Chapter 3

MLFS: A Novel MultiLevel Gene/MiRNA Feature Selection Approach

Chapter 3
MLFS: A Novel MultiLevel Gene/MiRNA Feature Selection Approach

3.1 Introduction

In this chapter, our proposed approach Multilevel Feature Selection (MLFS) is described in detail. The description starts by presenting an overview of the proposed approach, then discussing each of its components in detail and finally presenting our system extension to deal with miRNA feature selection problem. Experimental results are presented by applying our proposed approach on six cancer types, namely, breast cancer, HCC, lung cancer, prostate cancer, colon cancer and ovarian cancer.

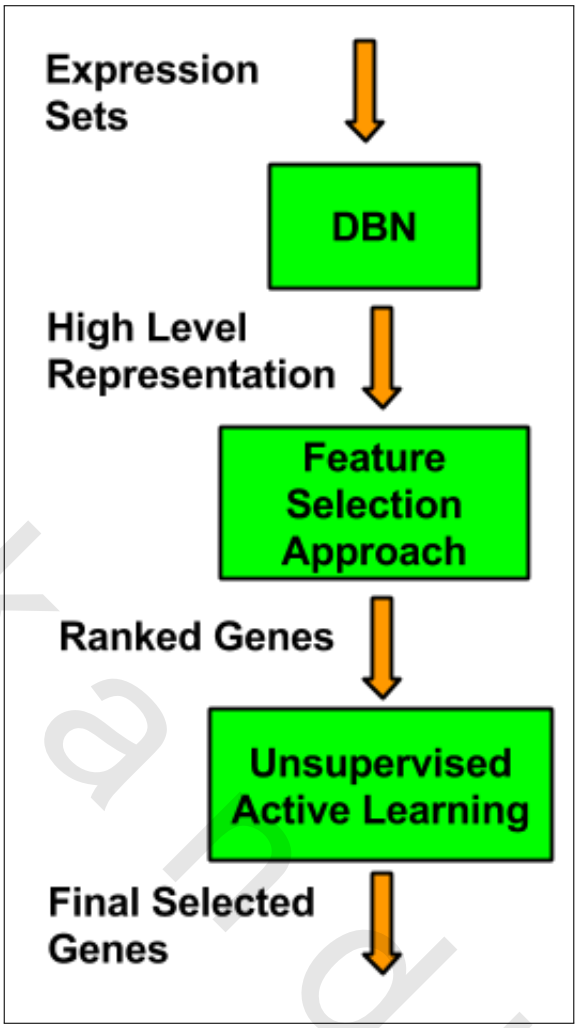

Figure 3.1: Multilevel Feature Selection Approach for Gene Expression Sets

3.2 Approach Overview

As shown in Figure 3.1, the proposed approach for gene feature selection is composed of three main components which are:

- Deep Belief Net: The DBN [20] is used to generate a high level representation of the genes that captures their interaction and behavior.
- Feature Selection: Feature selection is then applied on the high level rep-

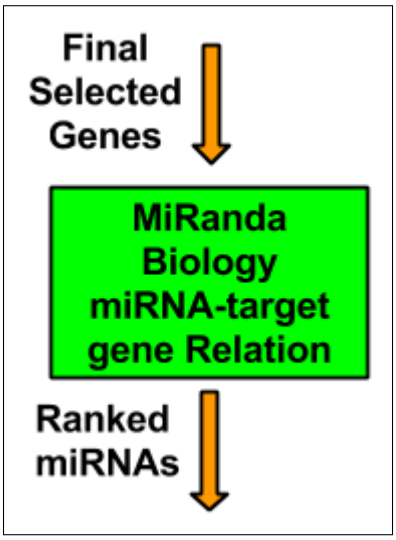

Figure 3.2: miRNAs Feature Selection based on Multilevel Gene Feature Selection Approach

resentation generated by the DBN to select a subset of the genes.

- Unsupervised Active Learning: Unsupervised active learning is then applied to reduce the subset of genes by selecting the least number of genes that increase or maintain the same accuracy of classification.

In addition, the approach is extended for miRNA feature selection as shown in Figure 3.2. The next subsections explain each component of the proposed approach in detail.

3.3 Deep Belief Nets (DBNs)

DBNs are defined as graphical models which can learn to extract a deep hierarchical representations of the training data. We use the training procedure proposed in [20] and the open source library implementation in [21]. The deep

learning library was used in this work with two hidden layers and the number of neurons at each layer is equal to the input layer original features.

3.4 Feature Selection

The feature selection step is applied on the high level features generated by the DBN. In this step, any appropriate feature selection method as statistical t-test [39] or relief-f [33] can be applied. Statistical t-test and relief-f were used in the experimental results section.

3.5 Unsupervised Active Learning

Traditional active learning [22] is a semi-supervised machine learning approach. It first trains a basic classifier using a small number of training data. Then, it selects the most informative data to the classifier and labels them according to a human. In order to select the most informative data, uncertainty sampling was explored and used in our experiments.

