Chapter 6

AN EFFICIENT AND ROBUST APPROACH FOR THE PREDICTION OF G-PROTEIN COUPLED RECEPTORS AND THEIR SUBFAMILIES

G-protein coupled receptors (GPCRs) are seven-transmembrane domain receptors that sense molecules outside the cell and activate inside signal transduction pathways for cellular responses. These are called seven-transmembrane receptors because they pass through the cell membrane seven times. GPCRs can be grouped into six classes based on sequence homology and functional similarity these are Class A (Rhodopsin-like), Class B (Secretin like), Class C (Metabotropic glutamate), Class D (cyclic AMP), Class E (Taste) and Class F (Vomeronasal) receptors. There are a larger number of G-protein coupled receptors are available in human in these some have been identified their function like growth factors, light, hormones, amines, neurotransmitters, and lipids etc. However, a large number of the GPCRs found in the human genome have unknown functions and so it is necessary to design an efficient approach to predict families and subfamilies of G-protein coupled receptors for the new drug discovery.

In this chapter, a weighted k- nearest neighbor in which inverse kernel function is applied to calculate weighted distance to improve the prediction

performance of G-protein coupled receptors and their subfamilies by using sequence derived properties. In this chapter, 1497 sequence derived features with eight features vectors such as amino acid composition, dipeptide composition, correlation, composition, transition, distribution, sequence order descriptors and pseudo amino acid composition are used to predict G-protein coupled receptors and their subfamilies. For non-redundant, relevant, robust, and optimal feature subset selection, a feature selection method based on fusion of five supervised filter based methods is proposed. These supervised feature selection methods include Fisher score based feature selection (Guyon *et al.,* 2003), ReliefF (Kira *et al.,* 1992), fast correlation based filter (FCBF) (Yu *et al.,* 2003), minimum redundancy and maximum relevancy (MRMR) (Peng *et al.,* 2005), and support vector machine based recursive feature elimination (SVM-RFE) (Guyon *et al.*, 2002). If we apply, these feature selection methods on the same dataset then each of them results in different feature subset where features are ranked according to their rank. Also the performance of a classifier for each feature subset selected by different method may be different. Therefore, here we address this problem by proposing a method for optimal feature selection by the fusion of feature subsets produced by these methods using UNION of the selected features by different feature selection algorithms. Further, in this chapter the proposed method used three level strategies to predict G-protein coupled receptors and their subfamilies. First, it is determined that protein sequence is G-protein coupled receptors or non-G-protein coupled receptors. Second, if protein is classified as G-protein coupled receptors then the method classify the subfamilies of G-protein coupled receptors. Third, if it is classified as class A G-protein coupled receptors then the method classify the subfamilies of class A G-protein coupled receptors. In this chapter, hold one out cross validation is used to find the best k between 1 and 30 the value specified to the k-NN parameter.

6.1. Background

Here, the sequence of G-protein coupled receptors with their properties and the proposed methods and model that are used to predict G-protein coupled receptors and their subfamilies are presented.

6.1.1. Material and methods

In this chapter, the sequences of the G-protein coupled receptors are extracted from GPCRDB (Horn *et al.,* 2003) (http://www.gpcr.org/7tm/). Here, all the 576 non- G-protein coupled receptors proteins are selected from Uniport database with the keyword NOT GPCRs. 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 datasets is shown in Table 6.1.

Table 6.1: Number of sequences belonging to each GPCRs and their subfamilies

<table>
<tr><th>Families</th><th>GPCR Families</th><th>Class A GPCR sub-families</th><th colspan="2">No. of Sequences</th></tr>
<tr><td rowspan="20">GPCR</td><td rowspan="13">Rhodopsin like</td><td>Amine</td><td>78</td><td rowspan="13">974</td></tr>
<tr><td>Peptide</td><td>65</td></tr>
<tr><td>Rhodopsin</td><td>61</td></tr>
<tr><td>Olfactory</td><td>64</td></tr>
<tr><td>Prostanoid</td><td>60</td></tr>
<tr><td>Nucleotide-like</td><td>99</td></tr>
<tr><td>Cannabinoid</td><td>66</td></tr>
<tr><td>Platelet activating factor</td><td>50</td></tr>
<tr><td>Gonadotropin-releasing hormone</td><td>92</td></tr>
<tr><td>Thyrotropin-releasing hormone</td><td>71</td></tr>
<tr><td>Viral</td><td>47</td></tr>
<tr><td>Lysosphingolipid and LPA</td><td>70</td></tr>
<tr><td>Leukotriene B4 receptor</td><td>59</td></tr>
<tr><td>Ecdysis triggering hormone receptor</td><td>25</td></tr>
<tr><td>Hydroxycarboxylic acid receptor</td><td>43</td></tr>
<tr><td>CAPA</td><td>24</td></tr>
<tr><td colspan="2">Secretin like</td><td colspan="2">621</td></tr>
<tr><td colspan="2">Metabotropic glutamate</td><td colspan="2">454</td></tr>
<tr><td colspan="2">cAMP receptors</td><td colspan="2">8</td></tr>
<tr><td colspan="2">Taste</td><td colspan="2">480</td></tr>
<tr><td colspan="2">Vomeronasal receptors</td><td colspan="2">425</td></tr>
<tr><td colspan="3">Non-GPCR</td><td colspan="2">576</td></tr>
</table>

6.1.2. Features extraction of protein sequences

In this chapter, to fully characterize protein sequence eight feature vectors are extracted from PROFEAT server (Rao *et al.,* 2011) to represent the protein sample, including amino acid composition, dipeptide composition, correlation, composition, transition, distribution of physiochemical properties, sequence order descriptors, and pseudo amino acid composition with total of 1497 number of features being calculated for the prediction of G-protein coupled receptors and their subfamilies. The brief description of sequence derived properties of G-protein coupled receptors is given in Table 6.2.

