CHAPTER 5
RE-ENGINEERING OF ENSEMBLES FOR REAL WORLD DATASETS

5.1 INTRODUCTION

Ranking and selection measures are traditionally used to extract the key features from high-dimensional feature space. These techniques can be used to improve the accuracy of data categorization and feature reduction. Feature selection is a challenging and critical issue for microarray data due to noise, missing values and clustering. ABC, PSO, and GA are classic methods in data classification that feature selection are used. However, these models are very expensive and computationally intensive. Traditional data partitioning based classification models require feature selection measures and ranking approaches in order to improve the classification accuracy. However, traditional ensemble learning models require high computational time if the number of instances in the training and test dataset increases. In order to estimate test data instances, traditional approaches primarily focus on detecting relevant patterns from the trained data. The robust partition-based classifier was implemented in this proposed work in order to find the top anomalies using attribute relationships. Along with uncertain characteristics with a high true positive rate, this model effectively identifies the anomaly characteristics. Experimental results show that the approach proposed works well against traditional optimization techniques for anomaly optimization. In the current heterogeneous data repositories such as gene and disease identification and prediction are the essential factors for content clustering and classification models.

The chances of a cyber-attack on the high-speed networks are rising as the high-speed networks continue to grow. Anomaly detection is an important part of a number of valuable functionality. Since there is a communication between the client and the server, there is a probability of an error in the distributed processing of the large distributed data. The large amount of features make anomaly detection a demanding task. To identify anomaly patterns or characteristics using association and classification techniques, several anomaly detection techniques have been implemented in the literature. Unfortunately, all the normal or abnormal features cannot be covered by some anomaly detection models via data mining.

5.2 AUTOMATIC DATA FEATURE EXTRACTION

Automatic data feature extraction approaches are one way to help internet users to generate concise information efficiently and effectively. Effective feature extraction approach should properly assume the following requirements.

Diversity: Each medical data filtered information should contain minimal redundant phrases as possible.

Coverage: Each summary should have essential key features from all the medical data, i.e., information loss in feature extraction model can be minimized.

Balance: Each summary should represent the different key concepts of the large feature data in a uniform manner.

Added to this, the demands of genomic data classification for dimension reduction i.e. derivation of reduced feature set either by extraction of features or by selection methods and further classification using effective classifiers. The classification question divides the unseen test instances into classes as the class label specifies. Chi-Square is the common statistical test which measures the divergence from the expected distribution if the assumed feature occurrence is in fact independent of the class value. When the Chi-square Statistics is greater than the critical value defined by the degrees of freedom, then the function and class are deemed dependent. The T-test is another common statistical method used to compare a sample mean or mean difference to the true mean or predicted mean difference. T – test is yet another basic approach used in the study of gene expression. The logarithm of expression levels was manipulated, requiring comparison calculation to the mean variance of both treatment control groups. The basic issue with T-test is that it involves repeated controlled trials with care, which are both repetitive and expensive. Relief-F is a feature selection technique that selects instances at random, adjusting feature-relevance weights depending on the nearest neighbor. By its merits Relief-F is one of feature selection's most effective strategies. Relief-F key concept is to calculate a score for each characteristic that calculates how well this feature distinguishes neighboring samples in the original room. The nearest neighbor version searches for the closest example of the opposite class (nearest miss) in the original function space from the same class (nearest hit). The score is then the difference (or ratio) between the averages for all samples from the distance to the nearest miss the average distance on that function to the nearest hit in prediction [88]. In statistics, the selection of features also known as variable selection, is the method of selecting a subset of appropriate variables for

statistical model construction. Feature selection techniques are used under the central assumption that several redundant or inappropriate features are present in the data. Redundant apps are those that do not include any more detail than the currently selected apps.

5.3 TRADITIONAL BIOMEDICAL DATA CLASSIFICATION MODELS

In order to find the highest likelihood evaluation variance between the gene and its associated diseases in biomedical data sets, conventional probability assessment techniques like Naïve Bayes, the Markov model and the Bayesian model (68) are employed. Classification is a process of searches for and extracts from distributed medical data sources the key conceptual definitions of genes or disease trends, which has become an integral part of daily work in any field like cloud, internet, social networking and repositories.

Automatic text the Classification of the broad medical sets fulfills those objectives by the use of the Classification techniques at the user end. Summaries of medical data represent instances or phrases derived from various sources without any arbitrary human interference or editorial influence. The Classification area is highly interdisciplinary and includes fields such as knowledge collection, text mining and data collection, natural language processing and medical databases.

Meyer et.al [89] proposed a model mining pattern and disease prediction classification model. They used a microarray dataset to grade the pathways in a small data size and find disease-related trends. Random forest classification models were used to filter and classify co-related patterns of disease into microarray datasets. This model allows large data sets to have high calculation storage and memory.

Singh et.al [90] has introduced a new Meta graph construction model for gene-based disease patterns. Protein associations and the keywords of the biological gene are extracted in this model in order to classify patterns of a disease via a meta graphic model. Small datasets with maximum 10k instances. This model is limited.

Raweh et.al [91] suggested the use of the phase diagram method for new gene selection and disease classification. For disease prediction and classification, they used various microarray gene data sets. For microarray data sets with reduced instance space, PHADIA method is powerful.

J. A model for identification of gene sequence patterns in biomedical repositories has been developed by [92]. As training dataset, they used gene bank data to find the

relationship between the two sequences. Moreover, the secret markov model suggested generates the Genes sequence with the Gene Bank database in two or three states. This model only requires two gene sequences, with each sequence rising in size or the sequence number growing, this model cannot be ranking-efficient. All were employed as the Decision Tree and k neighbors, linear bays, smallest squares of vector supporters and radial networks. It is better to use decision trees when the outcomes are easy to grasp. It is rational and simple to understand by applying a series of rules in the classification that decision-making bodies be superior to other methods. The decision-making mechanisms include the nodes, branches and leaves. The attributes are evaluated in accordance with the specified node rules. At the branches, the results of the tests are shown. Finally, the class is specified on the leaves. The checking at the top leads to the blade, which decides the class. Out of the medical data from various health reports secret patterns can be retrieved. Sadly, these medical documents are distributed across the world and not evenly formatted. The doctors can even exaggerate some of the documents. The diagnosis is based on these trends achieved. The data must be obtained first and then cleaned up to obtain trends from the documents. Diagnosis is an incredibly critical medical practice. False diagnosis can have negative effects [93].

