CHAPTER 3
PROBABILISTIC FEATURE ENGINEERING

3.1 INTRODUCTION

Regular data sets are increasing in scale. Therefore, it is becoming more difficult to find important characteristics in the wide area due to data volumes and sparsity. The ranking and classification of microarrays is one of the biggest challenges for scientific and biomedical scientists, due to its wide size range and small sample numbers. There are several similar DNA molecules used to classify the gene-related disease in each microarray. The key strategies used for classifying high dimensional data with a high positive rate are feature transformation, feature ranking and data classification. The method of normalization of feature values is the transformation feature within the restricted range. The transformation of features will help enhance the ranking process in the space of high dimensions. The majority of standard methods such as log transformation, min-max normalization etc. for feature transformation are independent from data distribution and outliers.

Most statistical anomaly detection models are used in the verification process with comparatively limited data size and depth. Statistical analyses are the commonly used approaches to outlier detection. The goal is to identify those objects that generate those that do not. Given a certain statistical distribution, algorithms calculate the random values assuming all data points are generated by the specified distribution (mean and standard deviation). Extremely low-probability events are outliers. These statistical anomaly detection models are difficult to find the uncertain instances because of irregular variations in the input distribution. To overcome these issues, a hybrid statistical anomaly detection method is devised on the training datasets, without limitations imposed by the data used.

3.2 STATISTICAL FEATURE EXTRACTION IN HIGH DIMENSIONAL DATA

Internet and technology are progressing and large quantity of text data is more easily processed, stored and exchanged. In the field of biomedical applications, a lot of work has been done to greatly improve the text mining process. Medical data clustering is often appropriate among many unsupervised techniques. The medical data clustering method can be used in many future applications. In order to achieve a perfect and streamlined algorithm, several approaches to medical data classification have been introduced [64]. This algorithm works in two phases:

1. Extraction of Feature.
2. Data Classification by feature extraction results.

When genes are being screened, they are filtered for minimum redundancy maximum validity (MRMR). To test these classifiers, NB and SVM have been used to generate the gene subsets in phase 2. Feature Selection is a process of selecting the optimal features based on certain criteria. The measuring attribute subsets are specified in the corresponding subsets. The criterion will be selected according to the purpose of selecting the features. An optimal subset can be a minimum subset. This can provide the most accurate estimate for predictions within a subset. In some cases a subset with the appropriate number that meets the criterion (number of features) can be found. Choosing a subset of M features means that data has been carved in the hypothesis space so that a learning algorithm is easier to learn. Another tool to reduce superfluous information is a feature reduction tool called Rough Sets Attribute Reduction (RSAR). The primary selection criterion for information processing sets is to have a minimum reduction and high grade in the feature selection. Additionally, you can use evaluation criteria to judge the rules. The relevance and redundancy are assessed. Predictive variables are common. The reduction in dimension is achieved by constructing several main variables that explain the main category of variation in the predictor variables. Variance is useful when it is large, and less useful otherwise. orthogonal linear combinations are designed to maximize means of explanatory variables. The approach eliminates uncorrelated variables from the data. Then they have added other additional components to the compound.

To study the performance of the classification models, Wilcoxon tests are combined with various methods of feature selection. An investigation is being carried out to see whether the feature of several function selection methods has any effect on the performance of classification. Here, Symmetric Uncertainty, RELIEFF, Random forest and SVM performed a separate function ranking as functional selectors and the ensemble were compiled by aggregating the single rankings using linear aggregation by weighted voting. PCA has the basic principle that a specific dataset, which has many dependent variables, can decrease its dimension. These variables are interconnected (generating from unique datasets). In terms of feature representation and reconstruction, PCA's algorithm also achieves an optimal result. The main aim of the LDA algorithm is to optimally discriminate vectors by increasing the distance ratio between classes and intra classes. This algorithm is responsible in a particular dataset for finding the optimal class discrimination. The main problems are:

1. The overall performance of the algorithm is significantly reduced due to its poor availability of dimension space.

2. In addition, there is another problem of uniqueness.

3.3 STATISTICAL CONTROL CHART MODEL

In this analysis initially statistical control model is used to remove extreme outliers. Later the filtered data set is classified more accurately using modified decision tree algorithm with modified entropy and modified information gain.

In statistical control model Mean, Standard deviation, upper control limit and Lower control limit for each and every attribute in the dataset are computed. Based on these parameters extreme outliers have been identified. The identified outliers show considerable improvement in accuracy. In this proposed methodology a statistical control chart algorithm is used in order to find the anomalies in different continuous datasets. We proposed dynamic 3 sigma based control chart to detect anomalies in the dataset. Essential stream design of the proposed calculation is appeared beneath.

3.3.1 Statistical Control Chart Algorithm's Basic Flow Structure

Algorithm:

Input: Continuous data set, Output: filtered data set.

Procedure:

1: Continuous attribute loading dataset.

2: As a real attribute, test each attribute in the dataset or not.

3: Calculate each attribute's mean and standard deviation.

4: Calculate each attribute's upper control limit (2)

5: Calculate each attribute's lower limit (3).

6: Calculate a lower control limit for each attribute (1)

7 Test to see if each item in the dataset falls under three categories, i.e. lower, upper or central.

8: If the object is out of bound, it will be excluded from the instances of the dataset.

9: Repeat this process until all data points have been completed.

10: Finally, the data set is stored in the file without outliers.

Lower Control limit: $\mu_X - \lambda\sigma_X$ --(1)

Upper Control Limit: $\mu_X + \lambda\sigma_X$ ---(2)

Control Limit: μ_X

Where λ (lambda) is Dynamic parameter, μ-Mean of the attribute and σ-is the standard deviation of the attribute

3.3.2 Modified Decision Tree Model

The filtered data set obtained from the above is analyzed to classify the dataset into normal and anomaly candidates using modified decision tree algorithm incorporating modified entropy and modified information gain. Decision trees are used for classification and they can quickly produce the rules of classification. According to its attribute values, an undefined tuple X can be identified by crossing the decision tree. Decision tree classifiers generally have good accuracy, but it is dependent on the data at hand for successful use. The performance of these trees was compared on the basis of the accuracy of classification. This method calculates the gain of information for each attribute A as the difference between the information needed to classify the dataset based on just the proportion and the information needed to classify it after partitioning it on A. C4.5 is an enhanced version of ID3 decision-making bodies based on the training data using the entropy principle. The training data is a set s1,s2,s3 .... represents data objects in the dataset S. Each si = xl,x2,... is a sample values where xl,x2,... represents features or attributes of the sample. The training data associated with a vector C = cl,c2,... where cl,c2,...cn represents the class to which each sample belongs to dataset. Every node of the decision tree C4.5 selects a data attribute that most efficiently divides the sample range into sub-sets in a single class. The attribute with the highest calculated information gain is selected to get the decision attribute. Its criterion is to improve the data gain by selecting an attribute to split the data. Our modified decision tree algorithm employs Modified entropy and Modified Information gain to improve accuracy in anomaly detection

Modified Decision tree Algorithm

```
1. Read non-anomaly filtered data.
2. If the attributes are null or empty
3. Return node as null
4. Else
```

5. To each attribute in the attributes list

6. Compute attribute selection measure for tree pruning.

7. To each attribute, compute the entropy measure for attribute selection process in decision tree construction.

8. Select the attribute with highest modified entropy measure for root node selection.

9. Repeat steps 6-8 to the remaining attributes to construct the rest of the tree in the decision tree construction.

Modified Entropy is calculated as

ModInfo(D) or Entropy $= -S_i \sum_{i=1}^{m} l\,\mathrm{og}\sqrt{S_i}$

where 'm' indicates number of different classes

ModInfo(D)= $-S_i \sum_{i=1}^{2} l\,\mathrm{og}\sqrt{S_i} = -S_1 \log\sqrt{S_1} + S_2 \log\sqrt{S_2}$

Where S_1 indicates set of samples which belongs to target class 'anomaly', S_2 indicates set of samples which belongs to target class 'normal'.

