Chapter 2

Background and Related Work

Chapter 2
Background and Related Work

2.1 Introduction

This chapter gives the necessary background about genes/miRNAs expression analysis, deep learning, feature selection and semi-supervised machine learning techniques. In addition this chapter discusses the related works to our proposed genes/miRNAs feature selection approach and our proposed cancer classification.

2.2 Gene/miRNA Expression Analysis

A gene is defined as the molecular unit of heredity of a living organism [5], while gene expression is defined as the process by which information from a gene is used in the synthesis of a functional gene product [6]. Deoxyribonucleic Acid (DNA) microarray and other techniques measure the expression level of large number of genes in different conditions like different organisms, different environment conditions or different time points. DNA microarray is a group of microscopic DNA spots attached to a solid surface. It is the widely used method to measure the level of expression for large number of genes simultaneously. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition or sample corresponds to one column. Each element of this matrix represents the expression level of this gene under a certain condition or for a certain sample. Figure 2.1 shows the gene

expression matrix where each row represents a gene, each column represents a sample and e_{ij} represents the expression level of gene i at sample j.

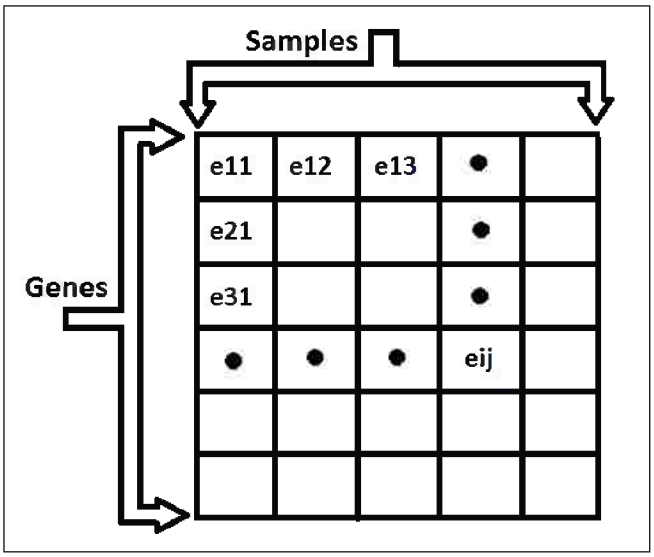

Figure 2.1: Genes-Expression Matrix

Ribonucleic Acid (RNA) is defined as a family of large biological molecules. It has important roles in coding, decoding, regulation, and expression of genes. RNA comprises the nucleic acids and is assembled as a chain of nucleotides. However, it is usually single-stranded [7]. MicroRNAs (MiRNAs) are short (1925 nucleotides) noncoding single-stranded RNA molecules [8]. MiRNAs regulate gene expression either at the transcriptional or translational level, based on specific binding to the complementary sequence in the coding or noncoding region of mRNA transcripts [8]. Also, miRNAs expression level can be measured by microarrays as mentioned in [9].

Recent research has pointed out the success of using miRNA and gene expression datasets in cancer classification; miRNA profiles were used recently to discriminate malignancies of the breast [10], lung [10] and [11], pancreas [10] and [12] and liver [13], [14] and [15].

2.3 Deep Learning

Deep learning has been widely used recently in several fields like image and audio applications. Deep learning algorithms attempt to discover good representations for the data, at multiple levels of abstraction. Each level is composed of non-linear transformations, which captures the similarities between the features and their relation to different classes. There has been a rapid progress in this area in recent years, both in terms of algorithms and in terms of applications. Deep Belief Net (DBN) is one of the algorithms used in applying deep learning and it has been widely used in [16], [17], [18] and [19]. It has also shown its ability to detect high level features and enhance the classification accuracy.