Bagging is another form of classification. In this case, the classifier will supply a category output. Any prediction is taken as a single vote. If a certain class earns the majority of votes, the results of the classification are considered. The packaging sums up the groups according to the number of votes. There are other classifications derived from enhancing this form of bagging. It's better done with trees. This is due to the simple understanding of the arrangement of the trees. Certain data can be unstructured or standardized. The best way to treat this sort of data is bagging. This is because this technique is successful. The used tree makes reading and depicting simpler. Unfortunately, the bagging technique does not work well for that type of data when the data is not noisy. Another method of classification is known as boosting. The exercise tuple is trained in the boosting process with weights. Then the tuples are repeatedly learned with the boost classifier. The weight of the tuples is recorded during each period of repetition. The correctly classified tuples are assigned less weight, but those which have been classified incorrectly are classified more. Since the tuple is given more weight, more attention is given. How correctly weights are allocated to the tuples is assessed to determine the output of this stimulation process. This boosting approach could be strengthened by depending on assumptions such that the large margins of error

would gradually be reduced to smaller margins. The AdaBoost.M1 and AdaBoost.M2 are two kinds of boosting methods. AdaBoost.M1 focuses on error reduction and therefore on the algorithms used for learning. In this case, it remains constant and not varied for the weights of the wrongly labeled tuples. In certain cases, this has been shown to lower the error to zero. This happens with a mistake of over 0.5 in the hypothesis. This drawback makes it unsuitable when the error in the hypothesis reaches 0.5. AdaBoost.M2 has the advantage that one of the labels with high probabilities is not considered, it focuses on labels which have been mistakenly identified. The minimized pseudo loss is important to achieve the exact and desired output. If more sophisticated algorithms can be used, large databases can efficiently use the boosting process. Using these complex algorithms reduces the classifier's errors.

The prediction process is very important in the decision-making process. This can be accomplished successfully using a statistical technique that is referred to as classifier or supervised pupil. The above prediction model is entirely liable to forecast the data of various applications that are invisible and unknown. The task of increasing predictive accuracy is very difficult and demanding. The accuracy of the classification depends entirely on the training data in the predictive model. Overall predictive accuracy is decreased by the unimportant, unrelated and redundant functions of training data. Therefore, to increase classification accuracy, it is a prudent decision to discard the above features. The predictive accuracy must be substantially improved in order to maximize the efficiency of the predictive model. After this, a reduction in building predictive model time may also increase performance.

5.4 ENSEMBLE BASED PARTITION CLASSIFIER MODELWITH LARGE NUMBER OF FEATURE SETS

Ensemble Model is the process by heterogeneous classifiers are combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the performance of a model, Ensemble modeling is the process of running two or more related but distinct l classification models and then synthesizing the results into a single score or spread to enhance the accuracy of applications for predictive analytics and data mining. Each training tuple has weights assigned. A series of k classifiers is learned iteratively. The weights are updated after a Mi classifier is learned to enable the subsequent classifier, Mi+1, to "pay more attention" to the training tuples misclassified by Mi. The final boosted classifier, M_, combines each individual

classifier's votes, where the weight of the vote of each classifier is a function of its precision. Due to the rapid growth of the high-speed network, the risk of cyber attacks on complex networks is also increased accordingly. Database anomaly discovery is a process of filtering uncertain features to allow a wide range of applications to be used.