Active learning is used mainly when there is a lot of unlabeled data and we need to select the minimum number of training data to label and use in training the classifier, typically when the labeling cost is high or difficult to do. In this thesis, an extension for the traditional active learning is proposed by using active learning in an unsupervised manner. Its objective is to select the smallest subset of genes that will yield the same high classification accuracy. This is done by considering the group behavior of genes instead of just selecting

genes based on their individual scoring function.

3.6 The Overall MLFS Approach

The proposed multilevel feature selection approach is used to perform feature selection for sample classifiers that are used to discriminate cancer types/subtypes. Our objective is to enhance classification accuracy using the least number of genes. The steps of the approach are described as follows:

1. Use DBN to extract high level representations (e') of the gene expression profiles (e).

2. Apply any classical feature selection method on the high level representations and rank the genes based on their scores (e.g., their p-values in statistical t-test).

3. Let n denote an initial random number of genes to use, g denote the whole set of genes, λ denote a certain percentage, c denote a classifier, a denote the accuracy, $Step$ denote number of genes to add to n at each iteration and $bestGenes$ denote the best set of genes in terms of accuracy. Select the number of genes to use as follows:

   - $x = n$
   - $bestGenes$ = top ranked x genes
   - While($x <= \lambda * |g|$):
     - $c \leftarrow$ TrainClassifier(top ranked x genes, $trainingSet$)
     - $a \leftarrow$ MeasureAccuracy(top ranked x genes, $testSet$)

- If ($a > bestAccuracy$):
  - $bestGenes \leftarrow$ top ranked x genes.
  - $bestAccuracy \leftarrow a$
- $x \leftarrow x + Step$.

4. Apply a reduction phase to reduce the number of genes in the *bestGenes* set using the proposed unsupervised active learning approach to choose the most informative genes from the *bestGenes* set, as follows:
   (a) Label all genes in the *bestGenes* set as either related to the cancer type if they have p-value < 0.05. Otherwise, label them as not related.
   (b) Construct a gene classifier using a random small subset of *bestGenes*. The classifier is used to tell if the given gene is related to the cancer type or not, based on its expression profile.
   (c) Let the classifier label the *bestGenes* set. Then, get the genes that are most informative to the classifier, which have a classification confidence close to 0.5, for example between 0.4 and 0.6.
   (d) Label the most informative genes using statistical t-test. Human judgment is usually used at this stage. However, no human judgment with strong biological background was available to us. So, statistical t-test labels were used.
   (e) Add the most informative genes to the training set, re-train the classifier and go back to step c.
   (f) Record the accuracy using the test set and stop when reaching close to or higher than the one produced using all *bestGenes* genes.

3.7 MLFS-miRNA: MiRNAs Feature Selection Extension

For efficient use of DBN, it should be given a huge training set of unlabeled data. However, the number of miRNAs is usually very small. That is why a new extended approach was proposed instead of using the same multilevel feature selection method. Our approach is extended by utilizing the biological relation of miRNA-target gene, as in databases such as miRanda [4]. MiRNAs are ranked based on the number of their target genes that exist in final gene feature selection set generated by the previous active learning phase. Finally, the number of miRNAs to select is tuned in the same way described in section 3.6.

3.8 Experimental Results

The results of our experiments are given in four subsections to show the effect of each of the proposed components on the classification accuracy. SVM classification is used to construct a gene classifier in the unsupervised active learning phase and a Random Forests (RFs) classifier is used to construct the sample classifier. The SVM and RF implementations were both used from Weka repository [40]. The evaluation measures used are:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp}$$

where tp is the true positive (number of samples correctly classified as positive class), fp is the false positive (number of samples predicted to be positive class, while they belong to the negative class), fn is the false negative (number of samples predicted to be negative class, while they belong to the positive class) and tn is the true negative (number of samples correctly classified as negative class). The code of our system was implemented in Java.

Table 3.1: Training and testing sample size using gene expression

Type	**BC (GSE20713)**		**BC (GSE20713)**		**HCC (GSE36 376)**		**Lung (GSE4 1271)**	
Sub Type	**ER+**	**ER-**	**HER2+**	**HER2-**	**NM**	**M**	**A**	**S**
Train	21	23	31	13	98	118	87	41
Test	21	22	31	12	98	120	86	40

Table 3.2: Training and testing samples size using miRNA expression

Type	**Breast Cancer (GSE15885)**		**Breast Cancer (GSE15885)**		**HCC (GSE6857)**	
Sub Type	**ER+**	**ER-**	**HER2+**	**HER2-**	**NM**	**M**
Train	9	7	13	3	193	62
Test	7	6	12	2	162	65