Information or Entropy to each attribute is calculated using

$$Info_A(D) = \sum_{i=1}^{v} |D_i| / |D| \times ModInfo(D_i)$$

The term Di /D acts as the weight of the jth partition. ModInfo(D) is the expected information required to classify a tuple from D based on the partitioning by A.

3.4 EXPERIMENTAL RESULTS

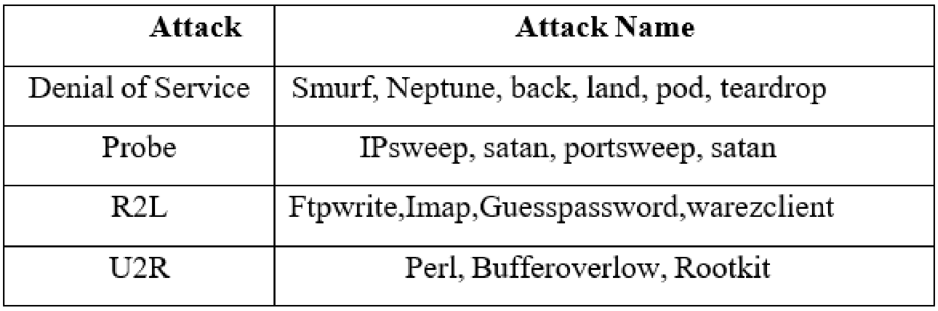
In this section, we have implemented our proposed cancer micro-array model and compared the results with conventional decision tree modeling. It summarizes the Microarray data set used for experimental evaluation. 10% of the training data are used as test data for performance evaluation in the experimental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

KDD99dataset: MIT Lincoln Laboratory conducted a reckon for evaluating intrusion detection tools in 1998-99. This research work was supported by DARPA, an intrusion detection community for further attacks evaluation. This KDDCup'99 data is most commonly used public intrusion detection dataset. The KDD99 data consists of both labeled and unlabeled instances.

In 1999, DARPA was accepted and approved as the traditional KDD Cup99 IDSbenchmark, which is available in http://www.kdd.ics.uci.edu/databases/kddcup99. In KDD99 data, different values of the attributes of labelled attack names are associated with each instance. The 5 types are classified as normal, slide attack, DOS attack, U2R attack and R2L attack. This class labels are classified into 5 types. A full KDD99 dataset includes close to 4 million labeled instances with 41 characteristics divided into 22 different kinds of attacks and summed up in Table 3.1.

Table 3.1 : Details of labeled attacks



Attack	Attack Name
Denial of Service	Smurf, Neptune, back, land, pod, teardrop
Probe	IPsweep, satan, portsweep, satan
R2L	Ftpwrite,Imap,Guesspassword,warezclient
U2R	Perl, Bufferoverlow, Rootkit

The logs were divided into 5 categories.

Denial of Service Attack (DoS): Network connections that try to deny legitimate users from accessing the network services in the victim’s targets system. Total number of instances of this category are 229853 records.

User to Root Attack (U2R): Network connections in which the victim’s system is already invaded, but the intruder tries to gain access with root privileges. Total number of instances of this category are 230 records.

Probe (Scanning): An intruder tries of in computing resources of the target system. Total number of instances of this category are 4166 records.

R2L (Root to Local): Network connections in which the intruder tries to get unauthorized access into a victim’s system. Total number of instances of this category are16187 records.

Normal: Network connections that fit the expected behavior in the military network.

Since the complete KDD99 data is too large to process, to minimize the computational time and memory for intrusion detection, we selected a subset of this data set, i.e. about 20% of the KDD99 data taken randomly from the original instances.

This, resulted 20% KDD99 dataset, is used to train the classification algorithm for intrusion detection system.

Table3.2 Description of IDS features

Feature Name	*Variable type	Description
Duration(F1)	C	Connection Duration
Type_protocol(F2)	D	Protocol Type (e.g., HTTP, TCP, etc.)
Packet_flag(F3)	D	Normal status or error status of the network connection
Network_service(F4)	D	Network service on the destination host connection,
source_bytes(F5)	C	Number of bytes of data transferred from source host to destination host
wrong_fragment(F6)	C	Count of ‘wrong’ fragments
destination_bytes(F7)	C	Number of bytes data transferred from destination host
Hot(F8)	C	Access count to system
Land(F9)	D	1: connection is to / from the same target system or port
urgent(F10)	C	Count of urgent packets
logged_in(F11)	D	1: success attempt
num_failed_logins(F12)	C	Count of failed logins
root_shell(F13)	C	1: root shell logged
number_compromised(F14)	C	Count of “compromised”
number_root(F16)	C	Number of super users
number_access_files(F17)	C	Count of access files