Proposed Ensemble Model

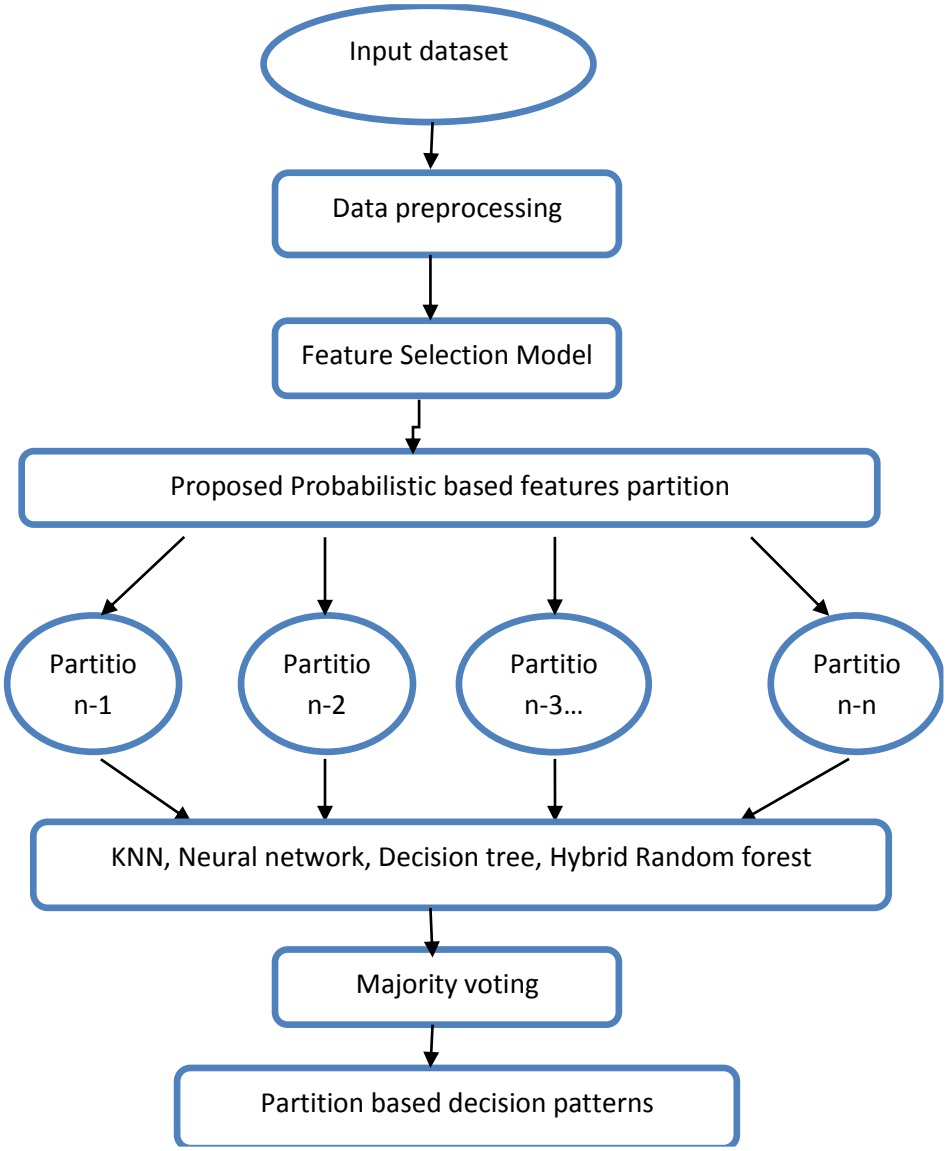


Figure 5.1: Proposed Partitioning based classification framework on high dimensional datasets

In the ensemble classification problem K-Nearest Neighbor (KNN), Neural Network and Hybrid Decision Tree approach are integrated to improve the classification accuracy on each partition. In the decision tree construction model, a novel attribute selection measure is proposed in order to improve the tree pruning on each partition for the majority voting process. Finally, majority voting is used to find the statistical analysis of the classification problem.

Since the online distributed data is the communication between the client and the server, it is difficult to predict the occurrence of an anomaly in the large distributed data. Anomaly detection on the complex data must take a long time due to the large number of features. Several anomaly detection techniques have been implemented in the literature to find the anomaly patterns or features using association and classification techniques. Unfortunately, some anomaly detection models over data mining cannot cover all the normal or abnormal features. Traditional approaches mainly focus on detecting relevant patterns from the trained data in order to estimate the test data instances. In this proposed work, the robust partition based classifier was implemented to find the topmost anomalies using attribute relationships. This model efficiently detects the anomaly features along with uncertain features with high true positive rate.

Figure, shows the block diagram of the proposed ensemble anomaly prediction model. It has three phases: 1) Data preprocessing 2) Feature selection process and 3) Classification model.

5.4.1 Algorithm:

Step 1: Input file (Filtered anomaly data)

Step 2: Preprocess anomaly data for missing values.

Step 3: Data transformation for unequal distribution as

For each attribute A_i in DB

Do

If(A_i.type==numerical)

Then

$$A_i.\text{value} = \frac{A_i.value + G.\text{M}(A_i)}{(A_i.value.Max - A_i.value.Min)} * (ScaleMax - ScaleMin)$$

End if

Else

Continue;

Step 4: For each randomized sample S_i

Do

$$Sim(S_i, S_j) = 0; if \text{ i=j}$$

$$Sim(S_i, S_j) = (1/2*\pi) * \sum (x_i - x_j)^2 * e^{-(|x_i - x_j|)^2/2*\sigma_x}; if \text{ i=j}$$

$$\sigma_x : S\tan dard \text{ deviation}$$

If($Sim(S_i, S_j) > 0$)

Then

$S' = add(\mathrm{Sim}(S_i, S_j))$;

Else

Continue;

End for

Step 5: Apply DT[1] and then use the partitioning approach using the probabilistic measure as

For each partition S_1' in S'

Do

Compute prior probability as

$$\mathrm{PProb} = \underset{i=1\ldots m}{\mathrm{argmax}} \frac{P(\mathrm{c} = \mathrm{c}_i) \prod_{j=1\ldots n} P(S_j(\mathrm{j}) / \mathrm{c}_i)}{P(\mathrm{A}_i = a_1, a_2 \ldots . a_k)}$$

Divide the partition into 2 classes as yes and no or true or false.

Let P and N are the two sample instances with positive and negative classes.

$P = \{x_1, x_2 \ldots . x_{N_1}\}$

$N = \{\mathrm{y}_1, \mathrm{y}_2 \ldots . \mathrm{y}_{N_2}\}$

True information entropy is computed as

E(P)=E($\{x_1, x_2 \ldots . x_{N_1}\}$)

Prior entropy function of true samples is given as

$$\mathrm{TProb} = \frac{-\sum P(x_i / \mathrm{c}_m)}{\sum P(x_i / c_m) + \sum P(y_j / c_m)} \{\log(\frac{\sum P(x_i / \mathrm{c}_m)}{\sum P(x_i / c_m) + \sum P(y_j / c_m)})\}$$

$$\mathrm{FProb} = \frac{-\sum P(\mathrm{y}_i / \mathrm{c}_m)}{\sum P(x_i / c_m) + \sum P(y_j / c_m)} \{\log(\frac{\sum P(\mathrm{y}_i / \mathrm{c}_m)}{\sum P(x_i / c_m) + \sum P(y_j / c_m)})\}$$

If(TProb<FProb)

Then

Display as anomaly

Else

Display as Normal

In the above ensemble based anomaly detection model, each attribute is checked against the data distribution. If the attribute is not uniform distributed then it was transformed to uniform format. For each attribute in the uniform distributed dataset,

instances are partitioned into set of sub partitions based on classes. After that, similarity computation was applied on the sub partitions to find the relevant relational anomaly features. True probability and False probability measures are used to find the high anomaly patterns in each partition.

5.5 Experimental Results

Results on Medical Datasets

In this section, we have implemented our proposed cancer micro-array model and compared the results with conventional decision tree modeling. It summarizes the microarray data set used for experimental evaluation. 10% of the training data are used as test data for performance evaluation in the experimental results.

Ensemble methods allow more accurate prediction of true positives of high dimensional datasets. The proposed model uses the entire training data for construction of decision patterns, therefore its prediction accuracy tends to be more accurate than the traditional ensemble models.

