Chapter 3

A MULTI STAGE APPROACH FOR THE PREDICTION OF ION CHANNELS AND THEIR SUBFAMILIES

Ion channels are membrane proteins that are responsible for electrical signaling by gating the flow of ions across the cell membrane. These are the prominent component of nervous systems. Ion channels are classified by gating that is used for opening and closing the channels. The voltage gated ion channels are open and close based on the voltage gradient across the cell membrane, while ligand gated ion channels open and close based on the ligand binding of the ion channels. The voltage gated ion channels play an important role in generation and propagation of the nerve impulse and in cell homeostasis (Bezanilla, 2005). The dysfunction of ion channels play an important role in the development of various diseases such as hypertension, defective insulin secretion, cardiac arrhythmias, neurological diseases such as epilepsy and even developmental defects such as osteoporosis (Jentsch *et al.,* 2004). So it is necessary to know about the structure and function of the ion channels to develop a new drug for these diseases. Ion channels play an important target for antiepileptic drug design, antihypertensive and antipsychotics disorder such as schizophrenia (Abernethy *et al.,* 1999; Yogeeswari *et al.,* 2004).

In this chapter, a random forest based method is proposed to predict ion channels and their types by using sequence derived properties of protein sequences. Here, 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 the ion channels and their types. The minimum redundancy maximum relevance (MRMR) based feature selection is used to improve the predictive accuracy. The proposed method used four level strategies to predict ion channels and their types. First, it is determined that protein sequence is ion channel or non-ion channel. Second, if protein is classified as ion channels then the method classify the protein into two groups *viz*. voltage gated ion channels or ligand gated ion channels. Third, it is classified into the subfamilies of voltage gated ion channels and ligand gated ion channels. Fourth, it determines the subfamilies of calcium, potassium, sodium and chloride voltage gated ion channels. The two parameter of random forest the size of random subset of features *(mtry)* and the number of trees in the forest *(ntree)* are used to decrease the error rate. Therefore each of the four levels is developed using a random forest classifier with optimized value of *ntree* and *mtry*.

3.1. Background

Here, the sequence of ion channels with their properties and the proposed methods and model that are used to predict the ion channels and their types are presented.

3.1.1. Material and methods

To predict the ion channels and their types the sequence of the ion channels are extracted from the Uniport (http://www.uniprot.org), Ligand gated ion channels database (http://lenoverelab.org/LGICdb/LGICdb.php), National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/protein), voltage-gated potassium channels database (http://vkcdb.biology.ualberta.ca/) and KChannelDB (http://www.receptors.org/KCN). Here, all the 722 non-ion channels proteins are selected from Uniport database with the keyword NOT ion channels. To avoid the homology bias the CD-HIT server (Huang *et al.,* 2010) is used to remove the homologous sequence using 70% sequence identity as the cutoff, because when we decrease the cutoff as 0.5 and 0.4 respectively then very small sequences are left for the evaluation of classifier which is associated with lower performance values in comparison to 70% cutoff. The description of the dataset is shown in Table 3.1.

Table 3.1: Number of sequences belonging to each ion channels and their subfamilies

Ion/ non-ion channels	Families of IC	Subfamilies of IC	Sub-subfamilies of VGIC	No. of seq	No. of seq.	No. of seq.	No. of seq.
Ion Channel	Voltage gated ion channels	Calcium	P-Type	190	634	1827	2141
			R-Type	51			
			L-Type	280			
			N-Type	25			
			T-Type	88			
		Potassium	Kv1	61	646		
			Kv2	51			
			Kv3	52			
			Kv4	63			
			Kv5	22			
			Kv6	51			
			Kv7	59			
			Kv8.2	42			
			Kv9	52			
			Kv10	55			
			Kv11	59			
			Kv12	52			
			Kv13	27			
		Sodium	Alpha subunits	250	401		
			Beta subunits	151			
		Chloride	ClC1	48	146		
			ClC2	18			
			ClC3	43			
			ClC4	7			
			ClC5	15			
			ClCk	9			
			ClC6	6			
	Ligand gated ion channels	GABAA receptors	27		314	314	
		Glycine receptors	34				
		glutamate receptors	184				
		Nicotinic acetylcholine receptors	69				
Non ion channel				722	----	--	722

3.1.2. Features extraction of protein sequences

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 classification of ion channels and their types. The description of total 857 number of features used for the prediction of ion channels and their types is shown in Table 3.2.

Table 3.2: Description of sequence derived features for the prediction of ion channels and their types

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

3.2. Proposed method and model

For the prediction of ion channels and their subfamilies a random forest based method have been proposed by using sequence derived properties of a protein. Here, the minimum redundancy maximum relevance (MRMR) based feature selection is used to improve the predictive accuracy. The proposed method used four level strategies to predict ion channels and their types. First, it is determined that protein sequence is ion channel or non-ion channel. Second, if protein is classified as ion channels then the method classify the protein into two groups *viz.* voltage gated ion channels or ligand gated ion channels. Third, it is classified into the subfamilies of voltage gated ion channels and ligand gated ion channels. Fourth, it determines the subfamilies of calcium, potassium, sodium and chloride voltage gated ion channels.

3.2.1. Feature subset selection

In this chapter, the minimum redundancy maximum relevance (MRMR) (Peng *et al.,* 2005) a filter method is used to select a feature subset. It has been already used by (Li *et al.,* 2010) for the classification of G-protein coupled receptors. The brief description of MRMR feature selection is as follows:

The minimum redundancy maximum relevance feature selection method select a feature subset in which each subset of feature has the minimal redundancy with other features and maximal relevance with target class. In this method the subset of features is obtained by calculating the mutual information between the features themselves and between the features and the class variables. For binary classification the class variable c_k is 1 or 2. The mutual information *MI(x,y)* of two features *x* and *y* is calculated as

$$MI(x,y) = \sum_{i,j \in N} p(x_i\,, y_j) log \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)} \tag{3.1}$$

where $p(x_i)$ and $p(y_j)$ is the marginal probability density and $p(x_i, y_j)$ is the joint probability. Similarly, the mutual information MI(x, c) of between class variable *c* and feature *x* is also calculated as

$$MI(x,c) = \sum_{i,k \in N} p(x_i\,, c_k) log \frac{p(x_i, c_k)}{p(x_i)\, p(c_k)} \tag{3.2}$$

The minimum redundancy condition is to minimize the total redundancy of all features selected by Min (Redundancy) where

$$Redundancy = \frac{1}{|S|^2} \sum_{x,y \in S} MI(x,y) \tag{3.3}$$

where *S* is the feature subset and |*S*| is the number of feature in *S*.

The maximum relevance condition is to maximize the total relevance between all features in *S* and class variable. It is calculated as Max (Relevance) where

$$Relevance = \frac{1}{|S|}\sum_{x \in S} MI(x, c) \qquad (3.4)$$

Here, first feature having the highest *MI(x, c)* is selected according to equation (3.4) and the rest of the features are selected in incremental way where earlier selected features are remains in the features set. The optimal subset of features is selected by optimizing the equations (3.3) and (3.4) simultaneously through mutual information difference criterion.

$$Max\ (Relevance - Redundancy) \qquad (3.5)$$

3.2.2. Classification of ion channels and their types

For the classification of ion channels and their types, here a random forest based classifier available in Weka 3.7.11 software tool (Hall *et al.,* 2009) has been used. Random forest classifier (Breiman, 2001) used an ensemble of random trees. Each of the random trees is generated by using a bootstrap sample data. At each node of the tree a subset of features with highest information gain is selected from a random subset of entire features. Thus random forest used bagging as well as feature selection to generate the trees. Once a forest is generated every tree participates in classification by voting to a class. The final classification is based on the majority voting of a particular class.

The error rate of random forest depends on the strength of each tree in the forest and the correlation between any two trees. Therefore increasing the strength of each tree and reducing the correlation between the trees may necessary to decrease the error rate of the forest. The two parameter of random forest the size of random subset of features (*mtry*) and the number of trees in the forest *(ntree*) are used to decrease the error rate. Increasing the value of *ntree* reduces the out-of-bag (OOB) error rate of random forest as well as correlation between trees but possibilities of over-fitting. The *mtry* value should be much smaller than total number of features. So it is necessary to obtain an optimal value of *mtry* and *ntree* to obtain the lowest out-of-bag error and higher accuracy. In this chapter, we have optimized the values of *ntree* and *mtry* at every level to improve the prediction accuracy.