number_outbound_cmds (F18)	C	Count of out bound Commands in a tcp or http
number_shells(F19)	C	Total count of shell attempts
number_file_creations(F20)	C	Count of file creations in the
Ishotlogin(F21)	D	1 - the login belongs to the ‘hot’ list (e.g. root, admin)

```
packet_service=ftp_data
|packet_flag= SF
    ||destinationhost_same_src_port_ratio<0.95
    |||number_root< 7.5
    ||||destinationhost_srv_diff_host_ratio<0.04
    |||||destinationhost_service_count<103.5 ==>normal
    | | | | | destinationhost_service_count>= 103.5
    ||||||destinationhost_service_count<106.5==> normal
    | | | | | | destinationhost_service_count>= 106.5==> normal


    | | | | destinationhost_srv_diff_host_ratio>=0.04
    |||||count < 10.5
    ||||||destinationhost_diff_srv_ratio<0.03
    |||||||destinationhost_same_service_ratio<0.22==>normal
    |  |    |    |    |    |    |destinationhost_same_service_ratio>=0.22==> normal
    |  |    |    |    |    |    destinationhost_diff_srv_ratio>=0.03
    |||||||source_bytes< 222==> normal
    |  |    |    |    |    |    |    source_bytes>=222==>normal
    |  |    |    |    |    count>=10.5==>normal
    |  |    |    number_root>=7.5==>normal
    |  |    destinationhost_same_src_port_ratio>=0.95
    |||destinationhost_srv_diff_host_ratio<0.05
    ||||source_bytes< 204.5==>normal
    |  |    |    |    source_bytes>=204.5==>normal

    |  |    |    destinationhost_srv_diff_host_ratio>=0.05

    ||||number_root< 2.5
```

```
|||||same_srv_ratio<0.75==>normal

| |   |   |   |   same_srv_ratio>=0.75

||||||destinationhost_count<2.5 ==> normal

| |   |   |   |   |   destinationhost_count>=2.5==>DDOS

| |   |   |   number_root>=2.5==>normal

|packet_flag=S0

||serror_ratio<0.97==>normal

| |   serror_ratio>=0.97

|||service_count<10.5==> DDOS

| |   |   service_count>= 10.5

||||destinationhost_service_count< 53

|||||destinationhost_service_count< 7

||||||count< 201.5==> DDOS

| |   |   |   |   |   count>=201.5==>normal

| |   |   |   |   destinationhost_service_count>=7==>normal

| |   |   |   destinationhost_service_count>=53==>normal

|packet_flag=REJ==>DDOS

|packet_flag=RSTR==>normal

|packet_flag=SH==>normal

|packet_flag=RSTO==>normal

|packet_flag=S1==>normal

|packet_flag=RSTOS0==>normal

|packet_flag=S3==>normal

|packet_flag=S2==>normal

|packet_flag=OTH==>normalpacket_service = other
|destinationhost_rerror_ratio<0.01
```

```
||destinationhost_same_src_port_ratio<0.99==>normal
| |    destinationhost_same_src_port_ratio>=0.99
|||source_bytes<73==> normal
| |    |    source_bytes>= 73
||||duration< 5019.5==>normal
| |    |    |    duration>=5019.5==>normal
|  destinationhost_rerror_ratio>=0.01
||destinationhost_count<97.5==>normal
| |    destinationhost_count>= 97.5
|||destinationhost_serror_ratio<0.07==>DDOS
| |    |    destinationhost_serror_ratio>=0.07==>normal
| |    |    |    |    |    count>=82.5==>normal
| |    |    |    |    destinationhost_rerror_ratio>=0.33
||||||serror_ratio<0.1 ==>DDOS
| |    |    |    |    |    serror_ratio>=0.1==>normal
||||packet_flag=RSTR==>DDOS
||||packet_flag=SH==>normal
||||packet_flag=RSTO==>normal
||||packet_flag=S1==>normal
||||packet_flag=RSTOS0 ==>normal
||||packet_flag=S3==>normal
||||packet_flag=S2==>normal
||||packet_flag=OTH==>normal
| |    service_count>= 9.5
|||same_srv_ratio<0.07
||||destinationhost_same_service_ratio<0.09==>DDOS
```

```
|  |    |    |    destinationhost_same_service_ratio>=0.09==>normal
|  |    |    same_srv_ratio>=0.07
||||destinationhost_service_count<14.5
|||||same_srv_ratio<0.1
||||||destinationhost_diff_srv_ratio<0.06==>normal
|  |    |    |    |    |    destinationhost_diff_srv_ratio>=0.06
|||||||diff_srv_ratio<0.07
||||||||count <222 ==> DDOS
| | | | | | | | count >= 222
|||||||||count <227 ==>normal
| | | | | | | | | count >= 227
||||||||||count < 241==> DDOS
| | | | | | | | | | count >= 241
|||||||||||count < 257==> normal
| | | | | | | | | | | count >=257 ==> DDOS
| | | | | | | diff_srv_ratio>=0.07
||||||||same_srv_ratio<0.08==>normal
| | | | | | | | same_srv_ratio>=0.08
|||||||||count< 117 ==>DDOS
```

The classifier efficiency is determined once the classifier is tested against the real time instances. The following factors have been used to determine the performance of different classifiers for intrusion detection system.

1. True Negative: Negative attacked patterns that are not correctly assigned to the positive class.
2. True Positive: Positive attacked patterns that are correctly assigned to the positive class.
3. False Positive: Attacked patterns that should be false but were predicted as true class.

4. False Negative: Attacked patterns that should be true class but were classified as negative class.

Experimental results are carried out with four classification models and also on different KDD99 test instances. After successfully analyzing the training data, testing of the classifier is performed over online DDOS experimental data.

Table3.3: Proposed vs Traditional accuracy details

Algorithm	Data Size	Accuracy	False positive
ACO-PSO	5000	96.1	0.225
ICA-NN	5000	96.12	0.31
PSO-NaïveBayes	5000	95.87	0.29
Proposed	5000	98.39	0.12

Table3.4 shows the accuracy of the proposed model and traditional models on KDD99 dataset. Using proposed model with partition, the classifiers give 98.39% of accuracy compared to traditional classifiers.

Table 3.4 Accuracy comparisons of Proposed model with other models on Different Datasets

Dataset	C4.5	Credal-C4.5	MID3	Proposed model (MODDT-Model-1)
Pima-diabetes	74.5	74.15	74.39	92.48
Glass	67.6	63.61	67.67	96.69
Pendigits	96.5	96.42	96.39	99.73
Credit-rating	85.6	85.43	84.03	88.63
German-credit	71.3	71.34	71.98	88.63

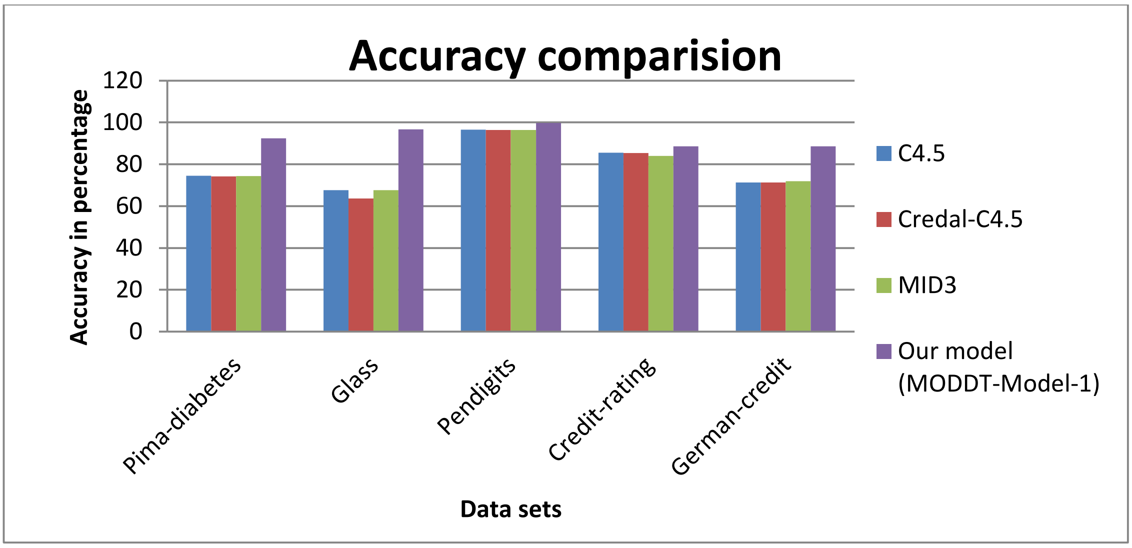


Fig 3.1: Comparison of various datasets with our model with other models.