Table 5.1: Datasets and its properties

Dataset Name	Features	Type
lung-Michigan	7000	Numeric
DLBCL-Stanford	4000	Numeric
Leukemia	6817	Numeric
lungCancer_train	12000	Numeric
Lymphoma	4026	Numeric

Ovarian Cancer dataset:

Ovarian cancer dataset is one of the high dimensional continuous attributes dataset. In this data a large number of feature space with limited number of instances are taken from the website(http://csse.szu.edu.cn/staff/zhuzx/Datasets.html). This dataset contains 15154 attributes and 253 instances for classification process.

Table 5.2 : Comparative study of present model to the traditional models on four microarray datasets by using accuracy measure

Datasets	PSO + SVM	PSO + FFNN	Statistical + Ensemble	FS + Ensemble	Proposed Model
Lung Cancer	0.8353	0.8525	0.9535	96.03	96.45
Lung Michigan	0.8353	0.845	0.9635	0.9903	0.993
Ovarian	0.8744	0.8574	0.97	0.984	0.991
Lymphoma	0.8835	0.8934	0.982	0.9845	0.989
Error Rate	0.3042	0.294	0.174	0.201	0.18
Runtime (ms)	6194	6364	3068	3465	3329

Table 5.3: Generated Patterns on leukemia dataset

```
32923_r_at < 4860.5
|  32000_g_at < 910
|  |  32279_at < 329.5
|  |  |  35794_at <399.5 : MLL (4/0)
|  |  |  35794_at >= 399.5
|  |  |  |  37403_at <1530 : ALL (6/0)
|  |  |  |  37403_at >= 1530 : MLL (1/0)
|  |  32279_at >= 329.5
|  |  |  39579_at < 75
|  |  |  |  38938_at < -5
|  |  |  |  |  32373_at < -215 : AML (1/0)
|  |  |  |  |  32373_at >= -215
|  |  |  |  |  |  36407_at < -442.5
|  |  |  |  |  |  |  35504_at < -1007.5 : ALL (1/0)
|  |  |  |  |  |  |  35504_at >= -1007.5 : MLL (5/0)
|  |  |  |  |  |  36407_at >= -442.5 : ALL (5/0)
|  |  |  |  38938_at >= -5
|  |  |  |  |  35544_at < 519.5
|  |  |  |  |  |  32437_at <34024.5 : ALL (6/0)
|  |  |  |  |  |  32437_at >= 34024.5 : MLL (1/0)
|  |  |  |  |  35544_at >= 519.5 : MLL (4/0)
|  |  |  39579_at >= 75 : AML (4/0)
|  32000_g_at >= 910
|  |  32473_at < -1367 : AML (1/0)
|  |  32473_at >= -1367
```

```
|  |  |  31423_at < -374.5 : AML (1/0)
|  |  |  31423_at >= -374.5
|  |  |  |  34560_at < 752.5
|  |  |  |  |  33089_s_at <198 : AML (1/0)
|  |  |  |  |  33089_s_at >= 198
|  |  |  |  |  |  34570_at <17171.5 : AML (1/0)
|  |  |  |  |  |  34570_at >= 17171.5
|  |  |  |  |  |  |  32000_g_at < 1662.5
|  |  |  |  |  |  |  |  36446_s_at <2823 : AML (1/0)
|  |  |  |  |  |  |  |  36446_s_at >= 2823
|  |  |  |  |  |  |  |  |  38417_at <4880.5 : ALL (6/0)
|  |  |  |  |  |  |  |  |  38417_at >= 4880.5 : MLL (5/0)
|  |  |  |  |  |  |  32000_g_at >= 1662.5 : AML (2/0)
|  |  |  |  34560_at >= 752.5 : AML (1/0)
32923_r_at >= 4860.5 : AML (9/0)
```

Size of the tree: 41

Table 5.4: Detailed Accuracy by Class wise on Cancer dataset

True Positive Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.167	0.000	0.000	0.000	0.000	?	?	ALL
0.000	0.000	0.000	0.000	0.000	0.000	?	?	MLL
0.833	0.000	1.000	0.833	0.909	0.000	?	1.000	AML
Weighted Average:0.833	0.000	1.000	0.833	0.909	0.000	0.000	1.000	

The robust partition based classifier gives the topmost anomalies using attribute relationships. This technique very reliably detects the anomalous features with high TRP. ensemble methods improve the effectiveness and usefulness of whole large data sets. The proposed model uses the entire training data for construction of decision patterns, therefore its prediction accuracy tends to be more accurate than the traditional ensemble models.

Table 5.5: Comparative study of present classification model to the conventional models for true positive rate and precision rate.

Average TP and precision		
Model	**True Positive Rate**	**Precision**
PSO + Ensemble	0.8144	0.8355
GPSO + Ensemble	0.9253	0.9083
Fuzzy PSO + Ensemble	0.9185	0.9034
IPSO + Ensemble	0.9674	0.9586
IABC + Ensemble	0.9725	0.9784
Proposed Approach	0.997	0.9898

The performance of the proposed model on all cancer datasets is described in Table 5.5. Here, all the cancer datasets are evaluated to find the average true positive rate and accuracy rate on the high dimensional datasets using the proposed model. It is visualized from the table that the model proposed has a high true positive rate and accuracy over the existing models.

Table 5.6: Average runtime(ms) of all training datasets.

Model	**Error Rate**	**Runtime(ms)**
PSO + Ensemble	0.1956	6963
GPSO + Ensemble	0.0895	6835
Fuzzy PSO + Ensemble	0.0815	6946
IPSO + Ensemble	0.0392	5793
IABC + Ensemble	0.0253	5195
IABC + IPSO + Ensemble	0.0153	4976

The comparison of the mean average runtime of all training datasets on the cancer datatypes as shown in Table 5.4. From the table, it is noted that, relative to existing models, the proposed model has a low error rate and runtime.

