Multiple Decision Techniques for RMSD prediction of Protein Structure

A Thesis

submitted in partial fulfillment of the requirements for the award of the degree of

Master of Engineering

in

Computer Science and Engineering Department

by

Jagmeet Kaur

(Reg no: 801532022)

Under the supervision of

Dr. Prashant Singh Rana

Thapar University

Patiala-147004, Punjab, India

July 2017

Candidate Declaration

I hereby certify that the work, which is being presented in the thesis, entitled **Multiple Decision Techniques for RMSD prediction of Protein Structure**, in partial fulfillment of the requirements for the award of the degree of **Master of Engineering** and submitted to the institution is an authentic record of my own work carried out during the period **July 2015** to **July 2017** under the supervision of **Dr. Prashant Singh Rana**. I have also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

The matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any institution.

Date: 17/7/2017

Jagmeet Kaur

Jagmeet Kaur
Candidate

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Date: 17/7/2017

Dr. Prashant Singh Rana
Supervisor

The M.E. Viva-Voice examination of **Jagmeet Kaur** has been held on 3 August, 2017.

Abstract

Protein sequences are converted into three dimensional tertiary designs to perform various biological processes. Physicochemical properties of amino acid remains and their ratio give rise to different associated forces which further lead the folding of a protein sequence into its distinct tertiary designs. A large amount of protein sequence data is storming as the outcome of different genomic and several other sequences projects. Due to inundation of such enormous amount of sequence data, there is the vital need to develop computational predictive approaches for prediction of protein structure from amino acid sequences.

The work presented in this thesis mainly focuses on the multiple decision techniques qualitative study of protein structure using supervised learning with six physicochemical properties. The objective is to predict the qualitative measure i.e. Root Mean Square Deviation (RMSD) of a protein structure in the absence of its true native state.

In this work, a performance study of classification machine learning models is carried out to classify the protein structure using Multiple Decision Techniques. The k-fold cross validation is used to measure the robustness of the proposed method.

Prediction of RMSD of the protein structure is the critical factor in order to differentiate the native protein structure or native like protein structures from the predicted structures. In this work, Principle Component Analysis (PCA) has been implemented in order to obtain independent and uncorrelated components to decrease the dimensionality of the feature space. PCA is very useful as it extract the relevant information from the dataset, analyze structure of observations and to represent it as a new set of principle components. The seventeen classification methods have been used which belong to different families of machine learning that makes a rigorous and least biased ensemble. Further, based on the several performance parameters of the particular classifier, Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) has been introduced to build a single performance score to rank the classifiers and based on the ranking given by TOPSIS to predict the RMSD of protein structure of protein sequence the ensembled model has been developed. Despite the simplicity of the technique used, the results obtained by these ensembles are found to be better in comparison to those produced by other methods. The empirical study indicated that the combination of performance score of individual classification algorithms increased the performance. That's why, TOPSIS based eight ensembles of classification algorithms have been generated to increase the performance. By intensive experimentation, it is found that ensemble of nine classification models

outperformed. There are several measures to evaluate performance and it is the critical undertaking to choose an outperforming classifier (or set of classifiers). Further, this work also introduced a rough set based ensembled approach which make rough sets of independent, uncorrelated and outperforming models. It is evident from the results that proposed novel rough set based ensemble has a high accuracy, Sensitivity, specificity, Area under the receiver operating characteristic curve (AUC), Positive Predictive Value (PPV), Negative Predictive value (NPV) and Detection Rate. The proposed model has been compared with available models and validated on benchmark dataset CASP 10. The k-fold cross validation has been used to check the robustness of proposed model.

Keywords: Protein Structure Prediction, MCDM, TOPSIS, Machine Learning Models, Rough set theory.

Acknowledgements

First, I would like to express my deep gratitude to my supervisor **Dr. Prashant Singh Rana** for their invaluable advice and encouragement at every step of my Master's program. Without their unfailing support and belief in me, this thesis would not have been possible. Their contribution to this thesis goes well beyond their role as an academic supervisor and includes constant support on a personal level without which this journey may never have been completed. And for this, I am truly grateful. They are great mentor for my life as well.

I am also thankful to **Dr. Ashutosh Mishra**, P.G. Coordinator for the motivation and inspiration that triggered me for the thesis work. He has always been supportive and provided us required information on regular intervals.

I would like to express my gratitude to Head of the Department, Computer Science and engineering **Dr. Maninder Singh** for setting good standards for his students and providing all the help and facilities that were essential throughout the journey. Your encouragement time and again has helped students to achieve the set goals.

Finally, I would like to express my sincere and deep gratitude to my parents and family member for their love, encouragement, care and support.

Jagmeet Kaur

Jagmeet Kaur

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter is an introductory of the work done in this thesis. Beside stating the definitions, it underlines the research orientation and thesis organisation.

Protein plays a very crucial role in every biological process and form the pivotal role in the cell metabolism. They are active elements used to catalyse the various biological processes and helps to control and direct several chemical reactions inside the living cell. They helps to control the flow of ions, transport oxygen in the blood and provide mechanical support to the cell as well as entire body. The structure of the protein is important for the proper functioning of the biological processes and several chemical pathways because these numerous functions are possible only if protein adopt several configurations. Proteins consists of a chain of amino acid of variable length. There are 20 amino acids that constitute the protein sequences and each amino acid perform different functions. When these amino acids are put together in some order, they interact and constitute a sequence that can defines the structure of protein uniquely. Thus, knowledge of the amino acid and their arrangement is important to predict the Root Mean Square Deviation (RMSD) of protein structure.

There are several methods available ranging from the purely ab-initio methods which predicts based on the physiochemical properties, to the homology methods which predicts based on the information available in sequences and genomic databases, and there are several other approaches which lies between these two extremes such as fold-recognition or threading which predicts by identifying the structural template. In Homology approaches, the focus is on the information from the sequence of the amino acids and predicts the RMSD of protein structure based on the linking of amino acids. In the ab-initio method, the physiochemical properties of the chain of sequence is used to predict structure of protein sequence. These properties of the sequences are the main factors that are responsible for the folding of protein sequence into its tertiary structures. Stoichiometry of amino acid also results in different energy contributing forces which helps in folding pathways. Many prediction techniques results in low quality structures because of the absence of the proper knowledge and inadequate observations. These structures obtained may look similar when monitored but in actual they may be away from their actual native states.

Due to inundation of such enormous amount of sequence data, there is the vital need to develop predictive approaches for protein structure prediction and hence to find difference between a structure and the native in the lack of its experimental structure.

The work presented in this thesis focuses on the Multiple Decision Techniques (MDT) for qualitative study of protein structure using supervised learning with six physiochemical properties. The objective is to predict the qualitative measure i.e. RMSD of a protein structure. In this work, a performance study of classification machine learning models is carried out to classify the protein structure using MDT. Seventeen classification algorithms belonging to different families are used in this work to predict the RMSD of protein structure. The considered models belong to different families of machine learning because they make a rigorous and least biased ensemble. Principle Component Analysis (PCA) is implemented in order to obtain independent components or uncorrelated components to decrease the dimensionality of the feature space. In this thesis, two methods are discussed to find the ensemble of models as follows:

- Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is utilised to build a single performance based on the number of performance parameters of the particular classifier score to rank the classifiers. With the use of this technique, this work systematises the information present within the scope of techniques of decision making and rank the models based on the single performance score, and then **ensembled model** is developed based on the ranking given by TOPSIS approach. Despite the simplicity of the technique used, the results obtained by the ensemble are found to be better to those produced by other methods.
- There are several measures to evaluate performance of the classification models and it is the critical undertaking to choose an outperforming classifier (or set of classifiers). This work has introduced **a rough set based ensembled approach** which choose the random subsets from the whole pool of models. The randomly chosen rough sets are then evaluated and ensemble of independent, uncorrelated and outperforming models is developed.

The consistency of the proposed method is checked with the help of k-fold validation. The k-fold cross validation is carried out to evaluate the performance metrics of the predictive model. The total sample is subdivided randomly into k equal size sub samples. The (k-1) samples are used for training purpose, and remaining one sample is used for testing purpose in order to validate the model. This process is then reiterated k folds by taking each sub sample as the validation sample at least once. Further, the single score is obtained by aggregating the results from k folds. All the above brief descriptions are discussed in subsequent chapters.

1.1 Research Orientation

This section describes the motivation to carry out this research and the key research areas selected for this research.

1.1.1 Research Motivation

Physiochemical properties as well as solvents of amino acids are the main factors in bending a amino acid sequences into their distinct tertiary state. Various types of energies that are produced by these key factors generate folding pathways. Initially, structure determination approaches have employed to develop a structure or a group of structures using possible candidates in the form of sequences. Nowadays, prediction of tertiary protein structure has turned into one of the greatest competitive problem in bioinformatics. A large amount of protein sequence data is continuously storming as the outcome of different genomic and several other sequences projects. Due to inundation of such enormous amount of sequence data, there is the vital need to develop predictive approaches for prediction of structure of protein sequences. There are several methods available ranging from the purely ab-initio methods, to the homology methods that can be used for the prediction of structure. In the homology approaches, the knowledge of the peptides, amino acids and their bonding with each other is essential. The ab initio approaches can be employed, the physical and chemical properties of the sequences are the main parameters for folding protein structures into their tertiary structures. All the approaches that are developed so far are very expensive, time consuming and need experts to operate them. Moreover, many prediction techniques results in low quality structures because of the absence of the proper knowledge and inadequate observations. These structures obtained may look similar to any high resolution structure when monitored but in actual they may be away from their actual native states. Hence, it is really advisable to develop a predictive model that can predict difference between a structure and the native in the lack of its experimental structure. Hence, computational and machine learning approaches are applied to develop a technique that can help to overcome limitations in the existing methods. In this thesis, MDT are implemented to develop the outperforming models or set of models to predict the RMSD of the protein structure of the amino acid sequence.

1.2 Thesis Organization

The thesis is organized into 7 chapters. A brief outline is given below:

- **Chapter 1**: This chapter gives the brief introduction of the work done in this thesis. The research orientation and research motivation are discussed in this chapter. Further, in the last section, the thesis organisation is also given.
- **Chapter 2**: This chapter discusses about the literature survey of work done in this domain of prediction of RMSD of the structure of the protein sequence of the amino acid. This chapter presents the several approaches that are already developed to predict the structure of the protein sequence of amino acid.
- **Chapter 3**: This chapter defines the objective of the research problem. It also discusses about the current gaps in prediction of the protein structure. Further, it discusses about the objectives of the work done in the thesis.
- **Chapter 4**: This chapter gives the description of the dataset. It explains the feature extraction, feature measurement and feature transformation of the protein dataset. Further it discusses about principle component analysis that is used for feature extraction and feature importance.
- **Chapter 5**: This chapter discusses about the Multi decision techniques that are developed in this thesis. It discusses about the TOPSIS algorithm and the technique introduced to find the ensemble based on this Multi Criteria Decision Analysis (MCDM) approach. Further, it discusses about the Rough set based technique to choose the classification models to develop the outperforming ensemble. In the last section of this chapter it discusses about the the classification methods used in this thesis.
- **Chapter 6**: This chapter discusses about the evaluation techniques and the results that are obtained from the two approaches. This chapter gives the brief discussion about the results. It gives the k fold validation of the proposed approaches to check the robustness.
- **Chapter 7**: This chapter summarizes the key findings and main contributions of the thesis and lists the possible future research directions.