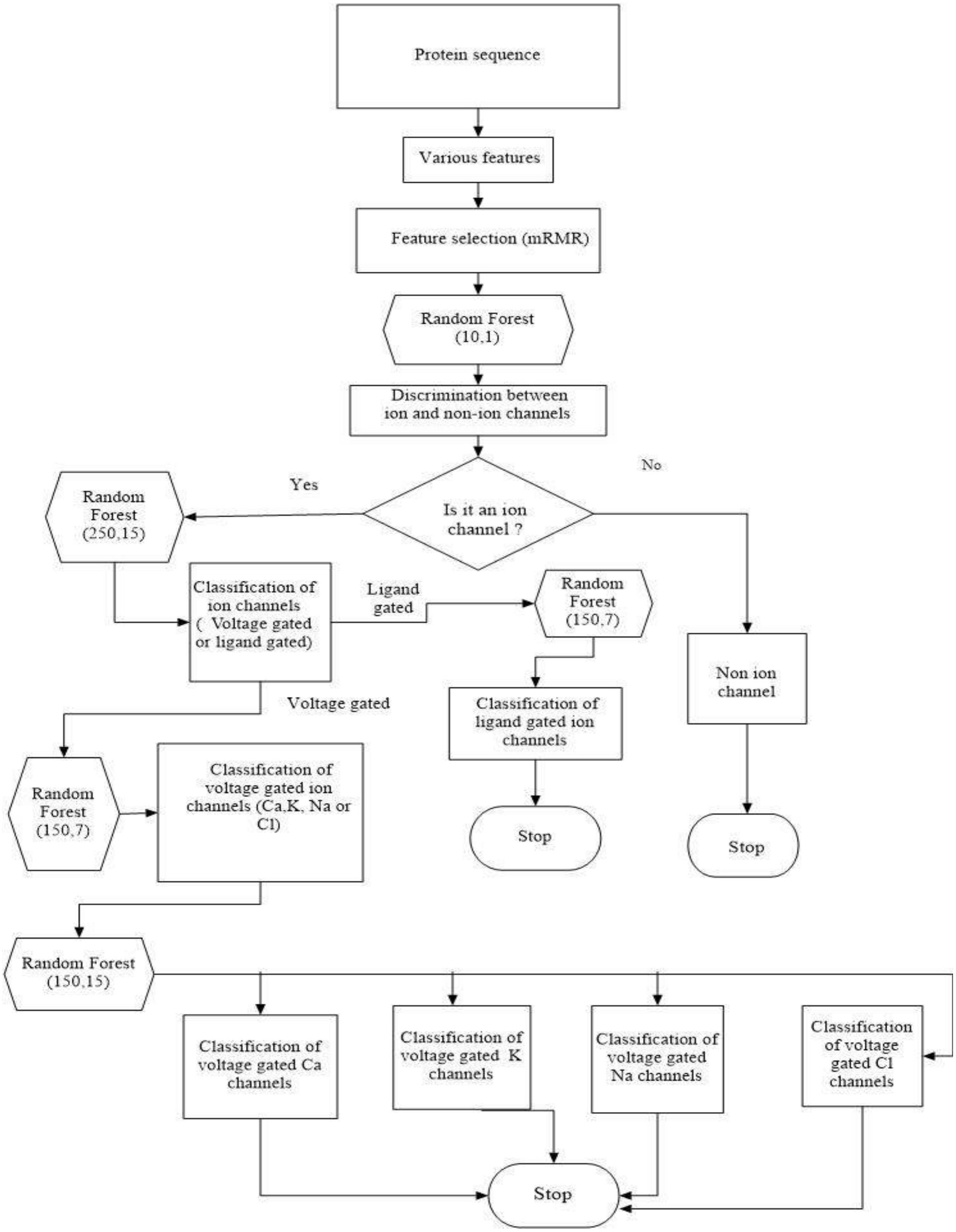

Protein sequence
Various features
Feature selection (mRMR)
Random Forest (10,1)
Discrimination between ion and non-ion channels
Yes
No
Is it an ion channel ?
Random Forest (250,15)
Classification of ion channels (Voltage gated or ligand gated)
Ligand gated
Random Forest (150,7)
Non ion channel
Voltage gated
Classification of ligand gated ion channels
Random Forest (150,7)
Classification of voltage gated ion channels (Ca,K, Na or Cl)
Stop
Stop
Random Forest (150,15)
Classification of voltage gated Ca channels
Classification of voltage gated K channels
Classification of voltage gated Na channels
Classification of voltage gated Cl channels
Stop

Figure 3.1: A flowchart for the proposed model for the prediction of ion channels and their types

In this chapter, the proposed method used four level strategies to predict ion channels and their types. The complete procedure of the proposed method with optimized parameter for the prediction of ion channels and their types is illustrated in Figure 3.1 and the steps are as follows:

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

Step 2: Select optimal number of features with minimum redundancy and maximum relevance (MRMR) algorithms.

Step 3: Apply random forest classifier with optimized value of *ntree* and *mtry* for each of the four levels for the prediction of ion channels families and their subfamilies are as follows:

Firstly, it is discriminated that protein sequence is ion channel or non-ion channel. Secondly, if protein is classified as ion channels then the method classify the protein into two group viz. voltage gated ion channels or ligand gated ion channels. Thirdly, it classifies the subfamilies of voltage gated ion channels and ligand gated ion channels. Finally, it also determines the subfamilies of calcium, potassium, sodium and chloride voltage gated ion channels.

3.3. Results and performance analysis

Here, the performance measures that are used to measure the performance of the proposed method and the analysis of the results obtained by the proposed method for the prediction of ion channels and their subfamilies are presented.

3.3.1. Performance measures

In this chapter, 10-fold cross validation is used to measure the performance of 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*), receiver operating characteristics (ROC) and Matthew's correlation coefficient (*MCC*) are used to measure the performance of the proposed method for the prediction of ion channels and their types.

3.3.2. Results and analysis

In this chapter, a random forest classifier is proposed to be used for the prediction of various ion channels and their subfamilies. The parameters *ntree* and *mtry* to be used by the random forest classifier are chosen experimentally in such a manner that minimizes the out-of-bag (OOB) error. The OOB is calculated with different values of *ntree* and *mtry* and it is observed that the random forest classifier is associated with minimum OOB error for *ntree* values of 200 and *mtry* values of 15 for the discrimination between voltage gated and ligand gated ion channels. For the classification of subfamilies of voltage gated ion channels and ligand gated ion channels the *ntree* and *mtry* values are 150 and 07 respectively. For the classification of subfamilies of voltage gated calcium, potassium, sodium and chloride ion channels the *ntree* and *mtry* values are 150 and 15 respectively (See Figure 3.2- 3.5). 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 ion channels and their types are presented and discussed. The performance analysis of the proposed model is proposed for different chosen features of the original datasets for each case as well as for the reduced datasets for each case after selecting minimum redundant and maximal relevant features after applies MRMR feature selection method.

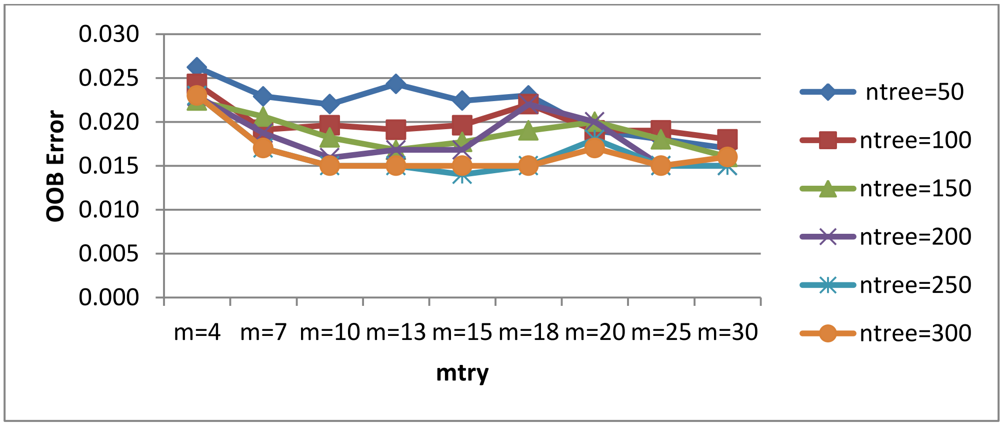


Figure 3.2: OOB Error for the different values of mtry and ntree for the discrimination between voltage gated and ligand gated ion channels

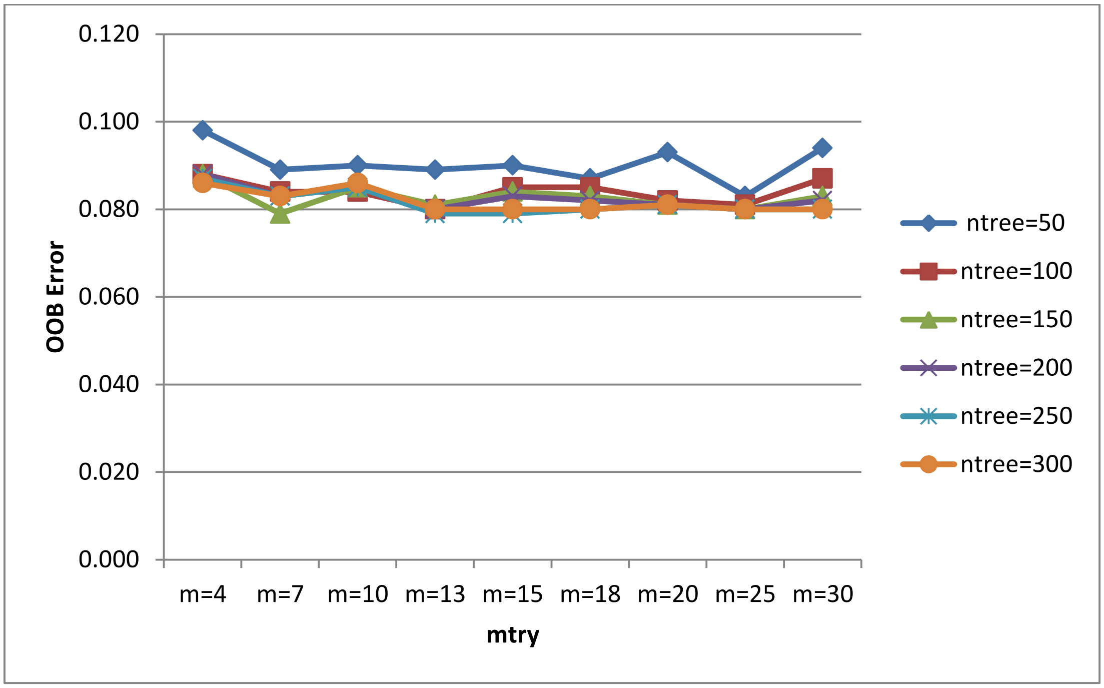


Figure 3.3: OOB Error for the different values of mtry and ntree for the classification of subfamilies of voltage gated ion channels

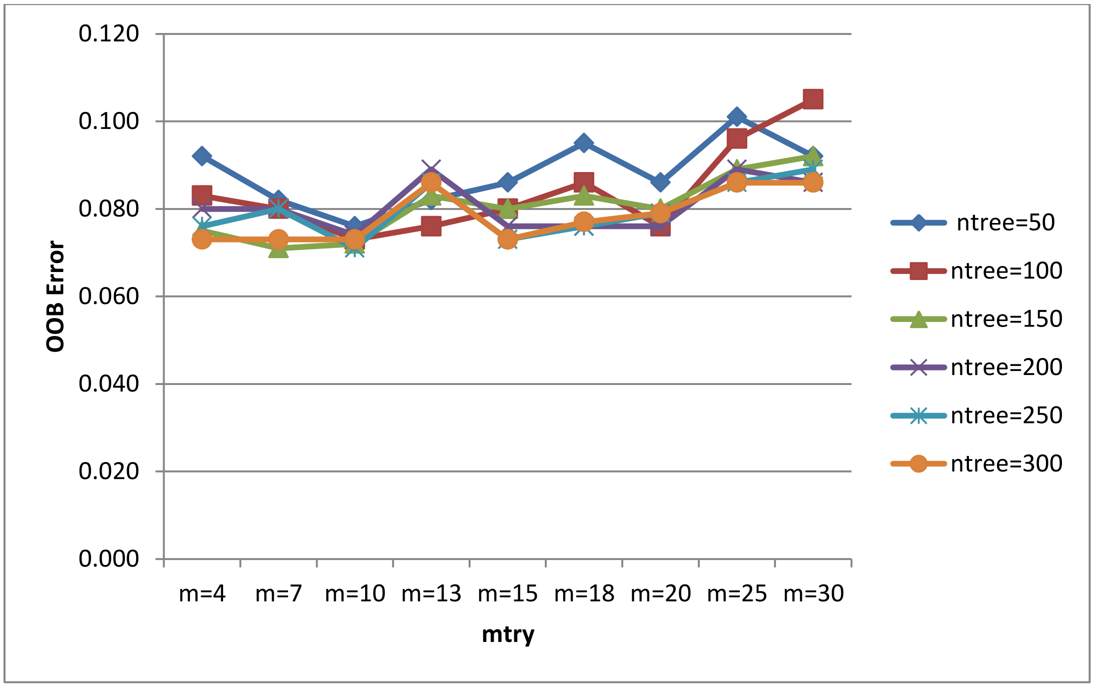


Figure 3.4: OOB Error for the different values of mtry and ntree for the classification of subfamilies of ligand gated ion channels

