Chapter 4

A TOP DOWN APPROACH FOR THE PREDICTION OF ENZYME FUNCTIONAL CLASSES AND SUBCLASSES

Enzymes are catalysts which speed up the rate of reaction without becoming the part of reaction. The Enzyme Commission (EC) has defined a hierarchical classification scheme for the enzyme proteins based on functional mechanism of the enzymes. Each enzyme is denoted by an EC number of the format a.b.c.d, where 'a' at the top of this scheme represents one of the six main classes, each main enzymes class further subdivided to three levels in the hierarchy. The six main classes are Oxidoreductases (EC1), Transferases (EC2), Hydrolases (EC3), Lyases (EC4), Isomerases (EC5), and Ligases (EC6). Here, we used the enzyme proteins that are classified to their functional classes and subclasses according to EC number. Enzyme proteins play an important role in metabolic pathways so it is necessary to develop an efficient and robust computational intelligence based approach to predict the enzyme functional classes and subclasses. Prediction of enzyme functional classes and subclasses play an important role into the research of the drugs design.

In this chapter, a rotational random forest is proposed to predict enzyme functional classes and subclasses by using sequence derived properties of a protein. In this chapter, 857 number of sequence derived features with seven features vectors such as amino acid composition, dipeptide composition,

correlation, and composition, transition, distribution and pseudo amino acid composition are used to predict enzyme functional classes and subclasses. In this chapter, the proposed method used three level strategies to predict enzyme functional classes and subclasses. First, it determines that protein sequence is enzyme or non-enzyme. Second, if protein is classified as enzyme then the method classifies it into the six functional classes of enzymes. Third, it is classified into the functional subclasses of enzymes.

4.1. Background

Here, the sequence of enzymes with their properties and the proposed methods and models that are used to predict functional classes and subclasses of enzymes are presented.

4.1.1. Material and methods

In this chapter, the sequences of enzymes are extracted from the enzyme repository such as SWISS-PROT (Boeckmann *et al.*, 2003) and universal protein resource (UniProt) (Bairoch *et al.*, 2005) database. Here, all the non-enzyme proteins are selected from Uniport database with the keyword NOT enzyme. To avoid the homology bias the CD-HIT server (Huang *et al.,* 2010) is used to remove the homologous sequence using 90% sequence identity as the cutoff. The description of the datasets is shown in Table 4.1.

4.1.2. Features extraction of protein sequences

In this chapter, to fully characterize protein sequence seven feature vectors are used to represent the protein sample, including amino acid composition (AAC), dipeptide composition (DC), correlation factors (CF), composition, transition, distribution (CTD) of physiochemical properties and pseudo amino acid composition (PAAC) with total of 857 number of features are extracted from the PROFEAT server (Rao *et al*., 2011) for the prediction of enzyme functional classes and their subclasses.

The description of total 857 number of sequence derived features of enzymes is shown in Table 4.2.

Table 4.1: Number of sequences belonging to each enzyme functional classes and sub-classes

Enzyme/non-enzyme	Classes of Enzymes	Subclasses of Enzymes	No. of sequences	No. of seq.	No. of seq.
Enzyme	EC1	EC1.1	99	533	2647
		EC1.2	88		
		EC1.3	97		
		EC1.4	88		
		EC1.5	98		
		EC1.6	63		
	EC2	EC2.1	98	553	
		EC2.2	91		
		EC2.3	72		
		EC2.4	97		
		EC2.5	97		
		EC2.6	98		
	EC3	EC3.1	56	420	
		EC3.2	90		
		EC3.4	99		
		EC3.5	93		
		EC3.6	82		
	EC4	EC4.1	89	330	
		EC4.2	97		
		EC4.3	61		
		EC4.4	83		
	EC5	EC5.1	90	446	
		EC5.2	99		
		EC5.3	62		
		EC5.4	97		
		EC5.5	98		
	EC6	EC6.1	59	365	
		EC6.2	73		
		EC6.3	40		
		EC6.5	98		
		EC6.6	95		
Non Enzyme					700