Deep Belief Nets (DBNs) are defined as graphical models which can learn to extract a deep hierarchical representation of the training data. It was shown in [20] that Restricted Boltzmann Machines (RBMs) [21] can be stacked and trained in a greedy manner to form DBNs. The description of the DBN as given in [3] is that mainly its joint distribution between observed vector x and the l hidden layers h^k follows the following equation:

$$P(x, h^1, ..., h^l) = (\prod_{k=0}^{l-2} P(h^k|h^{k+1}))P(h^{l-1}, h^l) \tag{2.1}$$

Where $x = h^0$,$P(h^{k-1}|h^k)$ is the conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{l-1}, h^l)$ is the visible-hidden joint distribution in the top-level RBM. Figure 2.2 shows the structure of the DBN in detail.

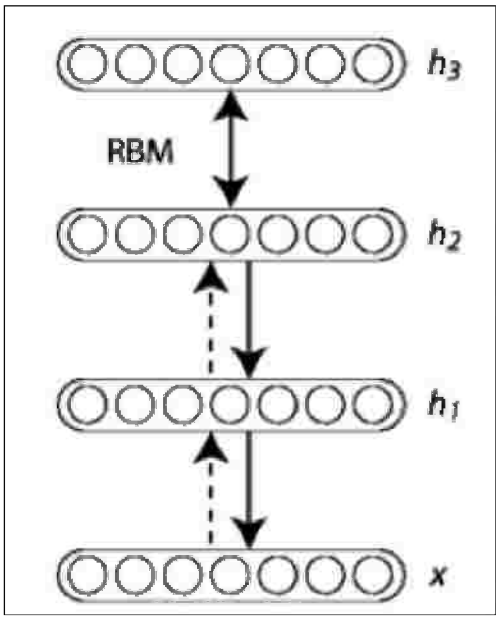

Figure 2.2: Deep Belief Nets Structure [3]

2.4 Semi-Supervised Machine Learning Techniques

2.4.1 Active Learning

Traditional active learning [22] is a semi-supervised machine learning approach in which a basic classifier is trained using a small number of training data. Then, the classifier training data is enriched by selecting the most informative data to the classifier and labeling them according to a human.

There are many strategies to select the most informative data to the classifier. The following is a summarization of the most common strategies:

- **Uncertainty Sampling:** Choose data that the classifier is not confident about its label, which can be identified when the classifier assigns a confusing classification confidence score (~ 0.5).

- **Query-By-Committee:** Choose data by asking committee of classifiers about its label and then the most informative data is considered to be the one about which they most disagree.
- **Expected Model Change:** Choose data that its addition will result in the greatest change to the current classification model.
- **Expected Error Reduction:** Choose data that its addition will result in the greatest expected error reduction.

Active learning is used in many applications, for example [23] discusses active learning applications in image/video annotation and content-based image retrieval. Also, [24] uses active learning in the text classification problem to reduce the required labeled data while maintaining same level of classification accuracy. Also, the work in [24] takes into account multiple label-information and uses active learning with expected loss reduction selection strategy. The results of the work on seven real-world data sets shows that the approach can obtain promising classification accuracy with much smaller labeled data.

2.4.2 Self-Learning

Self-learning [25] is a semi-supervised machine learning approach, in which the labeled set L is used to build the initial classifier. The unlabeled set U is utilized then to enhance its accuracy by adding the unlabeled samples with the highest classification confidence to the training set. Thus, resulting in making the classifier learns based on its own decision. Self-learning was used in many applications such as object detection [26] and word sense disambiguation [27].

As shown in [26], constructing object detection systems is a time consuming task that needs a large labeled training set. Object detection systems are constructed to detect instances of semantic objects of a certain class (for example: humans, buildings, or cars) in images or videos. [26] proposes a solution to the problem of the needed large labeled training set by using self-learning. It uses small number of fully labeled examples and an additional set of unlabeled or weakly labeled examples. The approach begins by training the object detector using the fully labeled examples, then it runs the detector on the unlabeled or weakly labeled examples and uses the output of the detector to re-label them. Finally, it selects some examples out of them to add to the detector training data according to some selection criteria and re-trains the detector. The paper explores different selection criteria to choose unlabeled examples to add. It first explores the usage of confident examples based on the classifier confident scores and then it explores the usage of the distance measure between corresponding unlabeled example (W_i) and all of the other examples in the training set (L_j). At the end, it chooses the unlabeled examples with the smallest score to add. The score is computed as follows [26]:

$$\text{Score}(W_i) = min_j \text{ Mahalanobis } (g(W_i), g(L_j), \Sigma)$$

Where $g(X)$ is a preprocessing function to normalize examples for scale, position and orientation and Σ is the weights for computing Mahalanobis distance. Mahalanobis distance is defined as follows:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$$