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 3.5: 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

Gene based cancer disease patterns using proposed decision tree classification model on cancer dataset.

```
[isFlanking=0]: 58 ==> [inExAct=true, dbSNP=true]: 30
[isFlanking=0]: 58 ==> [dbSNP=true]: 30
[fre=0.25]: 92 ==> [CNT=1, isSomatic=true]: 48
[CNT=1, fre=0.25]: 64 ==> [isFlanking=0]: 34
[fre=0.25, isFlanking=0]: 47 ==> [inExAct=false, dbSNP=false, isSomatic=true]: 25
[fre=0.25, isFlanking=0]: 47 ==> [inExAct=false, isSomatic=true]: 25
[fre=0.25, isFlanking=0]: 47 ==> [inExAct=false, dbSNP=false]: 25
[fre=0.25, isFlanking=0]: 47 ==> [inExAct=false]: 25
[inExAct=true]: 62 ==> [dbSNP=true, pattern=TC, isSomatic=false]: 33
[inExAct=true]: 62 ==> [dbSNP=true, pattern=TC]: 33
[inExAct=true]: 62 ==> [isFlanking=0]: 33
[isFlanking=0]: 58 ==> [pattern=TC]: 31
[CNT=1, isFlanking=0]: 43 ==> [pattern=TC]: 23
[inExAct=true, mutAss=neutral]: 41 ==> [polyphen=benign, isSomatic=false]: 22
[inExAct=true, mutAss=neutral]: 41 ==> [dbSNP=true, fre=0.25]: 22
[fre=0.25, mutAss=neutral]: 41 ==> [inExAct=true, dbSNP=true]: 22
[inExAct=true, mutAss=neutral]: 41 ==> [dbSNP=true, CNT=1]: 22
[fre=0.25, mutAss=neutral]: 41 ==> [dbSNP=true]: 22
[inExAct=true, mutAss=neutral]: 41 ==> [polyphen=benign]: 22
[inExAct=true, mutAss=neutral]: 41 ==> [CNT=1]: 22
[dbSNP=true]: 54 ==> [inExAct=true, fre=0.25, isSomatic=false]: 29
[inExAct=true, dbSNP=true]: 54 ==> [fre=0.25, isSomatic=false]: 29
[dbSNP=true]: 54 ==> [inExAct=true, CNT=1, isSomatic=false]: 29
[inExAct=true, dbSNP=true]: 54 ==> [CNT=1, isSomatic=false]: 29
[dbSNP=true]: 54 ==> [fre=0.25, isSomatic=false]: 29
[dbSNP=true]: 54 ==> [CNT=1, isSomatic=false]: 29
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, dbSNP=true, polyphen=benign]:
21
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, polyphen=benign]:
21
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, dbSNP=true, isFlanking=0]: 21
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, isFlanking=0]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, dbSNP=true, fre=0.25]: 21
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, fre=0.25]: 21
```

```
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, dbSNP=true, CNT=1]: 21
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, CNT=1]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, polyphen=benign]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, isFlanking=0]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, fre=0.25]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [dbSNP=true, CNT=1]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, CNT=1]: 21
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [CNT=1]: 21
[mutAss=neutral, isSomatic=false]: 39 ==> [CNT=1]: 21
[mutAss=neutral]: 57 ==> [polyphen=benign]: 31
[mutAss=neutral]: 57 ==> [pattern=TC]: 31
[dbSNP=true, isSomatic=false]: 53 ==> [inExAct=true, fre=0.25]: 29
[inExAct=true, dbSNP=true, isSomatic=false]: 53 ==> [fre=0.25]: 29
[dbSNP=true, isSomatic=false]: 53 ==> [inExAct=true, CNT=1]: 29
[inExAct=true, dbSNP=true, isSomatic=false]: 53 ==> [CNT=1]: 29
[dbSNP=true, isSomatic=false]: 53 ==> [fre=0.25]: 29
[dbSNP=true, isSomatic=false]: 53 ==> [CNT=1]: 29
[inExAct=true]: 62 ==> [fre=0.25, isSomatic=false]: 34
[isFlanking=0]: 58 ==> [inExAct=true, isSomatic=false]: 32
[isSomatic=false]: 58 ==> [inExAct=true, isFlanking=0]: 32
[inExAct=true, isSomatic=false]: 58 ==> [isFlanking=0]: 32
[isFlanking=0]: 58 ==> [isSomatic=false]: 32
[isSomatic=false]: 58 ==> [isFlanking=0]: 32
[isFlanking=0]: 58 ==> [mutAss=neutral]: 32
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, polyphen=benign, isSomatic=false]:
21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [polyphen=benign, isSomatic=false]:
21
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, isFlanking=0, isSomatic=false]: 21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [isFlanking=0, isSomatic=false]: 21
[inExAct=true, fre=0.25]: 38 ==> [dbSNP=true, mutAss=neutral, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, fre=0.25, isSomatic=false]: 21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [fre=0.25, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, CNT=1, isSomatic=false]: 21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [CNT=1, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [polyphen=benign, isSomatic=false]: 21
```

```
[dbSNP=true, mutAss=neutral]: 38 ==> [isFlanking=0, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [fre=0.25, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [CNT=1, isSomatic=false]: 21
[inExAct=true, fre=0.25]: 38 ==> [isFlanking=0, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, polyphen=benign]: 21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [polyphen=benign]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, isFlanking=0]: 21
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [isFlanking=0]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [polyphen=benign]: 21
[dbSNP=true, mutAss=neutral]: 38 ==> [isFlanking=0]: 21
[fre=0.25, isFlanking=0]: 47 ==> [dbSNP=false, isSomatic=true]: 26
[fre=0.25, isFlanking=0]: 47 ==> [isSomatic=true]: 26
[dbSNP=true]: 54 ==> [inExAct=true, isFlanking=0, isSomatic=false]: 30
[inExAct=true, dbSNP=true]: 54 ==> [isFlanking=0, isSomatic=false]: 30
[dbSNP=true]: 54 ==> [isFlanking=0, isSomatic=false]: 30
[dbSNP=true]: 54 ==> [inExAct=true, isFlanking=0]: 30