Table 4.2: Description of sequence derived features of enzymes

S. No.	Features of protein sequences	Total No. of features	Description
1	X_1 to X_{20}	20	Amino acid composition
2	X_{21} to X_{420}	400	Dipeptide composition
3	X_{421} to X_{660}	240	Correlation factors
4	X_{661} to X_{681}	21	Composition
5	X_{682} to X_{702}	21	Transition
6	X_{703} to X_{807}	105	Distribution of physiochemical properties
7	X_{808} to X_{857}	50	Pseudo amino acid composition

4.2. Proposed method and model

In this chapter, a rotational random forest is proposed to predict enzyme functional classes and subclasses by using sequence derived properties of an enzymes. The proposed method used three level strategies to predict enzyme functional classes and subclasses. Here, the pseudo code for the proposed method for the prediction of enzymes functional classes and subclasses is presented.

4.2.1. Prediction of enzyme functional classes and subclasses

In this chapter, a Rotation Random Forest (RRF), ensemble of rotation forest and random forest is proposed for the prediction of enzyme functional classes and subclasses. In literature various ensemble classifiers are available to solve these types of problems such as Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1996), Random Subspace Method (Ho, 1998), Random Forest (Breiman, 2001), and Rotation Forest (Rodriguez *et al.*, 2006). In bagging we take a bootstrap sample of objects and training classifier whereas the classifiers results are adapted by majority voting. Boosting enhance the functionality of a weak classifier by integrating it in an ensemble of classifiers. The random subspace method reduces dimensionality by randomly sampling subsets of dataset. It modifies the training dataset by generating k new training sets and by building classifiers using these modified training sets. The results are combined using majority voting. The random forest consists of decision trees in which case the node splitting the tree being considered as the best feature amongst the feature space. The rotation forest is an ensemble of classifiers based on feature extraction where principal component analysis (Wold *et al.*, 1987) is used to rotate the feature subsets. The rotation forest ensemble classifier is believed to outperform all other ensemble strategies as it retains all the principal components and can be used with different selections of base classifiers.

Therefore, to incorporate the advantages of rotation forest and random forest an ensemble classifier based on the combination of rotation forest and random forest a rotation random forest (RRF) is proposed for the prediction of enzyme functional classes and subclasses. Rotation forest is based on feature

extraction by using principal component analysis. Unlike other classification techniques, the training data is not just a proportion of the whole data set instead the features present in the data set are randomly spitted into k subsets. Consequently PCA is applied on each subset. The feature subsets could be disjoint or intersecting but here, in order to achieve diversity the disjoint subsets are selected. In order to make sure that all of the information lies in the data so all the feature sets having the principal components are kept under observation. Therefore, a completely new feature set is obtained which is fed as input to the random forest after undergoing k-axis rotations. This technique enhanced individual accuracy and diversity for the classifiers present in the ensemble.

4.2.2. Pseudo code for the rotation random forest

The pseudo code for prediction of enzyme functional classes and sub-classes using proposed method is given as follows:

Pseudo code for the rotation random forest

Let X be the dataset containing the training instances in a form of a matrix of order (N x n), where $x = [x_1 \ldots x_n]^T$ represents a particular instance being described by n features.

Let Y represents a vector of class labels for X in the form (N x 1) as $Y = [y_1 \ldots y_N]^T$.

The L number of classifiers in the ensemble is represented by $C_1 \ldots C_L$.

The feature set of the whole dataset is represented by F.

Step-I Training phase

Given

X: $X = [x_1 \ldots x_n]$ in the form of a (N x n) matrix as a training data set.

Y: (N x 1) Column matrix representing class labels for X.

F: The original feature set

k: Number of feature subsets for input to PCA

C: Classifiers in the ensemble, C_1 ....C_L

L: Number of classifiers in the ensemble

1. Use 10-fold cross validation on X and Y to obtain training and testing data sets along with their corresponding class labels

2. For i = 1 to L

{

Prepare the rotation matrix R_i

Divide F into k random subsets

For j = 1 to k

{

a) Select X_{ij} dataset for a random subset of features

b) Select a bootstrap sample from X_{ij} of size 50%

c) Apply PCA on the bootstrap sample of the training objects in order to obtain the coefficient matrix R_{ij}.