3.8.1 MLFS Evaluation

The first two phases of the proposed approach (will be referred to as the DBN) are first compared to five feature selection methods in three cancer types, namely, breast cancer, HCC and lung cancer. Tables 3.1 and 3.2 show the sizes of the datasets. As shown in the tables, the breast cancer (BC) set contains four subtypes which are Estrogen Receptor Positive (ER+), Estrogen Receptor Negative (ER-), Human Epidermal Growth Factor Receptor 2 Positive (HER2+) and Human Epidermal Growth Factor Receptor 2 Negative (HER2-). ER+ breast cancer means that the cancer has receptors for estrogen, while ER- means that no receptors are present. HER2+ breast cancer means that the cancer is associated with HER2 gene amplification or HER2 protein over expression, while HER2- means the opposite. In addition HCC and lung cancer sets contain two subtypes, which are Non-Metastatic (NM) and Metastatic (M) in HCC and Adenocarcinoma (A) and Squamous (S) in lung cancer. The used feature selection methods in the comparison are random (by randomly selecting k genes), statistical t-test, information gain, relief-f and chi-square, which are applied on the original expressions. Statistical t-test was used from [41] while information gain, chi-square and relief-f were used from Weka repository [40]. F-measure was used to compare the DBN approach to the classical approaches. As shown in Figures 3.3, 3.4, 3.5 and 3.6, the DBN approach was able to achieve the highest F-measure among all other methods using the least number of genes.

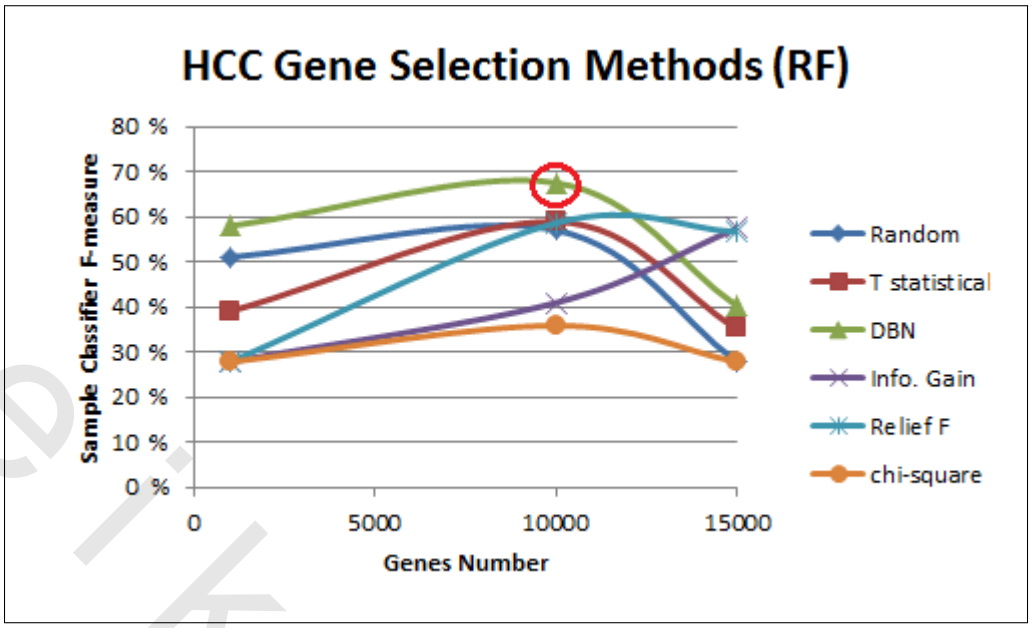

Figure 3.3: Comparison to Classical Feature Selection using Gene Expression Sets in HCC (Highest value is marked with a red circle)