[inExAct=true, dbSNP=true]: 54 ==> [isFlanking=0]: 30
[dbSNP=true]: 54 ==> [inExAct=true, fre=0.25]: 30
[inExAct=true, dbSNP=true]: 54 ==> [fre=0.25]: 30
[dbSNP=true]: 54 ==> [inExAct=true, CNT=1]: 30
[inExAct=true, dbSNP=true]: 54 ==> [CNT=1]: 30
[dbSNP=true]: 54 ==> [isFlanking=0]: 30
[dbSNP=true]: 54 ==> [fre=0.25]: 30
[dbSNP=true]: 54 ==> [CNT=1]: 30
[CNT=1, isFlanking=0]: 43 ==> [mutAss=neutral]: 24
[CNT=1]: 77 ==> [isFlanking=0]: 43
[inExAct=true, mutAss=neutral]: 41 ==> [isFlanking=0, isSomatic=false]: 23
[inExAct=true, mutAss=neutral]: 41 ==> [fre=0.25, isSomatic=false]: 23
[fre=0.25, mutAss=neutral]: 41 ==> [inExAct=true, isSomatic=false]: 23
[fre=0.25, mutAss=neutral]: 41 ==> [isSomatic=false]: 23
[mutAss=neutral]: 57 ==> [isFlanking=0]: 32
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, polyphen=benign]: 22
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [polyphen=benign]: 22
[mutAss=neutral, isSomatic=false]: 39 ==> [polyphen=benign]: 22
[inExAct=true]: 62 ==> [pattern=TC, isSomatic=false]: 35
[dbSNP=true, isSomatic=false]: 53 ==> [inExAct=true, isFlanking=0]: 30
```

```
[inExAct=true, dbSNP=true, isSomatic=false]: 53 ==> [isFlanking=0]: 30
[dbSNP=true, isSomatic=false]: 53 ==> [isFlanking=0]: 30
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [inExAct=true, polyphen=benign]:
21
[inExAct=true, dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [polyphen=benign]:
21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [inExAct=true, isFlanking=0]: 21
[inExAct=true, dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [isFlanking=0]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [inExAct=true, fre=0.25]: 21
[inExAct=true, dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [fre=0.25]: 21
[CNT=1, mutAss=neutral]: 37 ==> [inExAct=true, dbSNP=true, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [inExAct=true, CNT=1]: 21
[inExAct=true, dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [CNT=1]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [polyphen=benign]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [isFlanking=0]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [fre=0.25]: 21
[CNT=1, mutAss=neutral]: 37 ==> [dbSNP=true, isSomatic=false]: 21
[dbSNP=true, mutAss=neutral, isSomatic=false]: 37 ==> [CNT=1]: 21
[CNT=1, mutAss=neutral]: 37 ==> [inExAct=true, isSomatic=false]: 21
[CNT=1, mutAss=neutral]: 37 ==> [isSomatic=false]: 21
[polyphen=benign]: 44 ==> [inExAct=true, dbSNP=true, isSomatic=false]: 25
[polyphen=benign]: 44 ==> [dbSNP=true, isSomatic=false]: 25
[polyphen=benign]: 44 ==> [inExAct=true, dbSNP=true]: 25
[polyphen=benign]: 44 ==> [dbSNP=true]: 25
[isSomatic=false]: 58 ==> [inExAct=true, dbSNP=true, pattern=TC]: 33
[inExAct=true, isSomatic=false]: 58 ==> [dbSNP=true, pattern=TC]: 33
[isSomatic=false]: 58 ==> [dbSNP=true, pattern=TC]: 33
[isFlanking=0]: 58 ==> [inExAct=true]: 33
[inExAct=true, fre=0.25]: 38 ==> [pattern=TC, isSomatic=false]: 22
[inExAct=true, fre=0.25]: 38 ==> [dbSNP=true, mutAss=neutral]: 22
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, fre=0.25]: 22
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [fre=0.25]: 22
[dbSNP=true, mutAss=neutral]: 38 ==> [inExAct=true, CNT=1]: 22
[inExAct=true, dbSNP=true, mutAss=neutral]: 38 ==> [CNT=1]: 22
[dbSNP=true, mutAss=neutral]: 38 ==> [fre=0.25]: 22
[dbSNP=true, mutAss=neutral]: 38 ==> [CNT=1]: 22
```

```
[inExAct=true, fre=0.25]: 38 ==> [isFlanking=0]: 22
[inExAct=true]: 62 ==> [pattern=TC]: 36
[inExAct=true, mutAss=neutral]: 41 ==> [dbSNP=true, pattern=TC, isSomatic=false]: 24
[inExAct=true, mutAss=neutral]: 41 ==> [dbSNP=true, pattern=TC]: 24
[fre=0.25, mutAss=neutral]: 41 ==> [isFlanking=0]: 24
[inExAct=true, mutAss=neutral]: 41 ==> [isFlanking=0]: 24
[isFlanking=0]: 58 ==> [CNT=1, fre=0.25]: 34
[isSomatic=false]: 58 ==> [inExAct=true, fre=0.25]: 34
[inExAct=true, isSomatic=false]: 58 ==> [fre=0.25]: 34
[isSomatic=false]: 58 ==> [fre=0.25]: 34
[fre=0.25]: 92 ==> [inExAct=false, dbSNP=false, isSomatic=true]: 54
[fre=0.25]: 92 ==> [inExAct=false, isSomatic=true]: 54
[fre=0.25]: 92 ==> [inExAct=false, dbSNP=false]: 54
[fre=0.25]: 92 ==> [inExAct=false]: 54
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, isFlanking=0]: 23
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [isFlanking=0]: 23
[mutAss=neutral, isSomatic=false]: 39 ==> [inExAct=true, fre=0.25]: 23
[inExAct=true, mutAss=neutral, isSomatic=false]: 39 ==> [fre=0.25]: 23
[mutAss=neutral, isSomatic=false]: 39 ==> [isFlanking=0]: 23
[mutAss=neutral, isSomatic=false]: 39 ==> [fre=0.25]: 23
[CNT=1, mutAss=neutral]: 37 ==> [inExAct=true, dbSNP=true]: 22
[CNT=1, mutAss=neutral]: 37 ==> [dbSNP=true]: 22
[CNT=1, mutAss=neutral]: 37 ==> [inExAct=true]: 22
[fre=0.25, isFlanking=0]: 47 ==> [dbSNP=false]: 28
[inExAct=true]: 62 ==> [dbSNP=true, mutAss=neutral, isSomatic=false]: 37
[CNT=1]: 77 ==> [inExAct=false, dbSNP=false, fre=0.25, isSomatic=true]: 46
[CNT=1]: 77 ==> [inExAct=false, fre=0.25, isSomatic=true]: 46
[CNT=1]: 77 ==> [inExAct=false, dbSNP=false, isSomatic=true]: 46
[CNT=1]: 77 ==> [inExAct=false, dbSNP=false, fre=0.25]: 46
[CNT=1]: 77 ==> [inExAct=false, isSomatic=true]: 46
[CNT=1]: 77 ==> [inExAct=false, fre=0.25]: 46
[CNT=1]: 77 ==> [inExAct=false, dbSNP=false]: 46
[CNT=1]: 77 ==> [inExAct=false]: 46
[isSomatic=false]: 58 ==> [inExAct=true, pattern=TC]: 35
[inExAct=true, isSomatic=false]: 58 ==> [pattern=TC]: 35
[isSomatic=false]: 58 ==> [pattern=TC]: 35
```