}

Arrange the R_{ij} for j=1 to k in a rotation matrix R_i

Construct R_i by rearranging the column of R_i according to the

feature in F.

Generate the classification model for the random forest C_i using

(XR_i, Y) as a training set.

}

Step-II Classification phase

a) Generate number of trees using bootstrap samples from the (XR_i, Y) as a datasets.

b) For each of the bootstrap samples, grow an unpruned classification tree where at each node rather than choosing the best split among all predictors, randomly sample of the predictors and choose the best split from among those variables.

c) Predict test samples by aggregating the predictions of the number of trees i.e., majority votes for classification.

The proposed method used three level strategies to predict enzyme functional classes and sub-classes. The complete procedure of the proposed method for the prediction of enzyme functional classes and subclasses is illustrated in Figure 4.1 and the steps are as follows:

Step 1: Produce seven feature vectors with 857 number of features that represent a protein sequence.

Step2: Apply PCA on the bootstrap sample of the training datasets in order to obtain the coefficient matrix.

Step 3: Apply rotation random forest classifier for each of the three levels for the prediction of enzyme functional classes and sub-classes are as follows:

Firstly, it discriminates that protein sequence is enzyme or non-enzyme. Secondly, if protein is classified as enzyme then the method classifies the protein into six main functional classes of enzymes. Finally, it classifies the functional sub-classes of enzymes.