Table 6.2: Description of sequence derived properties of G-protein coupled receptors

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_{1140}	720	Correlation factors
4	X_{1141} to X_{1161}	21	Composition
5	X_{1162} to X_{1182}	21	Transition
6	X_{1183} to X_{1287}	105	Distribution of physiochemical properties
7	X_{1288} to X_{1447}	160	Sequence order descriptors
8	X_{1448} to X_{1497}	50	Pseudo amino acid composition

6.2. Proposed method and model

In this chapter, a feature selection method based on fusion of five supervised filter based methods is proposed for non-redundant, relevant, robust, and optimal feature subset selection for the prediction of G-protein coupled

receptor and their subfamilies. Here, a weighted k- nearest neighbor is proposed to use in which inverse kernel function is applied to calculate weighted distance to improve the prediction performance of G-protein coupled receptors and their subfamilies by using sequence derived properties.

6.2.1. Feature subset selection

In this chapter, five supervised filter based methods such as Fisher score based feature selection (Guyon *et al.*, 2003), ReliefF (Kira *et al.*, 1992), fast correlation based filter (FCBF) (Yu *et al.*, 2003), minimum redundancy and maximum relevancy (MRMR) (Peng *et al.*, 2005), and support vector machine based recursive feature elimination (SVM-RFE) (Guyon *et al.,* 2002) feature selection methods are used to obtain optimal number of features. If we apply, these feature selection methods on the same dataset then each of them results in different feature subset where features are ranked according to their rank. Also the performance of a classifier for each feature subset selected by different method may be different. Therefore, here we address this problem by proposing a method for optimal feature selection by the fusion of feature subsets produced by these methods using UNION of the selected features by different feature selection algorithms. The brief descriptions of these five supervised feature selection methods are as follows.

ReliefF

ReliefF is a simple and efficient supervised feature selection algorithm inspired by instance based learning to estimate quality of features in problems with strong dependencies between features. The ReliefF randomly select an instance R_i then searches for k of its nearest neighbors from the same class called nearest hits H_j, and also k nearest neighbors from each of the different classes, called nearest misses $M_j(C)$. It then updates the quality estimation $W[A]$ for an attribute based on their values for R_i, H_j, and $M_j(C)$. If instance R_i and H_j have different values of the attribute 'A' then the attribute 'A' separates two instances with the same class which is not desirable so the quality estimation $W[A]$ is decreased. On the other hand, if instance R_i and $M_j(C)$ have different values on the 'A' attribute then $W[A]$ is increased. This whole process is

repeated an *m* time which is defined by users and the quality of attributes are estimated as follows.

$$W[A] = W[A] - \sum_{j} \frac{diff(A, R_i, H_j)}{(m.k)} + \frac{\sum_{C \neq class(R_i)} \left[\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C)) \right]}{(m.k)} \quad (6.1)$$

Minimum redundancy maximum relevance

Minimum redundancy maximum relevance (MRMR) feature selection is a multivariate feature selection method which starts with an empty set uses mutual information to weight features and forward selection technique with sequential search strategy to find the best subset of features. 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

$$MI(x, y) = \sum_{i,j \in N} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)} \quad (6.2)$$

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

$$MI(x, c) = \sum_{i,k \in N} p(x_i, c_k) log \frac{p(x_i, c_k)}{p(x_i)\, p(c_k)} \quad (6.3)$$

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

$$Redundancy = \frac{1}{|S|^2} \sum_{x,y \in S} MI(x, y) \quad (6.4)$$

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

$$Relevance = \frac{1}{|S|}\sum_{x \in S} MI(x, c) \quad (6.5)$$

The optimal subset of features is selected by

$$Max\ (Relevance - \ Redundancy) \quad (6.6)$$

Fast correlation based filter (FCBF)

The fast correlation based filter is a multivariate feature selection method which starts with full set of features, uses symmetrical uncertainty to calculate dependences of features and finds best subset using backward selection technique with sequential search strategy. It consists of two stages: the first one is a relevance analysis that aimed at ordering the input variables depending on a relevance score, which is computed as the symmetric uncertainty with respect to the target output. This stage is also used to discard irrelevant variables, which are those whose ranking score is below a predefined threshold. The second stage is a redundancy analysis, aimed at selecting predominant features from the relevant set obtained in the first stage. It has an inside stopping criterion that makes it stop when there are no features left to eliminate. Symmetrical Uncertainty (SU) is a normalized information theoretic measure which uses entropy and conditional entropy values to calculate dependencies of features. If *X* is a random variable and *P*(*x*) is the probability of *x*, the entropy of *X* is

$$H(X) = -\sum p(x_i)\ log_2(p(x_i)) \quad (6.7)$$

Conditional uncertainty of X given random variable Y is the average conditional uncertainty of X over Y

$$H\left(\frac{X}{Y}\right) = -\sum_j p(y_j) \sum_i p\left(\frac{x_i}{y_i}\right)\ log_2(p\left(\frac{x_i}{y_i}\right) \quad (6.8)$$

Symmetrical uncertainty (SU) is defined as

$$SU(X,Y) = 2\left[\frac{H(X) - H\left(\frac{X}{Y}\right)}{H(X) + H(Y)}\right] \quad (6.9)$$

The symmetrical uncertainty value of 1 indicate that by using one feature value other feature value can be totally predicted and value 0 indicates two features are totally independent.

