A NOVEL MACHINE LEARNING FRAMEWORK FOR PHENOTYPE PREDICTION BASED ON GENOME-WIDE DNA METHYLATION DATA

A Thesis by

Vinay Vittal Karagod

Bachelor of Science, RYMEC, 2011

Submitted to the Department of Electrical Engineering and Computer Science
and the faculty of the Graduate School of
Wichita State University
in partial fulfillment of
the requirements for the degree of
Master of Science

December 2016

A NOVEL MACHINE LEARNING FRAMEWORK FOR PHENOTYPE PREDICTION BASED ON GENOME-WIDE DNA METHYLATION DATA

The following faculty members have examined the final copy of this thesis for form and content, and recommend that it be accepted in partial fulfillment of the requirement for the degree of Master of Science with a major in Computer Science.

Kaushik Sinha, Committee Chair

Hamid Lankarani, Committee Member

Debswapna Bhattacharya, Committee Member

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Kaushik Sinha, for his guidance and support throughout the process of this work.

ABSTRACT

DNA methylation (DNAm) is an epigenetic mechanism used by cells to control gene expression, and identification of DNAm biomarkers can assist in early diagnosis of cancer. Identification of these biomarkers can be done using CpG (Cytosine-phosphate guanine) sites, or particular regions in DNA. Previous machine learning methods known as MS-SPCA and EVORA have been used to link DNAm biomarkers to specific stages of cervical cancer using CpG data. In this work, it is shown that a proposed framework yields greater AUC accuracy than the MS-SPCA and EVORA for predicting stages of cervical cancer using CpG data. This framework appears promising in regards to the data examined herein as well as in future biological studies.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

α	Amplification constant

CHAPTER 1

INTRODUCTION

Even with the medical progress that has been made in recent decades, Cancer remains a global problem. Studies from the National Cancer Institute (NCI) estimated that in the year 2016 1,685,210 people will be diagnosed with cancer in United States, out of which 595,690 will die from cancer related diseases [27]. Studies from NCI show that the most common type of cancers are breast cancer, lung and bronchus cancer, prostate cancer, colon and rectum cancer, bladder cancer, melanoma of the skin, non-Hodgkin lymphoma, thyroid cancer, kidney and renal pelvis cancer, leukemia, endometrial cancer, and pancreatic cancer [27]. Data shows that for every 100,000 men and women the number of new cancer patients are 454.8 and the number of deaths due to cancer are 171.2 per year [27]. Cancer remains one of the leading causes of death worldwide, and in 2012 there are around 14 million new cases and 8.2 million cancer deaths [27]. It is estimated that the number of cancer patients will rise to 22 million within the next two decades [27].

As new technologies continue to evolve, it creates a better opportunity to understand and eventually solve cancer. In recent years, advancement in technology have made it possible to measure a very large number of various complex biological entities as well as their variations which are believed to be somehow related to cancer. With these measurements available a formidable task that remains is to find association between these complex biological measurements and their functional characteristics leading to cancer. One way to achieve this by designing various efficient predictive models that can predict the possibility of someone having cancer, with high accuracy, based on various complex biological measurements obtained from him/her.

Various biological measurements which are commonly used for this purpose by various researchers are, gene expression data, Single Nucleotide Polymorphisms (SNPs), DNA Methylations patterns etc. Gene expression ratio is the measurement used for determining the relative expression levels of the genes, where the expression levels are the behavior of the genes according to different conditions. In condition 'A' genes are colored with red dye and in condition B genes are colored with green dye. Samples from both the conditions are combined on one microarray slide, and a laser is emitted on the microarray slide to capture the amount of red and green wavelengths. These measurements of red and green wavelengths are compared in the gene expression ratio. Various studies have shown that gene expression ratios are associated with cancer and can predict various types of cancer [37] [39] [40].

Single nucleotide Polymorphism (SNP) are the most common genetic variations found in the human body. A single SNP defines the difference in the DNA building block called a nucleotide. Most of the SNPs are not associated with any types of diseases but some SNPs act as a biomarkers. These biomarkers help in locating the genes that are associated with the diseases, and occur near the genes which causes the disease by directly impacting on the genes to vary their functionality. SNPs promise to provide meaningful information about the cancer phenotype by using polymorphic markers [30]. Many studies have shown association between SNPs and various types of cancer [47] [48] [49]

