Chapter 4

miRNA and Gene Expression-based Cancer Classification using Semi-supervised Learning Techniques

Chapter 4
miRNA and Gene Expression-based Cancer Classification using Semi-supervised Learning Techniques

4.1 Introduction

In this chapter, our proposed approaches which are miRNA and gene expression-based cancer classification using self-learning and co-training techniques are described in detail. The discussion starts by presenting an overview of the proposed approaches and then discussing the adaptation of self-learning and co-training techniques to our problem in detail.

4.2 Approach Overview

In our adaptation to self-learning and co-training, the objective is to construct a classifier to discriminate between different cancer subtypes, given the following:

- The expression vector of a sample i, denoted by x_i, which is defined as follows:

  $x_i = \{e_{i1}, e_{i2}, \dots, e_{ij}, \dots, e_{iM}\}$

  Where e_{ij} is the expression value of the jth gene/miRNA, and M is the number of genes/miRNAs.

- N is the number of samples.

Two sets are used in both self-learning and co-training, which are defined as follows:

- A set of labeled samples L; $L=\{x_i, y_i\}_{i=1}^{N}$, where y_i is the cancer subtype label.
- A set of unlabeled samples U; $U=\{x_i\}_{i=1}^{N}$

The size of U is expected to be much larger than L ($|U| >> |L|$), which is expected to help enhancing the accuracy of the classifiers by adding more expression vectors to the training data. Increasing the number of unlabeled sets leads to higher enrichment in the training set. Moreover, increasing the overlap between the genes/miRNAs in the labeled and unlabeled sets leads to increasing the effect of adding the unlabeled sets.

The next subsections explain how the self-learning and co-training approaches are adapted to use the unlabeled set U to enhance the baseline classifier constructed based on the labeled set L.

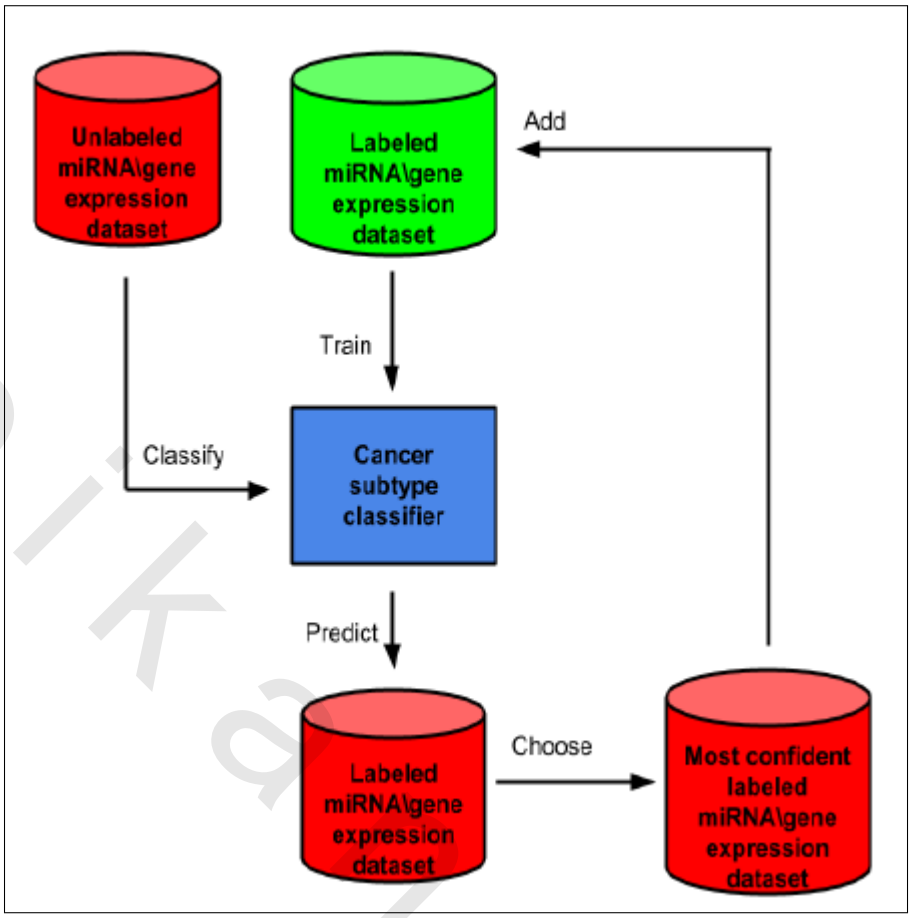

Figure 4.1: Self-Learning Approach Overview

4.3 Self-Learning Adaptation

Figure 4.1 shows an overview of the self-learning approach. The steps of adapting the self-learning approach [25] are described as follows:

1. Train an initial cancer subtype classifier using the labeled set L.
2. Use the initial classifier to identify the subtype labels of the unlabeled set U.
3. Choose the most confident subset of cancer samples (U'), i.e. samples classified with a confidence greater than a given threshold (α).

4. Append the set of most confident samples to the initial training dataset to form a new training set ($U' \bigcup L$) for re-training the classifier.

5. Use the classifier constructed at step 4 to perform several iterations over the unlabeled set(s). At each iteration, re-apply steps 2, 3 and 4.

The resulting classifier can then be used to classify new samples based on their gene/miRNA expression profiles. The confidence threshold α should be appropriately selected. Decreasing α can increase the false positives rate. On the other hand, increasing α can result in restricting the learning process to the highly confident samples, typically the ones that are most similar to the training data, thus losing the benefit of including more training samples to the labeled data. Tuning parameter α is thus important, since it affects the classifiers accuracy to choose the samples that will enhance the classifier. The next subsection explains the co-training idea and adaptation in details.

4.4 Co-Training Adaptation

The co-training approach [28] and [29] is adapted to classify cancer subtypes by training two different classifiers; the first is based on the gene expression view and the second is based on the miRNA expression view. Each view captures a different perspective of the underlying biology of cancer and integrating them using the co-training pipeline exploits this information diversity to enhance the classification accuracy. The following steps describe co-training in details:

1. Two initial cancer classifiers are separately constructed; one from the

miRNA expression dataset (L_{miRNA}) and the other one from the gene expression dataset (L_{gene}) using manually labeled cancer subtypes sets.

2. Let the initial classifiers separately classify the unlabeled cancer gene/miRNA expression datasets (U_{miRNA}/ U_{gene}) into cancer subtypes.

3. Choose the most confident labeled subtypes samples ($L\alpha_{miRNA}$& $L\alpha_{gene}$) that have classification confident scores greater than α.

4. Retrieve miRNA-gene relations using miRanda. For the classifiers to train each other, miRNA expression should be mapped to gene expression and vice versa. miRNAs and their target genes databases are used to map the datasets. In our case, miRanda [4] database is used.

5. Append the mapped miRNA expression sets to the gene expression training sets and the mapped gene expression sets to the miRNA expression training sets and re-train the classifiers.

6. Use the classifier constructed at step 5 to perform several iterations over the unlabeled set(s). At each iteration, re-apply steps 2, 3, 4 and 5.

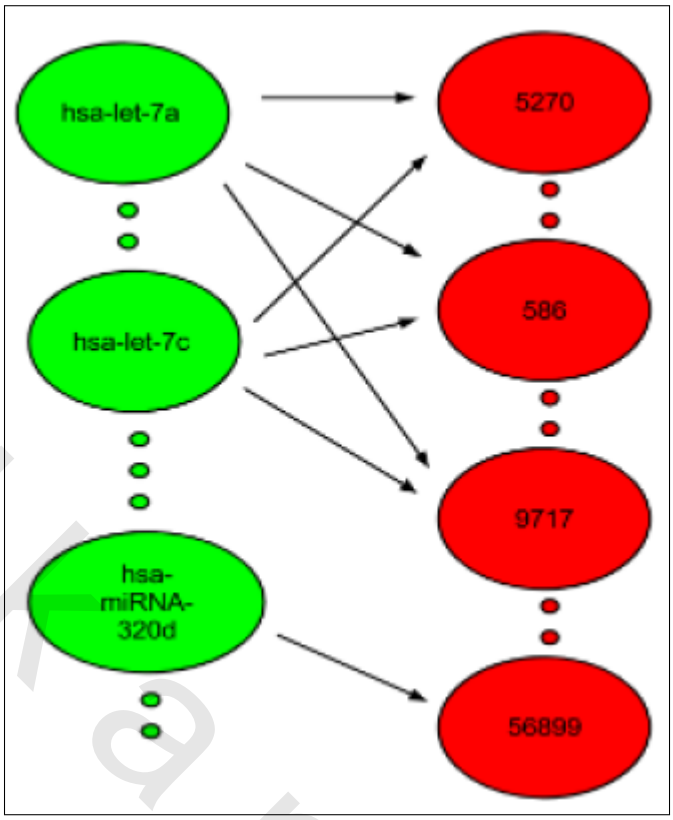

Figure 4.2: miRNAs and Their Target Genes are Related by a Many to Many Relationship. The first column represents miRNAs and the second column represents target genes ids.