Chapter 2

Literature Survey

This section discusses about the literature survey of work done in the domain of prediction of the structure of the protein sequence of the amino acid, TOPSIS approach and rough set based techniques. This chapter presents the several approaches that are already developed to predict the structure of the protein sequence of amino acid. In the last section, the machine learning methods have been described.

2.1 Prediction of Structure of Protein Sequence

Simons *et al.* (1999) predicted the structures of 21 of the 43 targets accessible in CASP3 that did not have evident homologous with known structures. Of these objectives, eighteen tentatively decided structures, which cover the array of secondary structure composition, are accessible for correlation with the predicted structures. At the point when the tentatively decided structures were made accessible before the CASP3 meeting, these predictions are assessed via scanning for native like sub structures utilizing the DALI server in both the substantial arrangement of 1,200 structures and the five submitted structures. Especially great expectations were made for four targets. The submitted structures were chosen on the premise of the quantity of different structures within 6 Å RMSD and the score. The outcomes about this were empowering: highlights of the forecasts incorporate a 99-deposit portion for MarA with a RMSD of 6.4 Å to the local structure, a 95-deposit (full length) forecast for the EH2 do-primary of EPS15 with a RMSD of 6.0 Å, a 75-deposit section of DNAB helicase with a RMSD of 4.7 Å, and a 67-buildup section of ribosomal protein L30 with a RMSD of 3.8 Å. These outcomes propose that abdominal muscle initio methods may soon end up noticeably valuable for low-determination structure expectation for proteins that do not have a nearby homologue of known structure [1].

Jones (1999) presented a two-arrange neural system that can be utilized to predict protein auxiliary structure in view of the position particular scoring frameworks created by PSI-

BLAST. Notwithstanding the effortlessness and comfort of the approach utilized, the outcomes are observed to be better than those delivered by different strategies, including the prominent PHD strategy as indicated by their own benchmarking comes about and the outcomes from the current Critical Assessment of Techniques for Protein Structure Prediction try (CASP3), where the technique was assessed by stringent blind testing [2].

Hua *et al.* (2001) presented another technique for protein secondary structure prediction which depends on the hypothesis of support vector machine (SVM). In this case, the performance of SVM either coordinates or is essentially superior to that of conventional machine learning approaches, including neural systems. The main utilization of the SVM way to deal with foresee protein secondary structure is depicted here. In the interim three-state general per-residue accuracy Q3 achieved 73.5 %, which is in any event similar to existing single prediction techniques. Moreover a helpful "dependability file" for the forecasts was produced. Also, SVM has numerous appealing features, including effective avoidance of overfitting, the capacity to deal with large feature spaces, data gathering of the given dataset, and so forth. The SVM technique is advantageously implemented on numerous other pattern recognition tasks in biology [3].

Pollastri *et al.* (2002) utilized ensembles of bidirectional recurrent neural network models, PSI-BLAST-determined profiles, and an huge non redundant training set to infer two new indicators: (a) the second form of the SSpro program for secondary structure classification into three types and (b) the principal version of the SSpro8 program for secondary structure classification into the eight classes created by the DSSP program. The outcomes of three distinctive test sets on which SSpro accomplished a sustained performance of around 78% amend prediction were described. SSpro and SSpro8 are additionally part of a more extensive suite of programs gone for predicting protein 3D structure by means of contact map prediction, and contact map prediction through forecast of basic features, for example, secondary structure, relative solvent accessibility (ACCpro), and contact numbers (CONpro) [4].

Steady advancement has been made in the field of initio protein folding. An variety of techniques now permit the determination of low-resolution structures of small proteins or protein parts up to roughly 100 amino acid residues in length. Such low-resolution structures might be adequate for the practical explanation of protein sequences on a all inclusive scale. Albeit no reliably dependable algorithm is right now accessible, the fundamental difficulties to building up a general hypothesis or way to deal with protein structure prediction are better understood. The energy landscapes coming about because of the structure forecast calculations are just mostly channeled to the local condition of

the protein. Corey Hardin *et al.* (2002) gave a survey that focuses on two ranges of late advances in ab initio structure prediction improvements in the vitality capacities and procedures to look through the caldera region of the energy landscapes [5].

Zhang *et al.* (2003) built up another combined approach for ab initio protein structure prediction. The protein compliance is portrayed as a lattice chain interfacing $c\alpha$, with joined $C\beta$ particles and side-chain centers of mass. The model compel field incorporates different short-run and long-extend learning based possibilities gotten from a factual examination of the regularities of protein structures. The mix of these energy terms is advanced through the boost of connection for $30 \times 60,000$ imitations between RMSD to native and energies, and in addition the vitality hole amongst local and the imitation gathering. To quicken the conformational search, a recently created parallel hyperbolic testing calculation with a composite development set is utilized in the Monte Carlo simulation processes. We exploit this procedure to effectively fold 41/100 small proteins $(36 \times 120 residues)$ with anticipated structures having a RMSD from local beneath 6.5Å in the main five bunch centroids. To overlay bigger size proteins also as to enhance the collapsing yield of small proteins, we join into the essential compel field side-chain contact forecasts from their threading program PROSPECTOR where homologous proteins were avoided from the information base. With these threading-based restraints, the program can overlay 83/125 test proteins $(36 \times 174 residues)$ with structures having a RMSD to local underneath 6.5 in the main five group centroids. This demonstrates the noteworthy change of collapsing by utilizing anticipated tertiary limitations, particularly when the precision of side-chain contact prediction is less than 20%. For native fold choice, we present quantities subject to the cluster density and the combination of energy furthermore, free energy, which demonstrate a higher discriminative energy to choose the native structure than the beforehand utilized group vitality or bunch size, and which can be utilized as a part of local structure recognizable proof in blind simulations. These techniques are promptly computerized and are being executed on a genomic scale [6].

Opricovic *et al.* (2004) displayed the relative analysis of the two Multiple Criteria Decision Analysis (MCDM) techniques : TOPSIS and VIKOR. In this work, similarity and the distinctions of these two strategies were represented using numerical cases. These two MCDM methods dispose of the units of benchmark functions by utilizing distinctive methods for standardization or normalisation. TOPSIS strategy utilizes linear and in addition vector standardization though VIKOR utilizes vector standardization. It is found from the relative examination that both of these strategies utilize diverse methods for standardization/normalisation and aggregating work [7].

The procedure of determination of the correct suppliers, who can give the right quality

items and in addition services to the buyer at the right cost, right time and justified amounts, is the most basic stride for the foundation of the perfect supply chain. On the other hand, it is a truly hard to choose a supplier since it includes different clashing options on which decision maker's information is normally insufficient and loose. In this way, it is commonly a multi criteria group decision making problem. Boran *et al.* (2009) proposed a TOPSIS procedure combined with intuitionistic fuzzy set for the determination of sufficient provider. Intuitionistic fuzzy weighted averaging (IFWA) operator is presented for amassing the individual thoughts of the decision makers to rate the parameters and accessible choices. Subsequently, keeping in mind the end goal to approve the proposed approach, a numerical case is proposed [8].

Mishra *et al.* (2014) built up a D2N (separation to local) calculation to determine the RMSD of protein structures [9]. Rana *et al.* (2015) actualized Self-adaptive Differential Evolution algorithm (SaDE) to discover significance of attributes, investigated different machine learning techniques in request to predict RMSD, TM-score and GDT TS-score and found that Random forest outperformed [10]. Pathak *et al.* (2016) proposed a technique in which the modelled protein structures were gathered. Artificial bee colony bee algorithm state calculation was executed to choose the centrality of components. ABC is likewise an iterative search algorithm like other population based algorithms. The accuracy is figured for every one of the models on various testing tests. It was discovered that the random forest have outperformed than different models [11].

2.2 Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

Green seller choice is a key stage in green supply chain management. The framework of assessment files for merchant choice was advanced in view of coal industry group corporation ltd. In the meantime, Li-Yun Wu *et al.* (2008) built up the multi-level model of TOPSIS, which depends on entropy weight. The subjectivity which lies in finding out variables weights in bring down chain of importance was evaded in this model. So the assessment result is more goal than other assessment techniques. The enhanced decision model was connected in vendor determination in Ping Dingshan coal industry group. The order preference of sellers is achieved. The fantastic vendor and perfect vendors are discovered for the enterprise. Whats more, the contrast among these sellers can be picked up too. The assessment result demonstrates that this strategy be more sensible and less demanding than different strategies. Thus, it is simpler to promote this assessment

technique in enterprises [12].

Biswas *et al.* (2016) proposed a technique that can help in the multi variate decision making issues by broadening TOPSIS strategy. Order of alternatives based on each parameter was considered as single valued neutrosophic set and henceforth reflect the choice in view of the available data. Three autonomous degrees (i.e. truth-membership degree (T), indeterminacy-membership degree (I), and falsity-membership degree (F)) described neutrosophic set to convey appropriate information/knowledge. Single-valued neutrosophic set-based weighted averaging operator makes a typical sentiment by collecting all choices to rate the criteria. Further, an illustrative case was given to express its viability what's more, reasonableness of the approach proposed [13].

Balcerzak *et al.* (2016) utilized TOPSIS strategy to break down the objects in light of the few monetary factors. The dynamic synthetic index that depict the relative level of sustainable development of the nations and proposes a rating of the nations and gather them into homogenous subsets was defined. The gathering was done with help of novel breaks procedure. The examination of the rate in the period 2004-2013 demonstrates that European Union have gained a noteworthy ground in executing the idea of sustainable development. This research was empowered to coordinate the nations that are the pioneers in the field [14].

Afsordegan *et al.* (2016) proposed a quantitative TOPSIS technique and showed utilizing a energy case study. Sustainable energy planning issues include taking fundamental choices in an variety of dynamic complexities as for clashing ecological, financial, social and specialized criteria. These methods includes power of choices and are fit to assess distinctive measure under vulnerability utilizing phonetic factors which include qualitative names by giving their opinions. Seven energy options under nine criteria were assessed by the supposition of three ecological and vitality specialists. The proposed approach is contrasted and an altered fuzzy TOPSIS strategy, demonstrating the benefits of the proposed approach when managing with linguistic assessments to demonstrate instability and imprecision. Despite the fact that the new approach requires less intellectual push to decision makers, it yields comparable outcomes [15].

2.3 Rough set based Technique

Pawlak (1998) gives essential thoughts of rough set theory, another way to deal with data analysis. The lower and upper estimation of a set, the fundamental operations of the hypothesis, are naturally clarified and formally de fined. A few uses of rough set theory are quickly illustrated. It gives productive algorithms to find concealed patterns in data and creates sets of decision rules from information. Many algorithms in light of the rough set theory are especially suited for parallel processing, yet keeping in mind the end goal to exploit this feature completely [16].

Xiaohua Hu *et al.* (2001) introduced a novel way to develop a ensemble utilizing rough set theory and database operations. Ensembles of classifiers are figured decisively under the system of rough set hypothesis and developed effectively by utilizing set-oriented database operations. The significant components of this technique when compared with different strategies for building group of classifiers are: (1) develops a hypothetical model to clarify the method of developing ensembles; (2) each reduct is a minimum subset of features and has an indistinguishable capacity from the whole parameters; (3) each reduct classifier that is built from the comparing reduct has a minimum set of rules; (4) the test demonstrates that the quantity of classifiers used to enhance the accuracy is substantially less than different strategies [17].