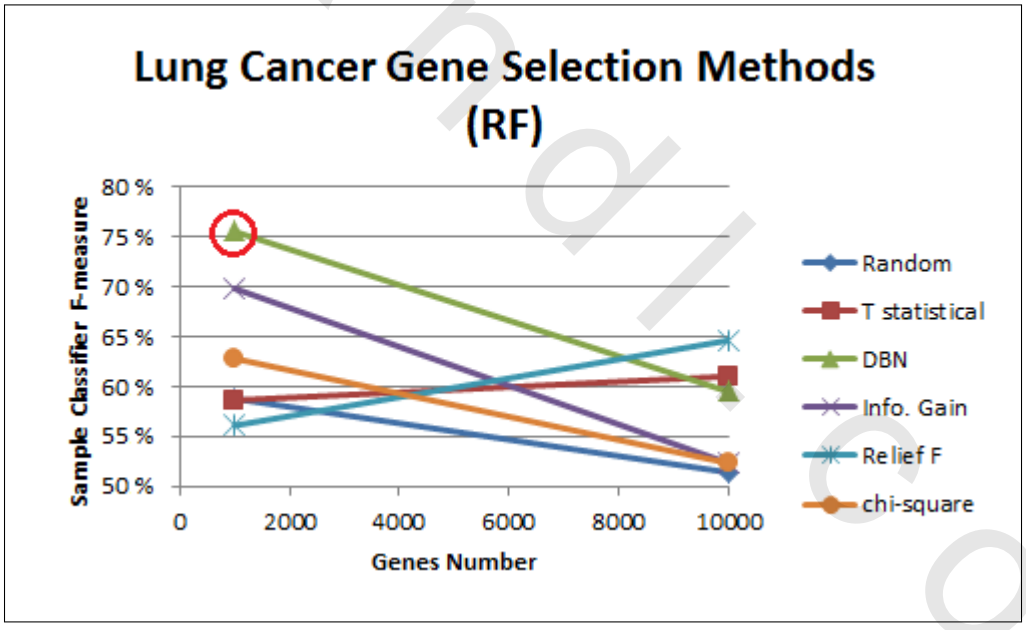

Figure 3.4: Comparison to Classical Feature Selection using Gene Expression Sets in Lung Cancer (Highest value is marked with a red circle)

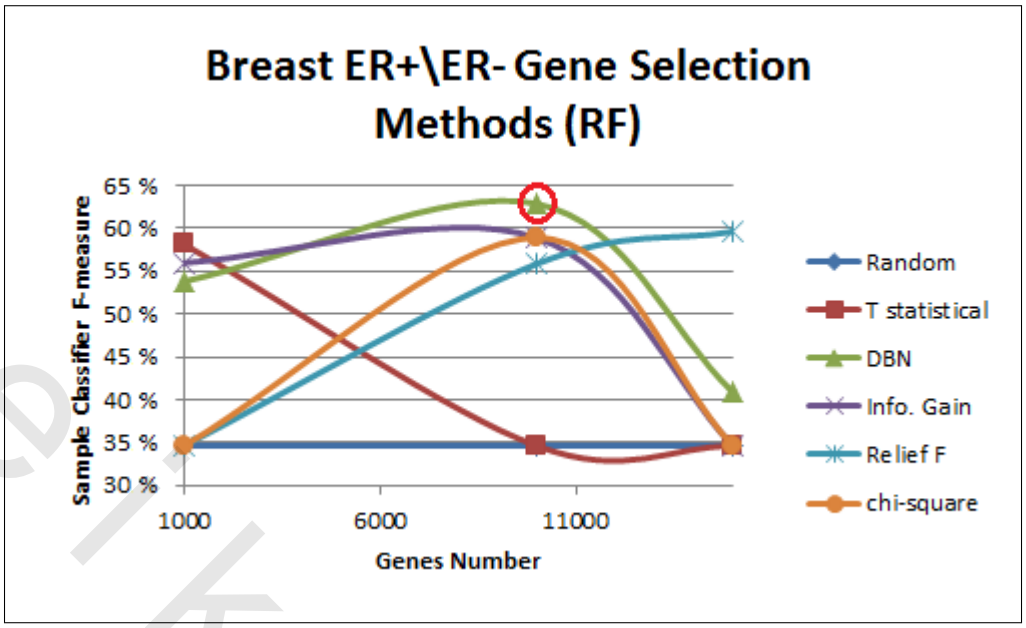

Figure 3.5: Comparison to Classical Feature Selection using Gene Expression Sets in Breast Cancer ER+/ER- (Highest value is marked with a red circle)

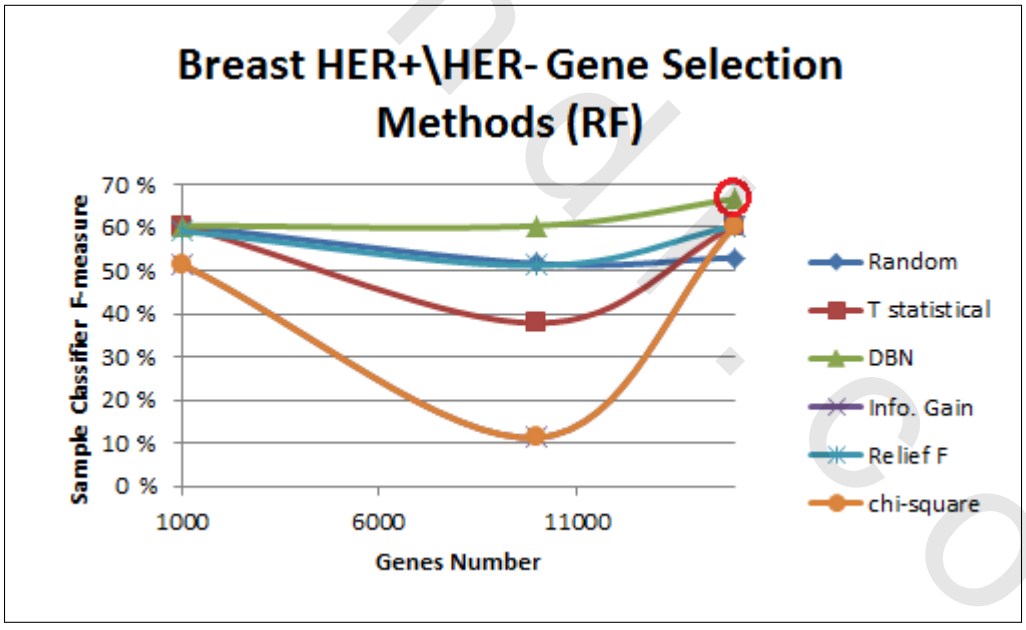

Figure 3.6: Comparison to Classical Feature Selection using Gene Expression Sets in Breast Cancer HER2+/HER2- (Highest value is marked with a red circle)