Where $\vec{x}$, $\vec{y}$ are two random vectors and S is the covariance matrix. Finally, the paper shows that the Mahalanobis distance measure improved the accuracy

of the classifier.

Moreover, the work in [27] uses self-learning and co-training for the word sense disambiguation task. The paper uses naive bayes classier for word sense disambiguation using several features like the word itself, its part of speech tagging, its k surrounding words and their part of speech tagging, first noun/verb before/after the word and verb-object/subject-verb relation to the word. The paper applies self-learning by first constructing a classifier using labeled examples, then it randomly selects a pool of unlabeled samples to label and labels them using the previously constructed classifier. Finally, it selects the most confident unlabeled examples, adds them to the classifier training set and re-trains the classifier.

2.4.3 Co-Training

Co-training [28] and [29] is also a semi-supervised machine learning approach, which requires two or more views of the data. Two or more classifiers are constructed separately for each view. Each classifier is then used to train the other one by classifying unlabeled samples and training the classifier with the other view using samples with the highest classification confidence.

Co-training was also used in many applications such as word sense disambiguation [27] and email classification [30]. In [27], co-training is used to enhance the word sense disambiguation task. The co-training relies on two different views. The first view is local features like the word itself and its part of speech tagging. While, the second view is topical features like the features that are extracted from a large context, mainly keywords that are specific to each

sense. Finally, each classifier is then used to train the other one by classifying unlabeled samples and train the other view using samples with the highest classification confidence.

For the email classification problem [30], co-training is used with two views which are the email subject and content. One classifier is trained based on the email subject and another classifier is trained based on the email content. Each classifier is then used to label the unlabeled emails based on its own features. Finally, the confident classified emails from one view (subject/body) are converted to the other view (body/subject) and added to the training data of the other classifier. This process is repeated until saturation is reached.

2.5 Gene Feature Selection Approaches

Many gene feature selection methods have been proposed in literature recently, e.g. [1], [2], [31], [32], [33], [34] and [35]. The work in [31] used three feature selection techniques which are Fast Correlation Based Filter (FCBF), reliefF and random feature selection. All the previous techniques depend only on gene individual behavior and didn't explore gene group performance. The work in [32] also proposed feature filtering algorithms that depend only on gene individual features.

[33] uses relief-f filtering feature selection method to select a small set of genes and then uses these genes with K-Nearest Neighbors (KNN) or Support Vector Machine (SVM) classifier to discriminate the samples. The work in this paper shows that relief-f performed better than chi-square, information gain and gain ratio feature selection methods in Acute Lymphoblastic

Leukemia (ALL)/Acute Myeloid Leukemia (AML), colon tumor and Mixed Lineage Leukemia (MLL) datasets. Moreover, the work in [35] uses an ensemble of feature selection methods to enhance feature selection stability and classification accuracy. The paper performs bootstrapping to generate different subsets of the data and then applies a feature selection method called Recursive Feature Elimination (RFE) to each subset. Finally it aggregates the scores from the different subsets. However, genes are evaluated independently regardless of how well they perform with other genes; this problem is solved in our proposed approach at the active learning phase as shown in the next chapter.

The genes grouping evaluation to select genes was addressed in [2] by not considering the gene alone but evaluating its performance jointly with other genes. Although a gene can contain essential information that no other gene have, it may perform badly when considered alone. However, by considering it with the suitable set of genes, the whole subset can perform better. [2] has proposed an algorithm that first divides genes into subsets recursively, then selects the most informative smaller ones. Finally, it merges the chosen genes with each other. This process is repeated until all subsets are merged into one informative subset. However, our approach evaluates joint gene performance differently by using DBN features and active learning.