Ahn *et al.* (2002) proposed a hybrid intelligent system that predicts the failure of firms in light of the past monetary performance data, joining rough set approach and neural network. Diminished information table, which infers that the quantity of performance criteria, for example, budget and other subjective factors are decreased with no data loss, can be acquired through rough set approach. And afterward, this diminished information is utilized to create classification rules and train neural network to deduce suitable features. The principles created by rough set analysis demonstrate the best prediction accuracy if a case matches any of the rules. The justification of this approach is done by utilizing rules created by rough sets. The adequacy of this strategy was confirmed by tests looking at conventional discriminant analysis and neural network approach with this hybrid method [18].

Wang *et al.* (2010) introduced a tumor classification approach based on the ensemble of probabilistic neural system (PNN) and neighborhood rough set model based gene reduction. Useful qualities were chosen at first by gene ranking using an iterative search margin algorithm and afterward were additionally refined by gene reduction to choose numerous minimum gene subsets. Finally, the proposed PNN classifiers were incorpo-

rated by majority voting system to build an ensemble. Results demonstrated that this approach can get both high and stable classification performance [19].

2.4 Machine Learning Methods

According to Arthur Samuel in 1959, Machine Learning is the field of artificial intelligence which train a machine/computer to learn without being explicitly programmed. It is evolved from the study of computational learning theory and pattern recognition in artificial intelligence. Machine learning basically explores development of those algorithms that get experience from and predict on data. It is employed in problems where there is huge data and program can not be implemented explicitly. Machine learning and statistics are firmly related fields. According to Michael I. Jordan, the ideas of machine learning have long pre-history in statistics whether it is related methodological principles or theoretical tools. Machine learning is typically divided into three types and these are given in Figure 2.1.

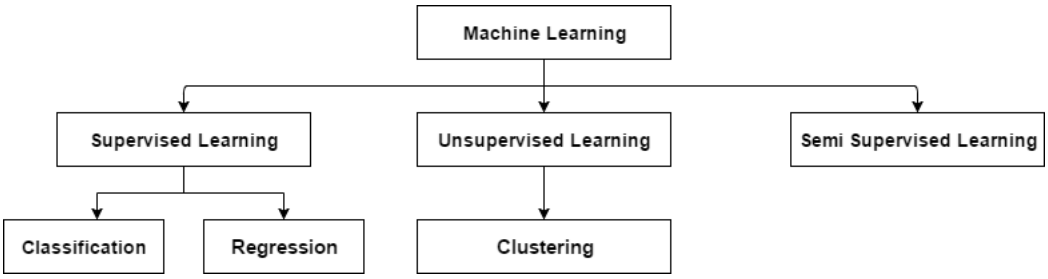


Figure 2.1: Classification of machine learning

1. **Supervised learning:** When the learning algorithm is presented with example inputs and associated output/labels. Supervised Learning is further classified into two categories:
    - **Classification Algorithms:** In this, labels are divided into two or more classes. Classes can be discrete values, string, character, etc. For example: Spam filtering
    - **Regression Algorithms:** In this, inputs are continuous rather than discrete.
2. **Unsupervised learning:** When the learning algorithm is given with no labels and its the goal of algorithm itself to find the pattern in the data or structure in inputs.

- **Clustering Algorithms:** In clustering, inputs is grouped into different sets. In this, the labels are known beforehand, hence its the goal of the algorithm itself to determine the group.

3. **Reinforcement learning:** When it is the goal of the learning algorithm to perform in perform in the dynamic environment in the certain way.

The Table 2.1 shows the recently developed classification and regression models. All the models are available in R open source software. R is licensed under GNU GPL.

Table 2.1: Machine learning models

S.No	Model	Model Type	Method	Package	Tuning Parameter
1	ada	Classification	ada	ada	maxdepth, iter, nu
2	avNNet	Dual Use	avNNet	caret	decay, size, bag
3	bag	Dual Use	bag	caret	vars
4	bdk	Dual Use	bdk	kohonen	xweight,topo,xdim,ydim
5	blackboost	Dual Use	blackboost	mboost	maxdepth, mstop
6	Boruta	Dual Use	Boruta	Boruta	mtry
7	bstTree	Dual Use	bstTree	bst	maxdepth, nu, mstop
8	C5.0	Classification	C5.0	C50	winnow, trials, model
9	cforest	Dual Use	cforest	party	mtry
10	ctree	Dual Use	ctree	party	mincriterion
11	cubist	Regression	cubist	Cubist	committees, neighbors
12	enet	Regression	enet	elasticnet	lambda, fraction
13	foba	Regression	foba	foba	lambda, k
14	GAMens	Classification	GAMens	GAMens	fusion, iter, rsm_size
15	gamLoess	Dual Use	gamLoess	gam	degree,span
16	gbm	Dual Use	gbm	gbm	trees, shrinkage,depth
17	gcvEarth	Dual Use	gcvEarth	earth	degree
18	glm	Dual Use	glm	stats	None
19	icr	Regression	icr	caret	n.comp
20	J48	Classification	J48	RWeka	C
21	JRip	Classification	JRip	RWeka	NumOpt
22	knn	Dual Use	knn	caret	k
23	lars	Regression	lars	lars	fraction
24	lda	Classification	lda	MASS	None
25	leapSeq	Regression	leapSeq	leaps	nvmax
26	Linda	Classification	Linda	rrcov	None
27	lm	Regression	lm	stats	None
28	logforest	Classification	logforest	LogForest	None
29	M5	Regression	M5	RWeka	smoothed,pruned,rules
30	nb	Classification	nb	klaR	usekernel, fL
31	neuralnet	Regression	neuralnet	neuralnet	layer1, layer2, layer3
32	nnet	Dual Use	nnet	nnet	size, decay

Table 2.1: Machine learning models (cont.)

S.No	Model	Model Type	Method	Package	Tuning Parameter
33	obliqueTree	Dual Use	obliqueTree	obliqueTree	splits, selection
34	OneR	Classification	OneR	RWeka	None
35	ORFsvm	Classification	ORFsvm	obliqueRF	mtry
36	pam	Classification	pam	pamr	threshold
37	parRF	Dual Use	parRF	randomForest	mtry
38	PART	Classification	PART	RWeka	pruned, threshold
39	partDSA	Dual Use	partDSA	partDSA	cut.off.growth, MPD
40	pcaNNet	Dual Use	pcaNNet	caret	decay, size
41	pcr	Regression	pcr	pls	ncomp
42	pda	Classification	pda	mda	lambda
43	plr	Classification	plr	stepPlr	cp, lambda
44	rbf	Dual Use	rbf	RSNNS	size
45	rbfDDA	Classification	rbfDDA	RSNNS	negativeThreshold
46	rda	Classification	rda	klaR	lambda, gamma
47	relaxo	Regression	relaxo	relaxo	phi, lambda
48	rf	Dual Use	rf	randomForest	mtry
49	rpart	Dual Use	rpart	rpart	cp
50	RRF	Dual Use	RRF	RRF	mtry,coefReg,coefImp
51	rvmLinear	Regression	rvmLinear	kernlab	None
52	sda	Classification	sda	sda	diagonal
53	slda	Classification	slda	ipred	None
54	smda	Classification	smda	sparseLDA	R, lambda, NumVars
55	svmLinear	Dual Use	svmLinear	kernlab	C
56	svmPoly	Dual Use	svmPoly	kernlab	degree, scale, C
57	svmRadial	Dual Use	svmRadial	kernlab	C, sigma
58	treebag	Dual Use	treebag	ipred	None

Chapter 3

Problem Formulation

3.1 Problem Statement

Biology would not have achieved anything without the discovery of the the laws of hereditary by Mendel. However, there are still several undetermined laws present in the nature that are very difficult to predict as compared to Mendel's laws. Hence, Computational techniques have emerged as very reliable, fast and effective techniques that can help in the discovery of such laws. With the help of these techniques, it will be easy to discover the hidden and undetermined laws to solve several complex biological problems and relieve the biologists from doing tedious work. The science of development and use of computational methods in complex biological problems is known as Bioinformatics.

Prediction of the structure of the amino acid sequence is one of the critical problem in biological environment and theoretical chemistry. The amino acid sequence (primary sequence) of a protein folds into some stable structures with minimum energy states. "Wet Lab" techniques that are used in order to determine these protein structures are expensive and time consuming. Hence, use of computational techniques to predict the protein structures are on the rise. There are several methods available ranging from the purely ab-initio methods which predicts based on the physiochemical properties, to the homology methods which predicts based on the information available in sequences and genomic databases. And there are several other approaches which lies between these two extremes such as fold-recognition or threading which predicts by identifying the structural template. In the study, ab-initio approach is used as physiochemical properties of the sequences are the key factors for folding protein into its tertiary protein structures. There are many physiochemical properties which are used to predict the Root Mean Square Deviation (RMSD) of the protein structure of the amino acid sequence which are listed below.

- Total surface area (Area)
- Euclidean distance (ED)
- Total empirical energy (Energy)

- Secondary structure penalty (SS)
- Sequence length (SL)
- Pair number (PN)

The techniques that are used these days to predict the RMSD of the protein structure of protein sequence are every expensive, time consuming and have high maintenance cost. However, this thesis is focused to predict protein structure with the help of computational approaches which are very efficient and cheap. It introduced a MDT approach to find the best set of classification models.

3.2 Research Gaps

Following are the gaps that are identified during Literature survey:

1. The Homologous approaches, which involves the study of the protein structure of the amino acid, are very expensive and time consuming [20].
2. Many prediction techniques results in low quality structures because of the absence of the proper knowledge and inadequate observations. The protein structures obtained may look same as that of any high resolution structure when monitored but in actual they may be away from their actual native states [21].
3. In ab initio approach, data is to be generated from the protein structure by calculating the physiochemical properties. The design of promising ab initio methods is a challenging task. The methods require tremendous amounts of computational power. For this reason, most of the present ab initio methods depend on a simplified representation of protein structures, rather than on an atomic level representation [22].
4. No one has applied rough set approach and used TOPSIS on this problem.

3.3 Research Objectives

The following research objectives are formulated:

1. Feature measurement of the protein structures of amino acid sequences.
2. To develop and analyse various machine learning algorithms and ranking of these algorithms using MCDM-TOPSIS.

3. To find the combination of the various machine learning algorithms belonging to different families which outperforms using rough set theory.
4. To validate the proposed approach using k fold cross validation.

Chapter 4

Data Set Description

There are 16382 decoys of protein with 4608 native structures and these sequences are collected from CASP and RCSB. Table 4.1 describe the physiochemical properties that are extracted from these sequences and used in this study. The dataset is multivariate, widely distributed and highly overlap. Table 4.2 shows the sample of data set.

4.1 Data set and its features

The modelled structures are taken from protein structure prediction center (CASP-5 to CASP-10 experiments) and public decoys database [23], while native structure from protein data bank (RCSB). There are total 500 protein structures consist of 67 native structures and remaining are modelled structures. Table 4.1 describes the physicochemical properties used in this study. Table 4.2 shows the sample of data set and correlation between each features is presented in Table 4.3.

4.2 Data transformation for Classification

Here, the value of RMSD of protein structure lies between the range 0Å-11Å. For classification purpose, RMSD value is transformed into discrete value with the help of equation

Table 4.1: Description of the features.

Feature	**Information**
Area	Total surface area.
ED	Euclidean distance.
Energy	Total empirical energy.
SS	Secondary structure penalty.
RL	Residue length
PN	Pair number

Table 4.2: Sample dataset.

RMSD	Area	ED	Energy	SS	RL	PN
0	43915.5	1566640	-96449	163	874	37040
5	8562.34	22305.6	8457.23	59	177	1313
4	8313.41	9194.25	1132.9	39	144	553
2	9224.05	20146.5	-3258.04	101	163	1162
5	5969.02	2528.08	378.82	9	81	169

Table 4.3: Correlation between each feature.