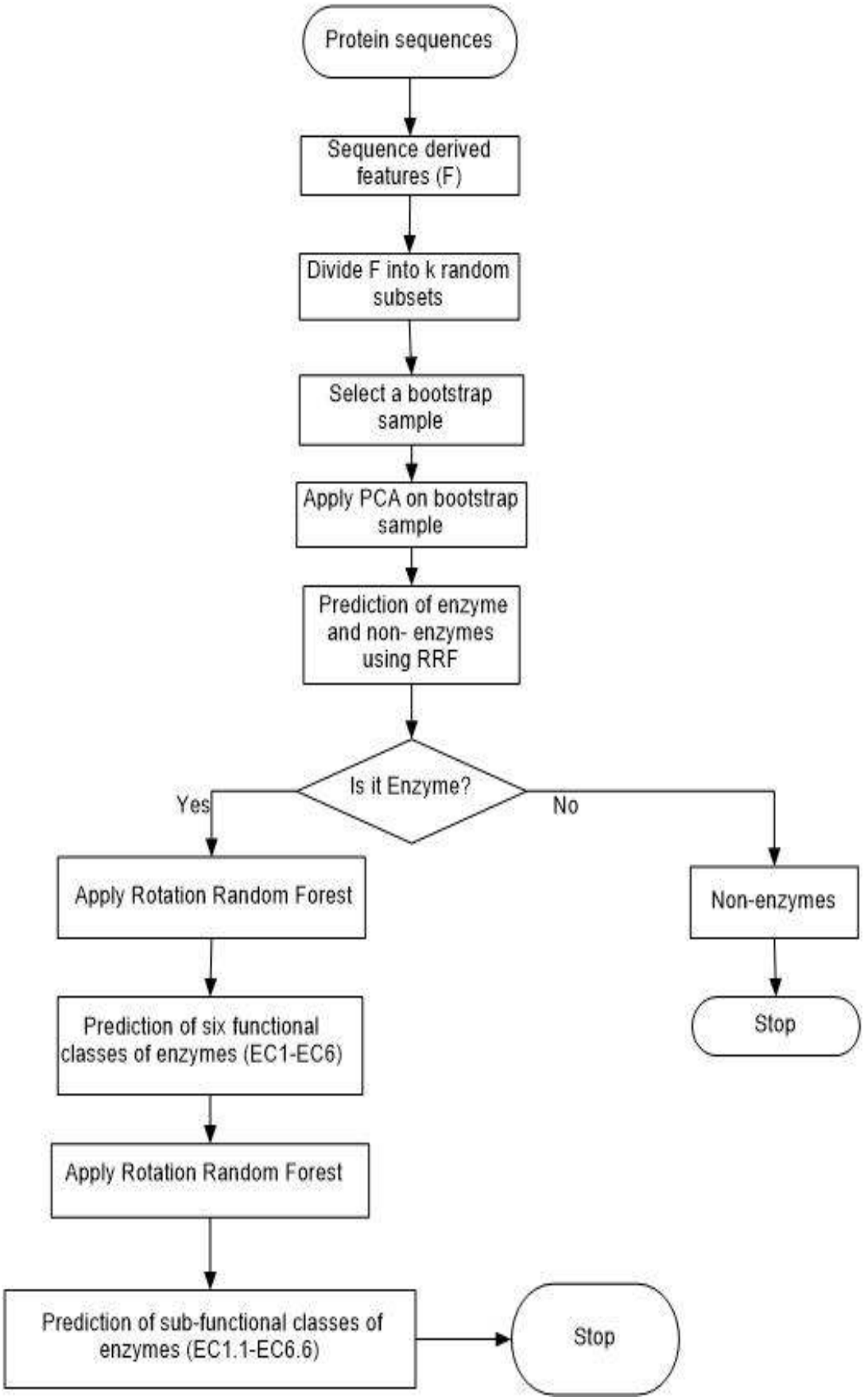

Figure 4.1: Prediction of enzyme functional classes and subclasses

4.3. Results and performance analysis

Here, an ensemble classifier Rotation Random Forest based on Rotation Forest and Random Forest classifiers is proposed for the prediction of enzyme functional classes and subclasses. For partitioning of the datasets into train and test sets and evaluating the performance of the proposed model the 10-fold cross validations are used. In subsequent subsections the results and performance analysis of the proposed model for the prediction of enzyme functional classes and subclasses are presented and discussed.

4.3.1. Performance measures

In this chapter, 10-fold cross validation is used to measure the performance of rotation random forest classifier. In *K*-fold cross validation the dataset of all proteins is partitioned into *K* subsets where one subset is used for validation and remaining *K-1* subsets is used for training. This process is repeated for *K* times so that every subset is used once as a test data. In this chapter, accuracy (*ACC*), precision, receiver operating characteristics (*ROC*) and Matthew's correlation coefficient (*MCC*) are used to measure the performance.

4.3.2. Results and comparative analysis

The proposed method used three level strategies to predict enzyme functional classes and sub-classes. This section presents the results and analysis of the proposed methods i.e. rotational random forest for the prediction of enzymes and non-enzymes, functional classes and subclasses of enzymes.

4.3.2.1. Prediction of enzymes and non-enzymes

To predict the enzymes and non-enzymes, a rotation random forest is evaluated using 10-fold cross validation and it is observed that the 100% accuracy is obtained for discrimination between enzymes and non-enzymes (See Table-4.3).

Table 4.3: Result analysis for the classification of enzymes and non-enzymes

Family	Proposed Method (Rotational Random Forest)			
	ACC	**MCC**	**ROC area**	**Precision**
Non-enzyme	100	1.00	1.00	100
Enzyme	100	1.00	1.00	100
Overall	**100**	**1.00**	**1.00**	**100**

4.3.2.2. Prediction of enzymes functional classes

To predict enzymes functional classes a rotation random forest is evaluated using 10-fold cross validation and it is observed that the overall 88.7% accuracy is obtained for the prediction of enzymes functional classes. The complete analysis of results is shown in Table 4.4.

