UNIVERSITE PARIS 13 NORD

UPMC PARIS UNIVERSITAS

Accuracy-Rejection Curves (ARCs) for Comparison of Classification Methods with Reject Option

M.Sajjad-Ahmed NADEEM

Jean-Daniel ZUCKER

Blaise HANCZAR

6-September-2009
MLSB 09 Ljubljana

Lim&Bio

Centre de Recherche des Cordeliers

Instituts thématiques Inserm
Institut national
de la santé et de la recherche médicale

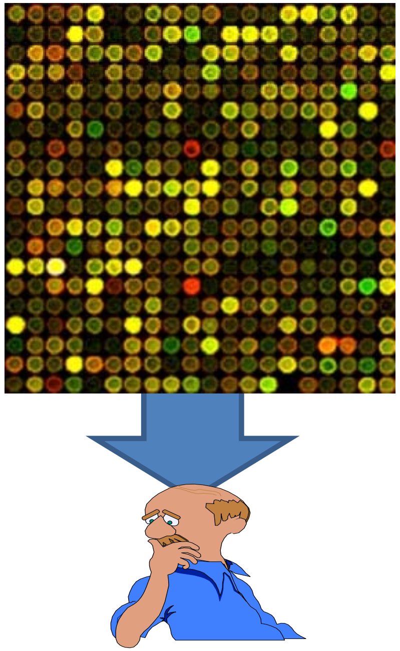
IRD
Institut de recherche
pour le développement

Outline:

- Introduction & Motivation
- State -of-art (Reject Option)
- Problem
- Comparing Classifiers with Reject Option
- Hypothesis
- Experiments
- Discussion & Conclusion

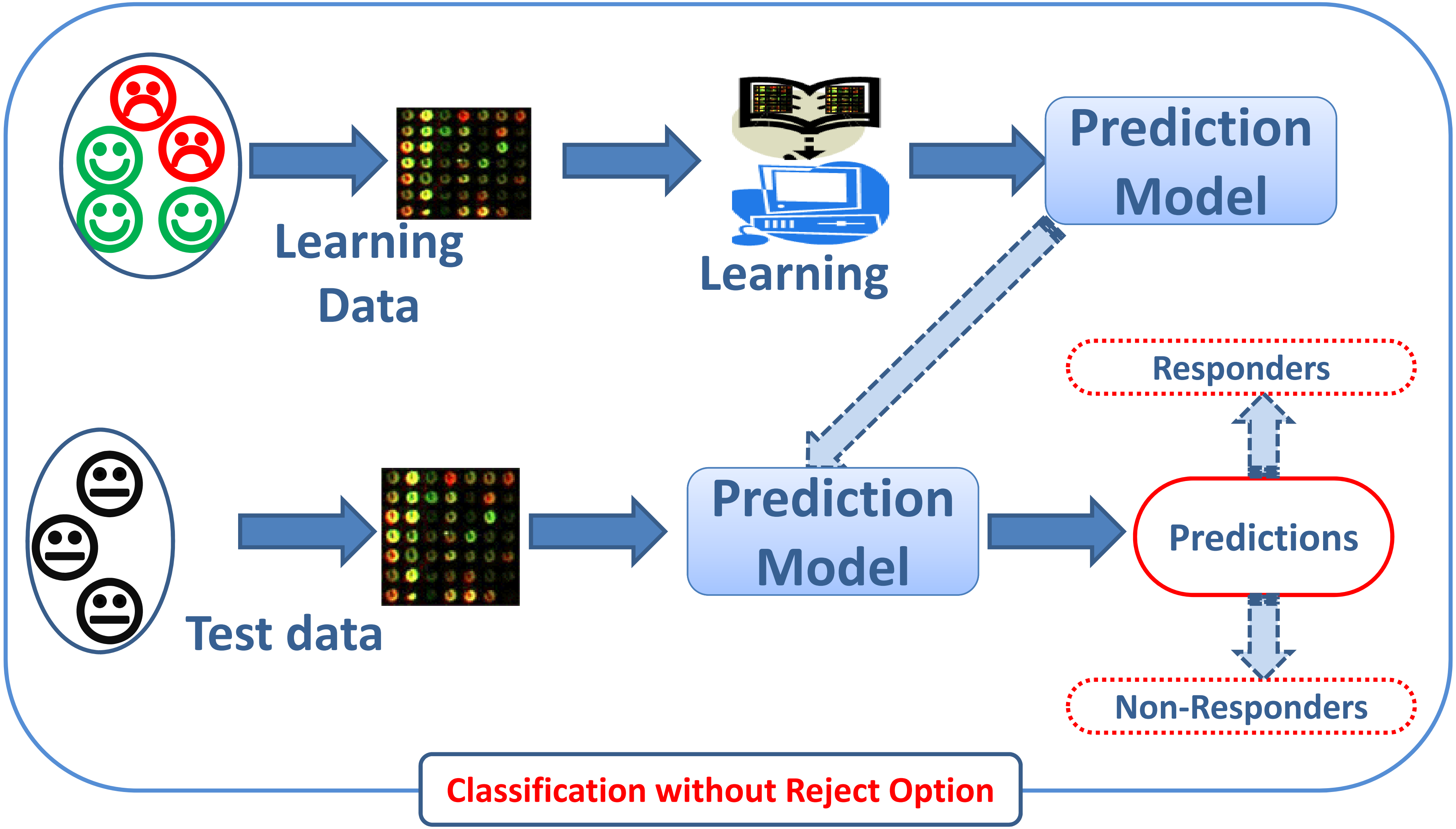
Introduction & Motivation: (1/4)

- Goal = classification with high accuracy.
- Thousands of genes.
- Few number of examples
  - Generally (50 to 100)



- Huge volumes of data in the form of microarrays.
- Humanly not possible to go-through and analyse the data.

Introduction & Motivation: (2/4)



Classification without Reject Option

Introduction & Motivation: (3/4)