	Area	ED	Energy	SS	SL	PN
Area	1.00	0.82	0.00	0.58	0.98	0.88
ED	0.82	1.00	0.00	0.42	0.85	0.98
Energy	0.00	0.00	1.00	0.00	0.00	0.00
SS	0.58	0.42	0.00	1.00	0.58	0.47
SL	0.98	0.85	0.00	0.58	1.00	0.91
PN	0.88	0.98	0.00	0.47	0.91	1.00

4.1, keeping in mind that the closer RMSD have similar structures.

$$Class = \begin{cases} 0 & \text{if } 0 \leqslant RMSD \leqslant 1.0 \\ 1 & \text{if } 1.0 < RMSD \leqslant 2.0 \\ 2 & \text{if } 2.0 < RMSD \leqslant 3.0 \\ 3 & \text{if } 3.0 < RMSD \leqslant 4.0 \\ 4 & \text{if } 4.0 < RMSD \leqslant 5.0 \\ 5 & \text{if } 5.0 < RMSD \leqslant 6.0 \\ 6 & \text{if } 6.0 < RMSD \leqslant 7.0 \end{cases} \tag{4.1}$$

4.3 Feature Measurement

Six physical and chemical features are selected for the empirical analysis of protein prediction are shown in Table 4.1. In this chapter, measurement of the six physiochemical features namely Area, ED, Energy, SS, PN and RL is briefly explained and discussed.

4.3.1 Root Mean Square Deviation (RMSD)

The RMSD of the protein structure can be calculated as the average distance between the atoms of the superimposed proteins. Usually, RMSD is one of the very important

factor for the prediction of similarity between the two or more proteins. The RMSD of the protein is measured mathematically using the equation below:

$$RMSD = \sqrt{\sum_{i}^{N}(d_p * d_q)/N} \tag{4.2}$$

where, d_p is the distance between matched pair p, N is the number of matched pairs of protein. RMSD is calculated using the freely available program at [24].

4.3.2 Total surface area (Area)

Degree of the external forces on the protein depends on the surface of protein that is exposed to the solvent, which coveys the strong dependency of free energy on solvent accessible surface area (SASA) [25]. SASA has been widely used as one of the key factor to assess the quality of structure of the protein sequence. Each amino acid shows a different affinity to be found on the surface of the protein, based on the functional groups present in its side chain [26].

4.3.3 Euclidean distance (ED)

The overall conformation of a protein is decided by spatial positioning of $C\alpha$ atoms. Recently, neighborhood profiles of $C\alpha$ atoms for each pair of residues have been characterized and observed to be invariant in 3618 native proteins [27] suggesting certain geometrical constraints in their positioning. Here, four aliphatic non polar residues are considered namely Alanine (ALA), Valine (VAL), Leucine (LEU), and Isoleucine (ILE). These residues form 10 unique pairs among each other. Cumulative inter-atomic distance of their respective $C\beta$ atoms were calculated for each residue pair. Euclidean distance is calculated by taking the cumulative difference between $C\alpha$ and $C\beta$.

4.3.4 Total empirical energy (Energy)

The total empirical energy are explained by Arora and Jayaram [28] and narang et al. [29] and comprised of (i) electrostatic force, (ii) van der Waals force, and (iii) hydrophobic force. It is computed as follows:

$$E_{elec}^{pq} = (332 * q_p * q_q) \,/\, (D * r_{pq}) \tag{4.3}$$

$$E_{vdW}^{pq} = \left(C_{12}^{pq}/r_{pq}^{12}\right) - \left(C_{6}^{pq}/r_{pq}^{6}\right) \tag{4.4}$$

$$E_{hyd}^{pq} = \left(M_{12}^{pq}/r_{pq}^{12}\right) - \left(M_{6}^{pq}/r_{pq}^{6}\right) \tag{4.5}$$

where, r_{pq} is the distance between pair of atoms p and q, $C_{12}^{pq} = \epsilon\sigma^{12}$, $C_{6}^{pq} = 2\epsilon\sigma^{6}$, σ is the van der Waals radii, ϵ is the well depth, $M_{12}^{pq} = \epsilon R^{12}$, $M_{6}^{pq} = \epsilon R^{6}$, R is the distance variable and ϵ is set to 1.

Finally total empirical energy is given as:

$$E_{total} = \sum_{p}^{n-1} \sum_{q=p+1}^{n} \left(E_{elec}^{pq} + E_{vdW}^{pq} + E_{hyd}^{pq}\right) \tag{4.6}$$

4.3.5 Secondary Structure penalty (SS)

Secondary structure prediction has reached to 82% accuracy [30] over the last few years. Therefore, deviation from ideal predicted secondary structures can be used as a measure to quantify the quality of a structure. Detailed measurement of the secondary structure penalty is described by Mishra et al. [31].

4.3.6 Pair Number (PN)

Pair number is the total number of pairs between the $C\beta$ carbon in the protein sequence.

4.3.7 Residue Length (RL)

Residue length is the total number of $C\alpha$ carbons in the protein sequence.

4.4 Dimensionality Reduction Using Principal Component Analysis(PCA)

PCA is a procedure which is used for dimensional reduction. It searches uncorrelated features from dataset and these uncorrelated features are known as principal components.

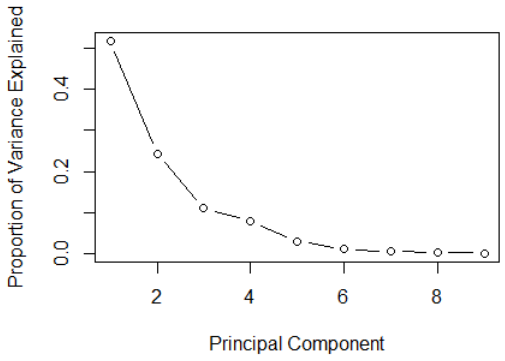
Table 4.4: Dataset sample after PCA.

RMSD	PC1	PC2	PC3	PC4	PC5	PC6
3	0.40	0.63	-0.01	-0.23	-0.08	0.09
2	-4.47	-0.53	-0.02	-2.52	0.04	-0.79
0	-0.27	1.91	-0.05	0.36	-0.13	-0.04
4	1.97	-0.88	0.06	0.52	0.07	-0.10
5	2.24	-1.83	0.09	0.33	0.26	-0.14

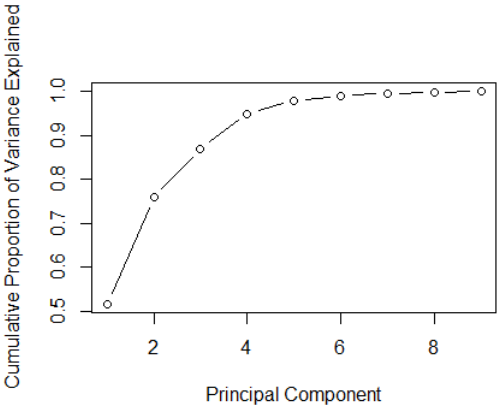
The objective of PCA is to expain maximum variance with least amount of features. Consider there are a set of predictors as P^1, P^2...,P^n. The principal components can be mathematically written as:

$$C^i = \Phi^{11}P^1 + \Phi^{21}P^2 + \Phi^{31}P^3 + ... + \Phi^{n1}P^n \tag{4.7}$$

where, C^1 is ith principal component. Φ^{ni} is the loading vector comprising of loadings (Φ^1 , Φ^2 ...) of ith principal component. $P^1..P^n$ are normalized predictors. Normalized predictors have zero mean and one standard deviation.



(a) Describing variance of different features using principle component analysis.



(b) Describing variance of different features/components using a cumulative variance plot.

Figure 4.1: Describing variance of different parameters using PCA and cumulative variance plot.

As one can observe in Figure 4.1a only 6 features out of 8 are enough to explain about 98.4% variance in the dataset. So we can use only these 6 features instead of 8 hence,reducing the dimensionality of our dataset. we can confirm our results be by plotting a cumulative variance plot. The plot in Figure 4.1b shows the that 6 features are enough to define variance close to 98%. Therefore, in this case, we will only be using these 6 features only. These 6 features are used for model training.

PCA is highly effective method for feature selection and extraction. PCA is used widely in many fields like data mining, pattern recognition, computer vision and data compression. PCA is more useful when dealing with higher dimensional samples [32].

In the next chapter, RMSD prediction of protein structure using TOPSIS based ensembled approach is introduced and classification algorithms are ranked based on the TOPSIS approach and RMSD prediction of Protein Structure using Rough Set Based Ensemble Approach is introduced.

Chapter 5

Proposed Methodology

5.1 Classification machine learning algorithms

The classification algorithms used in this paper include seventeen techniques from seventeen different families (discriminant analysis, random forests and other ensembles, neural networks, decision trees, rule-based classifiers, boosting, bagging, support vector machines, stacking, generalized linear models, nearest neighbors, partial least squares and principal component regression, Bayesian, logistic and multinomial regression, multiple adaptive regression splines and other methods). In order to have a comprehensible idea of the potential of each method and family, it would be better to develop a comparison of a large number of classifiers from different families and areas of knowledge [33]. The seventeen classification methods are PenalizedLDA, NaiveBayes (NB), Radial basis functions (RBF) neural network, Support vector machines (SVM), Decision tree, $C5.0$, AdaBoost, treebag, Deep Neural Network (DNN), Random Forest (RF), generalized linear model via penalized maximum likelihood (GLMNET), k nearest neighbor (KNN), Partial Least Squares (PLS), Support Vector Machine (KSVM), Functional Data Analysis (FDA), Partitioning Around Medoids (PAM), Multinomial logistic regression and Ordinal. In this section, the description of classification methods is presented in Table 5.1. This Table contains name of family to which the classification method belongs, their tuning parameters (which provide the maximum accuracy with the help of k - fold validation) and necessary packages.

5.2 MCDM-TOPSIS based ensembled approach for protein structure prediction

The problem of selecting the algorithm is a challenging task in the research area in almost every field, such as artificial intelligence, data mining, and machine learning. The evaluating the performance of classification algorithms usually involves more than one criterion, such as sensitivity, specificity, Area Under Curve, Positive Predictive Value,

S.no.	Family	Method	Tuning Parameters	Package
1	Discriminant analysis	PenalizedLDA	lambda = 1e-04 and K = 5	penalizedLDA
2	Bayesian (BY) approach	nb	fL = 0, usekernel = TRUE and adjust = 1.	nb
3	Neural networks	nnet	size=10, linout = TRUE, MaxNWts = 10000, trace = FALSE, maxit = 100	nnet
4	Decision trees	rpart	n = 12145	rpart
5	Rule-based methods	C5.0	trials = 20, model = tree and winnow = TRUE	c50
6	Boosting	adaboost	boos=TRUE and mfinal=5	adabag
7	Bagging	treebag		treebag
8	Randomforest	randomForest	Number of trees = 500, No. of variables tried at each split = 2	randomForest
9	stacking	dnn	layer1 = 3, layer2 = 0, layer3 = 2, hidden_dropout = 0 and visible_dropout = 0	dnn
10	Other ensembles	clm	link = logit, threshold = flexible, nobs = 12145, logLik = -9991.04, AIC = 20008.07, niter = 87(13), max.grad = 4.56e-07 and cond.H = 8.5e+20	ordinal
11	Generalized Linear Models	glmnet	alpha = 1 and lambda = 0.005744351	glmnet
12	Nearest neighbor method	knn	k = 9	knn
13	Partial least squares and principal component regression	pls	ncomp = 3	pls
14	Support vector machines	ksvm	kernel=“rbfdot“, prob.model=TRUE	kernlab
15	Multivariate adaptive regression splines	fda	degree = 1 and nprune = 12	mda
16	Logistic and multinomial regression	multinom	trace = FALSE, maxit = 1000	nnet
17	Other Methods	pam	threshold = 1.105954	pam

Table 5.1: A brief description of the classification machine learning methods used in this work.