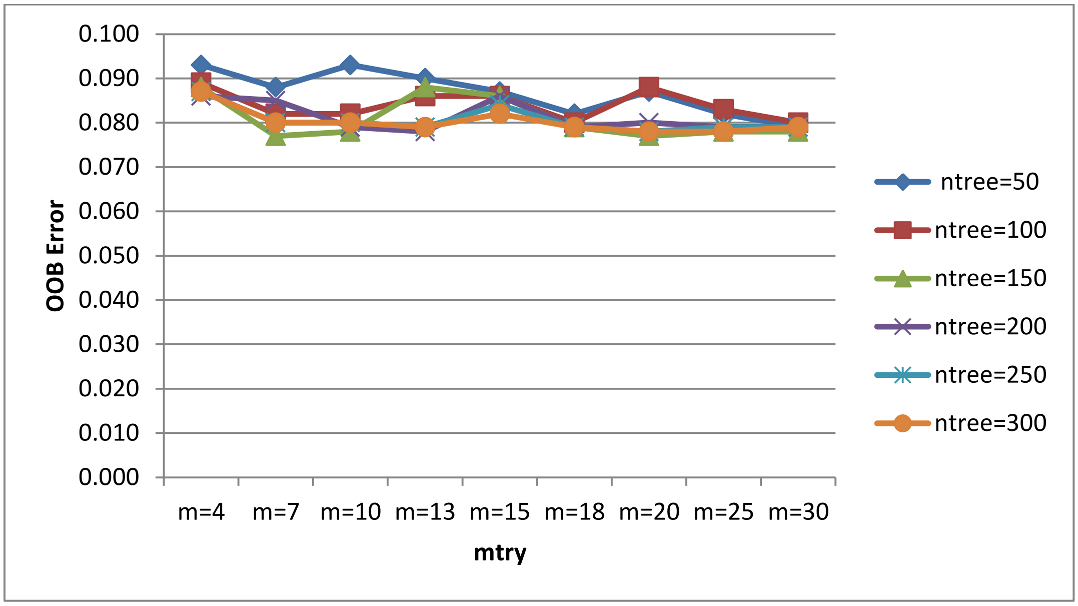


Figure 3.5: OOB Error for the different values of mtry and ntree for the classification of subfamilies of voltage gated calcium, potassium, sodium and chloride ion channels

3.3.2.1. Prediction of ion channels and non-ion channels

To predict the ion channels and non-ion channels, a 10-fold cross validation is used on a dataset containing 857 number of sequence derived features of 2141 number of ion channels and 722 number of non-ion channels protein sequences and hence a total of 2863 number of sequences . The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition; amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature respectively (See Table 3.3). composition, transition, distribution and pseudo amino acid composition feature vector are not affecting the performance discrimination between ion channels and non-ion channels but have an importance for discrimination between voltage gated ion channels and ligand gated ion channels and subfamily prediction of voltage gated and ligand gated ion channels (See Table 3.3).

Further, the minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier to discriminate ion channels and non-ion channels. The accuracy and MCC are evaluated for different number of features and it is observed that the 100% accuracy is obtained with best 50 features for discrimination between ion channels and non-ion channels (See Table 3.3). The best 50 features selected by minimum redundancy maximum relevance (MRMR) algorithm are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier to discriminate ion channels and non-ion channels. The accuracy and MCC are evaluated for different classifiers with best 50 features for prediction of ion channels and non-ion channels (See Table 3.3). The complete analysis of results is shown in Table 3.3.

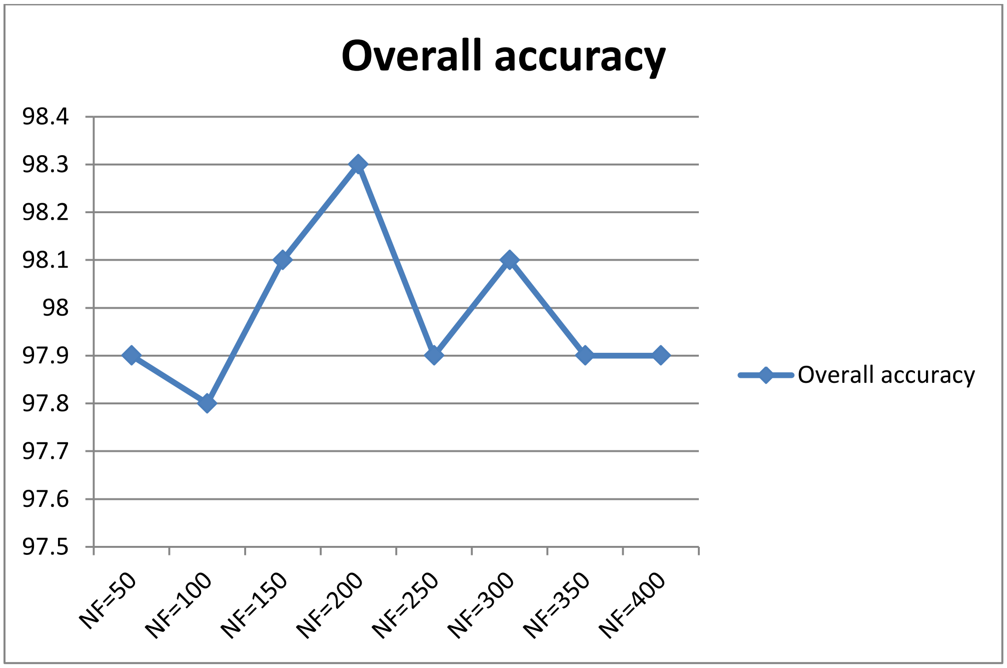
Table 3.3: The results of the prediction of ion channels and non-ion channels

Method		Non-ion	Ion channels	Overall
Random forest with AAC	ACC	98.3	99.8	99.4
	MCC	0.98	0.98	0.98
Random forest with AAC+DC	ACC	99.9	99.6	99.7
	MCC	0.99	0.99	0.99
RF with AAC+DC+CF	ACC	100	100	100
	MCC	1	1	1
RF with AAC+DC+CF+CTD	ACC	100	100	100
	MCC	1	1	1
RF with AAC+DC+CF+CTD+PAAC	ACC	100	100	100
	MCC	1	1	1
RF (with best 50 features)	**ACC**	**100**	**100**	**100**
	MCC	**1**	**1**	**1**
	ROC area	**1**	**1**	**1**
SVM (with best 50 features)	ACC	97	100	99.2
	MCC	0.98	0.98	0.98
	ROC area	0.98	0.98	0.98
kNN (with best 50 features)	ACC	100	100	100
	MCC	1	1	1
	ROC area	1	1	1
Naïve Bayes (with best 50 features)	ACC	100	100	100
	MCC	1	1	1
	ROC area	1	1	1

From the analysis of Table 3.3 it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides accuracy of 100%, MCC of 1.00 and ROC area of 1.00 for the prediction of ion channels and non-ion channels with the best 50 features selected by MRMR algorithm. Table 3.3 also shows that the proposed method perform better in comparison with SVM, k-NN, and Naïve Bayes classifier.

3.3.2.2. Prediction of voltage and ligand gated ion channels

To predict voltage gated ion channels (VGIC) and ligand gated ion channels (LGIC) a random forest with 10-fold cross validation is used on a dataset containing 857 number of sequence derived features of 1827 number of voltage gated ion channels and 314 number of ligand gated ion channels protein sequences and hence a total of 2141 number of sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature and so on (See Table 3.4). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier to classify voltage gated ion channel and ligand gated ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.6 it is observed that the overall 98.3% accuracy and MCC value of 0.93 is obtained with best 200 features for discrimination between voltage gated ion channel and ligand gated ion channels.



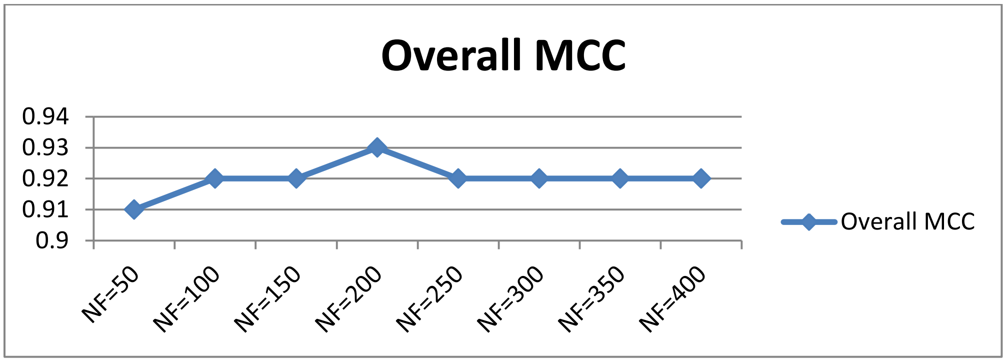


Figure 3.6: Accuracy and MCC for prediction of voltage and ligand gated ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 200 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier to classify voltage gated ion channel and ligand gated ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 200 features to classify voltage gated ion channel and ligand gated ion channels (See Table 3.4). The complete analysis of results is shown in Table 3.4.