Table 3.6: Leukaemia Data Results

In this table, Leukaemia data are taken as input for decision tree analysis. Initially, outlier detection approach is applied on the leukaemia dataset to remove the anomaly objects for classification process. Leukaemia dataset is taken from the URL(http://leo.ugr.es/elvira/DBCRepository/Leukemia/ALLAML.html).

In the following table, each attribute and its mean and standard deviation values are displayed for classification accuracy.

```
Attribute                Statistical Properties
========================================
AFFX-BioB-5_at
  mean                  -101.2368
  std. dev.               88.4871
AFFX- BioB-M_at
  mean                  -163.7105
  std. dev.              112.4806
AFFX-BioB-3_at
  mean                    -4.0789
  std. dev.              112.6976
AFFX-BioC-5_at
  mean                   199.5789
  std. dev.              113.5188
AFFX-BioC-3_at
  mean                  -265.3421
  std. dev.              117.838
AFFX-BioDn-5_at
  mean                  -390.8421
  std. dev.              148.1833


AFFX-BioDn-3_at
  mean                   -73.1316
  std. dev.              279.7015
AFFX-CreX-5_at
  mean                  -188.6842
  std. dev.              107.76
AFFX-CreX-3_at
  mean                    81.7368
```

std. dev.	97.2059
AFFX-BioB-5_st	
mean	126.5526
std. dev.	197.4741
AFFX- BioB-M_st	
mean	-7.9211
std. dev.	158.111
AFFX- BioB-3_st	
mean	-666.5526
std. dev.	286.5725
AFFX- BioC-5_st	
mean	-505.5263
std. dev.	273.1509
AFFX-BioC-3_st	
mean	-190
std. dev.	149.741
FFX- BioDn-5_st	
mean	110.0789
std. dev.	125.5154
AFFX- BioDn-3_st	
mean	169.9211
std. dev.	152.4278
AFFX-CreX-5_st	
mean	-71.8684
std. dev.	86.4755
AFFX-CreX-3_st	
mean	-330.1842
std. dev.	211.5202
hum_ alu_at	
mean	25185.9474
std. dev.	10636.1532
AFFX-DapX-5_at	
mean	-11.3158
std. dev.	147.0344
AFFX-DapX-M_at	
mean	134.5526

std. dev.	122.7517
AFFX-DapX-3_at	
mean	-87.5789
std. dev.	65.6869
AFFX-LysX-5_at	
mean	22.1579
std. dev.	43.3593
AFFX- LysX-M_at	
mean	-174.2368
std. dev.	270.3656
AFFX-LysX-3_at	
mean	21.7105
std. dev.	451.2567
AFFX-PheX-5_at	
mean	-90.5526
std. dev.	59.576
AFFX- PheX-M_at	
mean	-129.1053
std. dev.	59.2625
AFFX-PheX-3_at	
mean	-3.7632
std. dev.	59.7845
AFFX-ThrX-5_at	
mean	-24.5
std. dev.	49.6046
AFFX- ThrX-M_at	
mean	-41.6316
std. dev.	56.4467
AFFX-ThrX-3_at	
mean	-303.9474
std. dev.	154.8755
AFFX-TrpnX-5_at	
mean	9.5
std. dev.	56.4571
AFFX- TrpnX-M_at	
mean	-623.3158

std. dev.	438.8245
AFFX-TrpnX-3_at	
mean	-294.7105
std. dev.	288.4411
AFFX-HUMISGF3A/M97935_5_at	
mean	-258.6316
std. dev.	303.6769
AFFX-HUMISGF3A/M97935_MA_at	
mean	-60.4474
std. dev.	536.8284
AFFX-HUMISGF3A/M97935_MB_at	
mean	204.3158
std. dev.	321.5426
AFFX-HUMISGF3A/M97935_3_at	
mean	727.7895
std. dev.	654.4552
AFFX-HUMRGE/M10098_5_at	
mean	3000.5263
std. dev.	5364.5012
AFFX-HUMRGE/M10098_M_at	
mean	128.2105
std. dev.	162.4698
AF000231_at	
mean	64.6053
std. dev.	92.2926
AF000234_at	
mean	-202.6316
std. dev.	273.9225
AF000430_at	
mean	3.4737
std. dev.	48.623
AF000545_at	
mean	-161.8947
std. dev.	186.3579
AF000560_at	
mean	227.5789

```
  std. dev.              219.4322
AF000562_at
  mean                   954.6053
  std. dev.              370.6618
AF000573_rna1_at
  mean                   -66.3158
  std. dev.               68.8595
AF000959_at
  mean                  -425.5263
  std. dev.              206.4726
AF001294_at
  mean                    63.3947
  std. dev.              180.9755
AF001620_at
  mean                     -62.5
  std. dev.               83.5526
AF002020_at
  mean                   183.0789
  std. dev.               98.4849
AF002224_at
  mean                   377.6316
  std. dev.               201.034
AF002700_at
  mean                   435.0789
  std. dev.              249.4553
AF003743_at
  mean                   207.6316
  std. dev.              354.3835
PROPOSED PATTERNS
---------------------------------
[U31973_s_at=0]: 4 ==> [CLASS=AML]: 2
[U31973_s_at=0]: 4 ==> [CLASS=ALL]: 2
[U31973_s_at=-54]: 3 ==> [M59499_at=238, CLASS=AML]: 2
[U31973_s_at=-54, CLASS=AML]: 3 ==> [M59499_at=238]: 2
[L40379_at=-159]: 3 ==> [U88892_at=-234, CLASS=AML]: 2
[L40379_at=-159, CLASS=AML]: 3 ==> [U88892_at=-234]: 2
[HG2007-HT2056_s_at=41]: 3 ==> [CLASS=ALL]: 2
```

```
[X81333_at=103]: 3 ==> [CLASS=AML]: 2
[X81333_at=103]: 3 ==> [U31973_s_at=-51]: 2
[U06454_at=-47]: 3 ==> [CLASS=ALL]: 2
[U31973_s_at=-54]: 3 ==> [M59499_at=238]: 2
[M59499_at=0]: 3 ==> [CLASS=ALL]: 2
[L40379_at=-159]: 3 ==> [U88892_at=-234]: 2
[L27476_at=116]: 3 ==> [CLASS=ALL]: 2
[J02645_at=48]: 3 ==> [CLASS=AML]: 2
[U43527_at=550]: 4 ==> [CLASS=AML]: 3
[L27476_at=136]: 4 ==> [CLASS=ALL]: 3
[M59499_at=238]: 2 ==> [U31973_s_at=-54, CLASS=AML]: 2
[M59499_at=238, U31973_s_at=-54]: 2 ==> [CLASS=AML]: 2
[M59499_at=238, CLASS=AML]: 2 ==> [U31973_s_at=-54]: 2
[U88892_at=-234]: 2 ==> [L40379_at=-159, CLASS=AML]: 2
[L40379_at=-159, U88892_at=-234]: 2 ==> [CLASS=AML]: 2
[U88892_at=-234, CLASS=AML]: 2 ==> [L40379_at=-159]: 2
[AJ000480_at=592]: 2 ==> [X81333_at=91, CLASS=ALL]: 2
[X81333_at=91]: 2 ==> [AJ000480_at=592, CLASS=ALL]: 2
[AJ000480_at=592, X81333_at=91]: 2 ==> [CLASS=ALL]: 2
[AJ000480_at=592, CLASS=ALL]: 2 ==> [X81333_at=91]: 2
[X81333_at=91, CLASS=ALL]: 2 ==> [AJ000480_at=592]: 2
[U10689_f_at=198]: 2 ==> [CLASS=ALL]: 2
[U10689_f_at=134]: 2 ==> [CLASS=ALL]: 2
[U10689_f_at=131]: 2 ==> [CLASS=ALL]: 2
[U10689_f_at=118]: 2 ==> [CLASS=AML]: 2
[HG4677-HT5102_s_at=288]: 2 ==> [CLASS=ALL]: 2
[HG4677-HT5102_s_at=84]: 2 ==> [CLASS=ALL]: 2
[U31973_s_at=-27]: 2 ==> [CLASS=ALL]: 2
[M22348_s_at=425]: 2 ==> [CLASS=ALL]: 2
[U64315_s_at=371]: 2 ==> [CLASS=AML]: 2
[U64315_s_at=370]: 2 ==> [CLASS=ALL]: 2
[U64315_s_at=327]: 2 ==> [CLASS=AML]: 2
[U64315_s_at=305]: 2 ==> [CLASS=ALL]: 2
[M86757_s_at=-486]: 2 ==> [CLASS=ALL]: 2
[HG2007-HT2056_s_at=4]: 2 ==> [CLASS=AML]: 2
[HG2007-HT2056_s_at=-2]: 2 ==> [CLASS=ALL]: 2
```

```
[HG2007-HT2056_s_at=-3]: 2 ==> [CLASS=ALL]: 2
[HG2007-HT2056_s_at=-5]: 2 ==> [CLASS=ALL]: 2
[HG2007-HT2056_s_at=-12]: 2 ==> [CLASS=ALL]: 2
[HG2007-HT2056_s_at=-49]: 2 ==> [CLASS=ALL]: 2
[HG2271-HT2367_at=-139]: 2 ==> [CLASS=ALL]: 2
[HG2271-HT2367_at=-360]: 2 ==> [CLASS=ALL]: 2
[X80818_at=2244]: 2 ==> [CLASS=ALL]: 2
[U31973_s_at=-51]: 2 ==> [X81333_at=103]: 2
[X81333_at=91]: 2 ==> [CLASS=ALL]: 2
[X81333_at=69]: 2 ==> [CLASS=ALL]: 2
[X81333_at=60]: 2 ==> [CLASS=AML]: 2
[X81333_at=57]: 2 ==> [CLASS=ALL]: 2
[X54380_at=24]: 2 ==> [CLASS=ALL]: 2
[X54380_at=3]: 2 ==> [CLASS=ALL]: 2
[X54380_at=-42]: 2 ==> [CLASS=ALL]: 2
[X54380_at=-91]: 2 ==> [CLASS=ALL]: 2
[X52142_at=122]: 2 ==> [CLASS=ALL]: 2
[X52142_at=109]: 2 ==> [U31973_s_at=-68]: 2
[U31973_s_at=-68]: 2 ==> [X52142_at=109]: 2
[X52142_at=14]: 2 ==> [CLASS=ALL]: 2
[X15880_at=378]: 2 ==> [CLASS=AML]: 2
[X15880_at=232]: 2 ==> [CLASS=ALL]: 2
[U88892_at=-153]: 2 ==> [CLASS=ALL]: 2
[U88892_at=-234]: 2 ==> [CLASS=AML]: 2
[U88892_at=-301]: 2 ==> [CLASS=ALL]: 2
[U43527_at=706]: 2 ==> [CLASS=ALL]: 2
[U43527_at=609]: 2 ==> [CLASS=ALL]: 2
[U06454_at=-19]: 2 ==> [CLASS=ALL]: 2
[U06454_at=-26]: 2 ==> [CLASS=ALL]: 2
[U06454_at=-44]: 2 ==> [CLASS=AML]: 2
[U06454_at=-88]: 2 ==> [CLASS=ALL]: 2
[M59499_at=238]: 2 ==> [CLASS=AML]: 2
[M59499_at=238]: 2 ==> [U31973_s_at=-54]: 2
```