Negative Predictive Value, Prevalence, Detection Rate and Detection Prevalence. Therefore algorithm selection can be modeled as multiple criteria decision making (MCDM) problems [34]. Since single performance parameter can not be used to rank models, one plausible route is to apply mixes of performance metrics to the single score. A score/rank concurred by numerous performance metrics is more trustful than depending on single parameter. In this work, first performance parameters are calculated by aggregating individual results of all classes and then the classification methods are ranked according to the MCDM TOPSIS algorithm. And hence a TOPSIS based ensemble approach will be introduced. As every class gives their own values of performance metrics thus, it is important to first generate the single performance score for the particular model. So, in the second step single value of each performance metric of particular model are calculated by using MCDM-TOPSIS [35]. In order to rank classification algorithms, several criteria such as sensitivity, specificity, Area Under Curve, Positive Predictive Value, Negative Predictive Value, Prevalence, Detection Rate and Detection Prevalence are considered. Therefore algorithm choice can be demonstrated as MCDM problems.

Hwang and Yoon [36] proposed the TOPSIS method that can be used to rank alternatives based on the number of different criterion available. It is a one of the statistical and analytical method. It can findout the order of evaluation objects with the help of the Positive Ideal Solution (PIS) and the Negative Ideal Solution (NIS). It finds the best options by minimizing the distance to the PIS and maximizing the distance to the NIS [37] and hence, the final ranking is calculated by means of the closeness index.

- Calculate the normalized decision matrix. The normalized value N_{pq} is calculated as:

$$N_{pq} = \frac{y_{pq}}{\sqrt{\sum_{q=1}^{Q} y_{pq}^2}}, q = 1, ..., Qp = 1, ..., P \tag{5.1}$$

  where Q and P denote the number of alternatives (i.e. classification methods) and the number of criteria (i.e. performance metrics), respectively. For alternative A_q, the performance metric of the pth criterion C_p is represented by y_{pq}.

- Find a set of weights w_p for each criterion and calculate the weighted normalized decision matrix. The weighted normalized value v_{pq} is calculated as: $v_{pq} = w_i y_{pq}$, q = 1,..., Q; p = 1,..., P. where w_p is the weight of the pth criterion, and $\sum_{p=1}^{n} = 1$.

- Find the PIS S^+, which is calculated as:

$$S^+ = v_1^+, ..., v_n^+ = (\max_q v_{pq} | p \in I^{'}), (min_q v_{pq} | pI^{''}),$$

  where $I^{'}$ is associated with benefit criteria and $I^{''}$ is associated with cost criteria.

- Find the NIS S, which is calculated as:

$$S = v_1, ..., v_n = (\min_q v_{pq} | p \in I^{'}), (\max_q v_{pq} | p \in I^{''}).$$

- calculate the separation measures, using then-dimensional Euclidean distance. The separation of each alternative from the ideal solution is calculated as:

$$D_q^+ = \sqrt{\sum_{p=1}^{n} (v_{pq} - v_p^+)^2}, q = 1, ..., Q$$

  The separation of each alternative from the negative-ideal solution is calculated as:

$$D_q^- = \sqrt{\sum_{p=1}^{n} (v_{pq} - v_p^-)^2}, q = 1, ..., Q$$

- Calculate a ratio R_q^+ that measures the relative closeness to the ideal solution and is calculated as:

$$R_q^+ = \frac{D_q^-}{D_q^+ + D_q^-}, p = 1, ..., Q$$

- Rank alternatives by maximizing the ratio in Step 5.

In this approach, the seventeen classification machine learning algorithms are ranked with the help of the TOPSIS algorithm discussed above and then the ensembles are developed based on the ranking. The eight ensembles are made by taking odd number of models into account. The best model out of the eight models is chosen based on the evaluation parameters of these eight ensembles. The flow diagram of this approach is described in the Figure 5.1.

5.3 Rough Set Based Ensemble Approach for protein structure prediction

No Free Lunch (NFL) theorem states, if one algorithm performs better on some cost fuction, it does not convey that it will perform better on every cost function, there is the possibility that some other algorithm performs better. There is no single algorithm that can give best performance for all performance measures. To evaluate classification algorithms, normally many criterion are examined. Hence, in order to avoid the baised

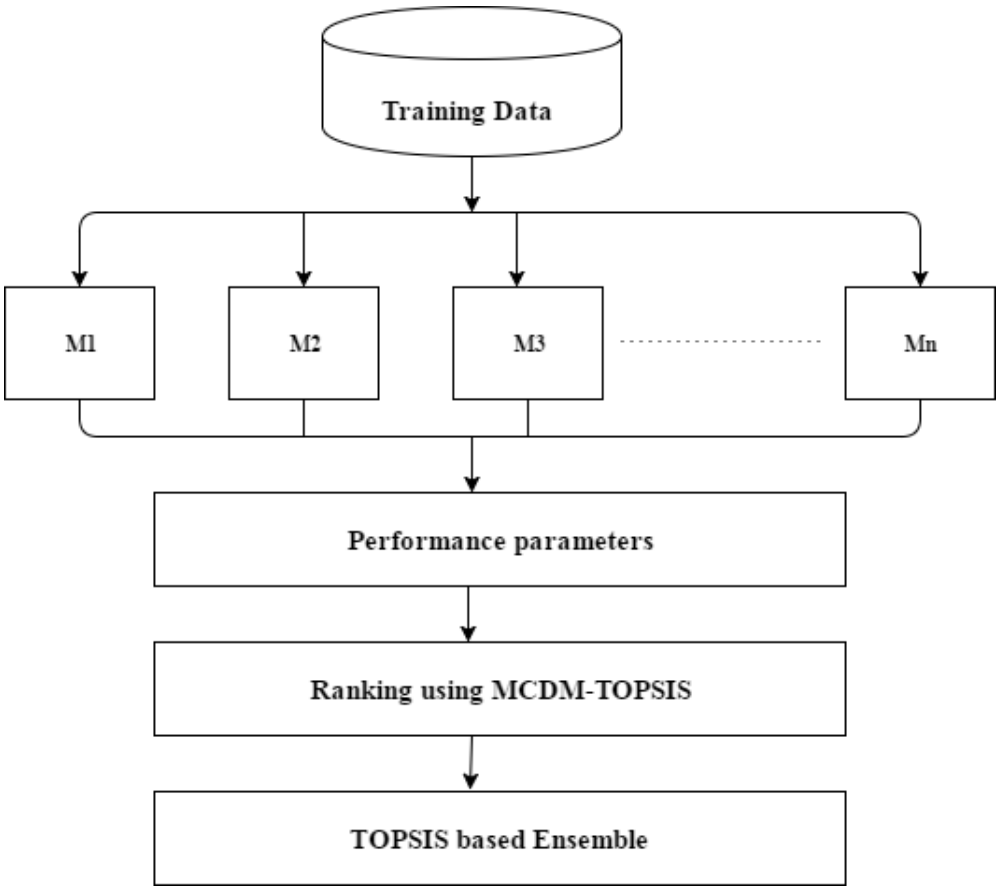
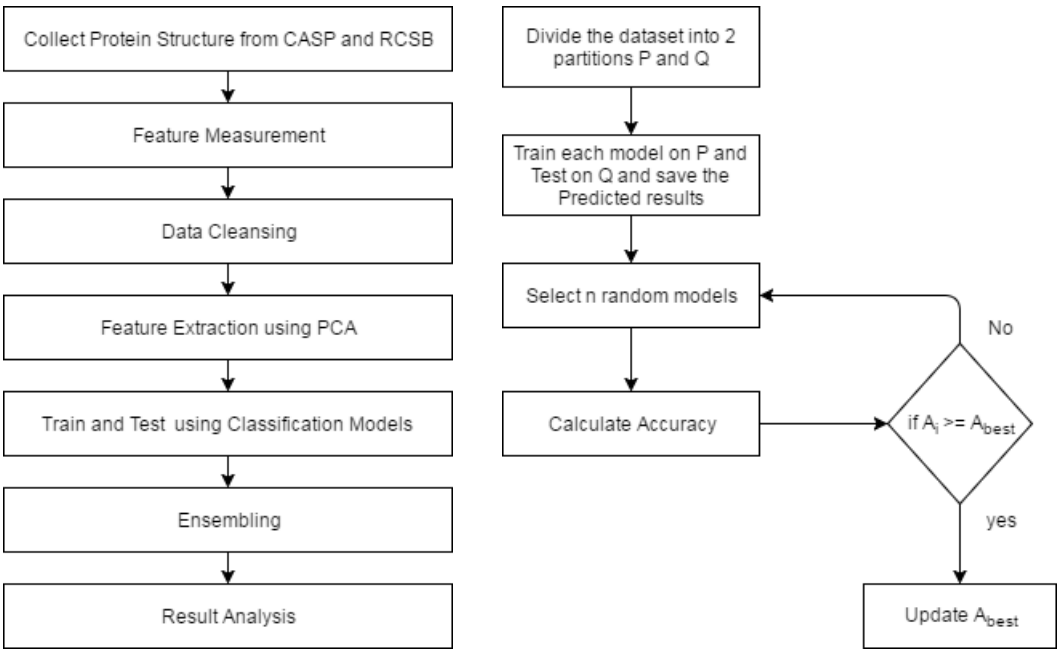


Figure 5.1: Flow diagram of the TOPSIS based ensemble approach.

results, the rough set based ensembling technique is used. The approach is explained in Figure 5.2a.

The rough set based ensemble technique is explained in the Algorithm 5.1 and Figure 5.2b. In this algorithm, first the dataset is divided into two parts P and Q and N is the number of models. Train each classification method on P and test the model on test data Q. The predicted values of all the classification methods obtained after testing are stored in their respective P_i. These steps are iterated k times (where k = 10,000). In each cycle of loop, random rough sets of classification models are selected and then the predicted values of the ensemble are calculated using majority voting ensembling and compared by calculating accuracy. If better accuracy is obtained than the previous accuracy, then it is stored in A_{best}. Hence, with the help of this algorithm, a less correlated, more accurate and robust ensembled model is obtained at the end.



(a) Methodology Used. (b) Flow diagram of rough set based ensembling.

Figure 5.2: Flow diagram of the methodology and rough set approach.