Table 3.4: The results of the prediction of voltage and ligand gated ion channels

Method		**VGIC**	**LGIC**	**Overall**
Random forest with AAC	ACC	99.5	83.1	97.1
	MCC	0.88	0.88	0.88
Random forest with AAC+DC	ACC	100	79.9	97.1
	MCC	0.88	0.88	0.88
RF with AAC+DC+CF	ACC	100	77.4	96.7
	MCC	0.86	0.86	0.86
RF with AAC+DC+CF+CTD	ACC	100	80.9	97.2
	MCC	0.89	0.89	0.89
RF with AAC+DC+CF+CTD+PAAC	ACC	100	82.8	97.5
	MCC	0.9	0.9	0.9
RF (with best 200 features)	**ACC**	**99.9**	**89.2**	**98.3**
	MCC	**0.93**	**0.93**	**0.93**
	ROC area	**0.99**	**0.99**	**0.99**
SVM (with best 200 features)	ACC	99.9	75.5	96.3
	MCC	0.85	0.85	0.85
	ROC area	0.93	0.93	0.93
kNN (with best 200 features)	ACC	97.9	93.9	97.3
	MCC	0.89	0.89	0.89
	ROC area	0.95	0.95	0.95
Naïve Bayes (with best 200 features)	ACC	44.8	93	51.8
	MCC	0.27	0.27	0.27
	ROC area	0.78	0.75	0.77

From the analysis of Table 3.4 it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides accuracy of 100% and MCC of 0.90 for the prediction of voltage gated ion channels and accuracy of 82.8% and MCC of 0.90 for the prediction of ligand gated ion channels with the complete datasets. The overall accuracy is increases from 97.5 % to 98.3% and overall MCC increases from 0.90 to 0.93 with best 200 features selected by MRMR algorithm (See Table 3.4). From Table 3.4 it is also observed that the proposed method may perform better in comparison with SVM, k-NN, and Naïve Bayes classifier.

3.3.2.3. Prediction of subfamilies of voltage gated ion channels

For the prediction of subfamilies of voltage gated ion channels, a random forest with 10-fold cross validation is used on a dataset containing 857 sequence derived features of 634 number of calcium, 646 number of potassium, 401 number of sodium and 146 number of chloride voltage gated ion channels protein sequences and hence a total of 1827 number of voltage gated ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature and so on (See Table 3.5). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier for prediction of subfamilies of voltage gated ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.7 it is observed that the overall accuracy of 92.1% and MCC value of 0.89 is obtained with best 150 features for the prediction of subfamilies of voltage gated ion channels.

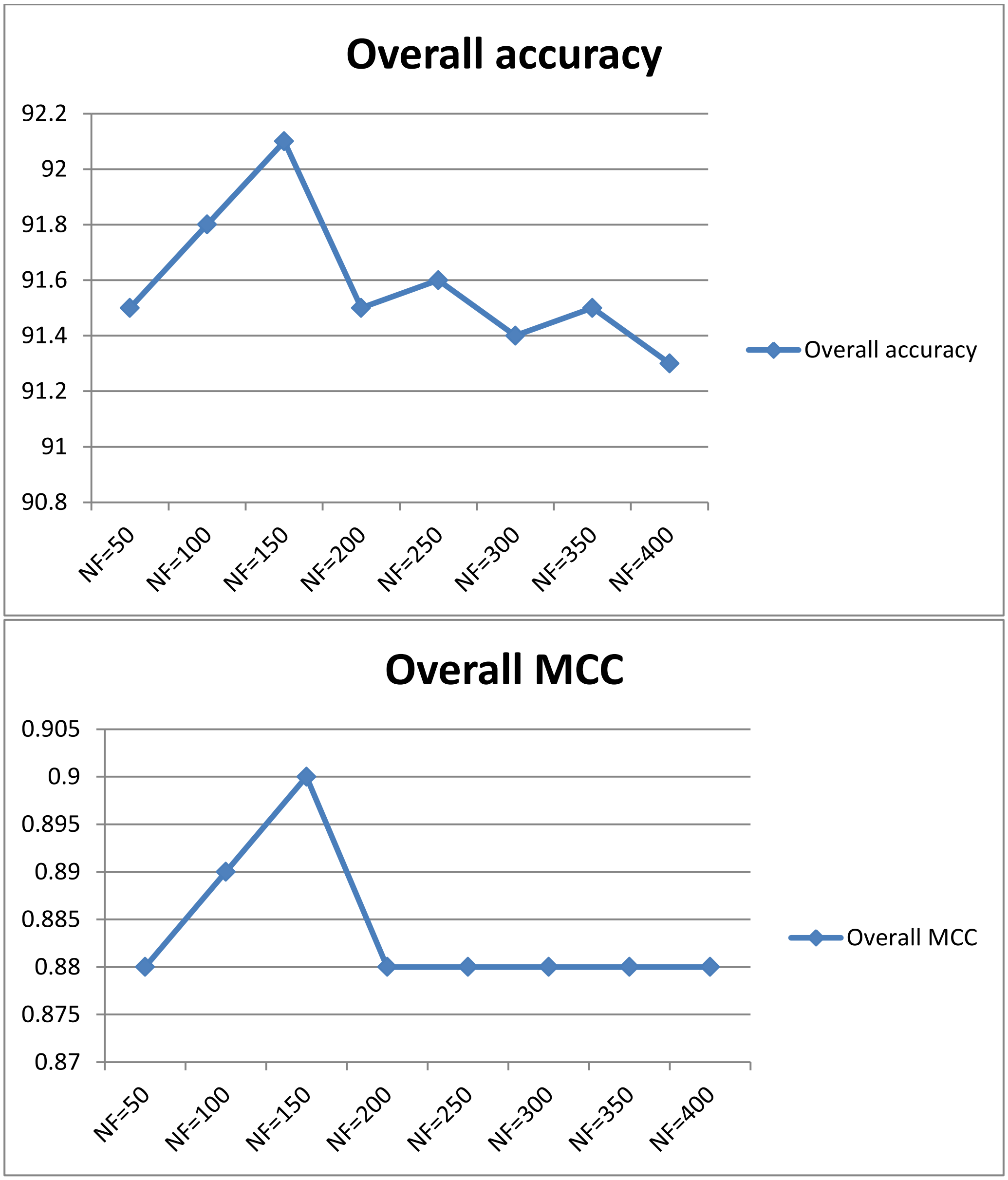


Figure 3.7 Accuracy and MCC for the prediction of subfamilies of voltage gated ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 150 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of subfamilies of voltage gated ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 150 features for the classification of subfamilies of voltage gated ion channel (See Table 3.5). The complete analysis of results is shown in Table 3.5.

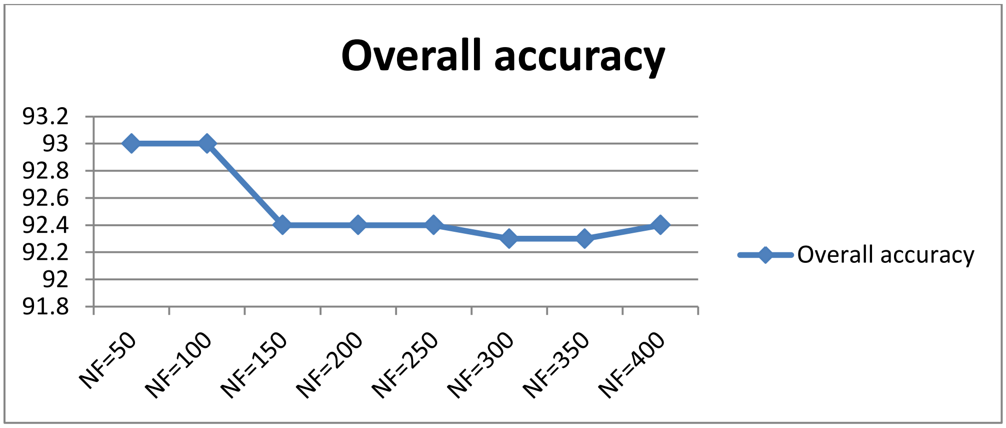
Table 3.5: The results of the prediction of subfamilies of voltage gated ion channels

Method		Sodium	Calcium	Potassium	Chloride	Overall
Random forest with AAC	ACC	77.1	90.5	99.4	81.5	90
	MCC	0.77	0.82	0.95	0.86	0.86
Random forest with AAC+DC	ACC	74.3	92.4	100	70.5	89.4
	MCC	0.76	0.82	0.94	0.83	0.86
RF with AAC+DC+CF	ACC	73.1	92.7	100	70.5	89.2
	MCC	0.78	0.83	0.92	0.81	0.85
RF with AAC+DC+CF+CTD	ACC	75.8	92.7	100	78.8	90.5
	MCC	0.79	0.83	0.96	0.86	0.87
RF with AAC+DC+CF+CTD+PAAC	ACC	77.3	93.8	100	79.5	91.2
	MCC	0.81	0.85	0.96	0.87	0.88
RF (with best 150 features)	**ACC**	**80.3**	**93.8**	**100**	**81.5**	**92.1**
	MCC	**0.82**	**0.87**	**0.98**	**0.87**	**0.90**
	ROC area	**0.97**	**0.98**	**1.00**	**1.00**	**0.99**
SVM (with best 150 features)	ACC	64.1	81.9	100	49.3	81.8
	MCC	0.73	0.78	0.76	0.68	0.75
	ROC area	0.83	0.87	0.97	0.87	0.89
kNN (with best 150 features)	ACC	82.3	88.6	93.7	87.7	88.9
	MCC	0.75	0.81	0.94	0.87	0.85
	ROC area	0.87	0.89	0.90	0.93	0.89
Naïve Bayes (with best 150 features)	ACC	86.3	80.0	96.6	24.0	57.8
	MCC	0.36	0.12	0.89	0.27	0.46
	ROC area	0.85	0.86	0.98	0.85	0.90

From the analysis of Table 3.5, it is observed that the performance of random forest classifier is continuously improved by using the mixture of the feature vectors. The proposed method provides accuracy of 77.3%, 93.8%, 100% and 79.5% and MCC of 0.81, 0.85, 0.96 and 0.87 for the prediction of sodium, calcium, potassium and chloride voltage gated ion channels respectively with the complete datasets. The overall accuracy is increases 91.2 % to 92.1% and overall MCC increases from 0.88 to 0.90 with best 150 features selected by MRMR algorithms (See Table 3.5). It is also observed that the Radom Forest provide accuracy of 92.1% , MCC values of 0.90 and ROC area of 0.99 that is better in comparison with SVM, k-NN, and Naïve Bayes classifier (See Table 3.5).