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.

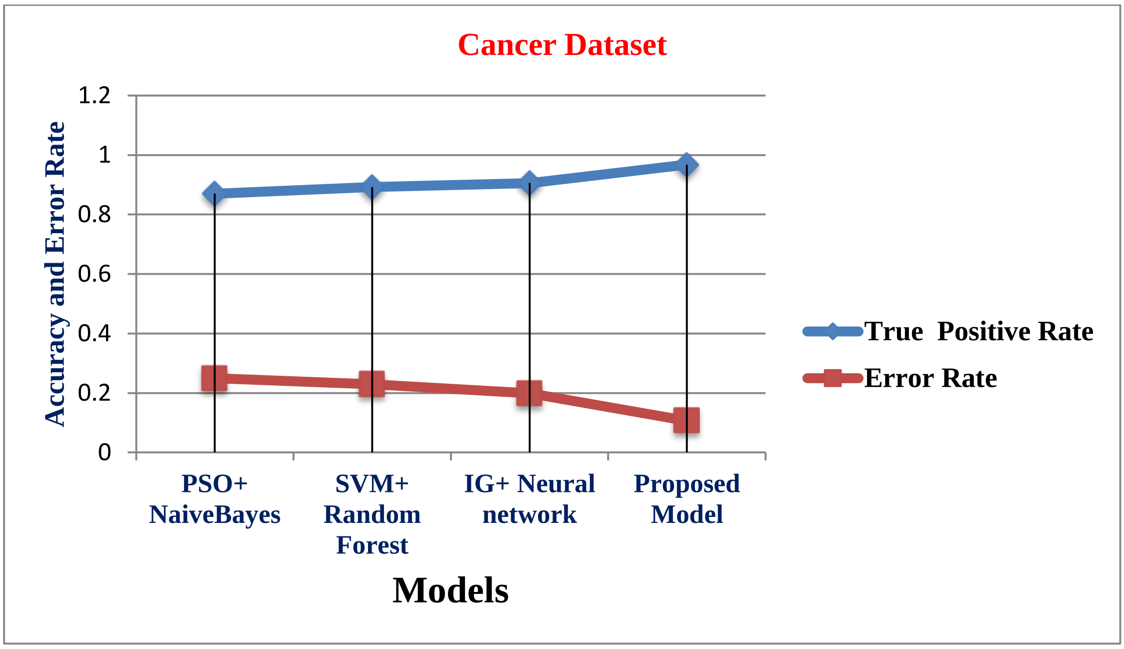
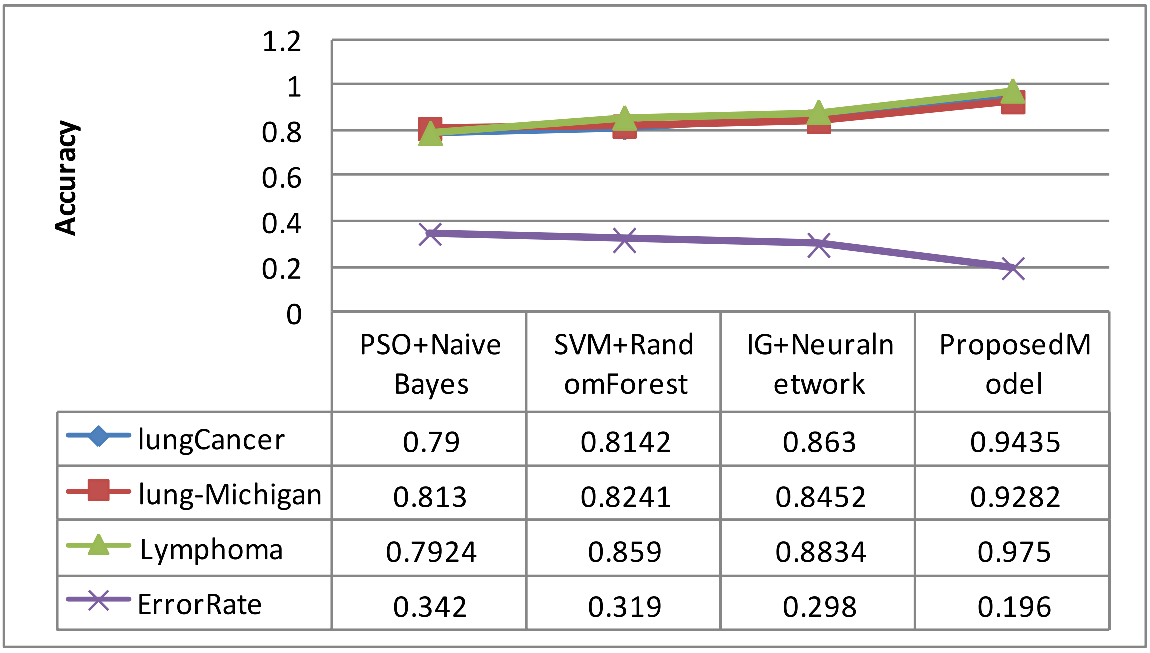


Figure 3.2: Performance of cancer disease using classification models

Figure 3.2 illustrates the improvement of proposed approach to the traditional weak classifiers by using error rate, runtime and true positive rate. Here, the performance was computed in terms of true positive rate, error rate and runtime on Ovarian dataset. From the figure, it is clearly observed that the present approach has high accuracy measure compared to traditional models.