[1] has used deep learning for feature selection by construing an auto-encoder to perform dimensionality reduction and then using DBN with the new dimensions to classify the samples. In our approach, genes were not mapped into principal components first as we want to keep track of the most discriminative genes for the biologist to analyze their biology behavior. Table 2.1 summarizes the comparison between the proposed gene feature selection approach

and the related work in literature.

Table 2.1: Comparison with related work in gene feature selection

Approach	**Genes group performance-based selection**	**DBN usage**	**Ability to Get Selected Genes Names**
[31] (FCBF, ReliefF, Random)			✓
[32] (Feature Filtering Algorithms)			✓
The Work in [33]			✓
The Work in [35]			✓
The Work in [2]	✓		✓
The Work in [1]		✓	
The Proposed Approach	✓	✓	✓

2.6 Cancer Classification

Using miRNA expression profiles to discriminate cancerous samples from normal ones, and to classify cancer into its subtypes, is an active research area and was applied to different cancer types as breast [10], lung [10] and [11], pancreas [10] and [12] and liver [13], [14] and [15]. The previous papers used one

of the following supervised machine learning techniques like SVM, Prediction Analysis of Microarrays (PAM) and compound covariate predictor.

Several attempts for enhancing cancer classifiers have been recently introduced [36], [37] and [38]. In [36], a number of feature selection methods, such as pearsons and spearmans correlations, euclidean distance, cosine coefficient, information gain, mutual information and signal-to-noise ratio are used to enhance cancer classifiers. Also different classification methods, namely, k-nearest neighbor methods, multilayer perceptrons, and support vector machines with linear kernel are used in [36]. The work has focused only on improving classifiers based on labeled samples of miRNA expression profiles and didn't use publicity available unlabeled sets, also, gene expression profiles were not used to enhance miRNA based cancer samples classifiers. Another work [38] has considered using Discrete Function Learning (DFL) method on the miRNA expression profiles to find the subset of miRNAs that shows strong distinction of expression levels in normal and tumor tissues and then uses these miRNAs to build a classifier. The paper didn't combine multiple miRNA datasets or use gene expression datasets to enhance the classifier.

Enhancing the classification accuracy by building two classifiers; one for miRNA data and another for Messenger RNA (mRNA) data were explored in [37]. It first applies feature selection using relief-F feature selection, then it uses bagged fuzzy KNN classifier and finally it combines the two classifiers using fusion decision rule. The drawback of the approach is that it assumes the existence of both miRNA and mRNA data for each patient and it just uses decision fusion rule to combine the classifiers decision without enhancing the classifiers themselves.

Semi-supervised machine learning approaches were introduced in classification using expression sets by using Low Density Separation (LDS) approach in [25] to enhance cancer recurrence classifiers. Semi-supervised machine learning approaches make use of the publicity available unlabeled sets to enrich the training data of the classifiers. However, the approach in [25] depends only on gene expression, and didn't combine both miRNA and gene expression sets. Table 2.2 summarizes the comparison between the proposed cancer classification approach and the related work in literature.

Table 2.2: Comparison with related work in cancer classification

Approach	**Unlabeled Sets Usage**	**miRNAs and Genes Expression Sets Integration**	**Ability to Apply Independently on Genes/miRNAs Sets**
The work in [10]			✓
The work in [11]			✓
The work in [13]			✓
The work in [14]			✓
The work in [15]			✓
The work in [36]			✓
The work in [37]		✓	
The work in [38]			✓
The work in [25]	✓		✓
The Proposed Approach	✓	✓	✓

2.7 Conclusion

In the this chapter, the background of genes/miRNAs expression analysis was provided and semi-supervised machine learning techniques (active learning, self-learning and co-training) were explained, in addition to deep learning. Moreover, the related works to our proposed work were discussed in detail. First, we discussed the related works to the proposed feature selection method. Then, we discussed the cancer classification related works. In the next chapters, our proposed work is explained in detail.