Algorithm 5.1 : Rough set Based Ensembling

Function RSBEnsemble():
Divide Dataset into two parts P and Q
for each model $M_i \in BucketB$ **do**
(i) Train M_i with P;
(ii) Test M_i on Q and store results in P_i;
end for
for Repeat following steps k times **do**
(i) Select n random models from B where n ∈ N;
(ii) Use majority voting to find new predicted values using n models;
(iii) Calculate new accuracy A_i;
if $A_i \geq A_{best}$ **then**
$A_{best} = A_i$;
Models = n
end if
end for

5.4 Model Evaluation Technique

This work uses Sensitivity, Specificity, Negative Predictive Value, Prevalence, Positive Predictive Value, Detection Rate, Detection Prevalence and Area under the receiver operating characteristic curve (AUC) to evaluate the classification methods. In this study, seventeen machine learning classification models belonging to different classes [33] are an-

alyzed using six physiochemical properties to determine RMSD of the protein structure in absence of its true native state. A confusion matrix (Kohavi & Provost, 1998) consists of information about actual and predicted classifications carried out using classification methods. Table 5.2 shows the confusion matrix or error matrix for a particular classifier. Sensitivity, specificity, Negative Predictive value, Positive Predictive Value, Detection Rate and Area under curve (AUC) can be defined mathematically by using the elements of the confusion matrix as

Actual	Predicted	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

Table 5.2: Confusion matrix representation

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.3}$$

$$PositivePredictiveValue = \frac{TP}{TP + FP} \tag{5.4}$$

$$NegativePredictiveValue = \frac{TN}{TN + FN} \tag{5.5}$$

$$DetectionRate = \frac{TP}{TN + FN + TP + FP} \tag{5.6}$$

Receiver operating characteristic (ROC) shows the ratio between True Positive rate and False Positive rate [38]. The area under the ROC curve (AUC) explains the performance of a classifier. Larger the area, higher the performance.

The k-fold cross validation is carried out to evaluate the performance metrics of the predictive model. The total sample is subdivided randomly into k equal size sub samples. The (k-1) samples are used for training purpose, and remaining one sample is used for testing purpose in order to validate the model. This process is then reiterated k folds by taking each sub sample as the validation sample at least once. Further, the single score is obtained by aggregating the results from k folds.

In the next chapter, the results of the two proposed methods are discussed briefly.

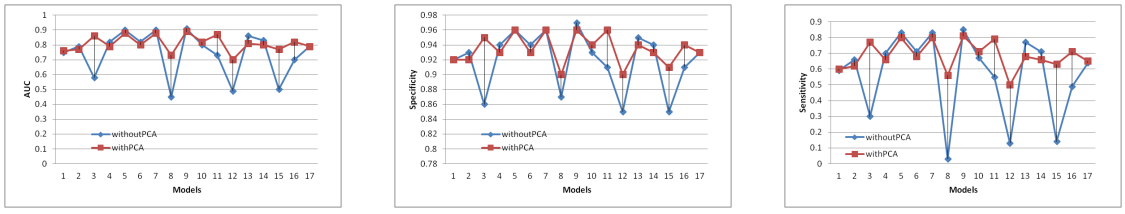
Chapter 6

Results Discussion

The prediction results of seventeen classification algorithms are analyzed on the testing dataset (CASP 10). The prediction results of models belonging to different families are analyzed using Principle component analysis as given in Table 5.1. It is observed from the table that Random forest outperformed. In the Figure 6.1, these results are compared with the results obtained using original features based on Area Under ROC curve, Sensitivity and Specificity. It is found that the performance of most of the models using PCA is much better than using original features.

Using a PCA we can easily identify what are the most important dimensions and just keep a few of them to explain most of the variance we see in out data. Hence we can drastically reduce the dimensionality of the data and make exploratory data analysis feasible again. Moreover, it will also enable us to identify what the most important variables in the original feature space are, that contribute most to the most important Principle components. Intuitively, one can imagine, that a dimension that has not much variability cannot explain much of the happenings and thus is not as important as more variable dimensions.

The seventeen methods are evaluated on the tuning parameters provided in Table 5.1. The model-wise performance parameters of all classes using equation $(5-9)$ is given in the Table 6.1. The aggregate results are given in Table 6.2.



(a) AUC (b) Specificity (c) Sensitivity

Figure 6.1: Comparison of results before using PCA and after using PCA

Table 6.1: Model-wise performance parameters of the seventeen classification machine learning models.

Model	Parameter	Class0	Class1	Class2	Class3	Class4	Class5
PenalizedLDA	Sensitivity	0.82	0.54	0.66	0.51	0.43	0.61
	Specificity	0.94	0.98	0.92	0.91	0.85	0.91
	Pos Pred Value	0.75	0.76	0.48	0.63	0.38	0.61
	Neg Pred Value	0.96	0.94	0.96	0.86	0.87	0.91
	Prevalence	0.18	0.11	0.1	0.24	0.18	0.18
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.88	0.76	0.79	0.71	0.64	0.76
nb	Sensitivity	0.88	0.7	0.63	0.5	0.46	0.53
	Specificity	0.97	0.98	0.93	0.9	0.82	0.94
	Pos Pred Value	0.89	0.82	0.58	0.6	0.16	0.78
	Neg Pred Value	0.97	0.97	0.94	0.86	0.95	0.84
	Prevalence	0.2	0.09	0.13	0.23	0.07	0.27
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.93	0.84	0.78	0.7	0.64	0.74
nnet	Sensitivity	1	0.88	0.77	0.65	0.57	0.73
	Specificity	1	0.99	0.96	0.93	0.89	0.93
	Pos Pred Value	1	0.9	0.76	0.72	0.55	0.67
	Neg Pred Value	1	0.99	0.96	0.91	0.9	0.94
	Prevalence	0.2	0.08	0.14	0.21	0.19	0.17
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	1	0.94	0.87	0.79	0.73	0.83
rpart	Sensitivity	0.86	0.74	0.59	0.51	0.48	0.75
	Specificity	1	0.98	0.91	0.91	0.88	0.89
	Pos Pred Value	0.99	0.73	0.46	0.66	0.52	0.46
	Neg Pred Value	0.96	0.98	0.95	0.85	0.86	0.97
	Prevalence	0.23	0.08	0.11	0.25	0.21	0.11
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.93	0.86	0.75	0.71	0.68	0.82
C5.0	Sensitivity	0.99	0.9	0.8	0.69	0.64	0.76
	Specificity	1	0.99	0.97	0.94	0.9	0.94
	Pos Pred Value	1	0.93	0.79	0.76	0.6	0.73
	Neg Pred Value	1	0.99	0.97	0.92	0.92	0.95
	Prevalence	0.2	0.08	0.14	0.21	0.18	0.18
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	1	0.95	0.88	0.82	0.77	0.85
adaboost	Sensitivity	0.78	0.8	0.65	0.54	0.49	0.79
	Specificity	1	0.98	0.91	0.91	0.89	0.89
	Pos Pred Value	1	0.71	0.44	0.63	0.58	0.49
	Neg Pred Value	0.93	0.98	0.96	0.87	0.85	0.97
	Prevalence	0.26	0.07	0.1	0.22	0.24	0.11

to be cont'd on next page

Table 6.1: Model-wise performance parameters of the seventeen classification machine learning models (cont.)

Model	Parameter	Class0	Class1	Class2	Class3	Class4	Class5
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.89	0.89	0.78	0.72	0.69	0.84
treebag	Sensitivity	0.99	0.89	0.8	0.69	0.64	0.78
	Specificity	1	0.99	0.96	0.94	0.91	0.94
	Pos Pred Value	1	0.92	0.79	0.75	0.62	0.72
	Neg Pred Value	1	0.99	0.97	0.92	0.91	0.95
	Prevalence	0.2	0.08	0.14	0.21	0.19	0.17
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	1	0.94	0.88	0.81	0.77	0.86
dnn	Sensitivity	0.6	1	0.65	0.4	0.26	0.47
	Specificity	1	0.92	0.87	0.84	0.81	0.98
	Pos Pred Value	1	0	0.14	0.31	0.16	0.92
	Neg Pred Value	0.84	1	0.99	0.89	0.89	0.76
	Prevalence	0.33	0	0.03	0.15	0.12	0.37
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.8	0.96	0.76	0.62	0.54	0.72
rf	Sensitivity	0.99	0.91	0.81	0.71	0.65	0.79
	Specificity	1	0.99	0.97	0.95	0.91	0.94
	Pos Pred Value	1	0.93	0.81	0.77	0.63	0.72
	Neg Pred Value	1	0.99	0.97	0.93	0.92	0.96
	Prevalence	0.2	0.08	0.15	0.21	0.19	0.17
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	1	0.95	0.89	0.83	0.78	0.86
glmnet	Sensitivity	0.98	0.84	0.68	0.55	0.5	0.7
	Specificity	1	0.98	0.94	0.91	0.87	0.93
	Pos Pred Value	1	0.82	0.64	0.62	0.48	0.67
	Neg Pred Value	0.99	0.99	0.95	0.88	0.88	0.93
	Prevalence	0.2	0.08	0.14	0.21	0.19	0.18
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.99	0.91	0.81	0.73	0.69	0.81
knn	Sensitivity	0.99	0.88	0.79	0.69	0.62	0.78
	Specificity	1	0.99	0.96	0.94	0.91	0.93
	Pos Pred Value	1	0.92	0.79	0.73	0.62	0.71
	Neg Pred Value	1	0.99	0.96	0.92	0.91	0.95
	Prevalence	0.2	0.08	0.14	0.2	0.2	0.17
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	1	0.93	0.88	0.81	0.76	0.86
pls	Sensitivity	0.45	1	0.31	0.42	0.41	0.41
	Specificity	1	0.92	0.86	0.83	0.82	0.96
	Pos Pred Value	1	0	0.01	0.18	0.17	0.86

to be cont'd on next page

Table 6.1: Model-wise performance parameters of the seventeen classification machine learning models (cont.)

Model	Parameter	Class0	Class1	Class2	Class3	Class4	Class5
	Neg Pred Value	0.69	1	1	0.94	0.94	0.72
	Prevalence	0.44	0	0	0.08	0.08	0.39
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.72	0.96	0.58	0.63	0.61	0.68
polr	Sensitivity	0.9	0.74	0.65	0.53	0.51	0.77
	Specificity	1	0.97	0.94	0.9	0.89	0.91
	Pos Pred Value	1	0.59	0.63	0.6	0.56	0.58
	Neg Pred Value	0.97	0.98	0.94	0.88	0.87	0.96
	Prevalence	0.22	0.06	0.14	0.21	0.22	0.14
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.95	0.85	0.79	0.72	0.7	0.84
fda	Sensitivity	0.85	0.73	0.64	0.57	0.51	0.68
	Specificity	0.99	0.97	0.94	0.9	0.87	0.92
	Pos Pred Value	0.97	0.67	0.62	0.59	0.46	0.67
	Neg Pred Value	0.96	0.98	0.94	0.89	0.89	0.93
	Prevalence	0.23	0.07	0.14	0.2	0.18	0.18
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.92	0.85	0.79	0.74	0.69	0.8
pam	Sensitivity	0.63	1	0.6	0.42	0.46	0.66
	Specificity	1	0.92	0.86	0.91	0.87	0.91
	Pos Pred Value	1	0	0	0.69	0.48	0.57
	Neg Pred Value	0.85	1	1	0.78	0.86	0.93
	Prevalence	0.32	0	0	0.31	0.2	0.16
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.81	0.96	0.73	0.67	0.66	0.78
multinom	Sensitivity	0.99	0.84	0.68	0.56	0.5	0.7
	Specificity	1	0.99	0.94	0.91	0.87	0.93
	Pos Pred Value	1	0.85	0.64	0.62	0.48	0.67
	Neg Pred Value	1	0.99	0.95	0.88	0.88	0.93
	Prevalence	0.2	0.08	0.14	0.21	0.19	0.18
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.99	0.91	0.81	0.73	0.69	0.81
$ordinal_clm$	Sensitivity	0.85	0.69	0.65	0.53	0.49	0.71
	Specificity	1	0.95	0.93	0.9	0.87	0.92
	Pos Pred Value	1	0.44	0.6	0.59	0.49	0.65
	Neg Pred Value	0.96	0.98	0.94	0.87	0.88	0.94
	Prevalence	0.24	0.05	0.13	0.21	0.19	0.17
	Detection Prevalence	0.2	0.08	0.15	0.19	0.2	0.19
	AUC	0.92	0.82	0.79	0.71	0.68	0.81
svm	Sensitivity	0.99	0.86	0.73	0.67	0.58	0.79

to be cont'd on next page

Table 6.2: Evaluation results of the seventeen classification methods used in this work.