In step 4, a mapping between the miRNA view and gene view is required. As shown in Figure 4.2, miRNAs and their target genes are related by a many to many relationship; multiple miRNAs target the same gene, and multiple genes are targeted by the same miRNA. For the classifier to exploit the two views, i.e. gene and miRNA sets, a miRNA expression vector is constructed from its target genes expression vector. Due to the many to many relationship between miRNAs and genes, it is suggested to use an aggregation of all expression vectors of the target genes to represent the miRNA expression vector. Similarly, a gene expression vector is constructed by aggregating the expression vectors of the miRNAs that target this gene. To map a gene to a miRNA, or the op-

posite, it is proposed to take the mean expression value of all miRNAs related to a gene, or the opposite, i.e. the mean expression value of all genes related to a miRNA. Experimental results show that taking the mean value of expressions has improved the classification accuracy. Part of the future work would be investigating the effect of using other methods as a mapping function.

After the co-training process, the two classifiers can be used independently, one on gene expression profile and the other on miRNA expression profile of cancer samples. Algorithm 1 shows the pseudo code of the co-training approach.

Inputs: miRNA expression profile of labeled set (L_{miRNA}), miRNA expression profile of unlabeled sets (U_{miRNA}), gene expression profile of labeled set (L_{gene}), gene expression profile of unlabeled sets (U_{gene}) and confident threshold (α).

Outputs: Two classifiers (C_{miRNA}) and (C_{gene}) that can separately classify cancer samples.

Begin

Repeat

i. C_{miRNA} = TrainClassifier (L_{miRNA})

ii. C_{gene} = TrainClassifier (L_{gene})

iii. L'_{miRNA} = Classify (C_{miRNA}, U_{miRNA})

iv. $L\alpha_{miRNA}$= ChooseMostConfident(L'_{miRNA}, α)

v. L'_{gene} = Classify (C_{gene}, U_{gene}).

vi. $L\alpha_{gene}$= ChooseMostConfident(L'_{gene}, α).

vii. $L_{miRNA} = L_{miRNA} \bigcup$ ConverToMiRNAs($L\alpha_{gene}$)

viii. $L_{gene} = L_{gene} \bigcup$ ConverToGenes($L\alpha_{miRNA}$)

Until (no improvement in accuracy OR reaching max iterations);

End

Algorithm 1: Pseudo code of co-training approach

4.5 Experimental Results

Two core classifiers of self-learning and co-training were used, which are Random Forests (RFs) and SVM. RF is a known classifier ensemble method [43] based on constructing multiple decision trees. Each decision tree is built on a

bootstrap sample of the training data using a randomly selected subset of features. For predicting a sample's label, a majority vote based on the classification obtained from the different decision trees is calculated. RF has been used in classifying cancer in [44], [45] and [46]. RF implementation from the Weka repository [40] was used, and the number of decision trees was set to 10. SVM implementation was also used from the Weka repository [40].

The approaches were evaluated using three cancer types, namely, breast cancer, HCC and lung cancer. MiRNA based classifiers were constructed for breast cancer and HCC sets, while gene based classifiers were constructed for all 3 sets. In addition, self-learning and co-training were compared to LDS used in [25] for breast cancer and HCC. LDS Matlab implementation was downloaded from [47]. Tables 4.1 and 4.2 show the size of the training and testing sets for each cancer type according to its subtypes. All miRNA and gene expression profiles were downloaded from NCBI [48]. Moreover, table 4.3 shows sample size and gene/miRNA numbers in the unlabeled sets. The evaluation measures used are precision, recall, f-measure and accuracy. The definitions of these measures where given in section 3.8. The code of our system was implemented in Java.