3.3.2.4. Prediction of subfamilies of ligand gated ion channels

For the prediction of subfamilies of ligand gated ion channels, a random forest with 10-fold cross validation is used on a dataset containing 857 sequence derived features of 27 number of GABA receptors, 34 number of Glycine receptors, 184 number of Inotropic glutamate receptors and 69 number of Nicotinic acetylcholine receptors (NAR) protein sequences and hence a total of 314 number of ligand gated ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature and so on (See Table 3.6). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classification of subfamilies of ligand gated ion channels. The accuracy and MCC are evaluated for different number of features and from Figure-3.8 it is observed that the highest overall accuracy of 93.0% and MCC of 0.88 is obtained with best 50 features for the prediction of subfamilies of ligand gated ion channels.



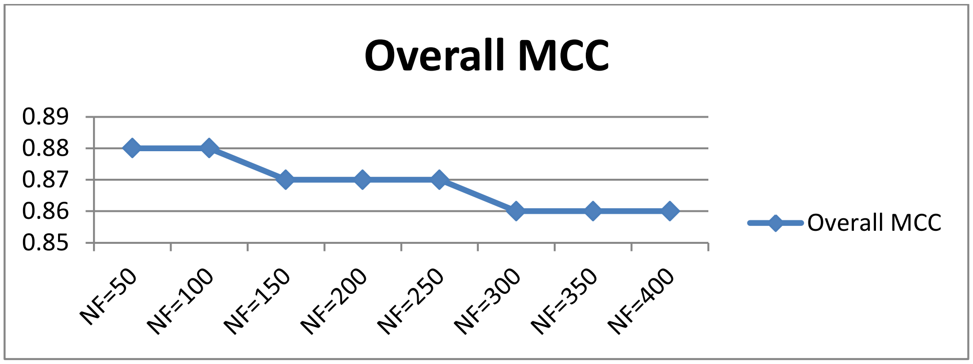


Figure 3.8: Accuracy and MCC for prediction of subfamilies of ligand gated ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 50 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of types of voltage gated ion channels. The accuracy, MCC and ROC area are evaluated for different-different classifier with best 50 features the classification of subfamilies of ligand gated ion channels (See Table 3.6). The complete analysis of results is shown in Table 3.5.

Table 3.6: The results of the prediction of subfamilies of ligand gated ion channels

Method		**GABAA**	**glutamate**	**Glycine**	**NAR**	**Overall**
Random forest with AAC	ACC	63	82.4	94.6	94.2	90.4
	MCC	0.64	0.85	0.86	0.89	0.84
Random forest with AAC+DC	ACC	55.6	82.4	96.7	94.2	91.1
	MCC	0.64	0.86	0.84	0.94	0.84
RF with AAC+DC+CF	ACC	63	82.4	96.7	95.7	92
	MCC	0.69	0.86	0.86	0.94	0.86
RF with AAC+DC+CF+CTD	ACC	66.7	85.3	96.2	95.7	92.4
	MCC	0.7	0.88	0.86	0.94	0.87
RF with AAC+DC+CF+CTD+PAA C	ACC	63	82.4	95.7	95.7	91.4
	MCC	0.67	0.86	0.85	0.93	0.85
RF (with best 50 features)	**ACC**	**66.7**	**88.2**	**97.3**	**94.2**	**93**
	MCC	**0.76**	**0.9**	**0.88**	**0.92**	**0.88**
	ROC	**0.96**	**0.99**	**0.99**	**0.99**	**0.99**
SVM (with best 50 features)	ACC	44.4	73.5	97.8	92.8	89.5
	MCC	0.57	0.83	0.8	0.93	0.81
	ROC	0.83	0.87	0.90	0.91	0.89
kNN (with best 50 features)	ACC	70.4	94.1	94.6	95.7	92.7
	MCC	0.66	0.93	0.89	0.93	0.88
	ROC	0.83	0.97	0.94	0.96	0.94
Naïve Bayes (with best 50 features)	ACC	81.5	17.6	69	21.7	54.1
	MCC	0.22	0.25	0.58	0.35	0.46
	ROC	0.86	0.78	0.92	0.93	0.90

From the analysis of Table 3.6 it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides accuracy of 63.0%, 82.4%, 95.7% and 95.7% and MCC of 0.67, 0.86, 0.85 and 0.93 for the prediction of GABAA, glutamate, Glycine and NAR ligand gated ion channels respectively with the complete datasets. The overall accuracy is increases 91.4 % to 93.0% and Overall MCC is increases from 0.85 to 0.88 with best 50 features selected by MRMR algorithms (See Table 3.6). It is also observed that the proposed method provide higher ACC, MCC and ROC area in comparison with the SVM, k-NN, and Naïve Bayes classifier (See Table 3.6).

3.3.2.5. Prediction of subfamilies of voltage gated calcium ion channels

For the prediction of subfamilies of voltage gated calcium ion channels a random forest with 10 fold cross validation is used on a dataset containing 857 sequence derived features of 190 number of P-Type, 51 number of R-Type, 280 number of L-Type, 25 number of N-Type and 88 number of T-Type protein sequences and hence a total of 634 number of voltage gated calcium ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature (See Table 3.7). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier for prediction of subfamilies of calcium ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.9 it is observed that the overall accuracy of 92.3% and MCC of 0.89 is obtained with best 400 features for the prediction of subfamilies of voltage gated calcium ion channels.

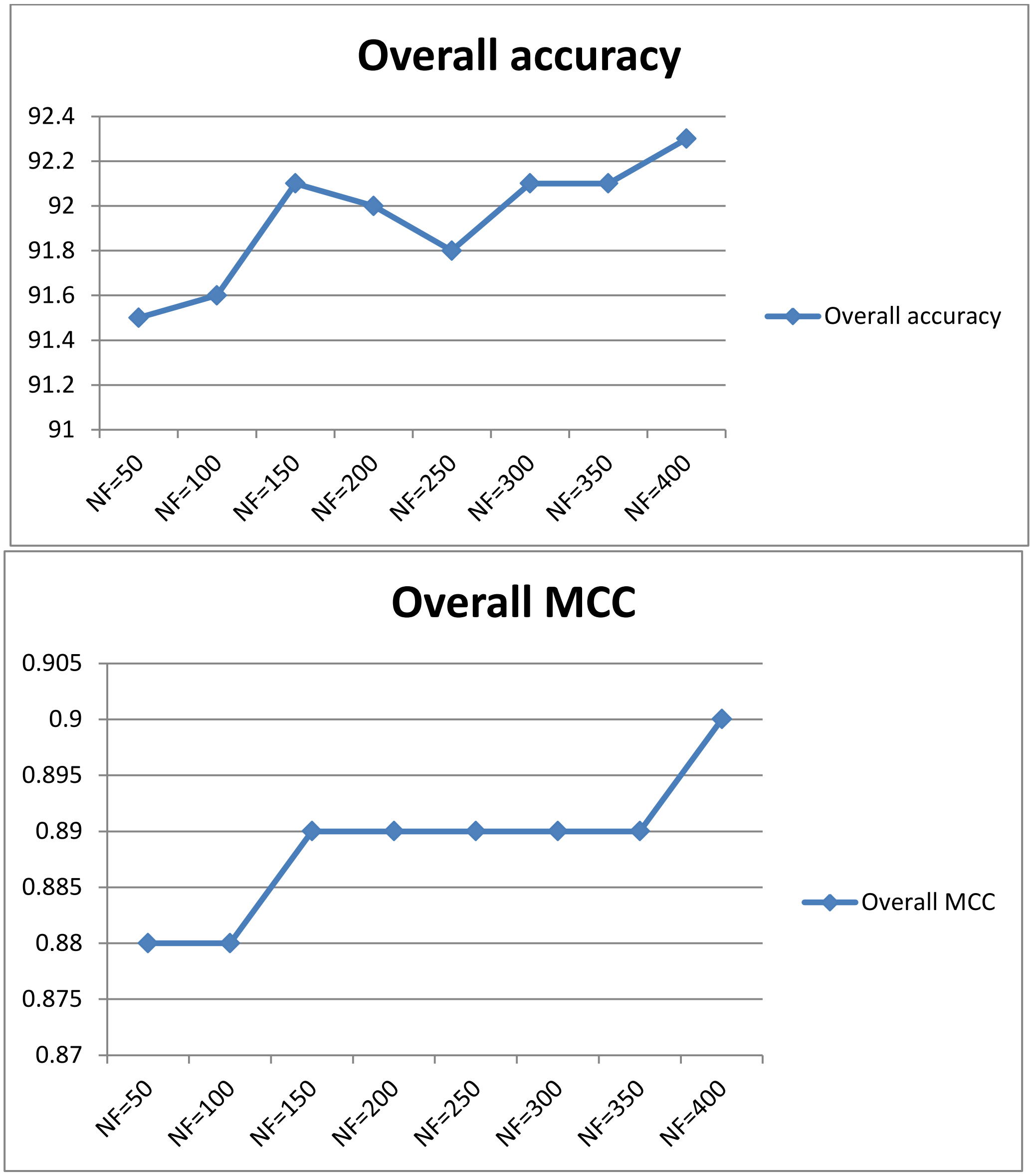


Figure 3.9: Accuracy and MCC for the prediction of subfamilies of voltage gated calcium ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 400 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of types of calcium ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 400 features for the prediction of types of calcium ion channels (See Table 3.7). The complete analysis of results is shown in Table 3.7.