3.8.2 Unsupervised Active Learning

Unsupervised active learning was used to reduce the *bestGenes* set, which was obtained from the first two phases of our approach. Tables 3.3 and 3.4 show the results after applying the active learning phase. The tables show that active learning was able to reduce the *bestGenes* set by 60% in lung cancer, 20% in HCC and 50% in Breast Cancer ER+/ER- while increasing or at least maintaining the same classification accuracy.

Table 3.3: HCC and lung cancer active learning results

	Precision	**Recall**	**F1-measure**
HCC Baseline (1k)	51.70%	50.46%	51.07%
HCC AL Iteration 1 (6k)	66.23%	59.17%	62.5%
HCC AL Iteration 2 (8k)	**78.73%**	**59.63%**	**67.86%**
Lung Baseline (100)	53.89%	59.84%	56.71%
Lung AL Iteration 1 (200)	50.37%	55.12%	52.64%
Lung AL Iteration 2 (400)	**79.75%**	**71.65%**	**75.48%**
Lung AL Iteration 3 (600)	60.08%	69.29%	64.6%

Table 3.4: Breast cancer ER+/ER- and HER2+/HER2- active learning results

	Precision	**Recall**	**F1-measure**
ER+/ER- Baseline (1k)	54.10%	53.49%	53.79%
ER+/ER- AL Iteration 1 (3k)	23.85%	48.80%	32.04%
ER+/ER- AL Iteration 2 (5k)	**78.15%**	**60.47%**	**68.18%**
HER2+/HER2- Baseline (1k)	51.97%	72.09%	60.40%
HER2+/HER2- Iteration 1 (7k)	51.97%	72.09%	60.40%
HER2+/HER2- Iteration 2 (13k)	51.97%	72.09%	60.40%
HER2+/HER2- AL Iteration 3 (14k)	**66.70%**	**72.09%**	**69.29%**

3.8.3 MLFS-miRNA Evaluation

The same five feature selection methods used for gene expression sets are applied to miRNA expression profiles and compared to our approach MLFS-miRNA. Moreover, the same five feature selection methods are applied differently by using the biology relation. The five feature selection methods are first applied directly to the genes expression sets instead of the miRNA expression sets and then miRNAs are ranked based on the number of their target genes that exist in gene feature selection set.

Figures 3.7, 3.8, 3.9, 3.10, 3.11 and 3.12 show a detailed comparison of miRNA extension approach and the five classical feature selection methods either applied directly on miRNA expression sets or applied on gene expression

sets and then using the biology relation to select. As shown in the figures, miRNA extension has performed better in all curves except for the last curve of the breast cancer HER2+/HER2- where it was outperformed by chi-square and information gain. In spite of that, the experimental results in general raise an important issue about the potential benefit of using miRNA feature selection methods that rely on the biology relations between genes and miRNAs since it was shown in all curves that it enhances the classification F-measure. Moreover, this idea was not explored before in literature.