Fisher score based filter

Fisher score is one of the most widely used supervised feature selection methods. The basis criterion for Fisher score based feature selection is that it is a univariate filter method which evaluates each feature individually. However, it selects each feature independently according to their scores under the Fisher criterion, which leads to a suboptimal subset of features. The fishers score is defined as the ratio of the variance of the between classes to the variance of within classes. It is defined as

$$\text{Fishers score} = \frac{\sigma^2\ between\ classes}{\sigma^2\ within\ classes} = \frac{maximum\ between\ class\ variance(difference\ of\ mean)}{minimum\ within\ class\ variance(sum\ of\ variances)} \quad (6.7)$$

The fisher score are calculated for each attributes and then high fisher score features are selected.

Support vector machine-recursive feature elimination (SVM-RFE)

Support vector machine-recursive feature elimination selects features in sequential backward elimination manner which starts with all the features and discards one feature at a time. It uses the weight magnitude as the ranking criterion. It performs following steps.

1. Train an SVM classifier on the training set
2. Compute the ranking criteria i.e. squared coefficients for all the features
3. Ranking the features using the weights of the resulting classifier
4. Eliminate features with the smallest weight
5. Repeat the process with the remaining features

6.2.2. Classification of G-protein coupled receptors and their subfamilies

For the classification of G-protein coupled receptors and their subfamilies the weighted k- neighbor classifier (wk-NN) (Hechenbichler *et al.,* 2004) has been used. The k-nearest neighbor (k-NN) classifier is a simple and effective instance based learning algorithms which are based on finding the k nearest neighbor and taking a majority vote among the classes of these k neighbors to assign a class for the given query. The distance functions affect the performance of the k-NN classifier. In this chapter, a weighted k-nearest neighbor is proposed to use in which inverse kernel function are applied to calculate weighted distance to improve the prediction performance of G-protein coupled receptors families and their subfamilies. Basically the traditional k-NN classifier is used each time a different k, starting from k = 1 to $\sqrt{n}$, where *n* is the size of the *training set*. Here, hold-one-out cross validation is used to find the best values of k between 1 and 30 the value specified to the wk-NN parameter. The weighted k-nearest neighbor classifier performs the following steps:

1. Let $L = \{(y_i, x_i),\ i = 1, \ldots\ldots\ldots, n_L\}$ be a training datasets where $y_i \in \{1 \ldots . c\}$ represents the class and $x_i' = (x_{i1}, \ldots\ldots\ldots, x_{ip})$ represents the predictor values. Let *x* be the test sample whose class level *y* has to be predicted.
2. Obtain k+1 nearest neighbor to *x* by using Manhattan distance function $d(x,x_i)$.Here

$$Manhattan\ distance = d(x, x_i) = \sum_{s=1}^{p} |x_s - x_{is}| \quad (6.8)$$

3. The $(k+1)^{th}$ neighbor is used for the standardization of the *k* smallest distance by using the equation

$$D(x, x_i) = \frac{d(x, x_i)}{d(x, x_{k+1})} \quad (6.9)$$

4. Transform and normalize distance by using inverse kernel function $K = \frac{1}{|d|}$ to obtain the weight $w_i = {}^{1}/_{D(x, x_i)}$ (6.10)

5. Assign a class, *y* of test sample, *x* which shows a weighted majority of the k-nearest neighbor.

In this chapter, the proposed method used three level strategies to predict G-protein coupled receptors and their subfamilies. The complete procedure of the proposed method for the prediction of G-protein coupled receptors and their subfamilies is illustrated in Figure 6.1 and the steps are as follows:

1. Produce eight feature vectors with 1497 number of features that represent a protein sequence.
2. Select best 50 number of features with Fisher score based, ReliefF, FCBF, MRMR, and SVM-RFE feature selection methods.
3. Fusion of feature subsets produced by these methods using UNION of the selected features by different feature selection algorithms.
4. Apply weighted k- nearest neighbor classifier for the prediction of G-protein coupled receptors and their subfamilies as follows:

First, it is determined that protein sequence is G-protein coupled receptors or non-G-protein coupled receptors. Second, if protein is classified as G-protein coupled receptors then the method classify the subfamilies of G-protein coupled receptors. Third, if it is classified as class A G-protein coupled receptors then the method classify the subfamilies of class A G-protein coupled receptors.

6.3. Result and performance analysis

Here, a weighted k- nearest neighbor is proposed to use in which inverse kernel function is applied to calculate weighted distance with UNION of features selected by five supervised filters, to improve the prediction performance of G-protein coupled receptors and their subfamilies.