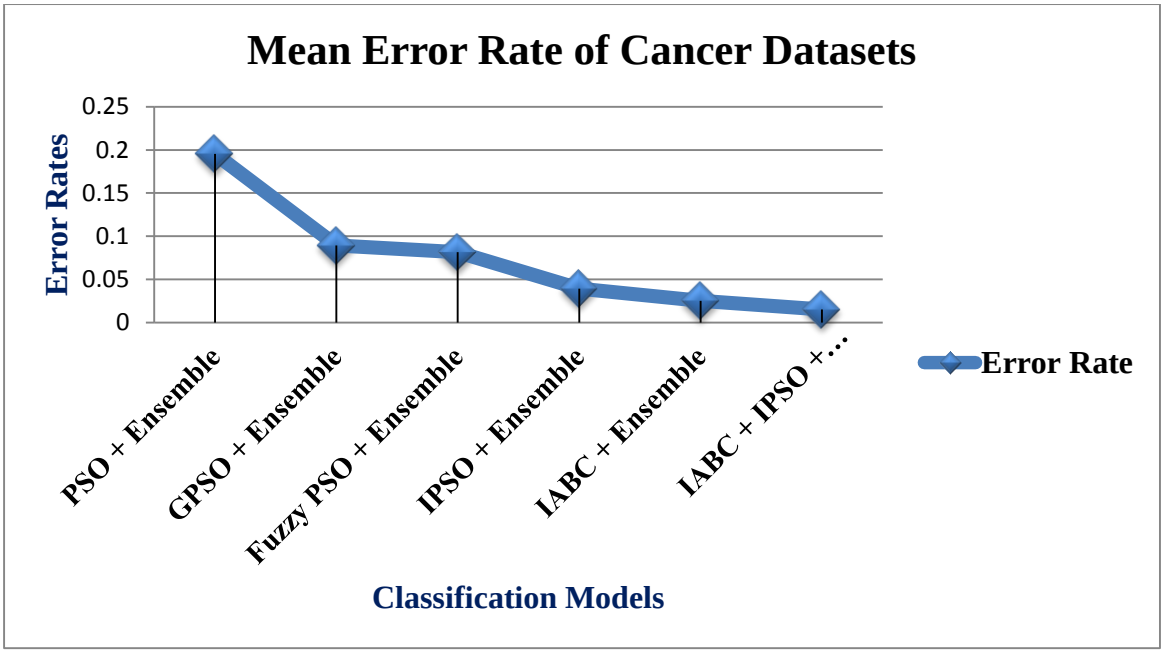


Figure 5.2: Mean average error rate of proposed model to conventional models on cancer datasets.

Figure 5.2 shows the average error rate of all microarray cancer datasets with all space features. It is clear from the figure that the proposed model has a low error rate compared to the existing models.

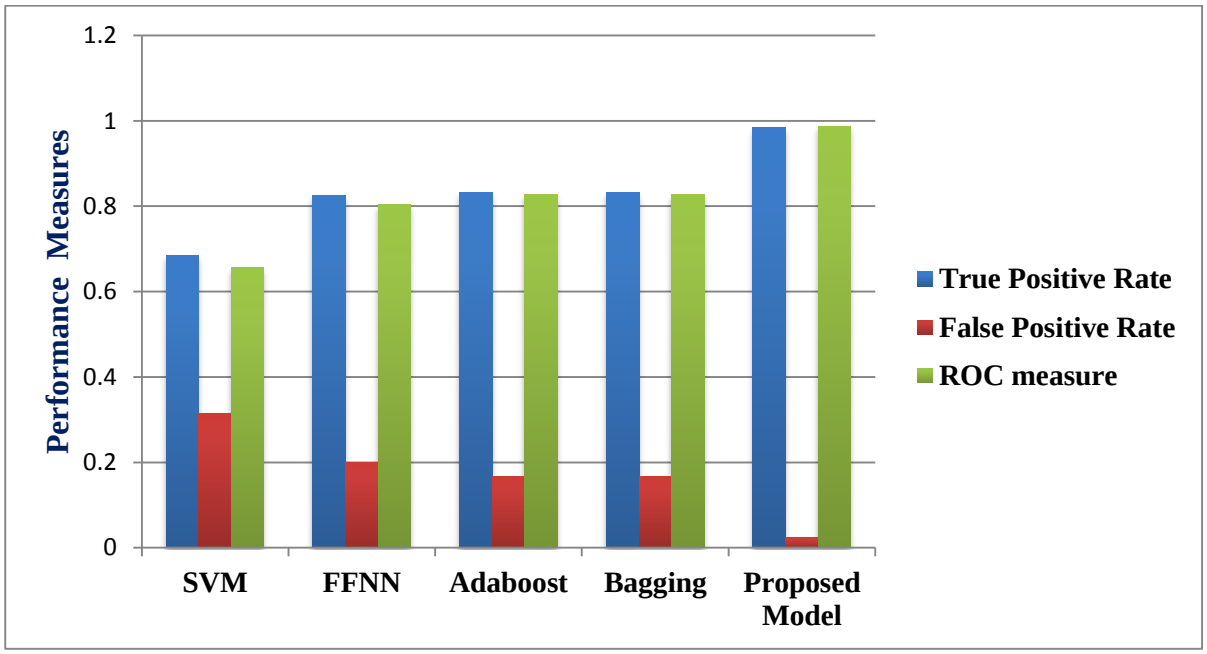


Figure 5.3: Comparison of Present approach to the conventional approaches on microarray disease dataset

Figure 5.3 describes the statistical comparison of the proposed model with the traditional disease prediction classification models. From the table, it is noted that true positive, false positive, F-measurement and accuracy improve on average by 5-10 per

cent when pre-processed with the proposed feature selection and ensemble classification model.

Table 5.7: Comparison of the proposed classifier to conventional classifiers for gene relationships, average classification rate and runtime (ms).

Model	Gene-Disease prediction rate	Classification Rate	Average Runtime (ms)
SVM	67.98	0.753	9184
Naïve Bayes	71.54	0.843	8744
Neural Networks	85.35	0.873	7925
PSO + NN	86.98	0.893	7253
Weighted Probability	95.17	0.974	6204
Proposed Model	98.91	0.9893	5794

Table 5.7 describes the comparison of the proposed model with the existing models in terms of gene prediction, rate of classification and duration of pattern detection. From the table, the proposed model has a high rate of genetic disease prediction, rate of classification and runtime(ms).