Table 3.7: The results of the prediction of subfamilies of voltage gated calcium ion channels

Method		P-Type	R-Type	L-Type	N-Type	T-Type	Overall
Random forest with AAC	ACC	96.3	69.8	97.7	68	84.1	92.1
	MCC	0.96	0.76	0.87	0.79	0.87	0.88
Random forest with AAC+DC	ACC	99.5	73.6	96.7	72	83	92.9
	MCC	0.96	0.79	0.86	0.84	0.89	0.89
RF with AAC+DC+CF	ACC	97.9	71.7	97.4	64	83	92.3
	MCC	0.96	0.8	0.85	0.79	0.89	0.88
RF with AAC+DC+CF+ CTD	ACC	96.3	71.7	97.4	68	81.8	91.8
	MCC	0.94	0.8	0.85	0.81	0.89	0.87
RF with AAC+DC+CF+ CTD+PAAC	ACC	97.4	71.7	96.7	68	83	92
	MCC	0.94	0.8	0.85	0.79	0.89	0.88
RF (with best 400 features)	**ACC**	**97.9**	**74.5**	**97.1**	**68**	**85.2**	**92.7**
	MCC	**0.95**	**0.84**	**0.88**	**0.92**	**0.9**	**0.9**
SVM (with best 400 features)	ACC	87.9	60.8	99.3	60	69.3	87.1
	MCC	0.91	0.75	0.77	0.77	0.81	0.82
kNN (with best 400 features)	ACC	97.9	82.4	93.6	80	87.5	92.6
	MCC	0.97	0.75	0.89	0.81	0.88	0.9
Naïve Bayes (with best 400 features)	ACC	92.1	29.4	72.1	16	76.1	73
	MCC	0.84	0.22	0.69	0.13	0.55	0.66

From the analysis of Table 3.7 it is observed that the performance of random forest classifier is affected by using the mixture of the feature vectors. The proposed method provides accuracy of 97.4%, 71.7%, 96.7% , 68.0% and 83.0% and MCC values of 0.94, 0.80, 0.85, 0.79 and 0.89 for the prediction of P-Type, R-Type, L-Type, N-Type and T-Type voltage gated calcium ion channels respectively with the complete datasets. The overall accuracy is increases 92.0 % to 92.7% and Overall MCC is increases from 0.88 to 0.90 with best 400 features selected by MRMR algorithms (See Table 3.7). It is also observed that the proposed method provide overall accuracy of 92.7%, MCC values of 0.90 and ROC area of 0.99 with best 400 features which is better in comparison with SVM, k-NN, and Naïve Bayes classifier (See Table 3.7).

3.3.2.6. Prediction of subfamilies of voltage gated potassium ion channels

For the classification of subfamilies of voltage gated potassium ion channels a random forest with 10 fold cross validation is used on a dataset containing 857 sequence derived features of 61 number of Kv1, 51 number of Kv2, 52 number of Kv3, 63 number of Kv4, 22 number of Kv5, 51 number of Kv6, 59 number of Kv7, 42 number of Kv8.2, 52 number of Kv9, 55 number of Kv10, 59 number of Kv11, 52 number of Kv12, and 27 number of Kv13 protein sequences and hence, a total of 646 number of voltage gated potassium ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is improved by using the amino acid composition, amino acid with dipeptide composition (See Table 3.8). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier for prediction of types of potassium ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.10 it is observed that the overall accuracy of 78.6% and MCC value of 0.79 is obtained with best 400 features for the prediction of subfamilies of voltage gated potassium ion channels.

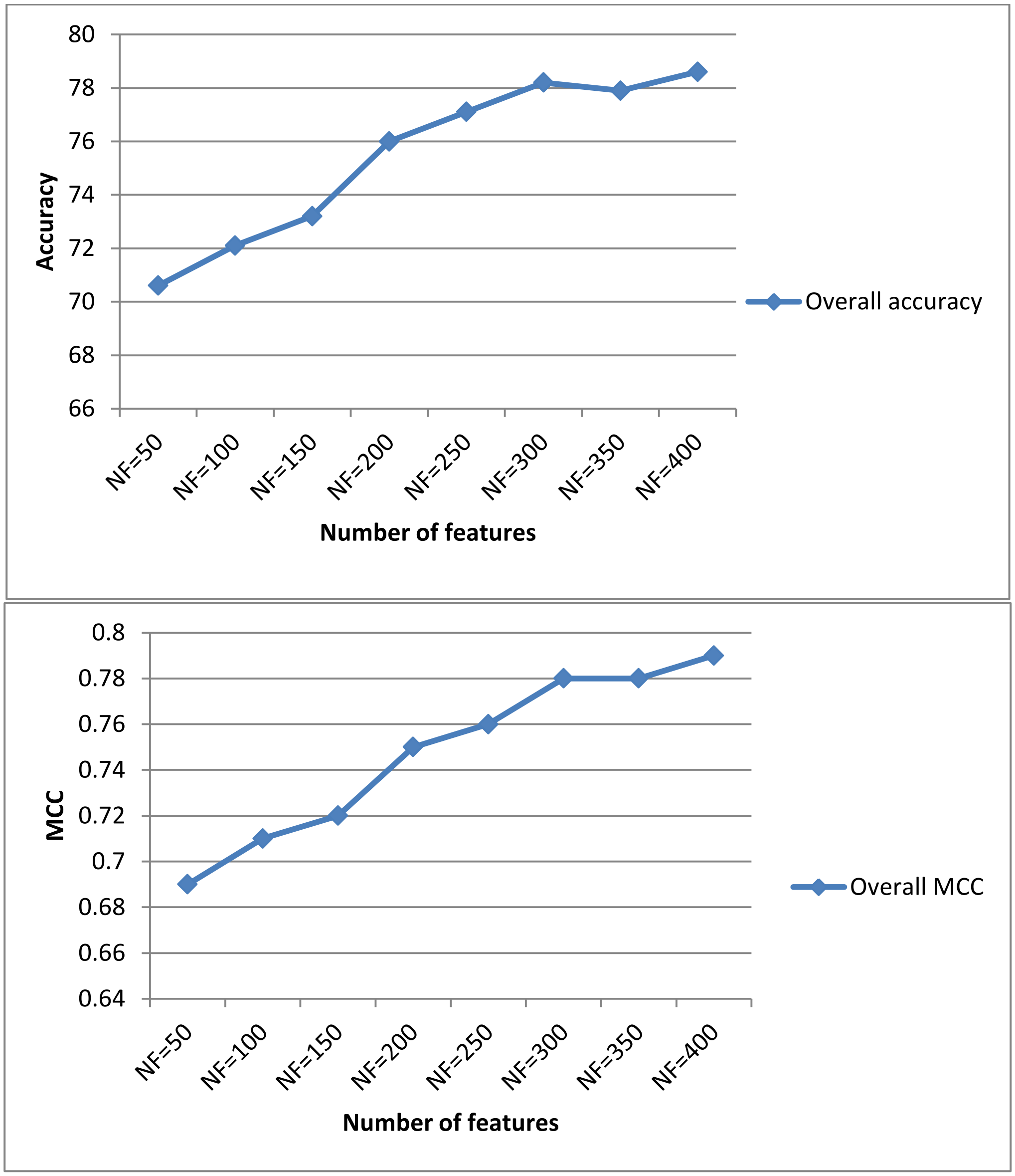


Figure 3.10 Overall accuracy and MCC for prediction of subfamilies of potassium ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 400 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of subfamilies of voltage gated potassium ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 400 features for the prediction of subfamilies of voltage gated potassium ion channels (See Table 3.8). The complete analysis of results is shown in Table 3.8.

Table 3.8: The results of the prediction of subfamilies of voltage gated potassium ion channels

Method		Kv1	Kv2	Kv3	Kv4	Kv5	Kv6	Kv7	Kv-8.2	Kv9	Kv-10	Kv-11	Kv-12	Kv-13	Ove-rall
Random forest with AAC	ACC	59	74.5	67.3	66.7	86.4	70.6	78	54.5	78.8	63.6	62.7	73.1	66.7	69
	MCC	0.48	0.73	0.6	0.59	0.93	0.63	0.69	0.65	0.78	0.67	0.69	0.74	0.7	0.67
Random forest with AAC+ DC	ACC	83.6	84.3	78.8	92.1	95.5	86.3	84.7	50	82.7	74.5	81.4	82.7	74.1	81.9
	MCC	0.65	0.89	0.81	0.88	0.98	0.83	0.64	0.7	0.88	0.85	0.85	0.86	0.86	0.82
RF with AAC+DC+CF	ACC	73.8	76.5	76.9	92.1	95.5	86.3	83.1	50	84.6	80	79.7	78.8	70.4	79.7
	MCC	0.6	0.79	0.76	0.85	0.98	0.78	0.64	0.67	0.9	0.85	0.85	0.86	0.83	0.79
RF with AAC+DC+CF+ CTD	ACC	72.1	76.5	78.8	87.3	95.5	80.4	79.7	50	82.7	74.5	81.4	78.8	66.7	77.9
	MCC	0.59	0.79	0.76	0.82	0.98	0.73	0.59	0.7	0.88	0.81	0.84	0.86	0.81	0.77
RF with AAC+DC+CF+ CTD+ PAAC	ACC	80.3	80.4	78.8	88.9	95.5	80.4	81.4	50	80.8	76.4	81.4	76.9	70.4	79.3
	MCC	0.62	0.84	0.76	0.85	0.98	0.77	0.64	0.7	0.86	0.83	0.81	0.86	0.81	0.79
RF with best 400 features	ACC	**83.6**	**78.4**	**78.8**	**88.9**	**95.5**	**80.4**	**84.7**	**54.5**	**84.6**	**78.2**	**78**	**76.9**	**70.4**	**80.2**
	MCC	**0.66**	**0.78**	**0.82**	**0.88**	**0.98**	**0.77**	**0.65**	**0.73**	**0.85**	**0.85**	**0.82**	**0.83**	**0.83**	**0.8**
SVM with best 400 features	ACC	100	27.5	17.3	31.7	9.1	13.7	16.9	0	44.2	41.8	35.6	19.2	0	31.4
	MCC	0.17	0.51	0.4	0.54	0.3	0.36	0.4	0	0.65	0.63	0.58	0.42	0	0.42
kNN with best 400 features	ACC	70.5	86.3	82.7	88.9	95.5	80.4	72.9	72.7	92.3	78.2	78	80.8	77.8	81
	MCC	0.66	0.84	0.84	0.89	0.98	0.76	0.74	0.78	0.87	0.74	0.8	0.8	0.67	0.8
Naïve Bayes with best 400 features	ACC	55.7	45.1	71.2	68.3	77.3	72.5	74.6	59.1	75	65.5	74.6	67.3	66.7	66.9
	MCC	0.41	0.56	0.6	0.7	0.88	0.59	0.58	0.71	0.75	0.75	0.74	0.75	0.63	0.66

From the analysis of Table 3.8 it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides overall accuracy of 79.3 and MCC of 0.79 for the prediction of subfamilies of voltage gated potassium ion channels with the complete datasets. The overall accuracy increases 79.3 % to 80.2% and overall MCC value increases from 0.79 to 0.80 with best 400 features selected by MRMR algorithms (See Table 3.8). It is also observed that the proposed method provides overall accuracy of 80.2%, MCC values of 0.80 and ROC area of 0.95 which is better in comparison with SVM, k-NN, and Naïve Bayes classifier (See Table 3.8).