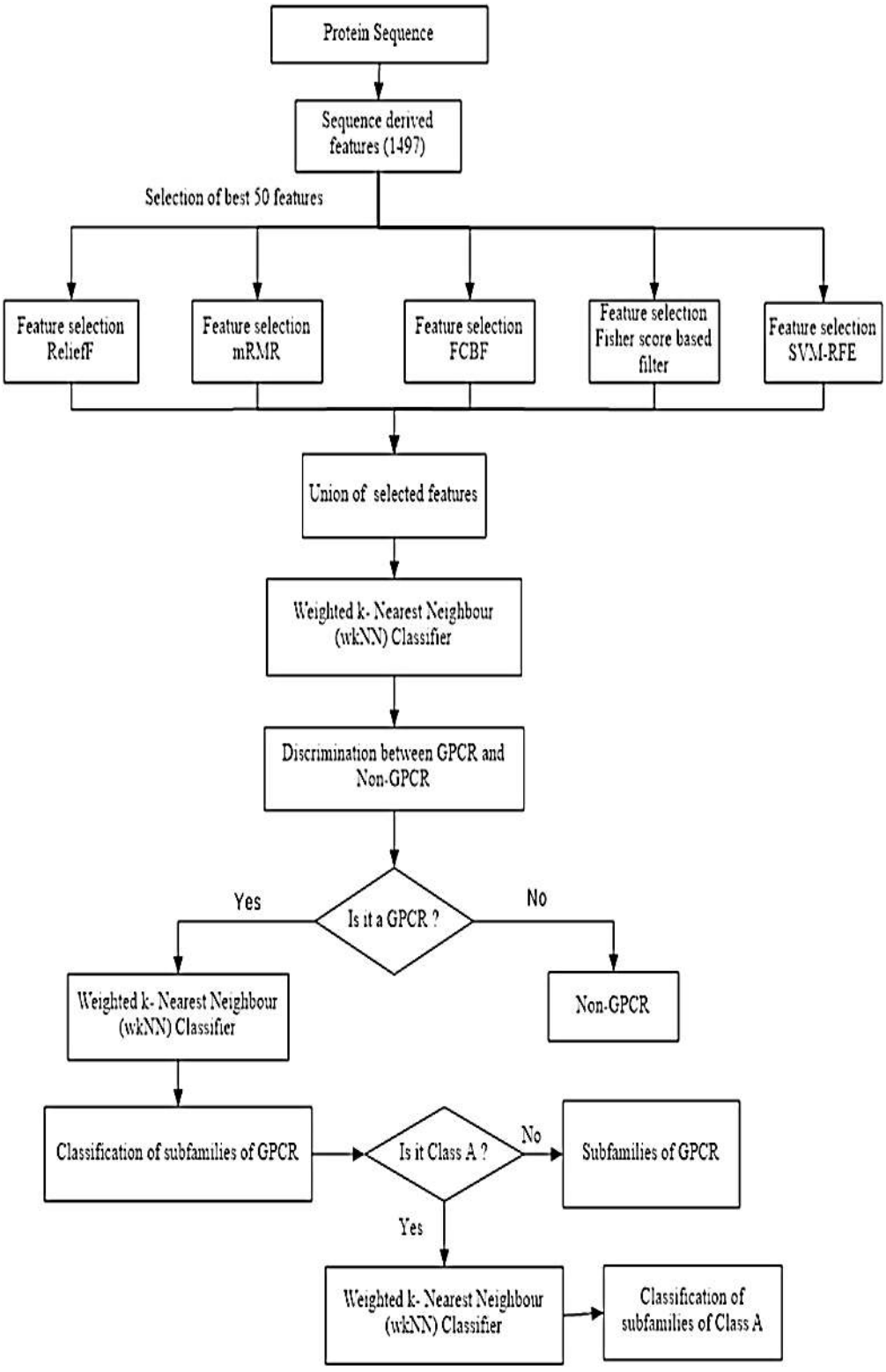


Figure 6.1: A flowchart for the proposed model for the prdiction of GPCR and their families

6.3.1. Performance measures

In this chapter, 10-fold cross validation is used to measure the performance of weighted k-NN 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.

6.3.2. Results and comparative analysis

In this chapter, a weighted k- nearest neighbor is proposed to be used with inverse kernel function to calculate weighted distance to improve the prediction performance of G-protein coupled receptors and their subfamilies by using sequence derived properties. Here, hold-one-out cross validation is used to find the best values of k between 1 and 30 used as the number of nearest neighbor value specified to the weighted k-NN classifier. 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 G-protein coupled receptors and their subfamilies are presented and discussed. The performance analysis of the proposed model is shown for the best 50 number of features selected among 1497 total number of features by using Fisher score based feature selection algorithms, ReliefF, FCBF, MRMR, and SVM-RFE feature selection methods, for each cases as well as for the UNION of these selected features.

6.3.2.1. Prediction of GPCRs and non-GPCRs

To predict the G-protein coupled receptors and Non-GPCRs, a weighted k-nearest neighbor is evaluated with for the best 50 number of features selected among 1497 total number of features by using Fisher score based feature selection algorithms, ReliefF, FCBF, MRMR, SVM-RFE feature selection

methods and the UNION of these selected features. It is observed that the performance of the classifier is improved by using the UNION of the best 50 features selected by five different feature selection algorithms (See Figure 6.2).