- Consider a binary classification problem with two classes

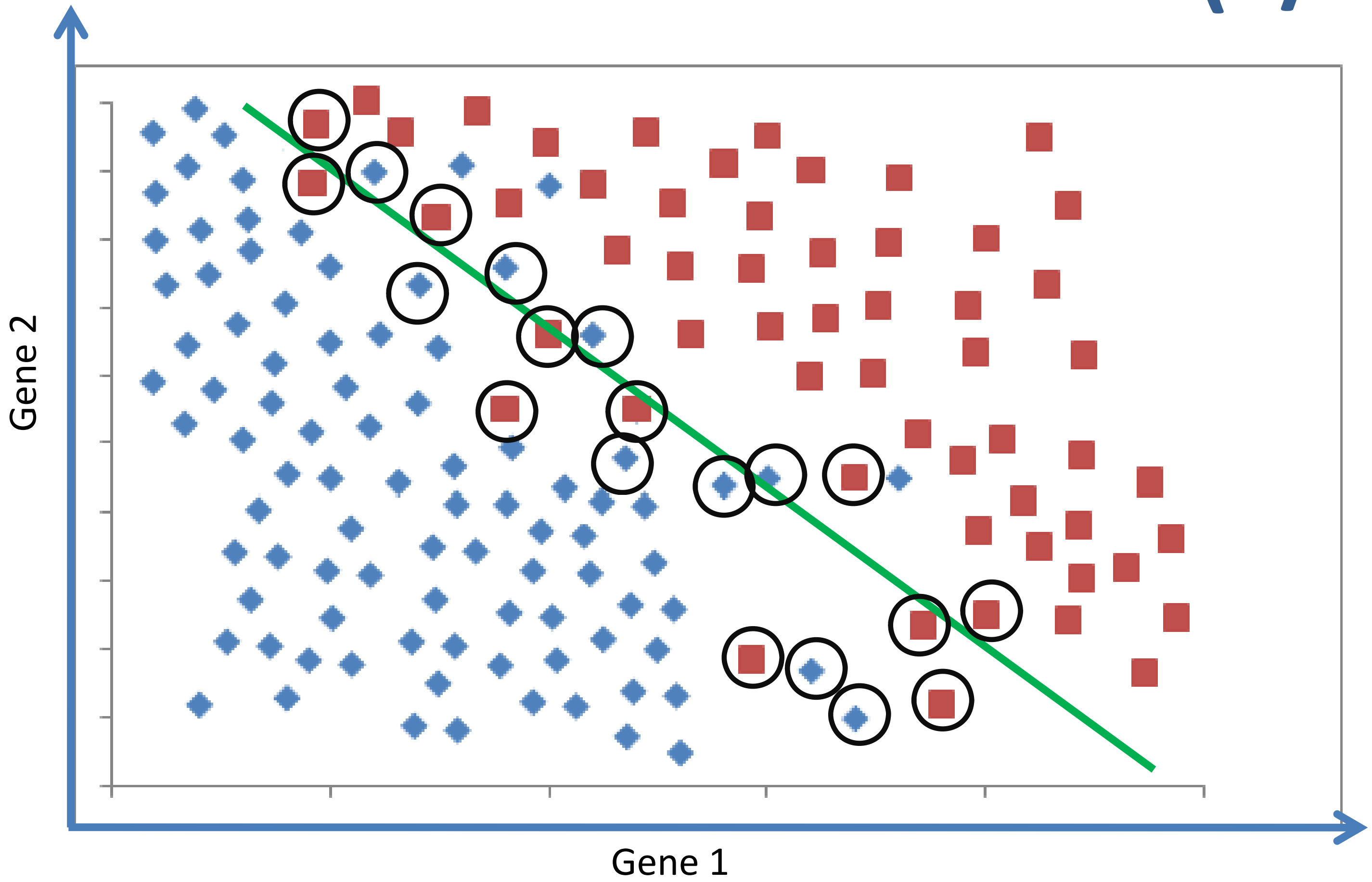
$$C = \{+1, -1\}$$

where an example is characterized by feature vector $z \in R_p$ and a label $y \in C$.

- An example x is classified as:

$$f(x) = \arg\max_{C_i}(p(C_i / x))$$

Introduction & Motivation: (4/3)



Low-confidence predictions cause high error rates.

Is improvement possible?

Reject Option (State-of-art):

Chow [Chow, 1970], Fumera et al. [Fumera et al., 2000],

Dubuisson and Masson [Dubuisson and Masson, 1993],

Landgrebe et al. [Landgrebe et al., 2006],

Li and Sehi [Li and Sethi, 2006],

Hanczar et al. [Hanczar et al., 2005]

Friedel et al. [Friedel et al., 2005]

and others worked on and proposed good methods of classification.

- **Are these methods applicable on biomedical data?**

Problem: (1/3)

- Existing data about fatal diseases like cancer etc. are available in the form of gene expression microarray.

- For a number of problems in biomedical field, existing methods of classification don't perform good enough to be used to make predictions.

- Making predictions about a person on the basis of his/her gene profile about a disease.

- Its crucial to separate patients and non-patients especially in cancer like diseases.

Problem: (2/3)

- Declaring a potential patient as non-patient and vice versa can be extremely harmful.

- High accuracy is required. Generally a system with 85% or more accuracy is acceptable.

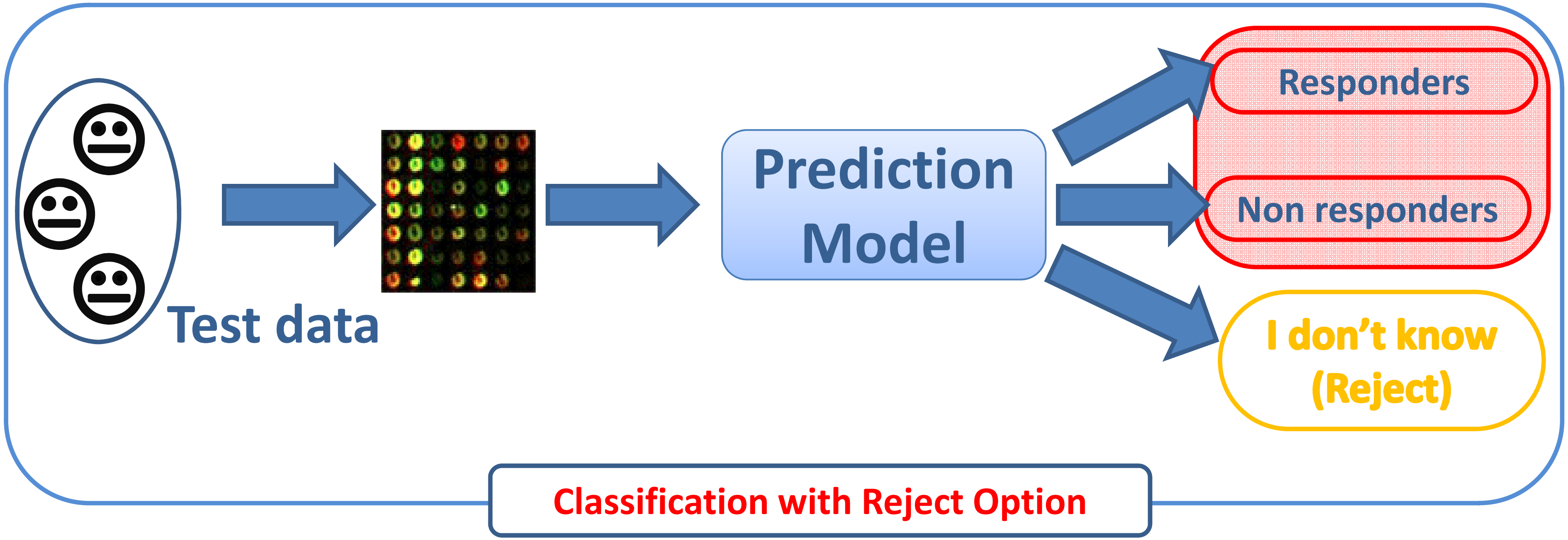
- Performance of a classifier depends heavily on data.

How to proceed in such cases?