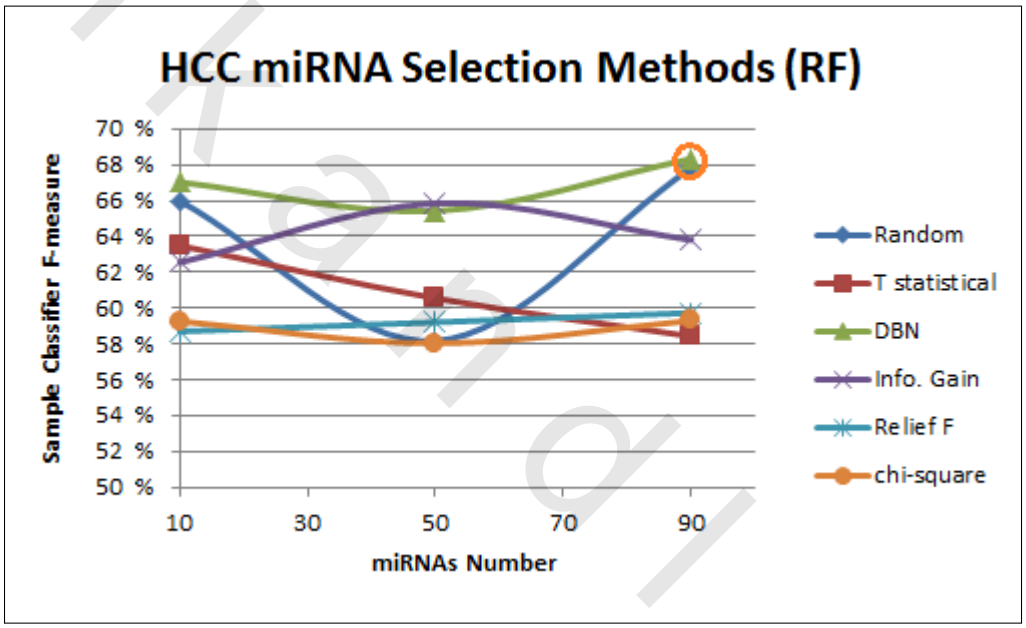

Figure 3.7: Comparison to Classical Feature Selection using MiRNA Expression Sets in HCC part 1 (Highest value is marked with a red circle)

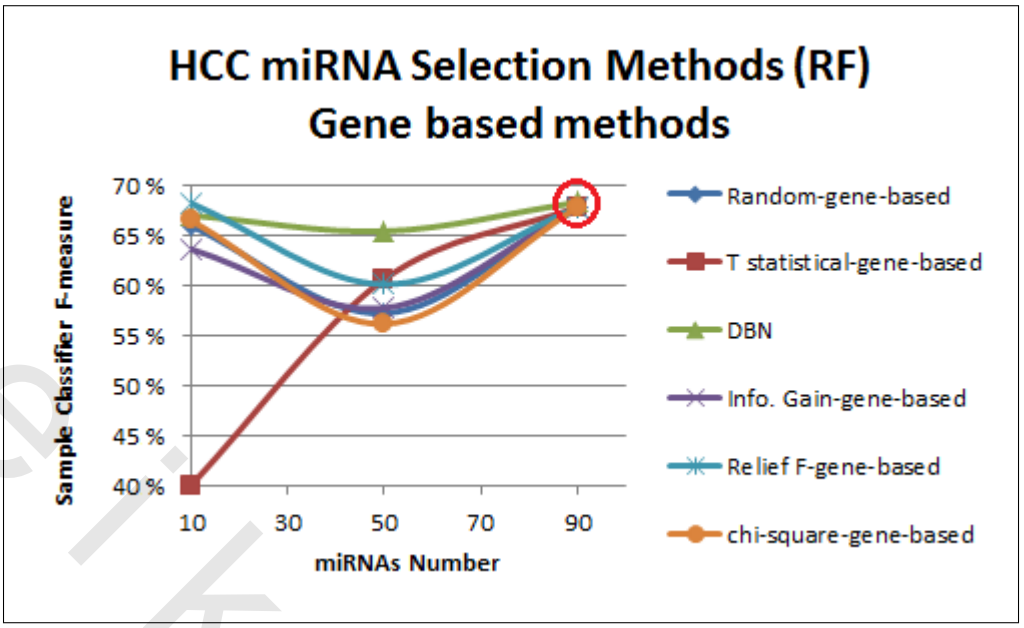

Figure 3.8: Comparison to Classical Feature Selection using MiRNA Expression Sets in HCC part 2 (Highest value is marked with a red circle)

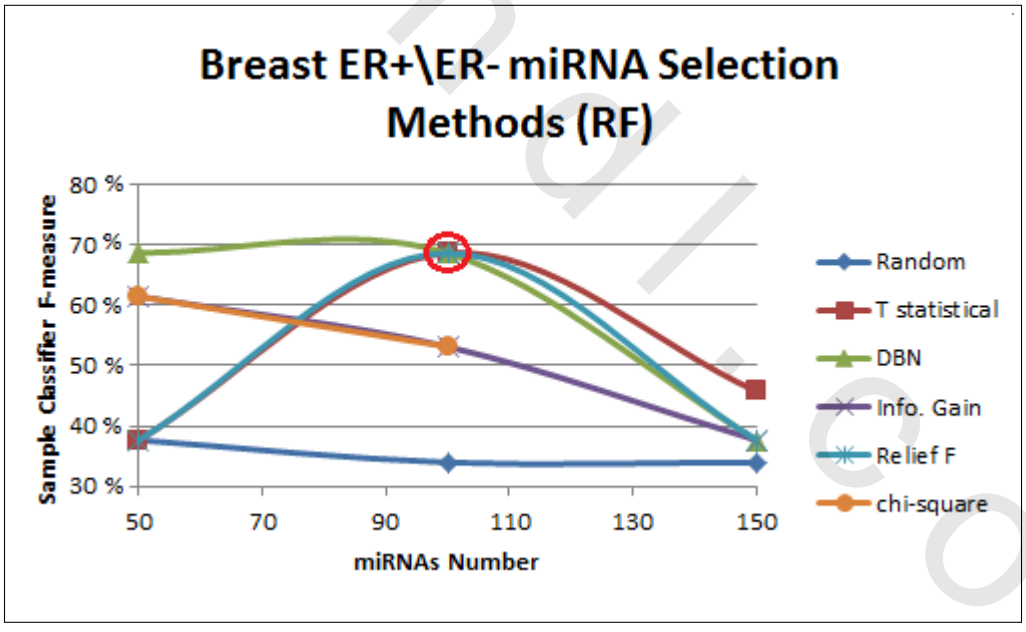

Figure 3.9: Comparison to Classical Feature Selection using MiRNA Expression Sets in Breast Cancer ER+/ER- part 1 (Highest value is marked with a red circle)