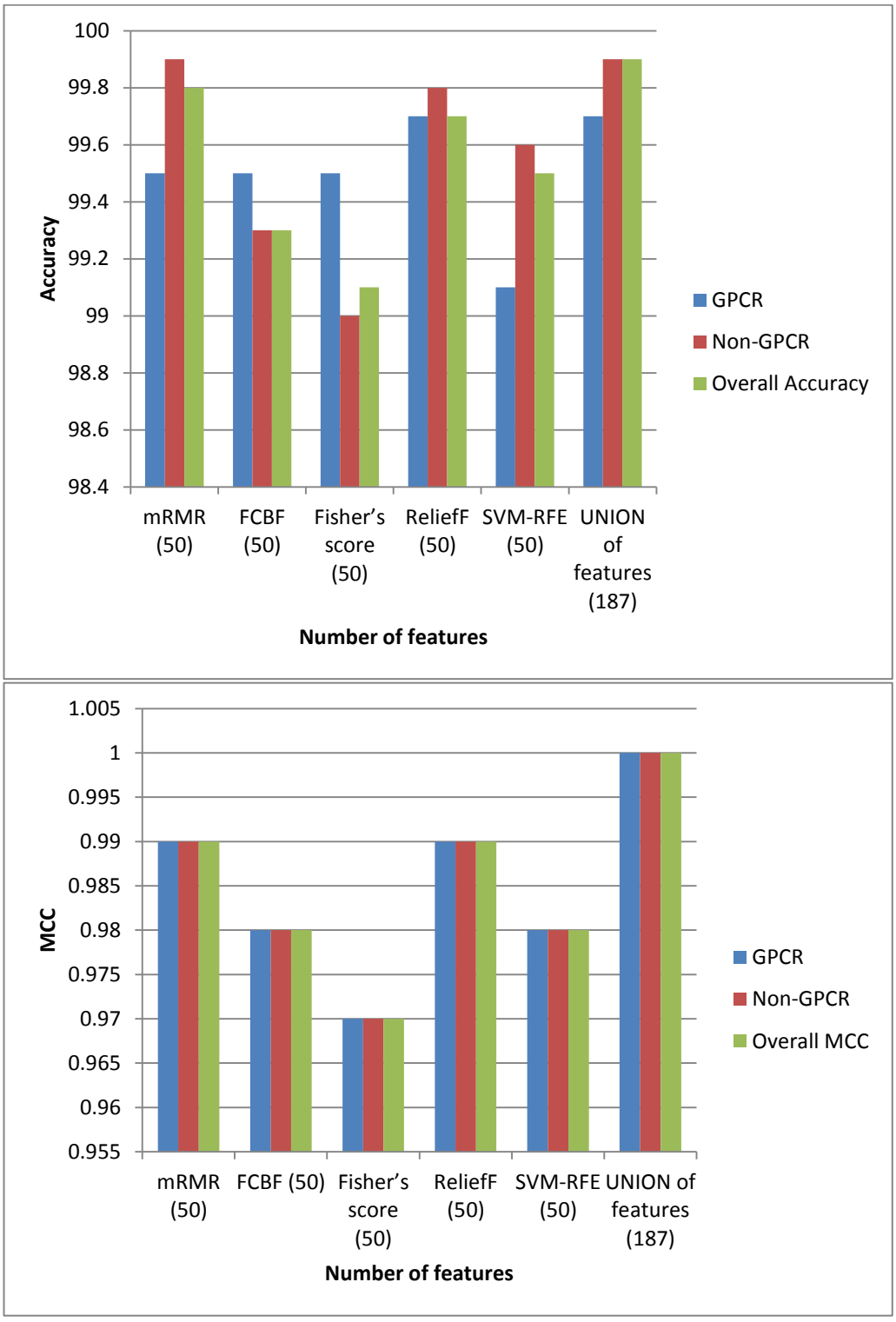


Figure 6.2: Accuracy and MCC for classification of GPCR and non-GPCR with different data sets

Table 6.3: Result analysis for the classification of GPCRs and non-GPCRs

Family	Proposed Method (Prediction results using proposed feature selection method and weighted kNN)				Prediction using proposed feature selection method and SVM, (Kernel =RBF, γ=200, C=100)			
	ACC	**MCC**	**ROC area**	**Precision**	**ACC**	**MCC**	**ROC area**	**Precision**
Non-GPCR	99.7	1.00	1.00	99.5	99.5	0.996	0.997	99.8
GPCR	99.9	1.00	1.00	99.9	100	0.996	0.997	99.9
Overall	**99.9**	**1.00**	**1.00**	**99.9**	99.9	0.996	0.997	99.9

From the analysis of Table 6.3 it is observed that the performance of a weighted k-nearest neighbor method provide overall accuracy of 99.9%, MCC of 1.00, ROC area of 1.00 and precision of 99.9%. Further, the support vector machine based classifier was also used to examine the prediction rate of G-protein coupled receptors and non-G-protein coupled receptors using the proposed feature selection method (with RBF kernel function and the tuning parameters γ=200, C=100) and it provides overall accuracy of 99.9%, MCC of 0.996, ROC area of 0.997 and precision of 99.9%. Out of the two classifiers viz. weighted k-NN and SVM, the weighted k-NN based classifier is performing better in conjunction with the proposed feature selection method. The performance of the proposed method is also compared with the (Bhasin, Manoj *et al.*, 2004) that used only dipeptide composition with SVM classifier and reported an accuracy of 99.5%, MCC of 0.99 for the prediction of G-protein coupled receptors and non- G-protein coupled receptors. It is observed that the proposed method is performing better in comparison to other methods available in literature.

6.3.2.2. Prediction of subfamilies of G-protein coupled receptors

To predict the subfamilies of G-protein coupled receptors, a weighted k-nearest neighbor is evaluated with for the best 50 number of features selected among 1497 total number of features by using Fisher score based feature selection algorithms, ReliefF, FCBF, MRMR, SVM-RFE feature selection methods and the UNION of these selected features. It is observed that the performance of the classifier is improved by using the UNION of the best 50 features selected by five different feature selection algorithms (See Figure 6.3).