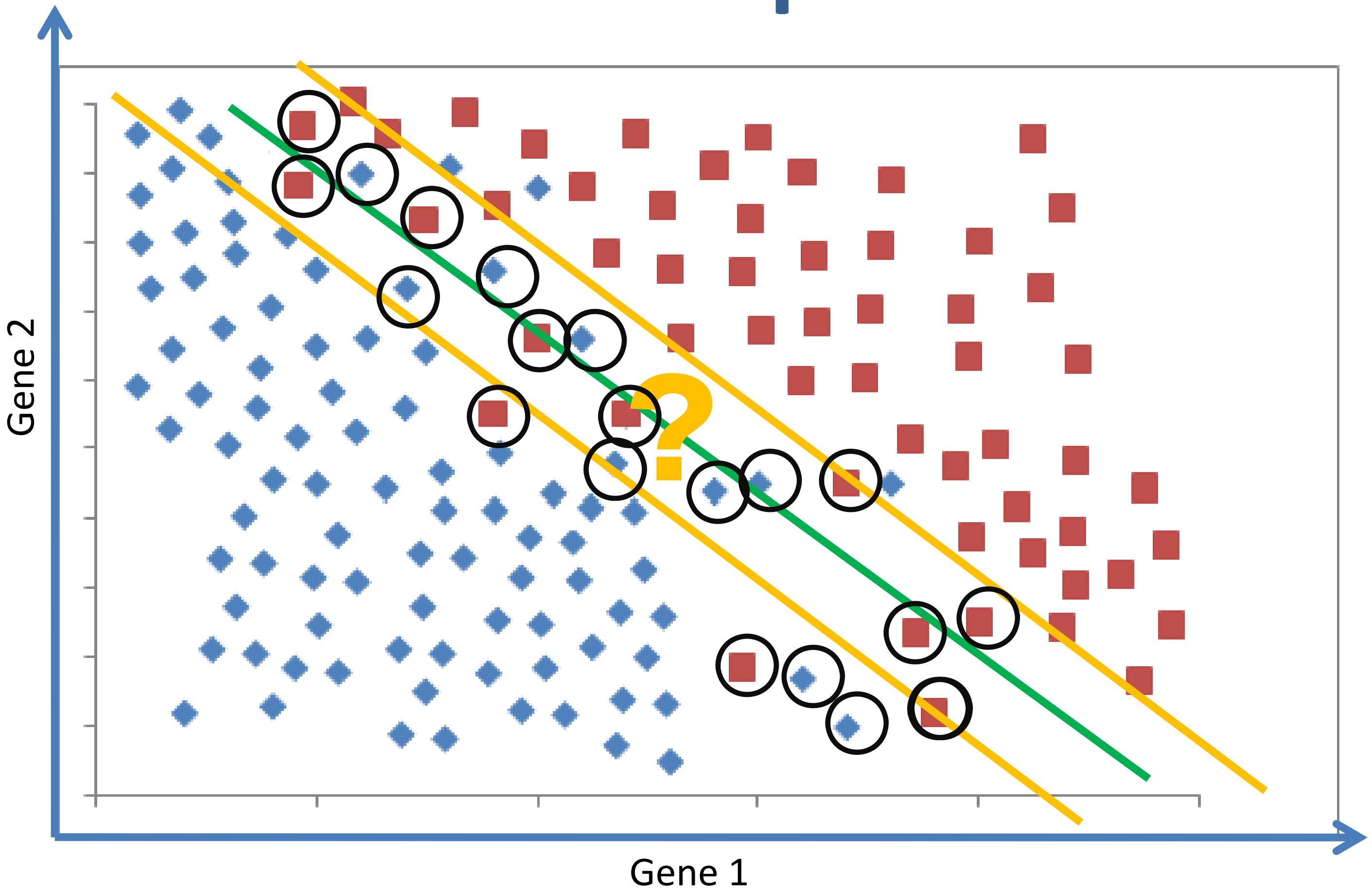
Problem: (2/3)

A physician refrains from therapy when (s)he is not confident enough in diagnosis.

This theory can be applied while making predictions on biomedical data.



Example:



Reject Option:

- Consider a binary classification problem with two classes

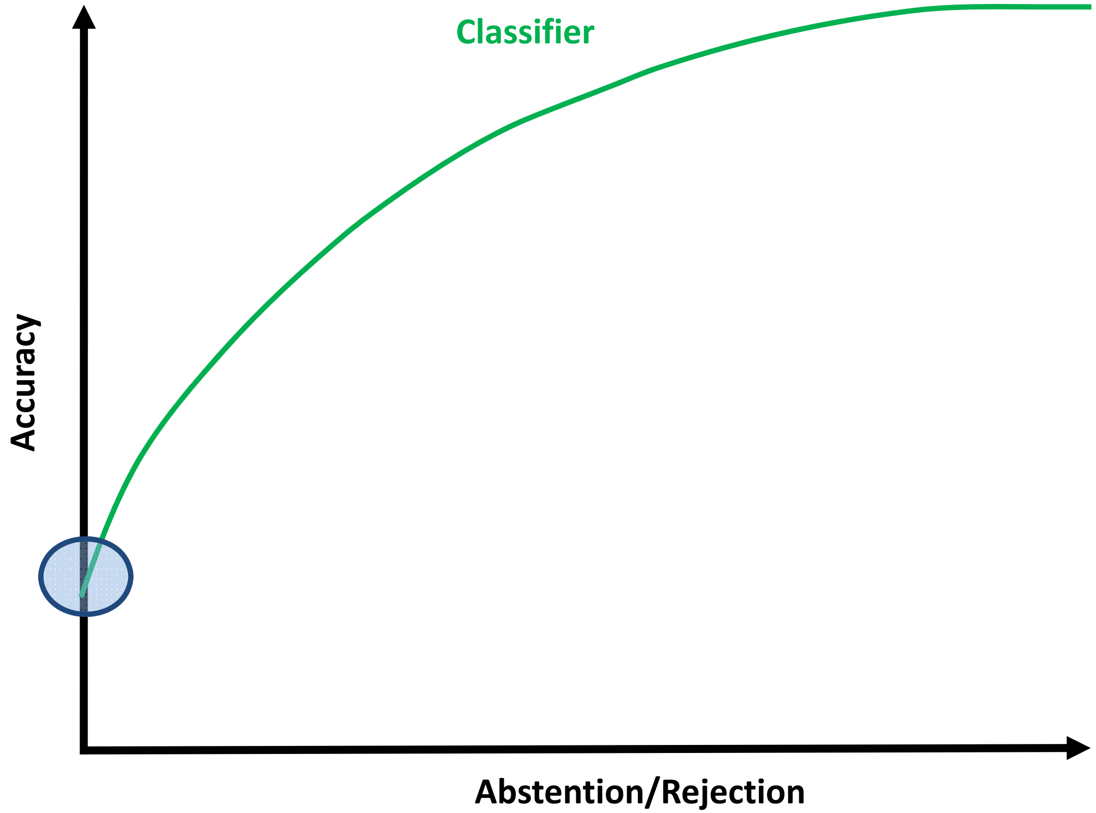
  $$C = \{+1, -1\}$$

  where an example is characterized by feature vector $z \in R_p$ and a label $y \in C$.

- A sample x is accepted only if the probability that x belongs to C_i is higher than or equal to a given probability threshold t

$$f(x) = \begin{cases} \arg\max_{C_i}(p(C_i/x)) \quad if \quad \max(p(C_i/x)) \geq t \\ reject \quad if \quad p(C_i/x) < t\forall_i \end{cases}$$

Tradeoff between rejection/accuracy:



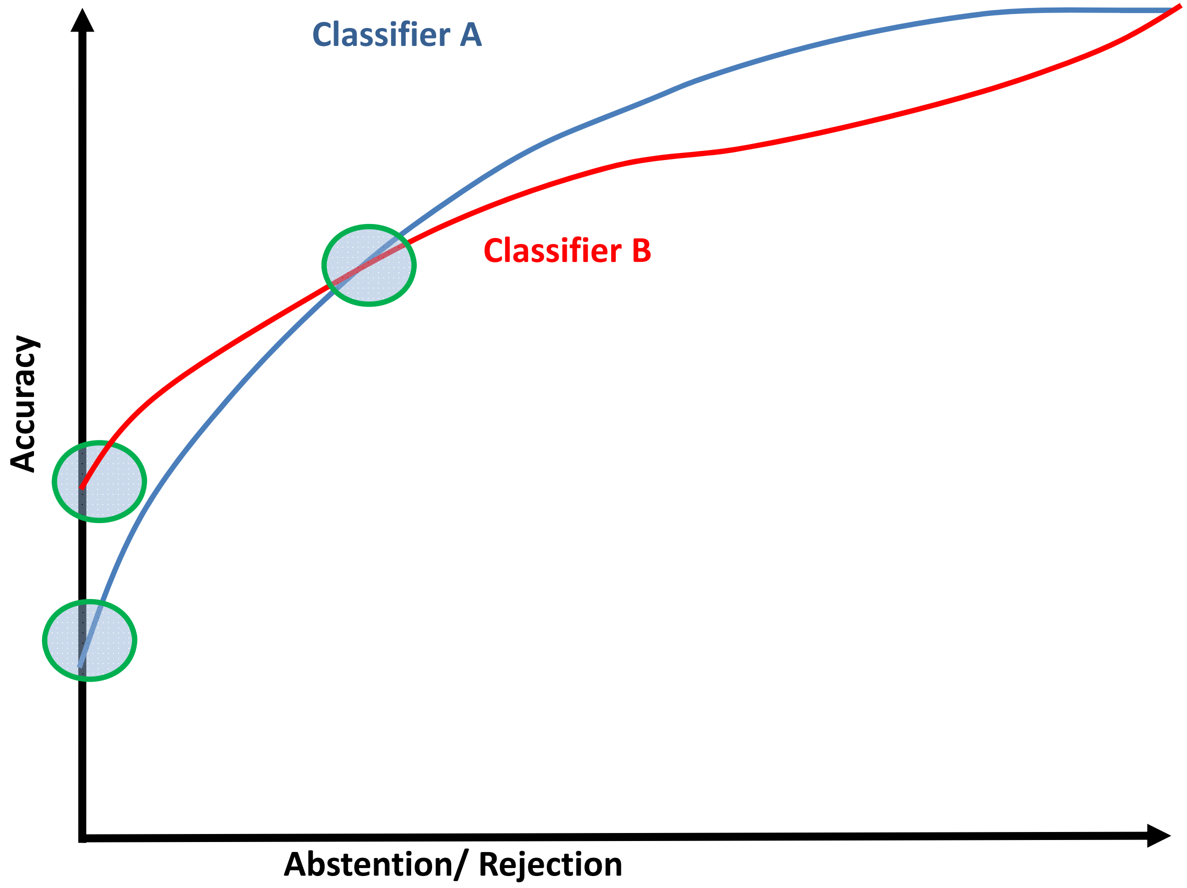
Comparing Classifiers with Reject Option: (1/3)

- Performances of classifiers are measured by their accuracy to predict the true class.

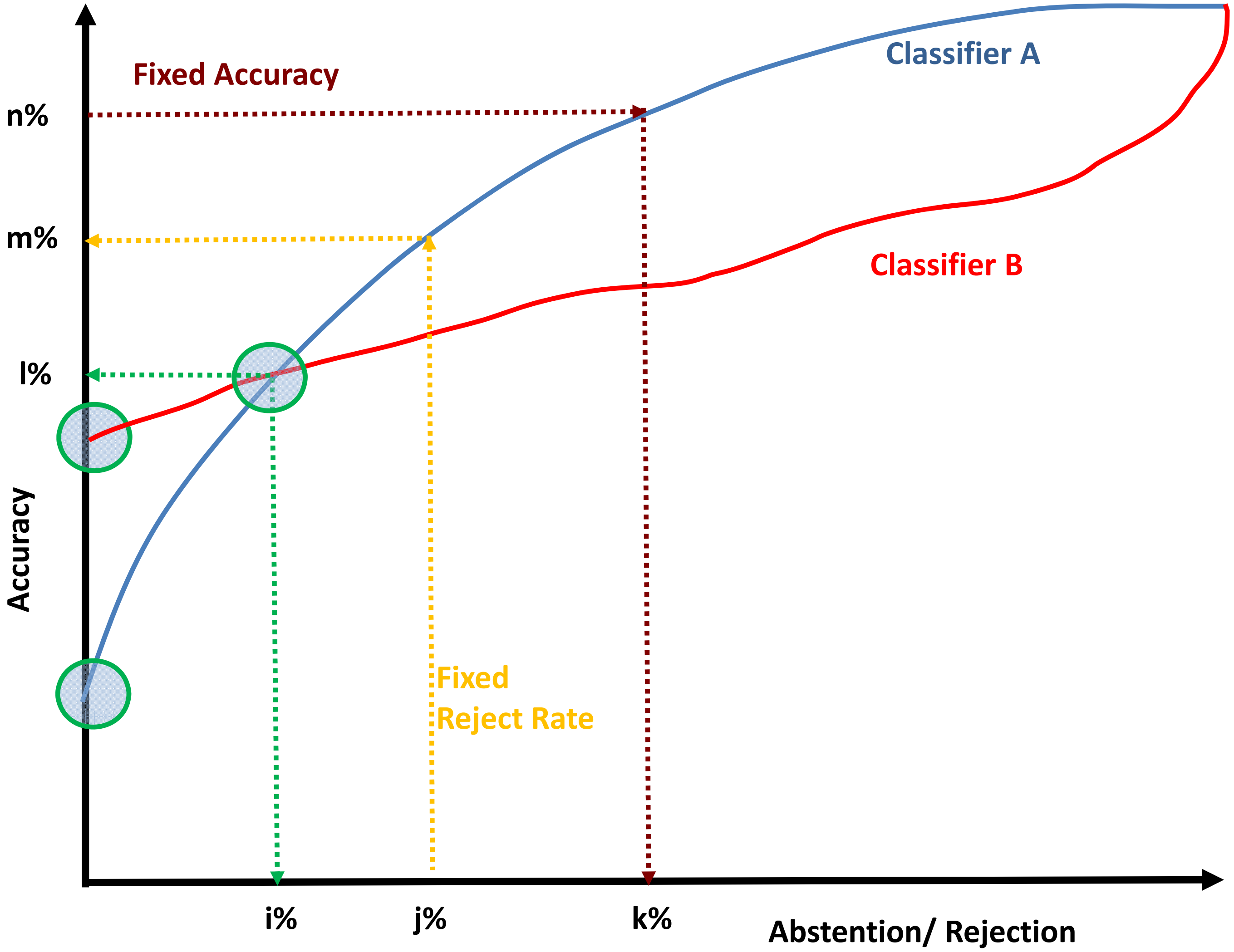
- Performance of a classifier depends heavily on the data.

- With reject option, the accuracy depends on the reject rate also. More rejection results in more better accuracy.

Comparing Classifiers with Reject Option: (2/3)

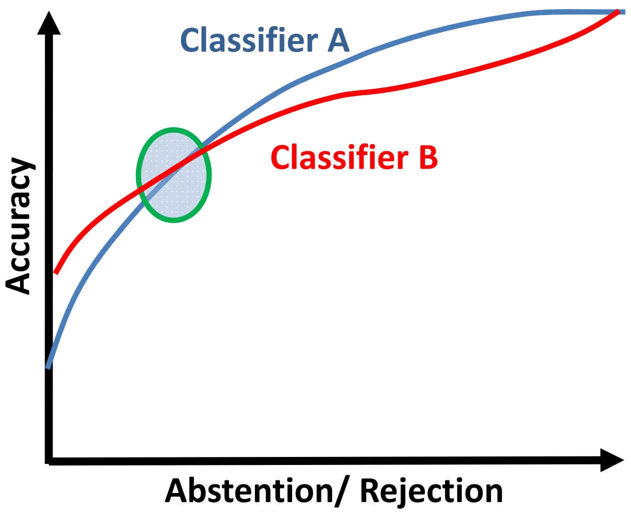


Comparing Classifiers with Reject Option: (3/3)

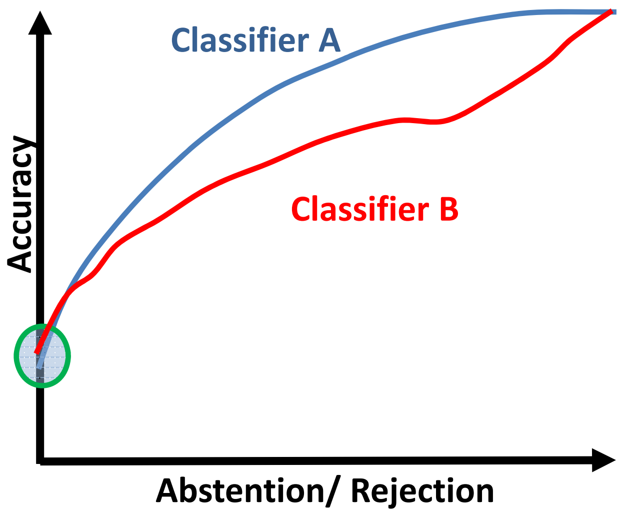


Hypothesis:

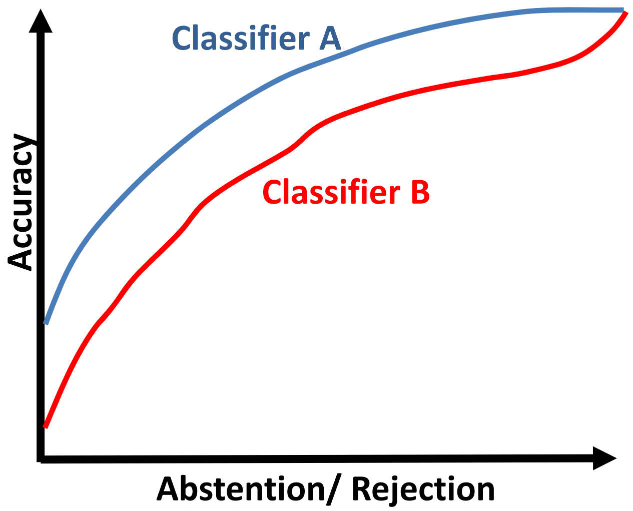
- Case-01 (T1)



- Case-02 (T2)



- Case-03 (T3)



Data:

- **Pure Synthetic data:**
  - Artificially generated data with user defined parameters .

- **Synthetic data:**
  - Artificially generated data with parameters computed from real microarray datasets.
    - Colon Cancer Data [Alon et al., 1999].
    - Lymphoid Malignancy [Shipp et al., 2002].
    - Leukemia [Golub et al., 1999].

Why Synthetic data:

- In real microarrays the number of samples remain very few.
- It becomes hard to effectively learn from few number of samples.
- Less number of test samples hinder to comprehensively test the built model.

Data Generation (1/2):

- **Pure Synthetic Data:**
  - User defined parameters.
  - 2 class classification problem where each class follows Gaussian distribution.
  - Equally likely class distribution.
  - Class conditional densities are $N(\mu_1;\sigma_1\Sigma)$ and $N(\mu_2;\sigma_2\Sigma)$ where $\mu_1=(-1,-1,-1,......)$ and $\mu_2=(1,1,1,......)$
  - For co-related data the covariance matrix of each class has a block structure like $\sum \quad B$.
  - Adding noise

- **Synthetic data from real Microarray data:**
  - Parameters are estimated from real data using Expectation Maximization (EM) algorithm.
  - 2 class classification problem.
  - Equally likely class distribution.
  - Adding noise

Data Generation (2/2):

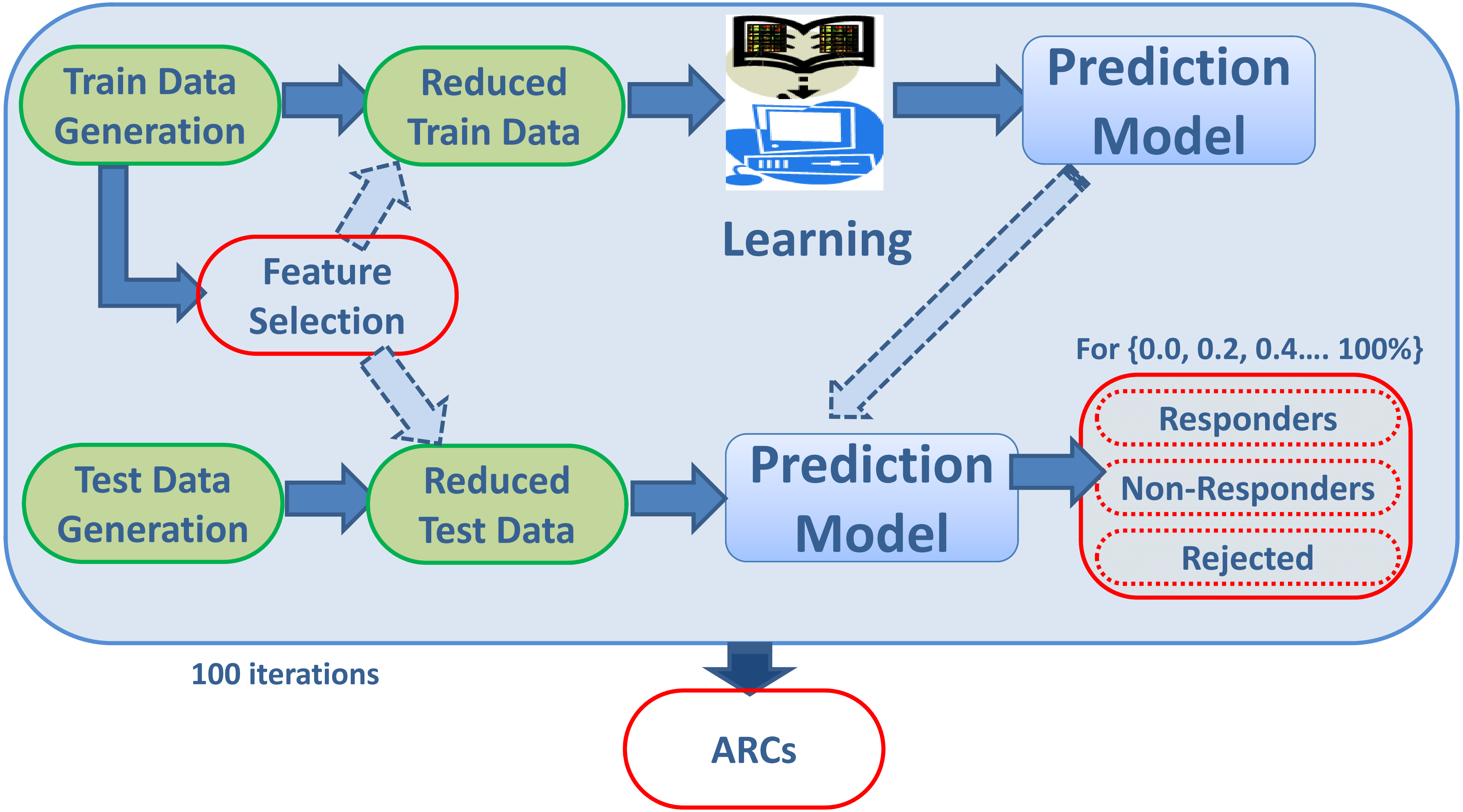
- **Parameters-Pure Synthetic data:**

Parameter description	Parameter	Numeric values used
Sample size train	n	50, 100, 200
No. of Gaussians per class	G	1, 2
No. of Boxes/cluster of features	B_{size}	1,2,4,5,10
Rejection Area	R_{win}	0.2%,0.4%,… 100%

- **Parameters- Synthetic data from real Microarray data:**

Parameter description	Parameter	Numeric values used
Sample size train	n	50, 100, 200
No. of Gaussians per class	G	1, 2
Rejection Area	R_{win}	0.2%,0.4%,… 100%
Mu and sigma	Calculated from real data	

Experimental Design:

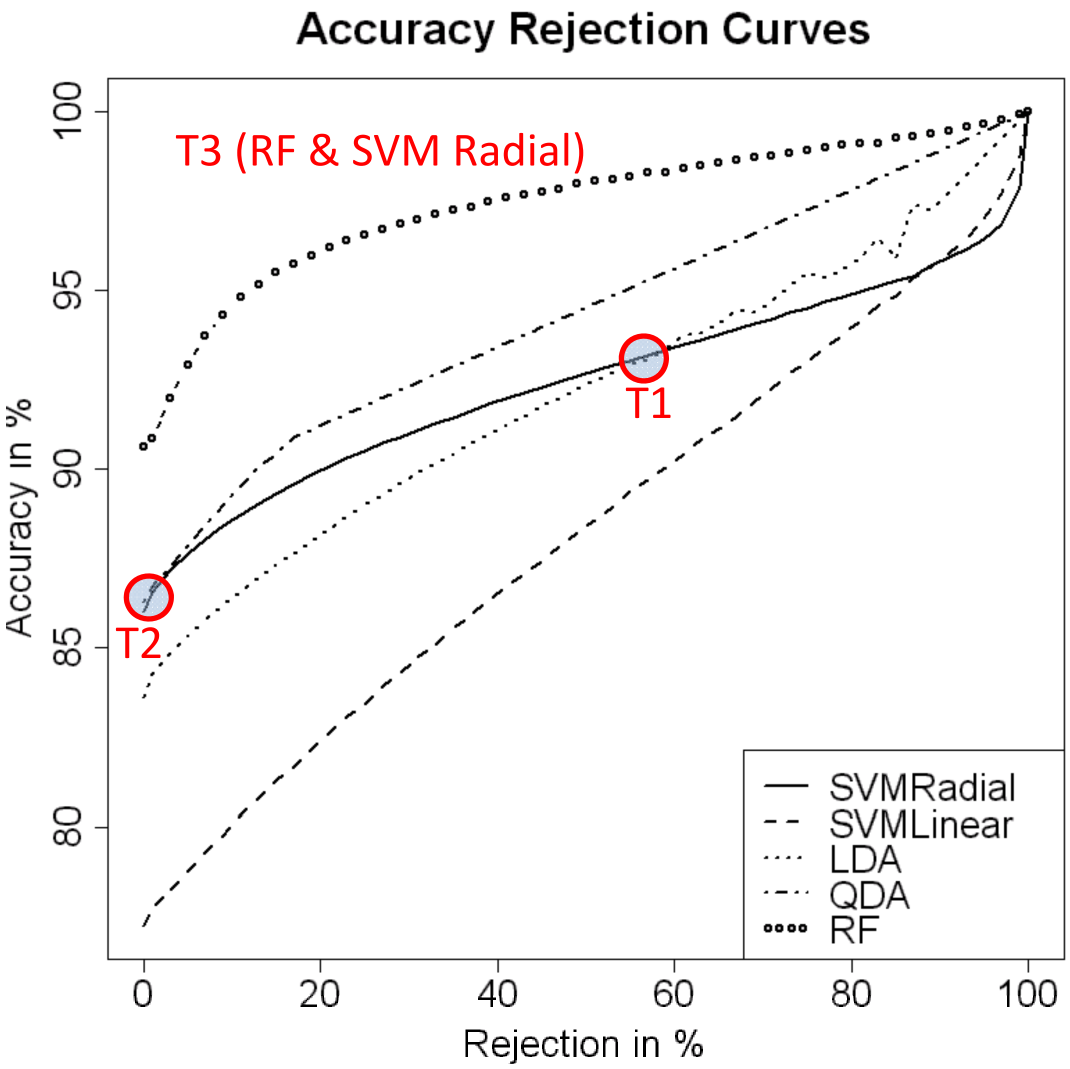


Results:
Synthetic Data from Colon Cancer

- Synthetic data from Colon Cancer.
- Gaussian per class = 5.
- Train =200
- Test= 10000
- Total Features = 400
- Noise Features = 390
- Noise free Features = 10
- Selected Features=40

From 3% abstention onwards QDA performs better than SVM Radial.

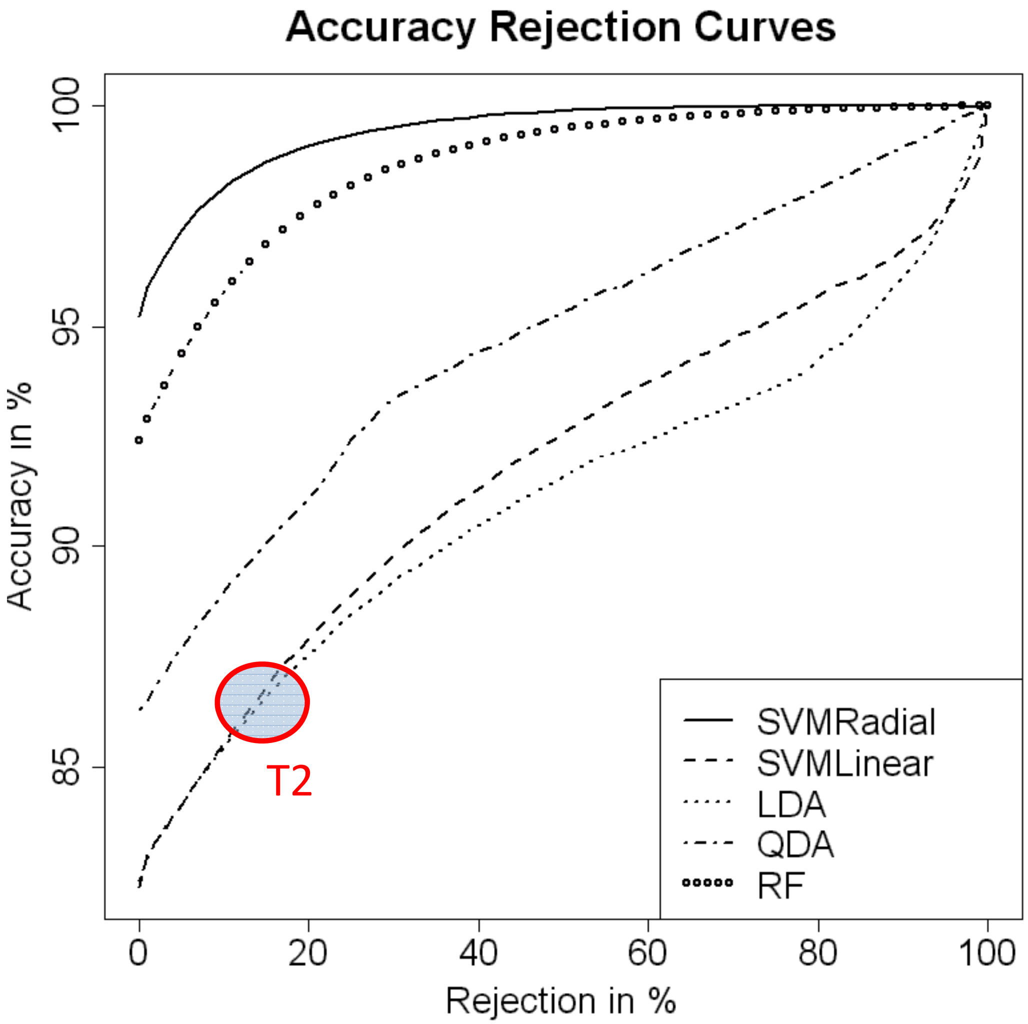
0 to 60% abstention SVM Radial performs better than LDA but after 60% vice versa.



Results: Synthetic data

- Non-Linear (SD2=SD1/2)
- Correlated Features
- Gaussians = 1
- Train = 100
- Test=10000
- Total Features = 400
- Noise Features = 380
- Noise free Features = 20
- Selected Features =20

From 19% abstention onwards SVM Linear performs better than LDA.