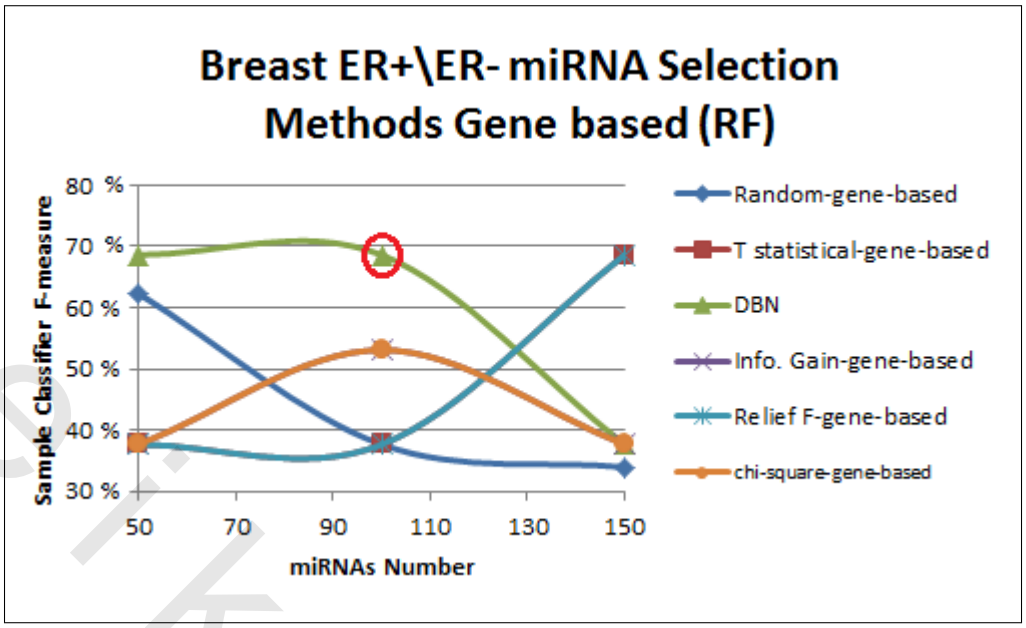

Figure 3.10: Comparison to Classical Feature Selection using MiRNA Expression Sets in Breast Cancer ER+/ER- part 2 (Highest value is marked with a red circle)

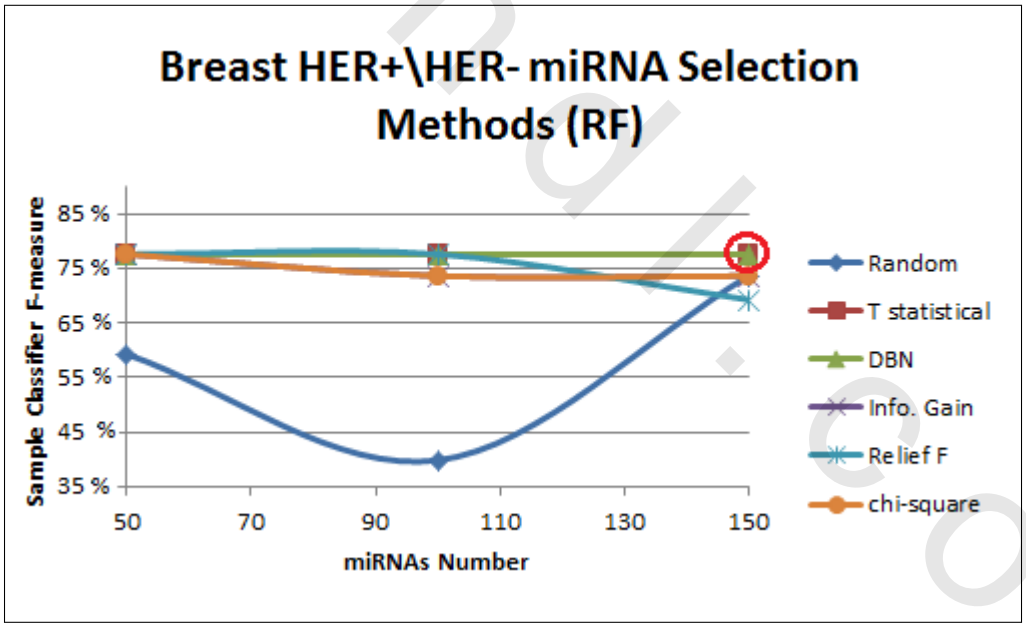

Figure 3.11: Comparison to Classical Feature Selection using MiRNA Expression Sets in Breast Cancer HER2-/HER2+ part 1 (Highest value is marked with a red circle)