Table 4.4: Result analysis for the classification of enzymes functional classes

Enzyme Classes	Proposed Method (Rotational Random Forest)			
	ACC	**MCC**	**ROC area**	**Precision**
Oxidoreductases (EC-1)	90.4	0.89	0.98	91.1
Transferases (EC-2)	92.6	0.80	0.97	77.1
Hydrolases (EC-3)	87.1	0.87	0.98	90.8
Lyases (EC-4)	82.1	0.90	0.99	100
Isomerases (EC-5)	85.9	0.86	0.99	91.2
Ligases (EC-6)	91.2	0.91	0.99	92.5
Overall	88.7	0.87	0.98	89.5

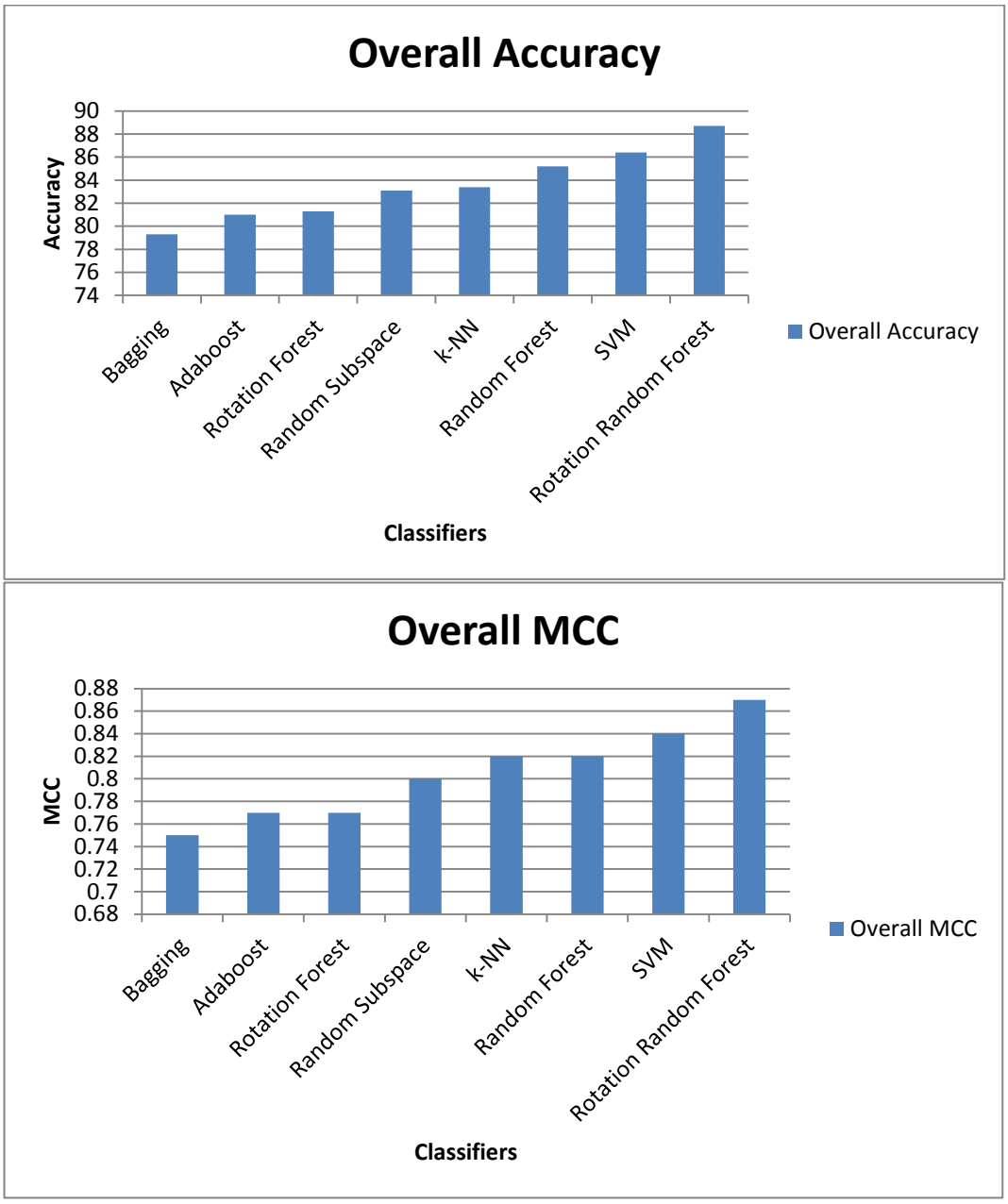

Figure 4.2: Accuracy and MCC obtained by different ensemble classifiers for the prediction of enzyme functional classes

The proposed method provides overall accuracy of 88.7%, MCC of 0.87 ROC area value of 0.98 and precision of 89.5% for the prediction of enzymes functional classes with the complete datasets (See Table 4.4). The Result obtained from proposed method is also compared with different classifiers and from Figure-4.2 it is observed that the proposed method may perform better in comparison with other classifiers such as Bagging, Adaboost, Rotation Forest, k- nearest neighbor (k-NN), Random Forest and support vector machine (SVM).