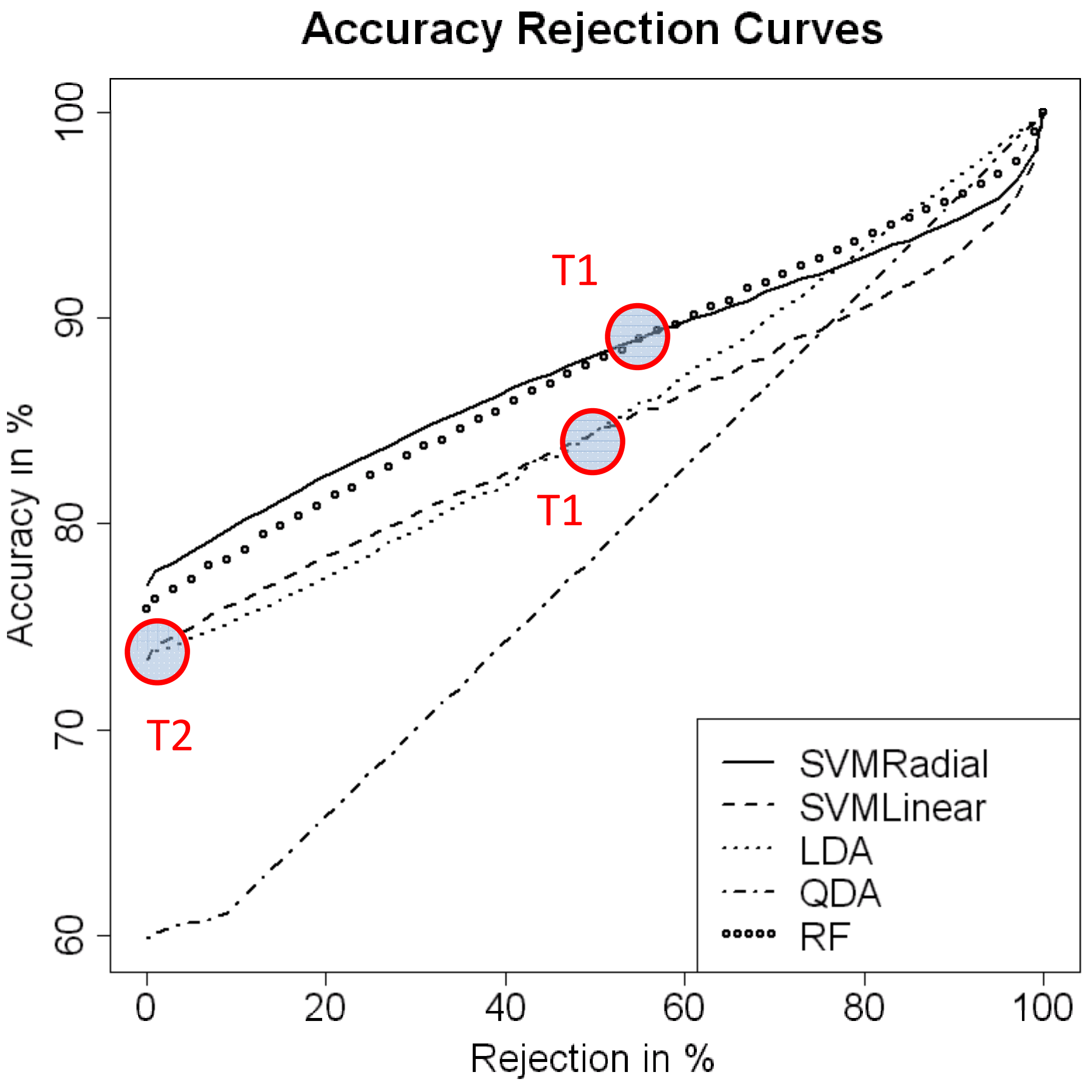


Results: Synthetic data

- Linear (SD1=SD1)
- Non Correlated Features
- Gaussians = 1
- Train = 50
- Test=10000
- Total Features = 400
- Noise Features = 380
- Noise free Features = 20
- Selected Features =20

From 58% abstention onwards LDA performs better than SVM Linear.

From 60% abstention onwards RF performs better than SVM Radial.



Results: Summary

Block Size (CF)	Train Size	No. of Gaussians			
		1		2	
		$\sigma_1 = \sigma_2$	$\sigma_2 = \sigma_1 / 2$	$\sigma_1 = \sigma_2$	$\sigma_2 = \sigma_1 / 2$
1	50	T1, T2	T3	T2	T2
	100	T1	T2	T1, T2	T1
	200	T1, T2	T2	T1, T2	T1, T2
2	50	T1, T2	T2, T3	T1, T2	T2
	100	T1, T2	T3	T1, T2	T1
	200	T1	T2	T1, T2	T2
4	50	T1, T2	T1, T2	T1, T3	T2
	100	T1, T2	T2	T1, T2	T1
	200	T2	T2, T3	T1, T2	T2, T3
5	50	T1, T2	T2	T1, T2	T1
	100	T1, T2	T2	T1, T2	T2
	200	T1, T2	T2, T3	T1, T2	T2
10	50	T2	T2	T1, T3	T1, T2
	100	T1, T2	T2	T1, T2	T2
	200	T1, T2	T1, T3	T1, T2	T2
No. Block (Non Correl)	50	T1, T2	T2, T3	T1, T2	T2
	100	T3	T2	T1, T2	T1
	200	T2, T3	T2	T2	T1, T2

Data	Train Size	No. of Gaussians		
		1	2	3
Golub	100	T3	T3	T3
	200	T2	T3	T3
Alon	100	T2	T2	T1
	200	T1, T2	T3	T1, T2
Shipp	100	T3	T3	xxx
	200	T3	T3	xxx

Exp. Types	Total Exp.	T1	T2	T3
PSD	72	40	59	12
PSD+SDR	90	43	64	22

xxx = Data N/A PSD = Pure Synthetic Data

SDR = Synthetic Data from Real Patients' data

Discussion & Conclusion:

- Obtaining T1,T2, T3 types of Accuracy-Rejection Curves may be beneficial in the selection of appropriate classification method for a given data.
- For a problem in hand, a measure (desired accuracy, acceptable rejection rate) should be known.
- For desired accuracy: move horizontally on ARCs plot and select the available classifier with least rejection rate.
- For fixed Rejection rate: Select the classifier with maximum prediction accuracy.
- Abstention considerably enhances prediction performance of some algorithms (LDA, KNN, RF) compared to others.

Future work:

- Experiments on real data
- Behavior of ARCs with Bagging , Boosting .
- ROC curves and ARC curves.

Questions

Thanks

Experimental Design:

1. Generate class-labeled train data {50, 100 or 200 examples}, test data {10000 examples} and a total of 400 features.
2. Apply t-test feature selection on train data and select 20 or 40 best features from train data and reduce train data to selected features.
3. Reduce test data to selected features.
4. Apply one of most widely used classification rule for microarray analysis to build a classification model based on train data.
5. Compute true error/rejection rates of the underlying model.
6. Repeat step 5 for all sizes of rejection windows {0.2; 0.4; 0.6; …100}
7. All steps 1-6 iterated 100 times.
8. Final result is averaged from all iterations.