Table 4.1: Training and testing samples size for breast cancer and HCC subtypes using miRNA expression

Type	**Breast Cancer (GSE15885)**				**HCC (GSE6857)**	
Sub Type	**ER+ /HER2-**	**ER- /HER2+**	**ER- /HER2-**	**ER- /HER2+**	**NM**	**M**
Train	8	2	5	1	193	62
Test	7	2	4	0	162	65

Table 4.2: Training and testing sample size for breast cancer, HCC and lung cancer subtypes using gene expression

Type	**Breast Cancer (GSE20713)**				**HCC**		**Lung**	
Sub Type	**ER+ /HER2-**	**ER- /HER2+**	**ER- /HER2-**	**ER- /HER2+**	**NM**	**M**	**A**	**S**
Train	17	9	14	4	98	118	87	41
Test	17	8	14	4	98	120	86	40

Table 4.3: Sample size and gene/miRNA numbers of unlabeled sets

	Sample Size	**gene/miRNA Numbers**
Breast Cancer (GSE26659)	94	237 miRNAs
Breast Cancer (GSE35412)	35	378 miRNAs
Breast Cancer (GSE16179)	19	54675 genes
Breast Cancer (GSE22865)	12	54675 genes
Breast Cancer (GSE32161)	6	54675 genes
HCC (GSE10694)	78	121 miRNAs
HCC (GSE15765)	90	42718 genes
Lung Cancer (GSE42127)	176	48803 genes

4.5.1 Breast Cancer

Breast cancer is a heterogeneous disease that has a range of phenotypically distinct tumor types. This heterogeneity has an underlying spectrum of molecular alterations and initiating events that was found clinically through a diversity of disease presentations and outcomes [49]. Due to the complexity of this type of tumor, there is a demand for an efficient classification of breast cancer samples.

For breast cancer, both adapted self-learning and co-training are used. Self-learning was applied for both miRNA and gene based classifiers. For sample classification using miRNA expression dataset, an initial breast cancer subtype labeled dataset (GSE15885) was used to build an initial cancer subtype classifier. The initial classifier was then used to predict the labels of the unlabeled breast

cancer subtypes (GSE26659 and GSE35412).Two iterations were performed over the two unlabeled datasets. The confident samples; the ones with classification confidence (α) greater than 0.9, were added to the training dataset and the subtype classifier was re-trained. The same operation was repeated for sample classification using gene expression dataset where the initial dataset (GSE20713) was used to build an initial classifier and the unknown subtype breast cancer (GSE16179) was used to enrich it. Table 4.4 shows the precision, recall and F1-measure enhancement against the RF classifier. The results show 12% improvement in F1-measure of breast cancer subtype classifier using miRNA expression profiles and 6% improvement in F1-measure of breast cancer subtype classifier using gene expression profiles. Moreover, table 4.5 shows the enhancement over SVM and LDS classifiers, only miRNA expression profiles were used in this comparison as LDS requires a lot of memory and thus we could not use it with large number of genes. The table shows that self-learning achieved 10% improvement in F1-measure over SVM classifier and 4% improvement in F1-measure over LDS classifier.

Co-training was evaluated for breast cancer subtypes in both miRNA expressions and gene expressions. To enhance miRNA expressions and gene expressions sample classification using miRNA expression, one labeled miRNA expression dataset (GSE15885) is used. One labeled gene expression dataset (GSE20713) and three unlabeled gene expression datasets (GSE16179, GSE22865 and GSE32161) are mapped into miRNA expression values. In addition, to enhance sample classification using gene expression, one labeled gene expression dataset (GSE20713) is used. One labeled miRNA expression dataset (GSE15885) and two unlabeled miRNA expression datasets (GSE26659 and GSE35412) are mapped into gene expression values and added to the gene training dataset.

Table 4.4 shows the significant improvements in F1-measure using co-training over RF classifier. Improvements up to 21% and 8% in F1-measure are observed when using miRNA expression profiles and gene expression profiles respectively. Moreover, table 4.5 shows the enhancement of co-training over SVM and LDS classifiers; co-training was able to enhance the F1-measure by around 25% compared to the LDS classifier.

