 30, pp. 275–349, 1995.

[4.4] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, “Understanding the recognition of protein structural classes by amino acid composition”, Proteins, vol. 29, pp. 172–185, 1997.

[4.5] D. L. Nelson, M. M. Cox, “Principles of Biochemistry”, seventh Edition, W. H. Freeman and Company, One New York Plaza, New York, NY, 760 USA, 2017.

[4.6] N. Jana, S. Das, J. Sil, “A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis”, Emergence, Complexity and Computation, Springer International Publishing, Cham, Switzerland, vol. 31, 2018.

[4.7] F. Campeotto, A. Dal Pal`u, A. Dovier, F. Fioretto, E. Pontelli, “A Constraint Solver for Flexible Protein Models”, Journal of Artificial Intelligence Research, vol. 48 (1), pp. 953-1000, 2013.

[4.8] D. H. Kalegari, H. S. Lopes, “An Improved Parallel Differential Evolution Approach for Protein Structure Prediction using both 2D and 3D off-lattice models”, in: IEEE Symposium on Differential Evolution (SDE), IEEE, Singapore, pp. 143 – 150, 2013.

[4.9] K. A. Dill, “Theory for the folding and stability of globular proteins”, Biochemistry, vol. 24 (6), pp. 1501 – 1509, 1985.

[4.10] B. Berger, T. Leighton, "Protein Folding in the Hydrophobic-hydrophilic (HP) is NP-complete", in: Proceedings of the Second Annual International Conference on Computational Molecular Biology, RECOMB'98, ACM, New York, NY, US, pp. 30 – 39, 1998.

[4.11] W. E. Hart, A. Newman, "Protein Structure Prediction with Lattice Models", 2005.

[4.12] F. H. Stillinger, T. Head-Gordon, C. L. Hirshfeld, "Toy model for protein folding", Phys. Rev. E 48, pp. 1469-1477, 1993.

[4.13] Birlutiu A, d'Alche-Buc F, Heskes T, "A Bayesian framework for combining protein and network topology information for predicting protein–protein interactions", IEEE Trans Comput Biol Bioinform, vol. 12(1), pp. 538–550, 2015.

[4.14] Song D, Chen J, Chen G, Li N, Li J, Fan J, Bu D, Li SC, "Parameterized BLOSUM matrices for protein alignment", IEEE Trans Comput Biol Bioinform, vol. 12(3), pp. 686–694, 2015.

[4.15] L. Hunter, "Artificial Intelligence and Molecular Biology", AAAI Press, Boston, USA, 1 edition, 1993.

[4.16] D.L. Nelson and M.M. Cox, "Lehninger Principles of Biochemistry",W.H. Freeman, 5th edition, 2008.

[4.17] H.S. Lopes, "Evolutionary algorithms for the protein folding problem: A review and current trends", In T.G. Smolinski, M.M. Milanova, and A-E Hassanien, editors, Computational Intelligence in Biomedicine and Bioinformatics, Springer-Verlag, Heidelberg, Germany, vol. I, pp. 297–315, 2008.

[4.18] A. Liwo, M. Khalili, and H. A. Scheraga, “Ab initio simulations of protein- folding pathways by molecular dynamics with the united-residue model of polypeptide chains”, Proceedings of the National Academy of Sciences, vol. 102(7), pp. 2362–2367, 2005.

[4.19] K.A. Dill, S. Bromberg, K. Yue, and K.M. Fiebig et al. “Principles of protein folding - a perspective from simple exact models”, Protein Science, vol. 4(4), pp. 561–602, 1995.

[4.20] F.H. Stillinger, T. Head-Gordon, and C. Hirshfeld, “Toy model for protein folding”, Physical Review E, vol. 48(2), pp. 1469–1477, 1993

[4.21] C.M.V. Ben´ıtez and H.S. Lopes, “Hierarchical parallel genetic algorithm applied to the three-dimensional HP side-chain protein folding problem”, In Proc. of IEEE International Conference on Systems, Man and Cybernetics, IEEE Computer Society, pp. 2669–2676, 2010.

[4.22] P. Crescenzi, D. Goldman, C. Papadimitrou, A. Piccolboni, and M. Yannakakis, “On the complexity of protein folding”, Journal of Computational Biolology, 5:423–446, 1998.

[4.23] J. Kennedy and R.C. Eberhart, “Particle swarm optimization”, In Proc. of the IEEE Int. Conf. on Neural Networks, pages 1942–1948, Piscataway, USA. IEEE Press, 1995.

[4.24] Tapas Si and Nanda Dulal Jana, “Particle swarm optimisation with differential mutation”, International Journal of Bio-Inspired Computation, vol. 11(3), pp. 212–251, 2012.

[4.25] A. Chatterjee, S.P. Ghoshal, and V. Mukherjee, “A maiden application of gravitational search algorithm with wavelet mutation for the solution of economic load dispatch problems”, International Journal of Bio-Inspired Computation, vol. 4(1), pp. 33–46, 2012.

[4.26] Esmat Rashedi, Hossein Nezamabadi-pour, and Saeid Saryazdi, "GSA: A gravitational search algorithm", Information Sciences, vol. 179(13), pp. 2232–2248, 2009.

[4.27] Sunil K. Patel, Atul K. Pandey, Ravi Roshan, Upendra K. Singh, "Application of PSO and GSA hybrid optimization method for 1-D inversion of magnetotelluric data", International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016.

[4.28] Shiquan Sun, Qinke Peng, "A hybrid PSO-GSA strategy for high-dimensional optimization and microarray data clustering", IEEE International Conference on Information and Automation (ICIA), 2014.