For Correlated data:

$$\begin{bmatrix} \sum_{B_{size;\rho}} & 0 & \dots & 0 \\ 0 & \sum_{B_{size;\rho}} & \dots & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & \dots & \sum_{B_{size;\rho}} \end{bmatrix}.$$

$$\sum_{B_{size;\rho}} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \rho & \rho & \dots & 1 \end{bmatrix}. \qquad \rho = 0.5$$

Constants parameters(PSD):

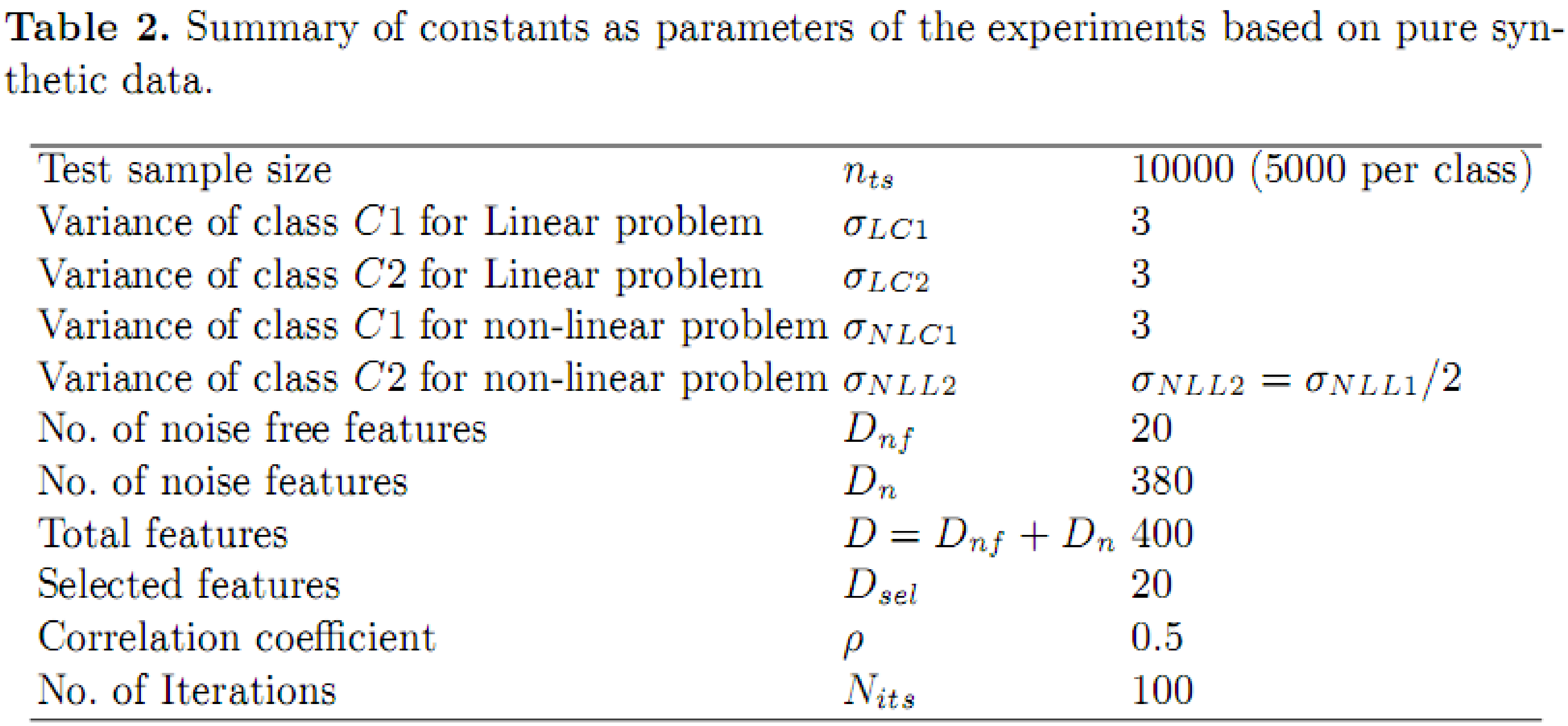


Table 2. Summary of constants as parameters of the experiments based on pure synthetic data.

Test sample size	n_{ts}	10000 (5000 per class)
Variance of class $C1$ for Linear problem	σ_{LC1}	3
Variance of class $C2$ for Linear problem	σ_{LC2}	3
Variance of class $C1$ for non-linear problem	σ_{NLC1}	3
Variance of class $C2$ for non-linear problem	σ_{NLL2}	$\sigma_{NLL2} = \sigma_{NLL1}/2$
No. of noise free features	D_{nf}	20
No. of noise features	D_n	380
Total features	$D = D_{nf} + D_n$	400
Selected features	D_{sel}	20
Correlation coefficient	ρ	0.5
No. of Iterations	N_{its}	100

Constants parameters(SDR):

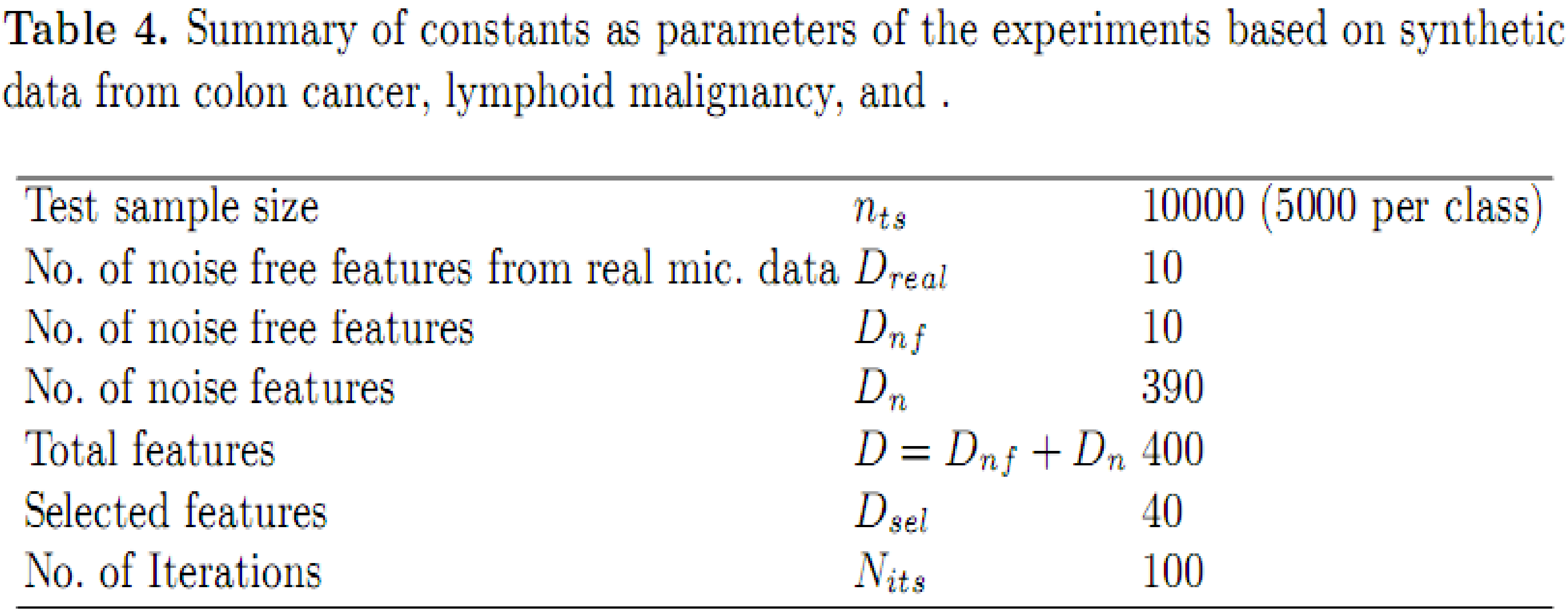


Table 4. Summary of constants as parameters of the experiments based on synthetic data from colon cancer, lymphoid malignancy, and .

Test sample size	n_{ts}	10000 (5000 per class)
No. of noise free features from real mic. data	D_{real}	10
No. of noise free features	D_{nf}	10
No. of noise features	D_n	390
Total features	$D = D_{nf} + D_n$	400
Selected features	D_{sel}	40
No. of Iterations	N_{its}	100

T-test score:

```
mC1 <- array of means of all features for class +1
sdC1 <- array of standard deviations of all features for class +1

mC2 <- array of means of all features for class -1
sdC2 <- array of standard deviations of all features for class -1


scores4AllFeature <- ( abs(mC1-mC2)/ (sdC1 + sdC2) )

sortedScores4AF <- sort (scores4AllFeature, decreasing=TRUE)
```