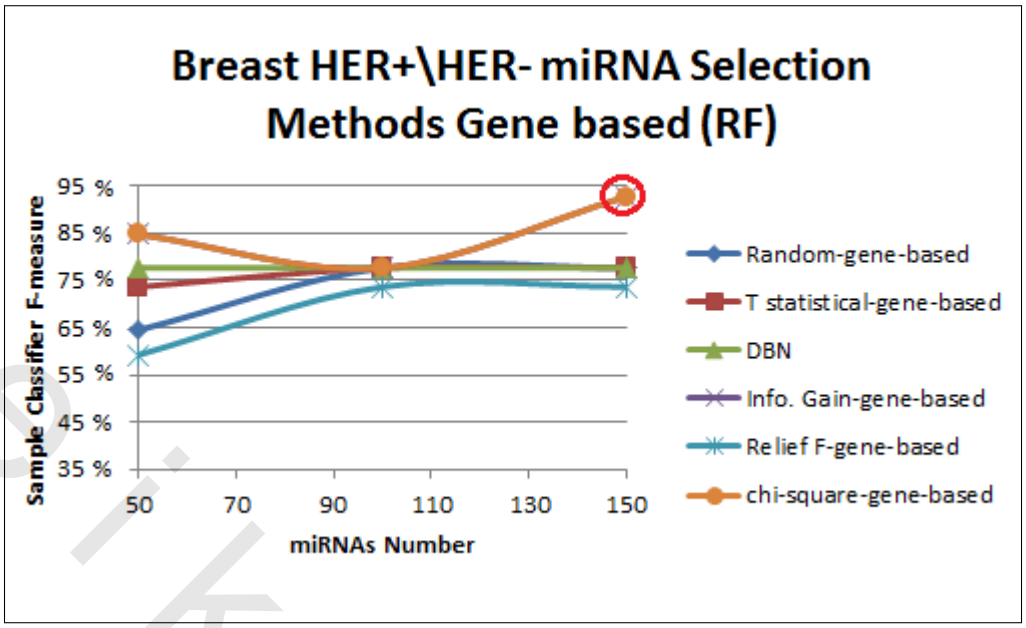

Figure 3.12: Comparison to Classical Feature Selection using MiRNA Expression Sets in Breast Cancer HER2-/HER2+ part 2 (Highest value is marked with a red circle)

Table 3.5: Dataset sizes for the related work in [1] and [2]

Dataset Name	Number of Genes	Training Samples	Testing Samples
Prostate Cancer	12600	102	34
Colon Cancer	2000	32	30
Ovarian Cancer	15154	153	100
SRBCT	2308	63	20
MLL	12582	57	15

3.8.4 Comparison to Related Work

Table 3.5 shows the datasets sizes used to compare our work to the related work in [1] and [2]. Table 3.6 shows the 10-fold cross-validation classification accuracy results compared to [1]. It is to be noted that two datasets, namely, Small Round Blue Cell Tumors (SRBCT) and MLL [42], were used to compare our approach to the approach proposed in [2]. As statistical t-test is limited to 2 classes, it was replaced by relief-f feature selection in Weka [40]. Tables 3.7 show the 10-fold cross validation results.

Table 3.6: Comparison to [1] using 10-fold cross-validation

	Stack Autoencoder	**Stacked Autoencoder with Fine Tuning**	**DBN (200 genes)**
Prostate Cancer	73.33%	73.33%	**97.06%**
Colon Cancer	66.67%	**83.33%**	73.33%
Ovarian Cancer	55.03%	99.00%	**100%**

Table 3.7: Comparison to [2] using 10-fold cross-validation

		DBN using top-4 genes in *bestGenes* set	**[2]**
SRBCT	**Precision**	**85.3%**	74.4%
	Recall	**84.1%**	73.0%
	F1-measure	**84.7%**	73.7%
MLL	**Precision**	**77.9%**	64.3%
	Recall	**66.7%**	**66.7%**
	F1-measure	**71.8%**	65.5%

3.9 Conclusion

This chapter has discussed our proposed feature selection approach MLFS in detail. First by giving an overview of the proposed approach, then describing each of its components and finally presenting its extension to be used for miRNA feature selection task. Experimental results were presented by applying our proposed approach on six cancer types, namely, breast cancer, HCC, lung cancer, prostate cancer, colon cancer and ovarian cancer. The results of MLFS system showed that it outperforms classical feature selection methods in terms of F1-measure by 9% in HCC, 6% in lung cancer and 10% in breast cancer. Moreover, it outperforms the recently related work in [1] and [2].