Model Name	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Detection Rate**	**AUC**
PenalizedLDA	0.60	0.92	0.60	0.92	0.10	0.76
nb	0.62	0.92	0.64	0.92	0.10	0.77
nnet	0.77	0.95	0.77	0.95	0.13	0.86
rpart	0.66	0.93	0.64	0.93	0.11	0.79
c5.0	0.80	0.96	0.80	**0.96**	0.13	0.88
adaboost	0.68	0.93	0.64	0.93	0.11	0.80
treebag	0.80	**0.96**	0.80	**0.96**	0.13	0.88
dnn	0.56	0.90	0.42	0.90	0.08	0.73
rf	**0.81**	0.95	**0.81**	**0.96**	0.13	**0.89**
glmnet	0.71	0.94	0.71	0.94	0.12	0.82
knn	0.79	**0.96**	0.8	**0.96**	0.13	0.87
pls	0.5	0.90	0.37	0.88	0.07	0.70
ksvm	0.77	0.95	0.77	0.95	**0.14**	0.86
fda	0.66	0.93	0.66	0.93	0.11	0.8
pam	0.63	0.91	0.46	0.9	0.09	0.77
multinom	0.71	0.94	0.71	0.94	0.12	0.82
ordinal	0.65	0.93	0.63	0.93	0.11	0.79

Table 6.1: Model-wise performance parameters of the seventeen classification machine learning models (cont.)

Model	**Parameter**	**Class0**	**Class1**	**Class2**	**Class3**	**Class4**	**Class5**
	Specificity	1.00	0.99	0.97	0.93	0.90	0.92
	Pos Pred Value	1.00	0.87	0.80	0.71	0.59	0.65
	Neg Pred Value	1.00	0.99	0.95	0.91	0.90	0.96
	Prevalence	0.20	0.08	0.16	0.21	0.19	0.15
	Detection Prevalence	0.20	0.08	0.14	0.20	0.19	0.19
	AUC	0.99	0.93	0.85	0.80	0.74	0.85

6.1 RMSD prediction of protein structure using MCDM-TOPSIS based ensembled approach

It is clear from the Table 6.2 that no single algorithm is good for all performance parameters. Hence, there is the need to determine a single performance score which can be

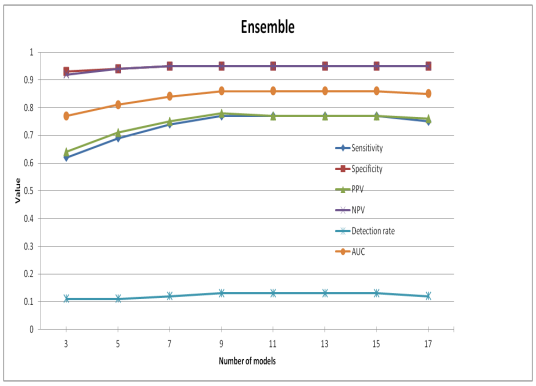
S.no.	Model	Ranking
1	penalizedLDA	14
2	nb	13
3	nnet	5
4	rpart	11
5	c5.0	2
6	adaboost	10
7	treebag	3
8	dnn	16
9	rf	1
10	glmnet	6
11	knn	4
12	pls	17
13	polr	8
14	fda	9
15	pam	15
16	multinom	7
17	ordinal	12

Table 6.3: Ranking of seventeen classification algorithms using MCDM-TOPSIS.

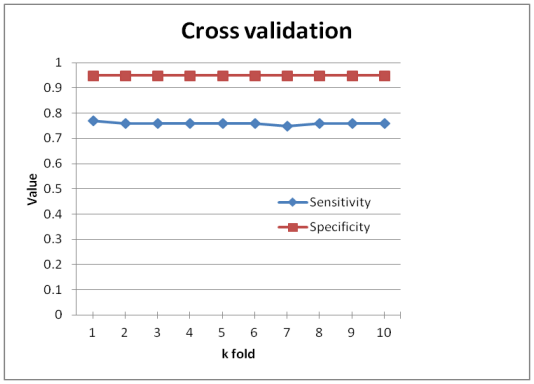
derived from these eight performance metrics that can be used as the measure to compare the classification algorithms. Hence, MCDM-TOPSIS is implemented to find the rank of seventeen algorithms by utilising TOPSIS algorithm in Section 5. In Table 6.3, ranking of the nine classification algorithms is shown. The eight ensembles of odd number of different classification models are generated. The performance of eight ensembles generated with the help of seventeen classification models are given in the Figure 6.2a. It is concluded from the figure that ensemble of nine classification models (i.e. rf, C5.0, treebag, multinom, fda, polr, knn, glmnet and nnet) outperformed with an sensitivity 0.77, specificity 0.96 and AUC of 0.86. 10 fold cross validation of the TOPSIS based ensemble is done to ensure the robustness of the proposed model.

6.2 RMSD Prediction of Protein Structure using Rough Set Based Ensemble Approach

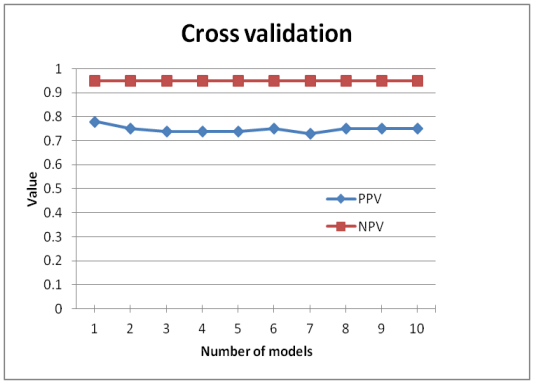
Further ensembling is performed using rough set Based Ensembling technique given in Chapter 5. The results obtained using the proposed RSB Ensemble algorithm are provided in the Figure 6.3a. It is evident from the figure that accuracy of the ensembled models increases upto iteration 2000, after iteration 2000, the is no prominent change in the accuracy value. Hence, the ensembled model of Random Forest, Support Vector



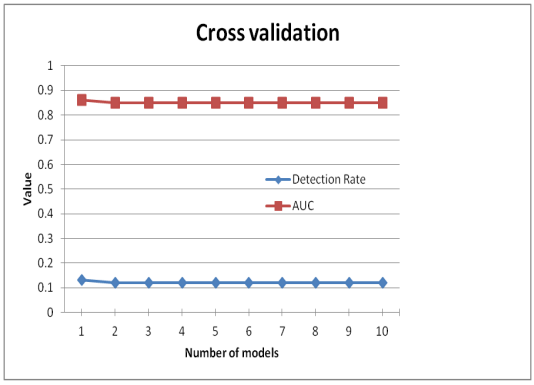
(a) Comparison of evaluation results of the eight ensembles developed based on TOPSIS ranking.



(b) k-fold cross validation of sensitivity and specificity using ensemble of nine models.



(c) k-fold cross validation of PPV and NPV using ensemble of nine models.



(d) k-fold cross validation of Detection rate and AUC using ensemble of nine models.

Figure 6.2: Evaluation results of eight ensembles and cross validation of the outperforming proposed ensemble.

Machine and $C5.0$ is the resultant ensembled model that outperformed with an accuracy of 84.94%. The k-fold cross validation is implemented in order to verify whether the proposed technique is robustness or not in Figure 6.3b.

6.2.1 Model Validation

Training and testing of the protein structure is done on the CASP (5-9) data. CASP 10 was not used in the whole process and kept for the validation purpose. Hence, this data is used on resultant ensemble model to validate the technique. These CASP 10 models are evaluated used proposed model and are compared with three already available models (ProQ2 Server [39], MetaMQAPII [40] and D2N: Distance to the native [9]). Table 6.5 shows the results.

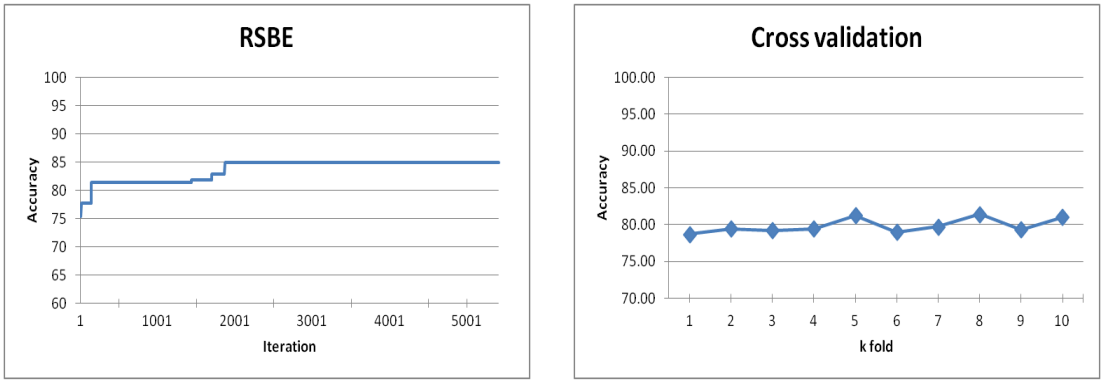
It is evident that rough set based ensembling (RSBE) outperformed with 9 out of 16

Iteration no.	**Accuracy**	**Set of models**
100	75.38	rpart, nnet, rf, fda, pls, polr, dnn, PenalizedLDA, C5.0, nb and knn
300	77.69	dnn, multinom, rf, treebag, polr, pls and fda
500	77.69	dnn, multinom, rf, treebag, polr, pls and fda
700	81.40	adaboost, treebag, pls, $C5.0$, ordinal, rf and nnet
900	81.40	adaboost, treebag, pls, $C5.0$, ordinal, rf and nnet
1100	81.40	adaboost, treebag, pls, $C5.0$, ordinal, rf and nnet
1300	81.40	adaboost, treebag, pls, $C5.0$, ordinal, rf and nnet
1500	81.83	treebag, fda, rf, $C5.0$ and adaboost
1700	82.94	rf, polr and $C5.0$
1900	82.94	rf, polr and $C5.0$
2100	84.94	rf, svm and $C5.0$

Table 6.4: Accuracy of ensembles generated at different iterations with RSBE algorithm.

S no	PDB id	Actual	RSBE	ProQ2	MetaMQII	D2N
1	$T0651\_RaptorX\_TS1$	5	5	4	4	5
2	$T0654\_BhageerathH\_TS2$	5	5	6	4	3
3	$T0654\_chunkTASSER\_TS3$	5	3	4	4	4
4	$T0654\_MATRIX\_TS2$	5	5	3	4	5
5	$T0654\_MufoldMD\_TS2$	3	2	2	4	3
6	$T0659\_AOBAserver\_TS4$	1	1	5	4	5
7	$T0659\_BhageerathH\_TS4$	4	2	3	6	3
8	$T0659\_FALCONTOPO\_TS2$	1	1	4	5	4
9	$T0659\_Jiang\_Fold\_TS4$	1	1	6	4	1
10	$T0659\_QUARK\_TS2$	4	5	4	5	2
11	$T0667\_MATRIX\_TS2$	4	4	7	5	5
12	$T0667\_MULTICOM-CLUSTER\_TS4$	5	5	4	4	5
13	$T0743\_PconsM\_TS1$	5	3	3	4	4
14	$T0743\_PMS\_TS3$	6	5	4	4	3
15	$T0753\_MULTICOM-CLUSTER\_TS3$	3	3	3	4	2
16	$T0644\_MufoldMD\_TS2$	5	5	3	4	5
			9/16	2/16	0/16	6/16

Table 6.5: Validation on CASP 10 dataset using RSBE technique.