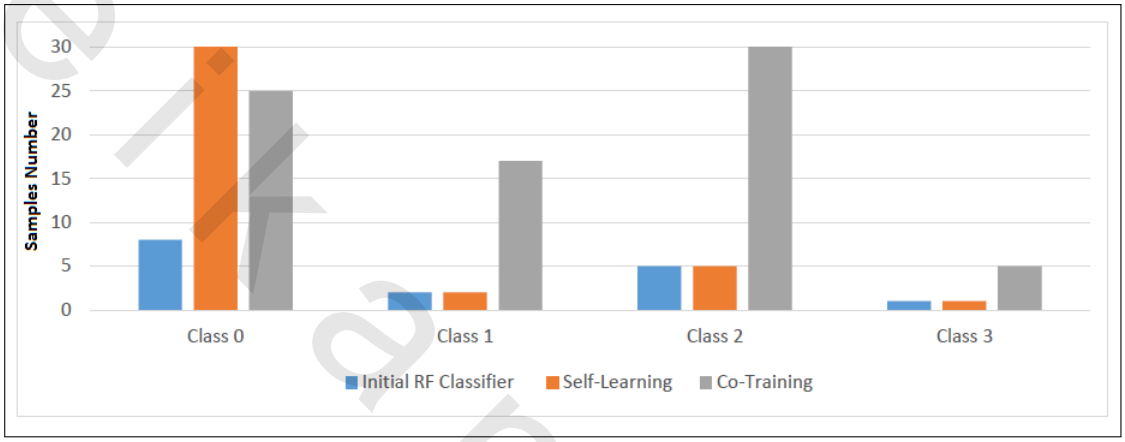

Figure 4.3: Training data size comparison of initial RF classifier compared to adapted self-learning and co-training for breast cancer using miRNA expression sets. (Class 0 is ER+/Her2-, class 1 is ER-/Her2+, class 2 is ER-/Her2- and class3 is ER+/Her2+)

To have a closer look on the behavior of the methods, the size of training data at each class is determined and shown at Figure 4.3. The figure shows that co-training was able to enrich the training data for all 4 classes which is reflected in the highest improvement in the results and self-learning was able to enrich that training set in class 0.

Moreover, table 4.8 shows that the higher the overlap between the miRNAs and genes of the initial set and those of the added sets, the higher the improvements become in breast cancer.

Table 4.4: Precision, recall and F1-measure for breast cancer subtypes RF classifiers using miRNA expression and gene expression dataset

Type	**miRNAs based classifier**			**Genes based classifier**		
	P	R	F1	P	R	F1
Baseline RF classifier	28.9%	53.9%	37.7%	40.8%	44.2%	42.5%
Self-learning iteration 1	31.4%	53.9%	39.7%	**54.6%**	**58.1%**	**56.3%**
Self-learning iteration 2	46.8%	61.5%	53.2%	-	-	-
Co-training	**56.9%**	**61.5%**	**59.1%**	44.7%	53.5%	48.7%

4.5.2 HCC

HCC represents an extremely poor prognostic cancer that remains one of the most common and aggressive human malignancies worldwide [50] and [51]. Metastasis is a complex process that involves multiple alterations [52] and [53], that is why discriminating metastasis and non-metastasis HCC is a challenging problem.

For HCC, both self-learning and co-training approaches were evaluated to discriminate between metastatic and non-metastatic HCC. The self-learning steps are applied using GSE6857 as an initial labeled miRNA expression dataset and GSE10694 as the unlabeled subtypes HCC samples. Also, GSE36376 was used as initial labeled gene expression datasets and GSE15765 as the unlabeled

Table 4.5: Precision, recall and F1-measure for breast cancer subtypes SVM and LDS classifiers using miRNA expression and gene expression dataset

	miRNAs based classifier		
	P	R	F1
Baseline SVM classifier	24.5%	38.5%	29.9%
LDS	42.3%	30.8%	35.6%
Self-learning	31.4%	53.8%	39.7%
Co-training	**62.2%**	**61.5%**	**61.9%**

subtypes HCC samples. For co-training, to enhance sample subtype classifier using miRNA expression, one labeled miRNA expression dataset (GSE6857) is used. One labeled gene expression dataset (GSE36376) and one unlabeled gene expression datasets (GSE15765) are mapped into miRNA expression values and added to the miRNA training datasets and the sample subtype classifiers are retrained. Also, in order to enhance the sample classification accuracy using gene expression, one labeled gene expression dataset (GSE36376) is used. One labeled miRNA expression dataset (GSE6857) and one unlabeled miRNA expression datasets (GSE10694) are mapped into gene expression datasets and added to the gene training dataset.