Table 3.7: Comparative study of present model to the traditional models on three cancer disease datasets by using accuracy measure

Datasets	PSO+ Naïve Bayes	SVM+ Random Forest	IG +Neural Network	MODDT (PROPOSED MODEL)
Lung Cancer	0.79	0.8142	0.863	0.9435
Lung Michigan	0.813	0.8241	0.8452	0.9282
Lymphoma	0.7924	0.859	0.8834	0.975
Error Rate	0.342	0.319	0.298	0.196
Runtime(ms)	6531	6294	5964	5193



	PSO+Naive Bayes	SVM+RandomForest	IG+Neuralnetwork	ProposedModel
lungCancer	0.79	0.8142	0.863	0.9435
lung-Michigan	0.813	0.8241	0.8452	0.9282
Lymphoma	0.7924	0.859	0.8834	0.975
ErrorRate	0.342	0.319	0.298	0.196

Figure 3.3: Performance of present approach to traditional approaches by using accuracy, error rate and run-time

Table 3.7 and Figure 3.3 describe the performance of the proposed ensemble model to the traditional models on microarray cancer disease datasets. From the table, it is clearly observed that the proposed model has high accuracy and less error rate compare to traditional models in terms of accuracy, error rate and runtime.

Table 3.8: Run-Time comparison of the Proposed Feature selection measure to the traditional feature selection measures

Feature selection Measures	**Diabetes (secs)**	**Leukemia (secs)**	**Cancer (secs)**	**Brain Disease (secs)**
Chi-square Rank	32.67	35.76	32.98	33.13
Mutual Info Rank	39.12	39.77	32.87	31.97
Probabilistic Rank	39.86	41.86	39.04	36.76
MODDT (Proposed Model)	27.654	26.87	26.98	24.87

Table 3.8 illustrates the runtime of the present model with the traditional models on biomedical data repositories. From the tabulated values, it is analyzed that the present feature ranking technique has low execution time compared to traditional techniques.

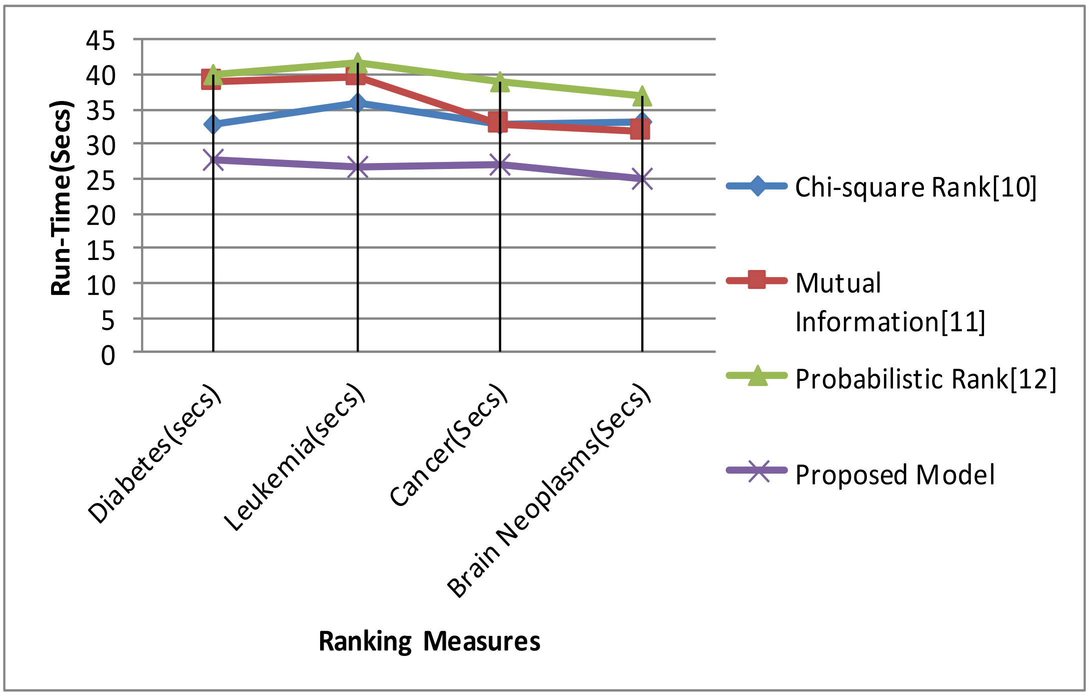


Figure 3.4: Run-Time comparison of the Proposed Ranking measure to the traditional ranking measures

Figure 3.4 illustrates the runtime of the present model with the traditional models on biomedical data repositories. From the tabulated values, it is analyzed that the present ranking technique has low execution time compared to traditional techniques.

Table 3.9: Average Feature Ranking Measure of The Present PSO Model to The Traditional Ranking Measures

Feature selection Measures	Diabetes (secs)	Leukemia (secs)	Cancer (secs)	Brain Disease (secs)
Chi-square Rank	0.76	0.69	0.83	0.719
Mutual Info Rank	0.83	0.879	0.9	0.858
Probabilistic Rank	0.792	0.87	0.848	0.869
MODDT (Proposed Model)	0.92	0.939	0.919	0.957

Table 3.9, describes the average ranking score of the present feature selection measure to the traditional approaches. From the table it was observed that the present rank algorithm has high average ranking score compared to the existing approaches.

Table 3.10: Comparative study of present model to traditional models by using average recall and precision rate on the training microarray datasets

Model	Recall	Precision
PSO+ NaiveBayes	0.845	0.864
SVM+ Random Forest	0.885	0.9054
IG+ Neural Network	0.9043	0.9053
MODDT (Proposed Model)	0.9735	0.9735

Table 3.10, illustrates the performance of the present model to the existing approaches by using average recall and precision rates. These average rates are simulated on the training microarray datasets.

Statistical control chart model aims to filter extreme anomalies using the filtered dataset, modified decision tree algorithm has been implemented with modified entropy and modified information gain.

3.5 SUMMARY

In this chapter, a hybrid statistical anomaly detection model is proposed to remove the noisy instances in the given datasets. Modified decision tree algorithm has improved classification accuracy in all the data sets considerably. Accuracy can be defined as the number of correct predictions. This filtered data is given to the classification problem to improve the accuracy and false positive rate on different datasets. Experimental results proved that the statistical anomaly detection-based classification problem has better efficiency on different conventional statistical anomaly model on different datasets. In Diabetes dataset the classification rate increased to 92%, which is better accuracy than the previous models (74.5%), In Glass identification dataset the classification rate increased to 96%, which is better accuracy than the previous models (68%). In Credit fraud dataset the classification accuracy rate is increased to 88% which is better than the previous models (71%). This model gives the better detection of diabetes, correct type of glass, better identification of network attacks, cancer decease and credit risks in the credit dataset.