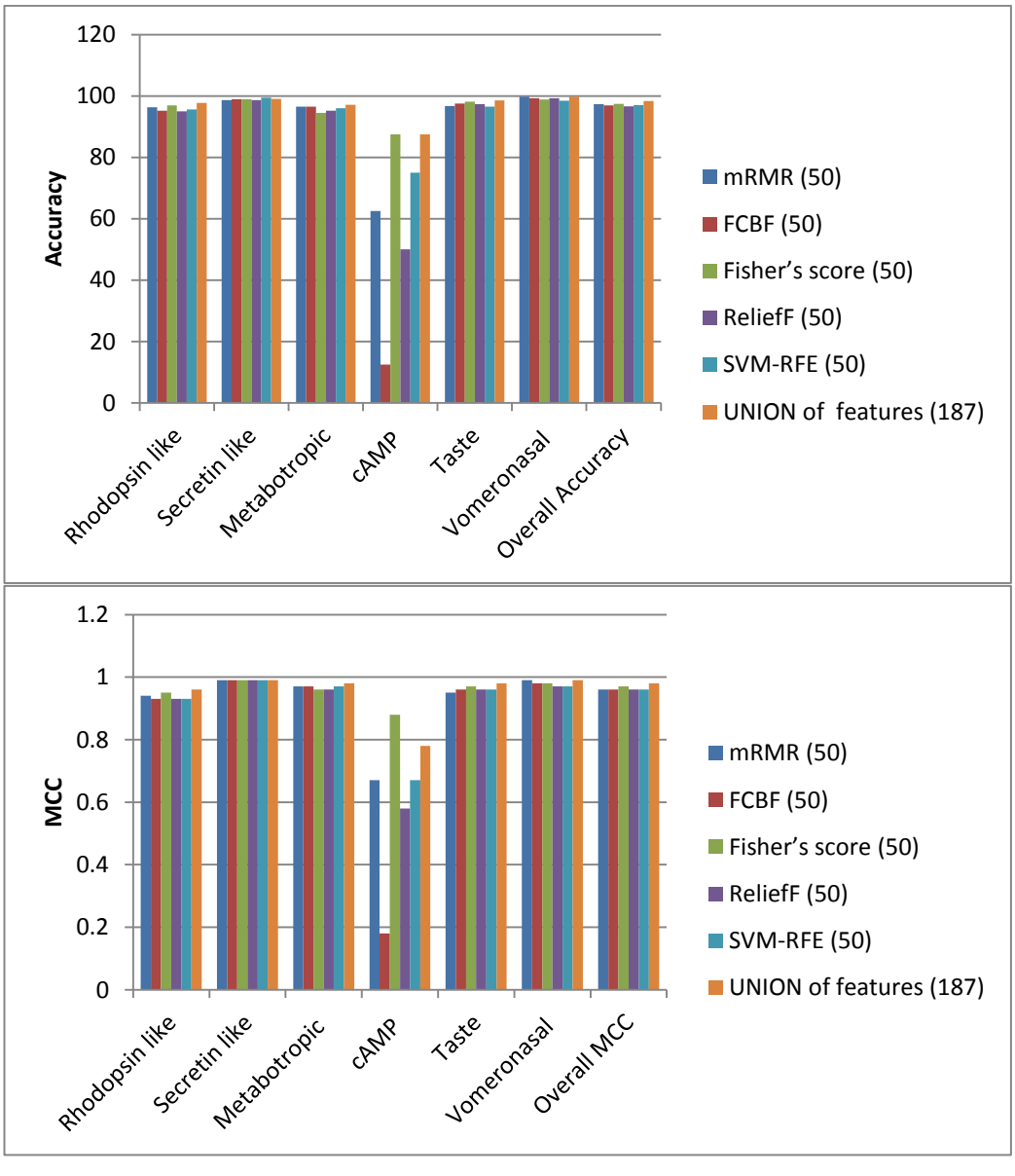


Figure 6.3: Accuracy and MCC for classification of families of GPCRs with different data sets

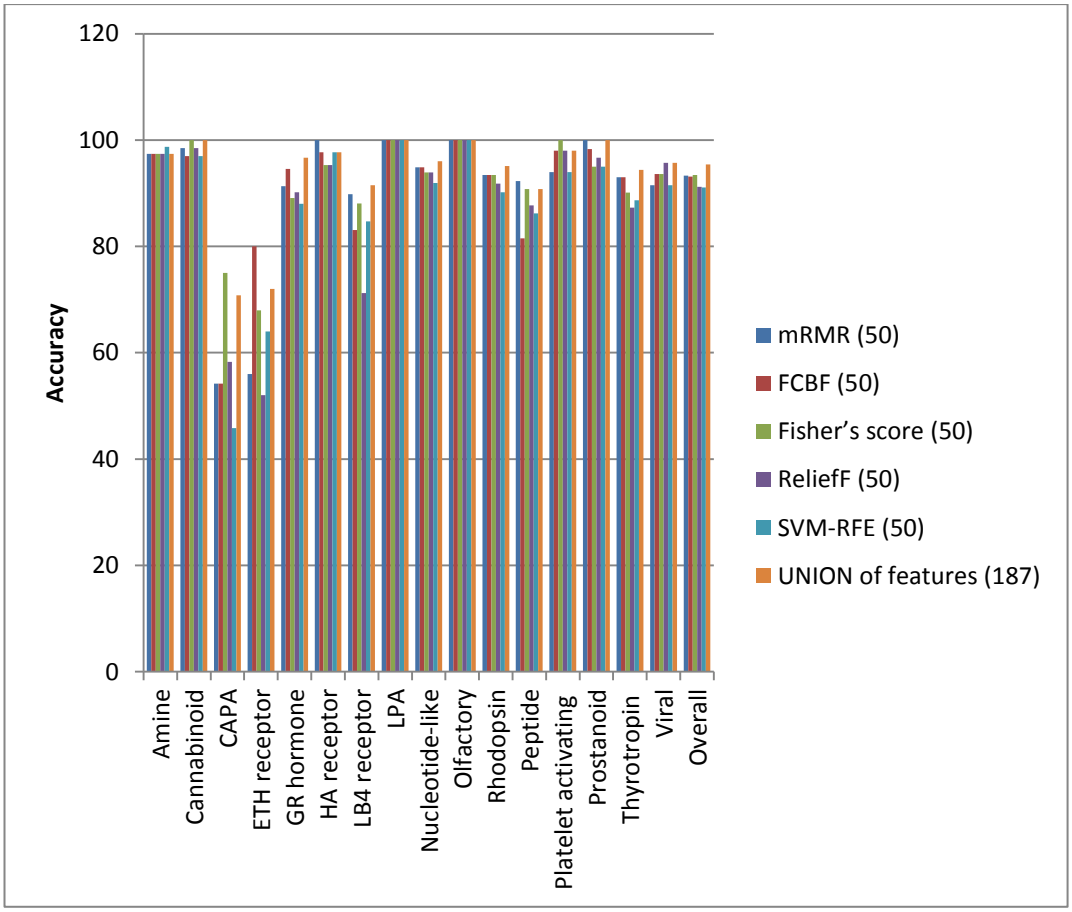
Table 6.4: Result analysis for the classification of subfamilies of GPCRs