4.3.2.3. Prediction of functional subclasses of enzymes

To predict functional subclasses of enzymes a rotation random forest is evaluated using 10-fold cross validation and it is observed that the overall 87.6% accuracy is obtained for the prediction of functional subclasses of enzymes. The complete analysis of results is shown in Table 4.5.

Table 4.5: Result analysis for the prediction of functional subclasses of enzymes

Subclasses of Enzymes	Proposed Method (Rotation Random Forest)			
	ACC	MCC	ROC area	Precision
EC1.1	92.9	0.95	0.99	97.9
EC1.2	96.6	0.97	1.00	96.6
EC1.3	94.8	0.87	0.99	80
EC1.4	80.7	0.86	0.97	93.4
EC1.5	82.7	0.80	0.96	78.6
EC1.6	66.7	0.80	0.94	97.7
EC2.1	83.7	0.79	0.97	75.9
EC2.2	95.6	0.93	1.00	91.6
EC2.3	72.2	0.81	0.96	91.2
EC2.4	75.3	0.73	0.95	73
EC2.5	90.7	0.88	0.98	86.3
EC2.6	85.7	0.84	0.99	83.2
EC3.1	80.4	0.88	0.98	97.8
EC3.2	93.3	0.93	0.99	92.3
EC3.4	85.9	0.84	0.98	83.3
EC3.5	91.4	0.94	0.97	97.7
EC3.6	80.5	0.88	0.96	97.1
EC4.1	92.1	0.94	0.99	96.5
EC4.2	94.8	0.94	0.99	93.9
EC4.3	88.5	0.94	1.00	100
EC4.4	83.1	0.88	0.98	94.5
EC5.1	85.6	0.83	0.97	81.9
EC5.2	86.9	0.86	0.99	86
EC5.3	93.5	0.97	0.99	100
EC5.4	87.6	0.88	0.99	89.5
EC5.5	86.7	0.90	0.98	94.4
EC6.1	100	1.00	1.00	100
EC6.2	91.8	0.93	0.99	94.4
EC6.3	77.5	0.87	0.94	96.9
EC6.5	98	0.76	1.00	60.4
EC6.6	89.5	0.85	0.99	82.5
Overall	87.6	0.88	0.98	88.8