3.3.2.7. Prediction of subfamilies of voltage gated sodium ion channels

For the prediction of subfamilies of voltage gated sodium ion channels a random forest with 10 fold cross validation is used on a dataset containing 857 sequence derived features of 250 number of alpha subunits and 151 number of beta subunit protein sequences and hence, a total of 646 number of voltage gated sodium ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature (See Table 3.9). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier for prediction of types of sodium ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.11 it is observed that the highest overall accuracy of 95.5 % and MCC value of 0.91 is obtained with best 200 features for the prediction of subfamilies of voltage gated sodium ion channels.

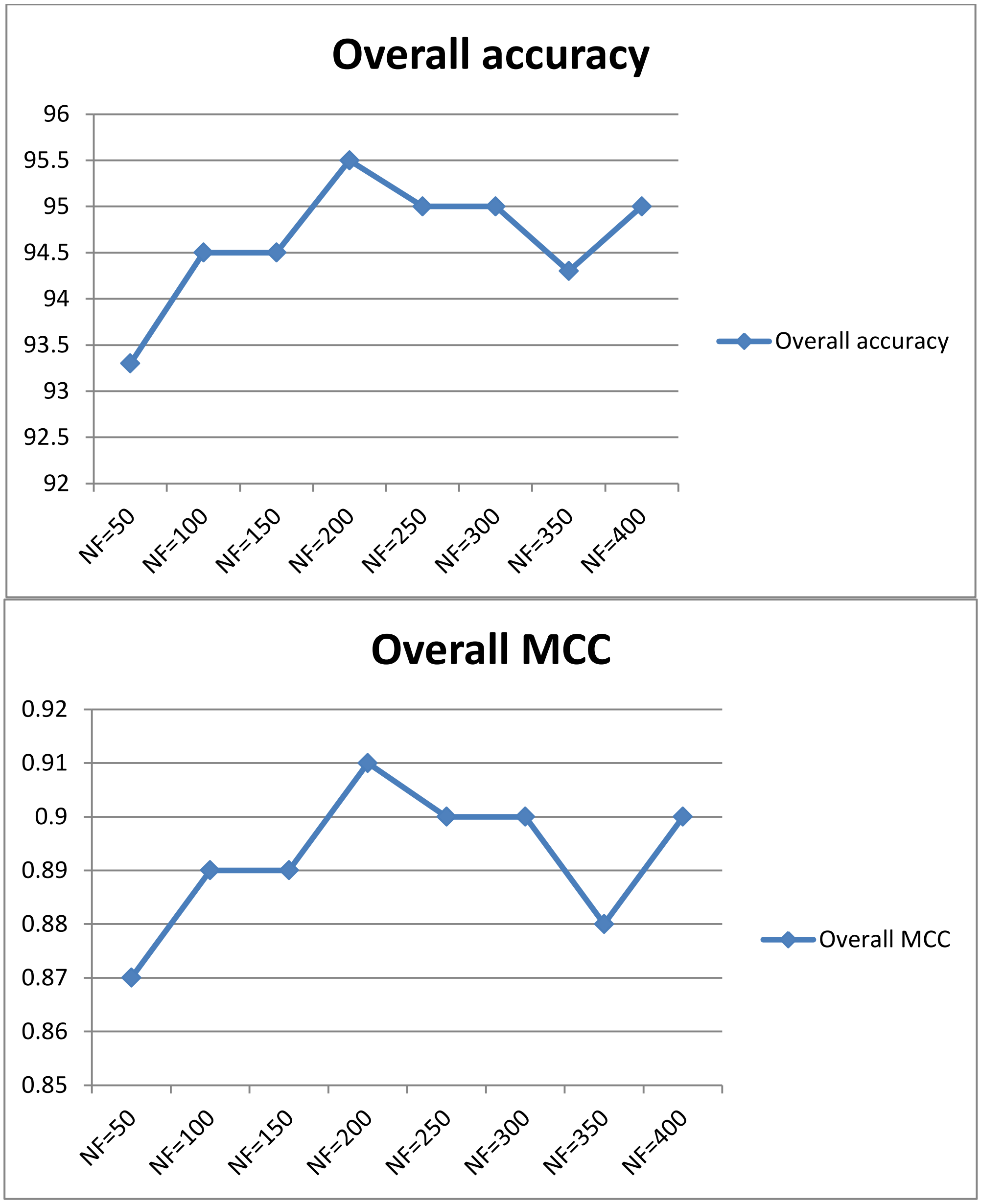


Figure 3.11: Accuracy and MCC for the prediction of subfamilies sodium ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 200 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of types of sodium ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 200 features for the classification of subfamilies of voltage gated sodium ion channels (See Table 3.9). The complete analysis of results is shown in Table 3.9.

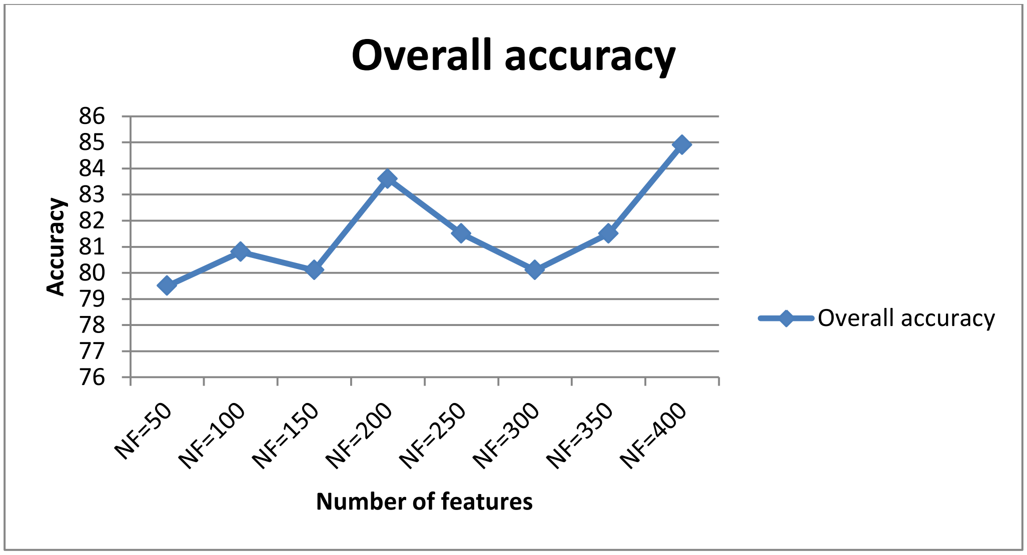
Table 3.9: The results of the prediction of subfamilies of voltage gated sodium ion channels

Method		Alpha	Beta	Overall
Random forest with AAC	ACC	90.8	92.7	91.5
	MCC	0.82	0.82	0.82
Random forest with AAC+DC	ACC	94.4	94.7	94.5
	MCC	0.88	0.88	0.88
RF with AAC+DC+CF	ACC	94	96	95
	MCC	0.89	0.89	0.89
RF with AAC+DC+CF+CTD	ACC	91.2	96	93
	MCC	0.86	0.86	0.86
RF with AAC+DC+CF+CTD+PAAC	ACC	94.8	94	94.5
	MCC	0.88	0.88	0.88
RF (with best 200 features)	**ACC**	**94.4**	**96**	**95**
	MCC	**0.9**	**0.9**	**0.9**
	ROC area	**0.99**	**0.99**	**0.99**
SVM (with best 200 features)	ACC	85.2	98	90
	MCC	0.81	0.81	0.81
	ROC area	0.88	0.88	0.88
kNN (with best 200 features)	ACC	95.2	94.7	95
	MCC	0.9	0.9	0.9
	ROC area	0.93	0.93	0.93
Naïve Bayes (with best 200 features)	ACC	87.2	91.4	88.8
	MCC	0.77	0.77	0.77
	ROC area	0.93	0.93	0.93

From the analysis of Table 3.9, it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides overall accuracy of 94.5% and MCC of 0.88 for the prediction of subfamilies of voltage gated sodium ion channels with the complete datasets. The overall accuracy is increases 94.5 % to 95.0% and overall MCC value increases from 0.88 to 0.90 with best 200 features selected by MRMR algorithms (See Table 3.9). It is also observed that the proposed method provides overall accuracy of 95.0%, MCC values of 0.90 and ROC area of 0.99 which is better in comparison with SVM, k-NN, and Naïve Bayes classifier (See Table 3.9).

3.3.2.8. Prediction of subfamilies of voltage gated chloride ion channels

For the classification of subfamilies of voltage gated chloride ion channels a random forest with 10-fold cross validation is used on a dataset containing 857 sequence derived features of 48 number of ClC1, 18 number of ClC2, 43 number of ClC3, 7 number of ClC4, 15 number of ClC5, 9 number of ClCk and 6 number of ClC6 protein sequences and hence a total of 146 number of voltage gated chloride ion channels sequences. The random forest is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature and so on (See Table 3.10). The minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier for the prediction of subfamilies of chloride ion channels. The accuracy and MCC are evaluated for different number of features and from Figure 3.12 it is observed that the overall accuracy of 84.9 % and MCC value of 0.81 is obtained with best 400 features for the prediction of subfamilies of chloride ion channels.