Subfamilies of GPCR	Proposed Method (Prediction results using proposed feature selection method and weighted kNN)				Prediction using proposed feature selection method and SVM, (Kernel =RBF, γ=200, C=100)			
	ACC	**MCC**	**ROC area**	**Pre-cision**	**ACC**	**MCC**	**ROC area**	**Pre-cision**
Rhodopsin like	97.7	0.96	0.997	97.3	96.7	0.942	0.973	95.5
Secretin like	99	0.99	0.999	99.8	98.6	0.987	0.992	99.4
Metabotropic glutamate	97.1	0.98	0.999	98.7	96.9	0.975	0.984	98.9
cAMP	87.5	0.78	0.977	70	37.5	0.612	0.688	100
Taste	98.5	0.98	0.998	98.3	97.1	0.963	0.982	96.7
Vomeronasal	99.8	0.99	1.00	98.6	98.4	0.974	0.989	97.2
Overall	**98.3**	**0.98**	**0.998**	**98.3**	97.3	0.964	0.981	97.3

From the analysis of Table 6.4 it is observed that the performance of a weighted k-nearest neighbor methods provides overall accuracy of 98.3%, MCC of 0.98, ROC area of 0.998 and precision of 98.3%. Further, the support vector machine based classifier was also used to examine the prediction rate of subfamilies of GPCRs using the proposed feature selection method (with RBF kernel function and the tuning parameters γ=200, C=100) and it provides overall accuracy of 97.3%, MCC of 0.964, ROC area of 0.981 and Precision of 97.3%. Out of the two classifiers viz. weighted k-NN and SVM, the weighted k-NN based classifier is performing better in conjunction with the proposed feature selection method (See Table 6.4). The performance of the proposed method is also compared with the (Bhasin, *et al.*, 2004) that used only dipeptide composition with SVM classifier and reported an accuracy of 91.2% for the prediction of subfamilies of GPCRs. It is observed that the proposed method is performing better in comparison to other methods.

6.3.2.3. Prediction of subfamilies of class A G-protein coupled receptors

To predict the subfamilies of class A G-protein coupled receptors, a weighted k-nearest neighbor is evaluated with for the best 50 number of features selected among 1497 total number of features by using Fisher score based feature selection, ReliefF, FCBF, MRMR, SVM-RFE feature selection methods and the UNION of these selected features. It is observed that the performance of the classifier is improved by using the UNION of the best 50 features selected by five different feature selection algorithms (See Figure 6.4).