(a) Running rough set based ensembling on 2100 iterations.

(b) 10 fold Cross validation for ensembled model.

Figure 6.3: Evaluation results of RSBE and cross validation of the outperforming ensemble.

correct results. Although other three models perform quite well in case of regression but in case of classification, RSBE gave impressive results.

In the next section, the conclusion of the whole thesis is discussed and some suggestions towards which the present work can be further extended are proposed.

Chapter 7

Conclusions and Future Works

This chapter is the concluding part of the thesis and also proposes some suggestions towards which the present work can be further extended. Section 7.1 brings out the overall conclusions of the research work carried out in this thesis and in section 7.2 suggestions regarding the future research directions and possible extensions of the work presented in the thesis are made. The figure 7.1 shows the brief summary of the research contributions.

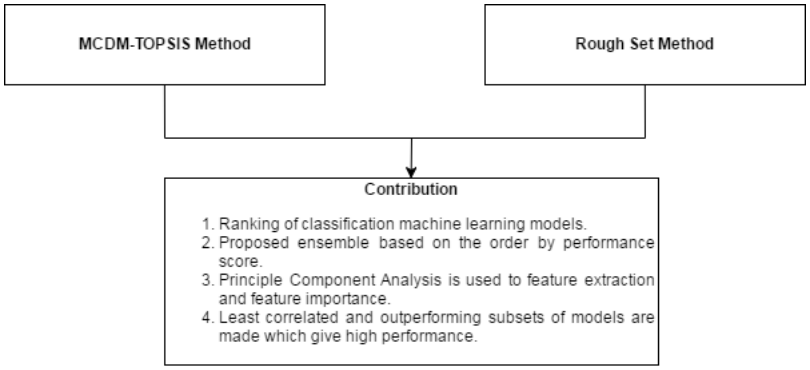


Figure 7.1: Research conclusion

7.1 Conclusion

In this thesis, an attempt has been made to solve the protein structure prediction problem computational techniques, MCDM-TOPSIS and rough set method. The main contribution of this thesis are done in several phases and they are as follows:

1. A comparative performance study is carried out for the classification of protein structure using machine learning models and six physicochemical properties.
2. Principle Component Analysis (PCA) is implemented for feature importance. It is found that data perform better by extracting features using PCA.
3. A MCDM-TOPSIS technique is used to rank the classification machine learning techniques and then combination of machine learning models which outperformed

is found (Chapter 6). The empirical study showed that the combination of performance scores of classification machine learning methods and promote the awareness of identified protein structures among decision makers. Hence, TOPSIS based three ensembles are generated.

4. A ensemble of classification machine learning which is least correlated, efficient and reliable is found with the help of rough set theory (Chapter 5). A novel rough set based ensemble is developed with high accuracy, Sensitivity, specificity, Area under the receiver operating characteristic curve (AUC), Positive Predictive Value (PPV), Negative Predictive value (NPV) and Detection Rate.

7.2 Scope for future work

Research is an iterative and continuous procedure. The work presented in the thesis focuses on the solving protein structure prediction problem using MCDM-TOPSIS and Rough set method. There are several directions in which this work could be expanded. Some of the suggestions for future work in this direction are:

1. Efficient modification can be done in the parallel implementation of the proposed RSBE algorithm that may improves its performance. A parallel implementation of the algorithms may be designed for a distributed and shared memory architecture.
2. The implementation of Nature Inspired Algorithms (NIA) (such as genetic algorithms, differential algorithm and particle swarm optimisation)
3. The protein structure prediction is done using four features such as total surface area, euclidean distance, total empirical energy and secondary structure penalty. More features need to be explored for more accurate prediction.
4. In this thesis, classification machine learning models are used for protein structure prediction. They need to be explored for accurate and fast predictions.

References

[1] Kim T Simons, Rich Bonneau, Ingo Ruczinski, and David Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):171–176, 1999.

[2] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292(2):195–202, 1999.

[3] Sujun Hua and Zhirong Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, 308(2):397–407, 2001.

[4] Gianluca Pollastri, Darisz Przybylski, Burkhard Rost, and Pierre Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002.

[5] Corey Hardin, Taras V Pogorelov, and Zaida Luthey-Schulten. Ab initio protein structure prediction. *Current opinion in structural biology*, 12(2):176–181, 2002.

[6] Yang Zhang, Andrzej Kolinski, and Jeffrey Skolnick. Touchstone ii: a new approach to ab initio protein structure prediction. *Biophysical journal*, 85(2):1145–1164, 2003.

[7] Serafim Opricovic and Gwo-Hshiung Tzeng. Compromise solution by mcdm methods: A comparative analysis of vikor and topsis. *European journal of operational research*, 156(2):445–455, 2004.

[8] Fatih Emre Boran, Serkan Genç, Mustafa Kurt, and Diyar Akay. A multi-criteria intuitionistic fuzzy group decision making for supplier selection with topsis method. *Expert Systems with Applications*, 36(8):11363–11368, 2009.

[9] Avinash Mishra, Prashant Singh Rana, Aditya Mittal, and B Jayaram. D2n: Distance to the native. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1844(10):1798–1807, 2014.

[10] Prashant Singh Rana, Harish Sharma, Mahua Bhattacharya, and Anupam Shukla. Quality assessment of modeled protein structure using physicochemical properties. *Journal of bioinformatics and computational biology*, 13(02):1550005, 2015.

[11] Yadunath Pathak, Prashant Singh Rana, PK Singh, and Mukesh Saraswat. Protein structure prediction (rmsd 5 å) using machine learning models. *International Journal of Data Mining and Bioinformatics*, 14(1):71–85, 2016.

[12] Li-Yun Wu and Yu-Zhong Yang. Topsis method for green vendor selection in coal industry group. In *Machine Learning and Cybernetics, 2008 International Conference*

on, volume 3, pages 1721–1725. IEEE, 2008.

[13] Pranab Biswas, Surapati Pramanik, and Bibhas C Giri. Topsis method for multi-attribute group decision-making under single-valued neutrosophic environment. *Neural computing and Applications*, 27(3):727–737, 2016.

[14] Adam P Balcerzak, Michal Bernard Pietrzak, et al. Application of topsis method for analysis of sustainable development in european union countries. *Chapters*, 1:82–92, 2016.

[15] Arayeh Afsordegan, M Sánchez, N Agell, Siamak Zahedi, and LV Cremades. Decision making under uncertainty using a qualitative topsis method for selecting sustainable energy alternatives. *International journal of environmental science and technology*, 13(6):1419–1432, 2016.

[16] Zdzislaw Pawlak. Rough set theory and its applications to data analysis. *Cybernetics & Systems*, 29(7):661–688, 1998.

[17] Xiaohua Hu. Using rough sets theory and database operations to construct a good ensemble of classifiers for data mining applications. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 233–240. IEEE, 2001.

[18] H.A. Abbass. The self-adaptive pareto differential evolution algorithm. In *IEEE Congress on Evolutionary Computation (CEC)*, volume 1, pages 831–836, 2002.

[19] Shu-Lin Wang, Xueling Li, Shanwen Zhang, Jie Gui, and De-Shuang Huang. Tumor classification by combining pnn classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology and Medicine*, 40(2):179–189, 2010.

[20] Yang Zhang. Progress and challenges in protein structure prediction. *Current opinion in structural biology*, 18(3):342–348, 2008.

[21] Johannes Söding, Andreas Biegert, and Andrei N Lupas. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(suppl_2):W244–W248, 2005.

[22] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. *Annual review of biophysics and biomolecular structure*, 30(1):173–189, 2001.

[23] www.scfbio-iitd.res.in/software/pcsm/dataset/Public_Decoys. 2012.

[24] http://zhanglab.ccmb.med.umich.edu/TMscore/RMSD.f. 2012.

[25] Elizabeth Durham, Brent Dorr, Nils Woetzel, René Staritzbichler, and Jens Meiler. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of molecular modeling, Springer*, 15(9):1093–1108, 2009.

[26] JOEL Janin. Surface and inside volumes in globular proteins. 1979.

[27] Aditya Mittal and B Jayaram. Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *Journal of Biomolecular Structure and Dynamics, Taylor*

& Francis, 28(4):443–454, 2011.

[28] Nidhi Arora and B Jayaram. Energetics of base pairs in B-DNA in solution: an appraisal of potential functions and dielectric treatments. *The Journal of Physical Chemistry B, ACS Publications*, 102(31):6139–6144, 1998.

[29] Pooja Narang, Kumkum Bhushan, Surojit Bose, and B Jayaram. Protein structure evaluation using an all-atom energy based empirical scoring function. *Journal of Biomolecular Structure and Dynamics, Taylor & Francis*, 23(4):385–406, 2006.

[30] Taner Z Sen, Robert L Jernigan, Jean Garnier, and Andrzej Kloczkowski. GOR V server for protein secondary structure prediction. *Bioinformatics, Oxford University Press*, 21(11):2787–2788, 2005.

[31] Avinash Mishra, Satyanarayan Rao, Aditya Mittal, and B Jayaram. Capturing Native/Native like Structures with a Physico-Chemical Metric (pcSM) in Protein Folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, Elsevier*, 2013.

[32] Q Guo, W Wu, DL Massart, C Boucon, and S De Jong. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1):123–132, 2002.

[33] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res*, 15(1):3133–3181, 2014.

[34] Salvatore Greco, J Figueira, and M Ehrgott. Multiple criteria decision analysis. *Springer's International series*, 2005.

[35] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[36] Jih-Jeng Huang and K Yoon. *Multiple attribute decision making: methods and applications*. Chapman and Hall/CRC, 2011.

[37] David L Olson. Comparison of weights in topsis models. *Mathematical and Computer Modelling*, 40(7-8):721–727, 2004.

[38] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[39] A Ray, E Lindahl, and B Wallner. Improved model quality assessment using ProQ2. *Bioinformatics*, 13:224, 2012.

[40] Djillali Tahri, Mohammed Seba, and Khedidja Benarous. Structure homology modeling of thaumetopoein, an urticating protein from thaumetopoea pityocampa schiff, using swiss-model workspace. *Chemical Informatics*, 2015.

List of Publications

1. Jagmeet Kaur, Akash Shrivastava, Abhishek Kapoor and Prashant Singh Rana, *"RMSD prediction of protein structure using MCDM-TOPSIS based ensembled approach"*, RACCCS-2017: 2nd International Conference on Recent Advancements in Computer, Communication and Computational Sciences. [UnderReview]

2. Jagmeet Kaur, Abhishek Kapoor, Akash Shrivastava and Prashant Singh Rana, *"RMSD Prediction of Protein Structure using Rough Set Based Ensemble Approach"*, National Conference on "Breaking Barriers through Bioinformatics and Computational Biology". [UnderReview]

3. Pratibha Sharma, Jagmeet Kaur, Vinay Arora, and Prashant Singh Rana. *"Information Retrieval in Web Crawling Using Population Based and Local Search Based Meta-heuristics: A Review"*, Springer, pages 87-104, 2017.