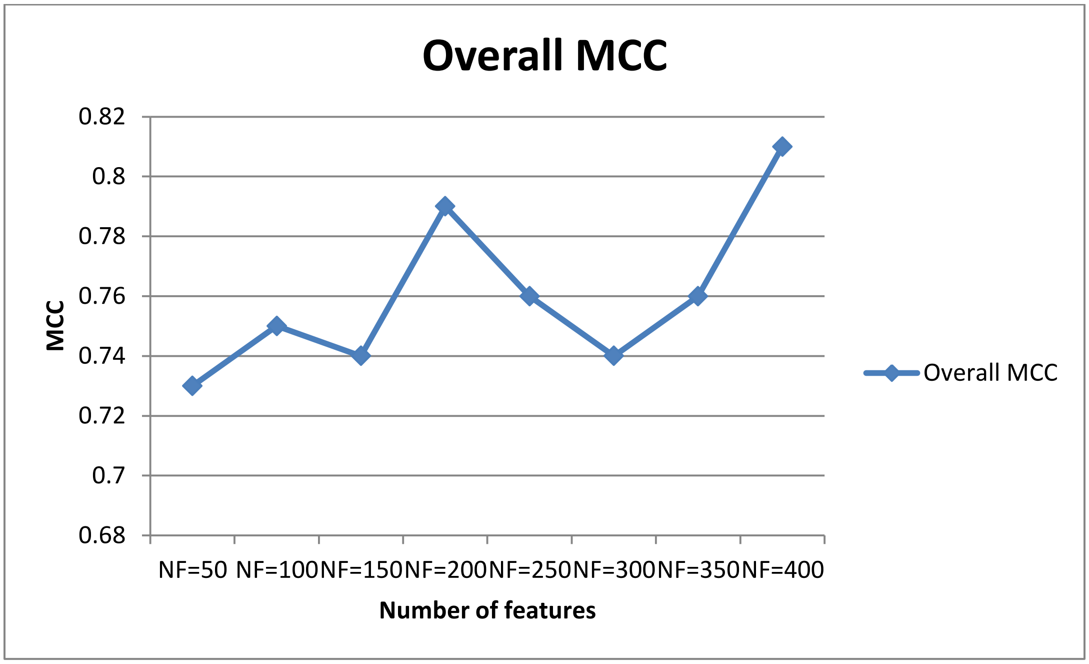


Figure 3.12 Overall accuracy and MCC for the prediction of subfamilies of chloride ion channels with different number of features (NF) selected by MRMR algorithms

Further, the best 400 features selected by MRMR are used with random forest (RF), support vector machine (SVM), k-nearest neighbor (k-NN) and Naïve Bayes classifier for the prediction of subfamilies of voltage gated chloride ion channels. The accuracy, MCC and ROC area are evaluated for different classifiers with best 400 features for the prediction of subfamilies of voltage gated chloride ion channels (See Table 3.10). The complete analysis of results is shown in Table 3.10.

Table 3.10: Results of the prediction of subfamilies of voltage gated chloride ion channels

Method		**ClC1**	**ClC2**	**ClC3**	**ClC4**	**ClC5**	**ClCk**	**ClC6**	**Overall**
Random forest with AAC	ACC	93.8	72.2	76.7	57.1	60	88.9	66.7	79.5
	MCC	0.74	0.68	0.72	0.66	0.71	0.94	0.81	0.73
Random forest with AAC+DC	ACC	93.8	72.2	76.7	57.1	60	88.9	66.7	79.5
	MCC	0.74	0.68	0.72	0.66	0.71	0.94	0.81	0.73

RF with AAC+DC+CF	ACC	97.9	55.6	86	57.1	73.3	88.9	83.3	83.6
	MCC	0.77	0.68	0.79	0.75	0.84	0.94	0.91	0.79
RF with AAC+DC+CF+CTD	ACC	100	55.6	88.4	57.1	73.3	88.9	83.3	84.9
	MCC	0.8	0.68	0.82	0.75	0.84	0.94	0.91	0.81
RF with AAC+DC+CF+CTD+PAAC	ACC	100	55.6	90.7	57.1	73.3	88.9	66.7	84.9
	MCC	0.79	0.68	0.85	0.75	0.84	0.94	0.81	0.81
RF (with best 400 features)	**ACC**	**100**	**55.6**	**90.7**	**57.1**	**73.3**	**88.9**	**66.7**	**84.9**
	MCC	**0.84**	**0.68**	**0.81**	**0.66**	**0.84**	**0.94**	**0.81**	**0.81**
	ROC area	**0.98**	**0.97**	**0.96**	**0.84**	**0.95**	**0.96**	**1.00**	**0.96**
SVM (with best 400 features)	ACC	100	50	46.5	57.1	66.7	88.9	50	69.9
	MCC	0.56	0.64	0.57	0.75	0.8	0.94	0.7	0.64
	ROC area	0.89	0.79	0.89	0.76	0.86	0.94	0.91	0.87
kNN (with best 400 features)	ACC	70.8	72.2	90.7	71.4	80	88.9	100	80.1
	MCC	0.67	0.68	0.77	0.7	0.85	0.79	1	0.74
	ROC area	0.83	0.82	0.91	0.90	0.90	0.96	1.00	0.88
Naïve Bayes (with best 400 features)	ACC	72.9	59.8	41.9	57.1	66.7	77.8	50	61.6
	MCC	0.56	0.48	0.39	0.55	0.5	0.88	0.7	0.52
	ROC area	0.88	0.82	0.80	0.72	0.82	0.94	0.73	0.83

From the analysis of Table 3.10, it is observed that the performance of random forest classifier is improved by using the mixture of the feature vectors. The proposed method provides overall accuracy of 84.9% and MCC value of 0.81 for the prediction of subfamilies of voltage gated chloride ion channels with the complete datasets and that of with best 400 features selected by MRMR algorithms (See Table 3.10). It is also observed that the proposed method provides overall accuracy of 84.9%, MCC values of 0.81 and ROC area of 0.96 which is better in comparison with SVM, k-NN, and Naïve Bayes classifier (See Table 3.10).

3.4. Comparative analysis

In this chapter, the performance of the classifier is evaluated at optimal number of features selected by MRMR feature selection algorithms and the results are compared with the previous approaches proposed by the authors of the papers (Chen *et al.*, 2012; Saha *et al.*, 2006; Lin *et al.,* 2011). It is observed that the proposed method improve the performance for the prediction of ion channels and their subfamilies. The comparative analysis is shown in Table 3.11, 3.12, 3.13, and 3.14.

Table 3.11: Result comparison for the prediction of ion channels and non-ion channels among existing methods and proposed method

Family	Proposed Method with best 50 features			SVM with RBF kernel, gamma=60, C=100, and threshold value=0.3 (with best 50 features) (**Saha *et al.*, 2006)**			SVM with One Vs. One strategy, kernel= RBF with best 50 features **(Lin *et al.*, 2011)**		
	ACC	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**
Non-ion channel	100	1	1	97	0.98	0.98	100	1	1
Ion channel	100	1	1	100	0.98	0.98	100	1	1
Overall	100	1	1	99.2	0.98	0.98	100	1	1

Table 3.12: Result comparison for the prediction of voltage and ligand gated ion channels among existing methods and proposed method

Subfamilies of ion channel	**Proposed Method with best 200 features**			**SVM with One Vs. One strategy, kernel= RBF with best 200 features (Lin *et al.*, 2011)**		
	ACC	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**
Voltage gated ion channels	99.9	0.93	0.99	99.9	0.85	0.93
Ligand gated ion channels	89.2	0.93	0.99	75.5	0.85	0.93
Overall	98.3	0.93	0.99	96.3	0.85	0.93

Table 3.13 :Result comparison for the prediction of subfamilies of voltage gated ion channels among existing methods and proposed method

Subfamilies Of voltage gated ion channels	Proposed Method with best 150 features			SVM, RBF kernel, gamma=50, C=10 with best 150 features (**Saha *et al.*, 2006)**			SVM with One Vs. One strategy, kernel= RBF with best 150 features (**Lin *et al.*, 2011)**		
	ACC	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**
Sodium	80.3	0.81	0.97	64.1	0.73	0.83	73.1	0.78	0.92
Calcium	93.8	0.86	0.98	81.9	0.78	0.87	93.5	0.84	0.93
Potassium	100	0.98	1.00	100	0.76	0.97	100	0.94	0.98
Chloride	81.5	0.86	1.00	49.3	0.68	0.87	77.4	0.83	0.93
Overall	92.1	0.89	0.99	81.8	0.75	0.89	90.0	0.86	0.95

Table 3.14: Result comparison for the prediction of subfamilies of voltage gated potassium ion channels among existing methods and proposed method

Subfamilies of voltage gated potassium ion channels	Random Forest with best 400 features			SVM with One Vs. One strategy, kernel= RBF with best 400 features **(Chen *et al.*, 2012)**		
	ACC	**MCC**	**ROC area**	**ACC**	**MCC**	**ROC area**
Kv1	83.6	0.66	0.94	83.6	0.60	0.89
Kv2	78.4	0.78	0.96	78.4	0.82	0.93
Kv3	78.8	0.82	0.94	69.2	0.70	0.90
Kv4	88.9	0.88	0.98	79.4	0.86	0.95
Kv5	95.5	0.98	1.00	95.5	0.98	0.98
Kv6	80.4	0.77	0.96	82.4	0.77	0.94
Kv7	84.7	0.65	0.95	74.6	0.59	0.90
Kv8.2	54.5	0.73	0.93	61.9	0.73	0.91
Kv9	84.6	0.85	0.96	80.8	0.83	0.92
Kv10	78.2	0.85	0.96	78.2	0.85	0.91
Kv11	78	0.82	0.94	79.7	0.78	0.92
Kv12	76.9	0.83	0.94	73.1	0.82	0.91
Kv13	70.4	0.83	0.94	63	0.76	0.94
Overall	80.2	0.8	0.95	76.9	0.77	0.92

3.5. Conclusion

In this chapter, random forest based approach has been proposed to predict ion channels and their subfamilies by using sequence derived features. The minimum redundancy and maximum relevance feature selection was used to find the optimal number of features for improving the prediction performance. The results shows that the MRMR feature selection algorithm reduced the number of input feature vectors by selecting the important features and improve the overall accuracy and MCC. In the 10-fold cross validation the proposed method has achieved an overall accuracy of 100%, 98.01%, 91.5%, 93.0%, 92.2%, 78.6%, 95.5%, 84.9%, MCC values of 1.00, 0.92, 0.88, 0.88, 0.90, 0.79, 0.91, 0.81 and ROC area values of 1.00, 0.99, 0.99, 0.99, 0.99, 0.95, 0.99 and 0.96 to predict ion channels and non-ion channels, voltage gated ion channels and ligand gated ion channels, four types (calcium, potassium, sodium and chloride) of voltage gated ion channels, ligand gated ion channels and predict subfamilies of voltage gated calcium, potassium, sodium and chloride ion channels respectively. The high accuracies, MCC and ROC area values indicate that the proposed method may be useful for the prediction of ion channels and their subfamilies.