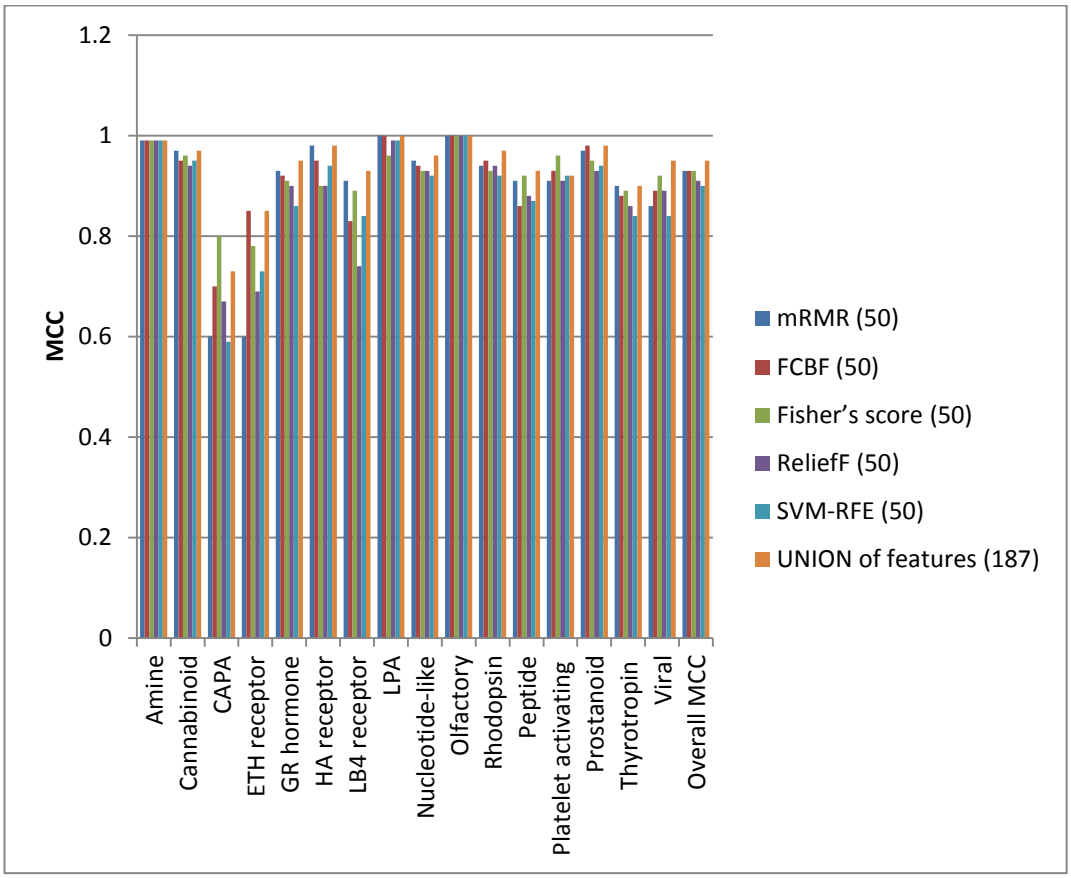


Figure 6.4: Accuracy and MCC for classification of Subfamilies of class A GPCRs with different data sets

From the analysis of Table 6.5 it is observed that the performance of weighted k-nearest neighbor methods provides overall accuracy of 95.4%, MCC of 0.951, ROC area of 0.996 and precision of 95.5%. Further, the support vector machine based classifier was also used to examine the prediction rate subfamilies of class A GPCRs using the proposed feature selection method (with RBF kernel function and the tuning parameters γ=200, C=100) and it provides overall accuracy of 91.5%, MCC of 0.909, ROC area of 0.955 and precision of 91.4%. Out of the two classifiers viz. weighted k-NN and SVM, the weighted k-NN based classifier is performing better in conjunction with the proposed feature selection method (See Table 6.5). Bhasin et al. (2004) used only dipeptide composition with SVM classifier and reported an accuracy of 92.6% for the prediction of subfamilies of class A G-protein coupled receptors.

Table 6.5: Result analysis for classification of subfamilies of class A GPCRs with different data sets

Subfamilies of Class A GPCR	Proposed Method (Prediction results using proposed feature selection method and weighted kNN)				Prediction using proposed feature selection method and SVM, (Kernel =RBF, γ=200, C=100)			
	ACC	**MCC**	**ROC area**	**Pre-cision**	**ACC**	**MCC**	**ROC area**	**Pre-cision**
Amine	97.4	0.986	0.996	100	98.7	0.993	0.994	100
Cannabinoid	100	0.969	1.00	94.3	95.5	0.936	0.975	92.6
cAPA	70.8	0.734	0.962	77.3	58.3	0.648	0.789	73.7
Ecdysis triggering hormone receptor	72	0.845	0.987	100	56	0.617	0.777	70
Gonadotropin-releasing hormone	96.7	0.953	0.997	94.7	93.5	0.922	0.963	92.5
Hydroxycarboxylic acid receptor	97.7	0.976	1.00	97.7	95.3	0.94	0.975	93.2
Leukotriene B4 receptor	91.5	0.927	0.996	94.7	78	0.751	0.882	75.4
Lysosphingolipid and LPA	100	1.00	1.00	100	100	1.00	1.00	100
Nucleotide-like	96	0.96	0.997	96.9	93.9	0.943	0.967	95.9
Olfactory	100	1.00	1.00	100	100	1.00	1.00	100
Rhodopsin	95.1	0.974	0.998	100	91.8	0.929	0.957	94.9
Peptide	90.8	0.933	0.994	96.7	87.7	0.891	0.936	91.9
Platelet activating factor	98	0.922	0.997	87.5	96	0.911	0.976	87.3
Prostanoid	100	0.983	1.00	96.8	96.7	0.956	0.982	95.1
Thyrotropin-releasing hormone	94.4	0.899	0.994	87	84.5	0.814	0.915	81.1
Viral	95.7	0.945	0.997	93.8	93.6	0.903	0.965	88
Overall	**95.4**	**0.951**	**0.996**	**95.5**	91.5	0.909	0.955	91.4

6.4. Conclusion

The G-protein coupled receptors are the largest superfamilies of membrane proteins and important targets for the drug design. Here, eight feature vectors were used to represent the protein sample, including amino acid composition, dipeptide composition, correlation features, composition, transition, distribution, sequence order descriptors and pseudo amino acid composition. In this chapter, at first an optimal feature subset selection method is proposed which provides the non-redundant, relevant, and robust feature subset by taking UNION of best 50 features selected by various supervised feature selection methods such as Fisher score based feature selection, ReliefF, FCBF, MRMR and SVM-RFE feature selection methods. In next stage, we proposed to use a weighted k -nearest neighbor classifier to predict the G-protein coupled receptors and their subfamilies. Using the 10-fold cross validation the proposed method achieved an overall accuracy of 99.9%, 98.3%, 95.4%, MCC values of 1.00, 0.98, 0.95, ROC area values of 1.00, 0.998, 0.996 and precision of 99.9%, 98.3% and 95.5% to predict the G-protein coupled receptors and non-G-protein coupled receptors, subfamilies of G-protein coupled receptors and subfamilies of class A G-protein coupled receptors respectively. The high accuracies, MCC, ROC area values and precision values indicate that the proposed method is better for the prediction of G-protein coupled receptors and their subfamilies.