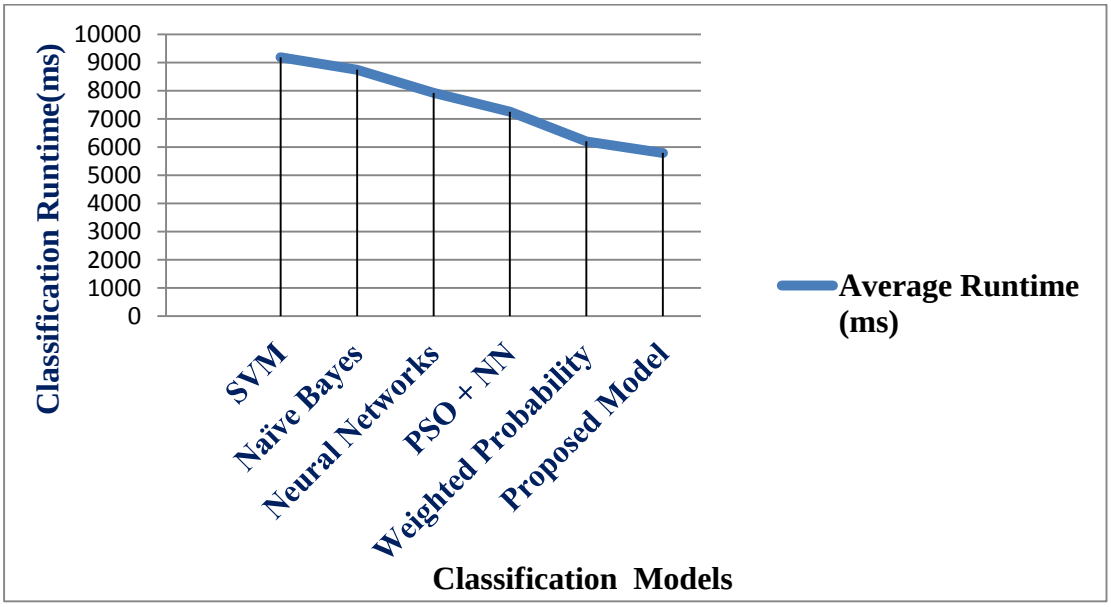


Figure 5.4: Performance of the proposed probabilistic gene-disease classification model on different distributed medical data sets

Figure 5.4 represent the average runtime graphical comparison of the proposed model to the traditional models on distributed biomedical repositories.

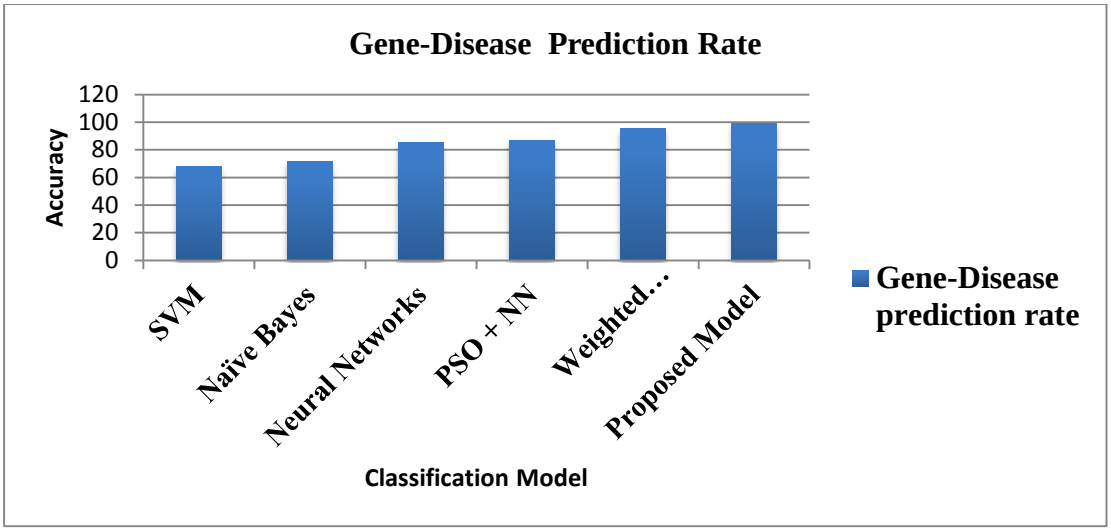


Figure 5.5: Graphical comparison of the gene to disease prediction rate for various prediction algorithms

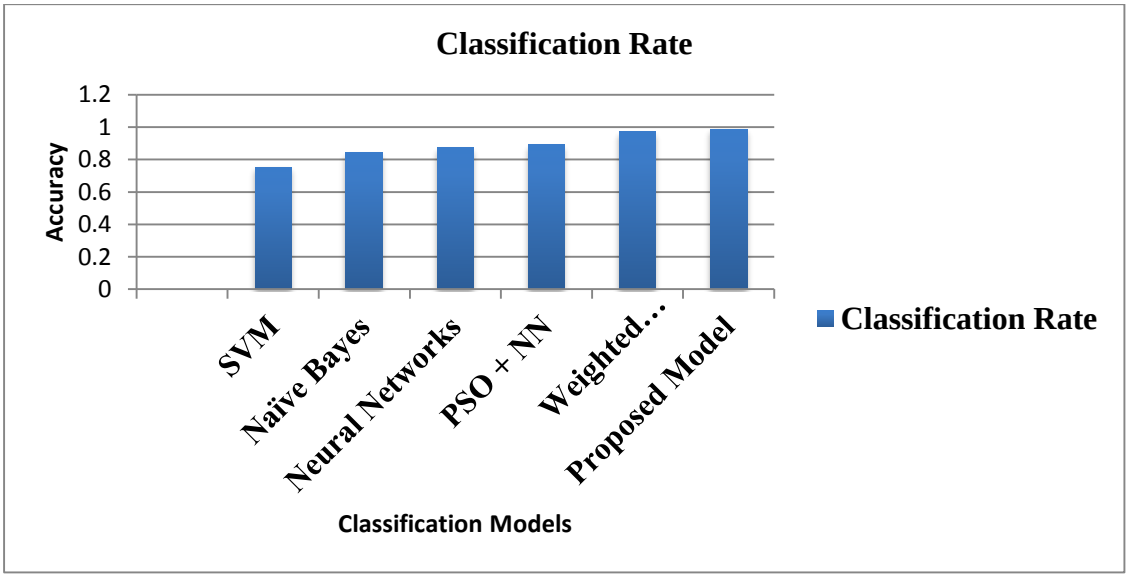


Figure 5.6: Graphical comparison of the classification rate on various classification algorithms