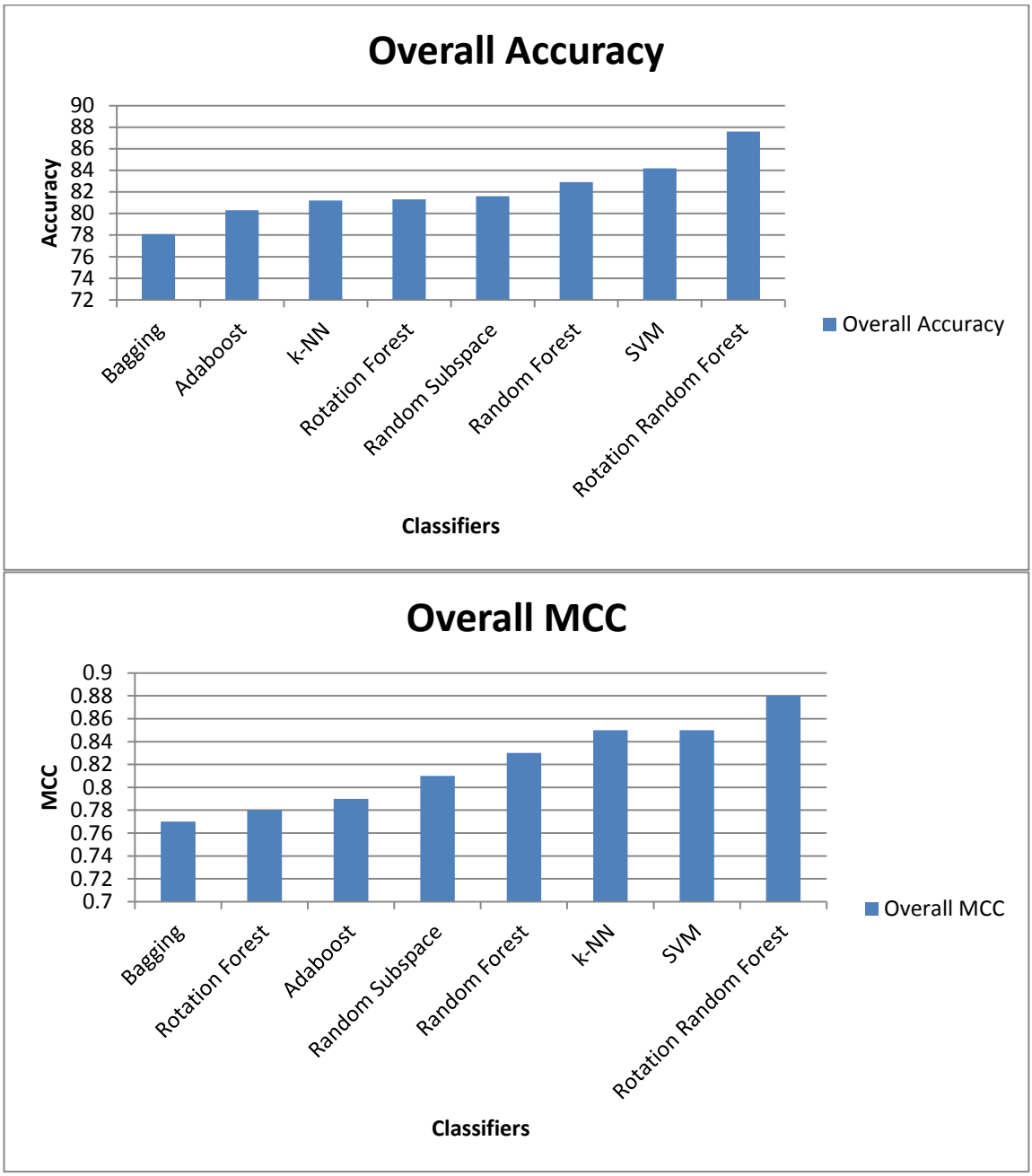

Figure 4.3: Accuracy and MCC obtained by different ensemble classifiers for the prediction of enzyme functional subclasses

The proposed method provides overall accuracy of 87.6%, MCC of 0.88 ROC area value of 0.98 and precision of 88.8% for the prediction of functional subclasses of enzymes with the complete datasets (See Table 4.5). The result obtained from proposed method is also compared with different classifiers and from Figure-4.3 it is observed that the proposed method may perform better in comparison with other classifiers such as Bagging, Adaboost, Rotation Forest,

k- nearest neighbor (k-NN), Random Forest and support vector machine (SVM) for the prediction of functional subclasses of enzymes.

4.4. Conclusion

In this chapter, an ensemble classifier Rotation random forest based of rotation forest and random forest has been proposed to predict the functional classes and sub-classes of enzymes by using sequence derived features. Here, seven feature vectors were used to represent the protein sample, including amino acid composition, dipeptide composition, correlation features, composition, transition, distribution and pseudo amino acid composition. In the 10-fold cross validation the proposed method has achieved an overall accuracy of 100%, 88.7%, 87.6% , MCC values of 1.00, 0.87, 0.88, ROC area values of 1.00, 0.98, 0.98 and precision of 100%, 89.5% and 88.8% to predict the enzymes and non-enzymes, functional classes and subclasses of enzymes respectively. The high accuracies, MCC, ROC area and precision values indicate that the proposed method is useful for the prediction of functional classes and subclasses of enzymes.