[5.1] Michael V. Boland, Mia K. Markey, and Robert F. Murphy, "Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images", Cytometry, vol. 33, pp. 366–375, 1998.

[5.2] Michael V. Boland and Robert F. Murphy A Neural "Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells", BIOINFORMATICS, vol 17 no 12, pp 1213-1223, 2001.

[5.3] Amina Chebira, Yann Barbotin, Charles Jackson, Thomas Merryman, Gowri Srinivasa1, Robert F Murphy and Jelena Kovacevic, "A multiresolution approach to automated classification of protein subcellular location images", BMC Bioinformatics, vol. 8:210, 2007.

[5.4] Kai Huang and Robert F Murphy, "Boosting accuracy of automated classification of fluorescence microscope images for location proteomics", BMC Bioinformatics, vol. 5:78, 2004.

[5.5] Loris Nanni, Alessandra Lumini. “A reliable method for cell phenotype image classification”, Artificial Intelligence in Medicine, vol. 43, pp. 87—97, 2008.

[5.6] Geert Litjens, Thijs Kooi , Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. Vander Laak, Bram Van Ginneken, Clara I. Sánchez, “A survey on deep learning in medical image analysis”, Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[5.7] Oren Z Kraus, Ben T Grys, Jimmy Ba1, Yolanda Chong, Brendan J Frey, Charles Boone & Brenda J Andrews. “Automated analysis of high-content microscopy data with deep learning.” Mol Syst Biol., vol. 13: 924, 2017

[5.8] Tanel Parnamaa, Leopold Parts. “Accurate classification of protein subcellular localization from high throughput microscopy images using deep learning”, G3:Genes|Genomics|Genetics, 2017.

[5.9] Kaisa Liimatainen, Mira Valkonen, Leena Latonen, Pekka Ruusuvuori. “Cell organelle classification with fully convolutional neural networks”, 1st Conference on Medical Imaging with Deep Learning, 2018.

[5.10] Mengli Xiao, Xiaotong Shen, Wei Pan. “Application of deep convolutional neural networks in classification of protein subcellular localization with microscopy images”, Genet. Epidemiol, 1–12, 2019.

[5.11] Alexander Kensert, Philip J. Harrison, and Ola Spjuth. "Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes", SLAS Discovery, Vol. 24(4), 2019.

[5.12] R. H. Garrett, C. M. Grisham, "Biochemistry", 6th Edition, Cengage Learning, Boston, MA, US, 2017.

[5.13] D. L. Nelson, M. M. Cox, "Principles of Biochemistry", seventh Edition, W. H. Freeman and Company, One New York Plaza, New York, NY, USA, 2017.

[5.14] B. Alberts, A. D. Johnson, J. Lewis, D. Morgan, M. Ra_, K. Roberts, P. Walter, "Molecular Biology of the Cell", 6th Edition, Garland Science, Third Avenue, New York, NY, US, 2014.

[5.15] N. Jana, S. Das, J. Sil, A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis, Emergence, Complexity, and Computation, Springer International Publishing, Cham, Switzerland, vol. 31, 2018.

[5.16] Xin-She Yang, "Firefly algorithm, stochastic test functions, and design optimization", International Journal of Bio-Inspired Computation archive, vol. 2, no. 2, March 2010.

[5.17] Nadhirah Ali, et al., "A Review of Firefly Algorithm", ARPN Journal of Engineering and Applied Sciences, vol. 9, no. 10, October 2014.

[5.18] Bin Wang, et al., "A modified firefly algorithm based on light intensity difference", Journal of Combinatorial Optimization, vol. 31, no. 3, pp. 1045-1060, 2016.

[5.19] Mengli Xiao, Xiaotong Shen, Wei Pan. “Application of deep convolutional neural networks in classification of protein subcellular localization with microscopy images”, Genet. Epidemiol, pp. 1–12, 2019.

[5.20] Alexander Kensert, Philip J. Harrison, and Ola Spjuth, “Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes”, SLAS Discovery, vol 24(4), 2019.

[5.21] Kaiming, He Xiangyu, Zhang Shaoqing, Ren Jian Sun, “Microsoft Research. “Deep Residual Learning for Image Recognition”, arXiv:1512.03385 [cs.CV], 2015.

[5.22] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, CoRR, vol: abs/ 1704.04861, 2017.

[5.23] Padmavathi Kora, K.SriRama Krishna. “Hybrid firefly and Particle Swarm Optimization algorithm for the detection of Bundle Branch Block”, Volume 2, Issue 1, March 2016, Pages 44-48.

LIST OF PUBLICATIONS

Sarneet Kaur, Babita Pandey, "Intelligence Computing Methods Deployed for Protein Structure Prediction: A review" in an International Conference on Future and challenges of computational and integrated sciences,7th & 8th November 2014.

Sarneet Kaur, Babita Pandey, "Protein Secondary Structure Prediction using Feed Forward Artificial Neural Network and Perceptron", International Journal of Control Theory and Applications ISSN: 0974-5572 Volume 9, Number 45, 2016.

Sarneet Kaur, Ashok Sharma, "Multi-Classifiers Comparison for Protein Secondary Structure Prediction", IEEE International Conference ICCCIS-2019, 18-19th Oct 2019.

Sarneet Kaur, Ashok Sharma, Parveen Singh, "Hybrid of PSO-GSA based Clustering Approach for Predicting Structural Class Prediction using Random Forest Method", European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 10, Pages 17-32.

Sarneet Kaur, Ashok Sharma, Parveen Singh, "Protein Structural Classes Prediction Based on Convolutional Neural Network Classifier with Feature Selection of Hybrid PSO-FA Optimization Approach", European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 10, Pages 252-265.