Results on KDD network intrusion dataset

In Network (KDD) dataset the classification rate increased to 98.7%, which is better accuracy than the previous models (94%), In credit dataset the classification rate increased to 98.5%, which is better accuracy than the previous models (82%) In Diabetes dataset the classification rate increased to 98.7%, which is better accuracy than the previous models (74.5%)

The results are shown below

5.5.1 Anomaly Patterns

employment != >=7 ->foreign_worker != no

personal_status != female single -> housing != for free

job != high qualif/self emp/mgmt AND own_telephone != yes ->credit_usage>= 4.0

purpose != other ->personal_status != female single

foreign_worker != no ->property_magnitude != no known property

personal_status != female single ->other_parties != guarantor

other_payment_plans != none -> purpose != other

over_draft != no checking AND housing != for free AND personal_status != male mar/widANDresidence_since>= 1.0 ->foreign_worker != no

num_dependents<= 2.0 -> housing != for free

purpose != other -> location <= 4.0

employment != >=7 -> purpose != other

housing != for free ->property_magnitude != no known property

property_magnitude != no known property AND housing != for free AND own_telephone != yes AND residence_since>= 1.0 ->personal_status != male mar/wid

personal_status != male mar/wid AND purpose != other AND other_payment_plans != none AND own_telephone != yes ->other_parties != guarantor

other_parties != guarantor ->personal_status != female single

other_payment_plans != none AND own_telephone != yes ->over_draft != no checking

property_magnitude != no known property ->other_payment_plans != none

property_magnitude != no known property AND personal_status != male mar/wid -> class != bad

other_payment_plans != none -> employment != >=7

property_magnitude != no known property ->num_dependents>= 1.0

other_parties != guarantor AND personal_status != male mar/wid -> class != bad

other_parties != guarantor -> job != high qualif/self emp/mgmt

own_telephone != yes ->other_parties != guarantor

other_payment_plans != none ->foreign_worker != no

foreign_worker != no -> class != bad

foreign_worker != no AND class != bad ->other_payment_plans != none

personal_status != male mar/wid ->foreign_worker != no

housing != for free -> class != bad

own_telephone != yes ->other_payment_plans != none

purpose != other ->other_parties != guarantor

own_telephone != yes ->other_parties != guarantor

other_parties != guarantor -> employment != >=7

foreign_worker != no AND own_telephone != yes ->other_parties != guarantor

own_telephone != yes AND personal_status != female single ->other_payment_plans != none

foreign_worker != no ->other_payment_plans != none

own_telephone != yes -> housing != for free

class != bad AND purpose != other AND other_payment_plans != none AND num_dependents>= 1.0 ->other_parties != guarantor

class != bad ->residence_since<= 4.0

current_balance<= 18424.0 AND housing != for free ->other_payment_plans != none

other_payment_plans != none ->foreign_worker != no

own_telephone != yes ->residence_since<= 4.0

other_payment_plans != none -> class != bad

own_telephone != yes AND class != bad -> purpose != other

class != bad ->other_parties != guarantor

class != bad AND own_telephone != yes ->other_parties != guarantor

purpose != other ->current_balance<= 18424.0

other_parties != guarantor ->over_draft != no checking

job != high qualif/self emp/mgmt AND num_dependents>= 1.0 ->other_payment_plans != none

own_telephone != yes AND class != bad ->other_payment_plans != none

class != bad AND housing != for free AND own_telephone != yes AND residence_since>= 1.0 ->personal_status != male mar/wid

personal_status != female single ->over_draft != no checking

employment != >=7 ->num_dependents>= 1.0

housing != for free ->other_parties != guarantor

housing != for free ->residence_since<= 4.0

Elapsed time: 11.689s

Classification Accuracy 0.97715

Number of Iterations :9

F-Measure: 0.87324

Recall : 0.85359

TP rate : 0.96992

FP rate : 0.030079999999999996

Classification Accuracy 0.98503

Results on Credit Dataset

Table 5.8: Comparative results for existing and proposed models for Credit Dataset

Datasize	Patterns	Existing Model	Proposed Runtime(ms)
500	25	3255	2467
1000	35	4234	3177
1500	58	5634	3277
2000	71	6267	4275
2500	157	9555	6335

In the above table, as the data size increases proposed model has low runtime compare to traditional model.

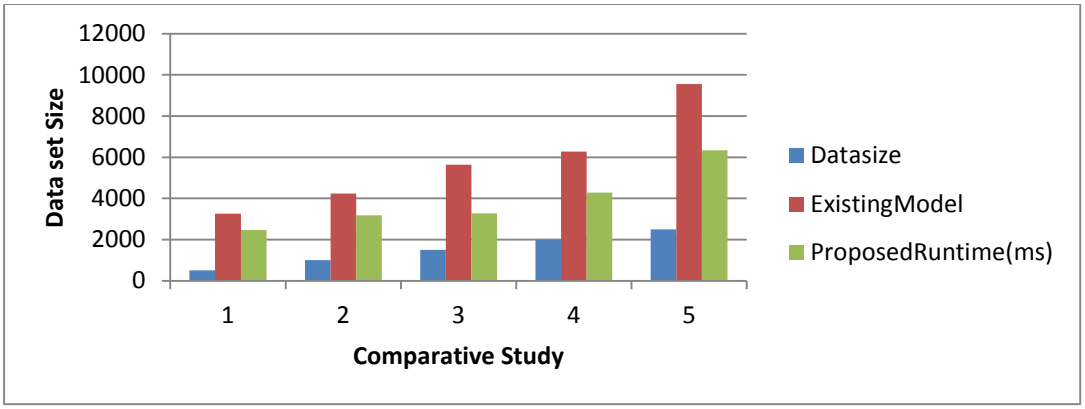


Figure 5.7: Comparative runtime study of present model to conventional models for high dimensional datasets.