Table 4.6 shows detailed results for HCC subtype classification using RF core classifier, there is around 10% improvement in precision of HCC metastasis class using miRNA expression sets and around 2% in F1-measure using gene expression sets. Moreover, table 4.7 shows the improvement of the techniques over SVM and LDS classifiers. Co-training achieved 5% enhancement in recall over

SVM classifier and 6% enhancement in F1-measure over LDS classifier. The improvement in HCC is less than breast cancer since in breast cancer the number of used unlabeled sets are larger. Also, the overlapping between the miRNAs and genes between the initial set and the added sets is an important factor. In order to understand why enhancements in breast cancer were more significant, the number of overlapping miRNAs and genes is calculated. Table 4.9 shows that the higher the overlap between the miRNAs and genes of the initial set and those of the added sets, the higher the improvements become in HCC.

Table 4.6: Results of HCC RF subtype classifiers using gene/miRNA expression dataset

	Class NM			**Class M**			**Weighted Evaluation**		
	P	R	F1	P	R	F1	P	R	F1
miRNA initial classifier	75%	93%	83%	59%	24%	34%	70%	73%	72%
miRNA self-learning	76%	**95%**	84%	**69%**	24%	36%	**74%**	75%	74%
miRNA co-training	**76.6%**	95%	**84.9%**	69%	27%	**39%**	**74%**	**75%**	**75%**
Genes initial classifier	95%	**98%**	96%	**98%**	95%	97%	96%	96%	96%
Genes self-learning	**100%**	96%	**98%**	97%	**100%**	**98%**	**98%**	**98%**	**98%**

Table 4.7: Results of HCC subtype SVM and LDS classifiers using miRNA expression dataset

	P	R	F1
Baseline SVM classifier	66.5%	66.1%	66.3%
LDS	61.2%	66.9%	63.9%
Self-learning	62.2%	61.5%	61.9%
Co-training	**67.7%**	**71.4%**	**69.5%**

Table 4.8: The number of overlapping miRNAs and genes between initial datasets and added datasets in breast cancer

	miRNAs initial dataset (GSE15885)	Genes initial dataset (GSE20713)
GSE15885	336	7
GSE26659	124	7
GSE35412	183	7
GSE20713	157	54676
GSE16179	157	54676

Table 4.9: The number of overlapping miRNAs and genes between initial datasets and added datasets in HCC

	miRNAs initial dataset (GSE6857)	Genes initial dataset (GSE36376)
GSE10694	36	-
GSE36376	52	47323
GSE15765	52	37282

4.5.3 Lung Cancer

Lung cancer is the leading cause of cancer-related death in both men and women worldwide, it results in over 160,000 deaths annually [54]. Only self-learning using gene expression dataset was evaluated in lung cancer as no labeled miRNA expression dataset was found on the web. The aim of the cancer subtype classifier is to discriminate between adenocarcinoma and squamous lung cancer subtypes. The labeled gene expression dataset (GSE41271) was used to build an initial classifier and the unlabeled gene expression dataset (GSE42127) was used to enhance it. Table 4.10 shows the enhancement achieved by self-learning, which is around 3% improvement in F1-measure of squamous lung cancer class.

Table 4.10: Results of lung cancer RF subtypes classifiers using gene expression dataset

	Class A			**Class S**			**Weighted Evaluation**		
	P	R	F1	P	R	F1	P	R	F1
Genes initial classifier	83%	95%	89%	83%	54%	65%	83%	83%	83%
Genes self-learning	**84%**	**96%**	**90%**	**87%**	**56%**	**68%**	**85%**	**85%**	**85%**

4.6 Conclusion

This chapter discussed our proposed miRNA and gene expression-based cancer classification using the adapted self-learning and co-training approaches in detail. First, it gives an overview of the proposed approaches, then it describes the adapted self-learning and co-training techniques in detail. Experimental results were presented by applying our proposed approaches on three cancer types, namely, breast cancer, HCC and lung cancer. The results showed up to 20% improvement in F1-measure in breast cancer, 10% improvement in precision in metastatic HCC cancer and 3% improvement in F1-measure in squamous lung cancer. Co-Training also outperformed the Low Density Separation (LDS) approach used in [25] by around 25% improvement in F1-measure for breast cancer.