In the above figure, as the data size increases proposed model has low runtime compare to traditional model.

Table 5.9: Comparison between true positive and false negative rate on credit card dataset.

Data size	Tprate	Fprate
500	0.9755	0.042
1000	0.9844	0.0156
1500	0.9854	0.175
2000	0.9785	0.098
2500	0.988	0.154

In the above table, as the data size increases proposed model has high true positive rate compare to traditional ensemble model. Also, as the data size increases, proposed model has low false positive rate compared to traditional model.

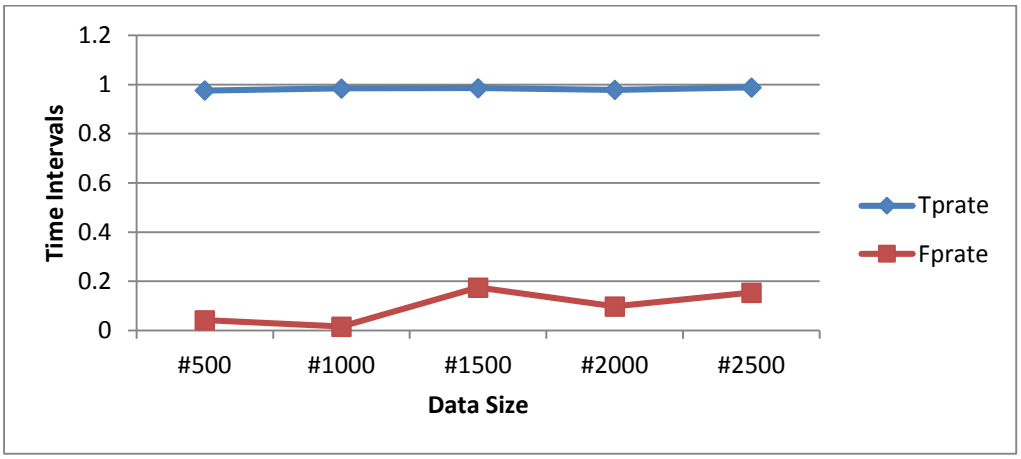


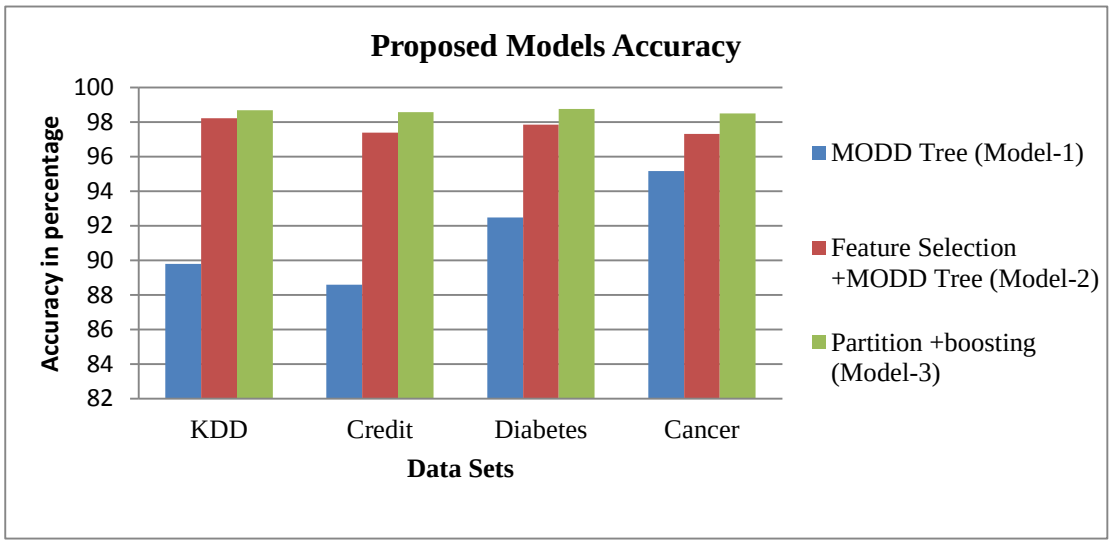
Figure 5.8: High true positive rate and low False positive rate

In the above figure, as the data size increases proposed model has high true positive rate compare to traditional ensemble model. Also, as the data size increases, proposed model has low false positive rate compared to traditional model.

Table 5.10: Proposed Models Accuracy Comparison of Datasets using three proposed models

Data set	KDD	Credit	Diabetes	Cancer
MODD Tree (Model-1)	89.8	88.6	92.48	96.34
Feature Selection +MODD Tree (Model-2)	98.2	97.39	97.86	97.93
Partition+boosting (Model-3)	98.7	98.58	98.76	98.81

Figure 5.9: Performance analysis of the Proposed Models Accuracy Comparison of Datasets



Proposed methods for the ensemble improve the effectiveness and accuracy of true positive values on large data sets as a whole. The proposed model uses the entire training data set to build decision patterns, so the accuracy of the prediction of each cross validation tends to be more precise than the traditional models of ensemble classification.

5.6 SUMMARY

This feature selection-based ensemble model is developed for high dimensional microarray dataset. Most of the traditional transformation approaches have data adjustment methods independent of data distribution and outliers. Under the traditional models, the features are chosen in a static way or limited number of reasons. Feature selection algorithm is used on different classification models to improve the true positive rate and error rate. In Network (KDD) dataset the classification rate increased to 98.7%, which is better accuracy than the previous models (94%),In credit dataset the classification rate increased to 98.5%, which is better accuracy than the previous models(82%) In Diabetes dataset the classification rate increased to 98.7%, which is better accuracy than the previous models (74.5%),This model gives the better discovery of diabetes, right sort of glass, better recognizing of network attacks, cancer decease and credit hazards in the credit dataset. Experimental results are simulated on different microarray data and it turns out that the hybrid feature selection ensemble approach has high accuracy compared to the traditional models.