DNA methylation is an epigenetic mechanism used by cells to control gene expression [1]. DNA methylation patterns are measured at so called CpG sites (Cytosine-phosphate guanine) which are regions consisting of cytosine nucleotides and guanine nucleotides which are in a linear sequence of bases along its 5' → 3' direction [34]. CpGs are promoters of high methylation levels, where a promoter region is a particular region in DNA which is responsible

for initiating the first steps of gene expression. If the methylation level of the CpG corresponds to either hypermethylation or hypomethylation, described below, then this CpG is identified as a DNAm biomarker. According to [22], hypermethylation CpGs typically occur at CpG islands in the promoter region and are associated with gene inactivation. Hypermethylation refers to the increase in epigenetic methylation of cytosine and adenosine residues in DNA [24], and CpG islands refer to the locations in the DNA where large numbers of CpGs are located. In contrast Hypomethylation is the decrease in epigenetic methylation of cytosine and adenosine residues in DNA [26]. Even though hypomethylation CpGs are not considered a potential cause of cancer, hypomethylation CpGs are responsible for the development of the cancer by various mechanisms [23], so hypomethylation CpGs are still identified as DNAm biomarkers. DNA methylation plays a crucial role in the development of nearly all types of cancer [5]. DNA methylation is a necessary part of normal development as it is associated with many processes such as genomic imprinting, X-chromosome inactivation, and repression of repetitive elements, aging and carcinogenesis [3]. Identifying DNAm biomarkers related to cancer can lead to early diagnosis and effective treatment. Cancer and tumor suppressor genes are the two major areas researchers are currently focusing on, since hypermethylation, or the increase in epigenetic methylation of cytosine and adenosine residues in DNA [24], often results in the silencing of tumor suppressor genes in cancerous cells [1].

A major problem in finding association between these biological measurements and their functional characteristics is that number of measurements are often very large while the number of samples is much smaller. For example, a typical gene expression dataset consists of measurement of gene expression ratios of tens of thousands of genes ratios per sample. A typical SNP dataset consists of mutation information of millions of SNP locations per samples. A typical

DNA methylation dataset consists of measurement of methylation patterns at tens of thousands of CPG sites per sample. While in most cases number samples is in tens or in hundreds. For statistical analysis of these datasets this leads to a major problem known as "small n, large p" problem. That is, number of samples n is small but the number of features/attributes/measurements per sample p is large. Due to law of large numbers, any consistent statistical analysis technique works well when n >> p, i.e., number samples is larger than number of attributes. However, when n<< p, as is the case in most biological datasets, this leads to curse of dimensionality problem (i.e. dimension of dataset measured by number attributes per sample is extremely high), which is a sub-problem of "small n, large p". When this happens, it becomes extremely difficult to draw meaningful conclusions from a dataset due to high dimensionality of the data and few samples. The field of machine learning plays a prominent role in these kind of situations by providing various dimension reduction techniques that helps to reduce data dimensionality. Typically, machine learning methods consists of two kind of techniques: supervised learning and un-supervised learning. In supervised learning (predictive modeling), models are developed to classify the data by training using labeled data with prior information. This technique is useful when the class labels are to be distinguished: for example, classification of cancer tumors. In un-supervised learning, the relations of the dataset are developed without using any prior labeling of the data. This technique is useful when exploratory analysis is required.

This thesis study the problem of designing an efficient predictive model for predicting cancer using DNA methylation patterns by partially addressing the small n large p problem. In particular, it significantly improves the cancer prediction accuracies in later stages of cancer using DNA methylation data compared to a recently proposed machine learning method called

Model Selection Supervised Principle Component Analysis (MS-SPCA) [4]. In [4], MS-SPCA is used to analyze publicly available genome-wide DNA Methylation data of cervical screening samples. The dataset used in [4] consisted of 27,578 CPGs corresponding to 14,495 genes [12]. The cases considered correspond to three different stages of cervical cancer development:

(i) Women who initially had no indications of a tumor but later developed a cervical intraepithelial tumor of grade 2 or higher (denoted by Normal in [4]) as determined from a subsequent screen after three years.

(ii) Two independent case–control datasets with tumor of grade 2 or higher (denoted in [4] as CIN2+ (a) and CIN2+ (b)).

(iii) Fully developed cancer [4].

MS-SPCA is the essentially an advanced version of another method called Supervised Principal Component Analysis (SPCA) [25], and works by building a large number of models (often in thousands) and *hand picking and reporting* the prediction accuracy of the best performing model, which is not very efficient. Moreover, in many test cases, the prediction accuracies of MS-SPCA are very low.

In this thesis a novel framework is proposed to address the shortcoming of MS-SPCA. The proposed framework is flexible and is designed in three modules. The first module essentially performs dimension reduction i.e., reduces the number of attributes by applying multiple general statistical tests, i.e. Cox test, t-test, cosine test, and univariate regression, and selecting statistically significant features/attributes based on p-vales and threshold cut-offs. This leads to a large number of new datasets each having same number of samples n as before, but now number of reduced attributes is comparable to n. The second module consists of

applying generic predictive supervised classifiers to these large number of new reduced datasets. This module is extremely flexible since one can use his/her favorite supervised classifier in this module. In this thesis, the choice of generic classifier was restricted to Support Vector Machine (SVM) classifier with different kernels as it worked well for this particular DNA methylation dataset. Note that by applying various generic classifiers to each of the large number of reduced datasets creates a large number of models, not necessarily each of these model have good prediction accuracy. Therefore, to reduce the number of models, only those models with prediction accuracy greater that 50% on the training dataset were chosen and the rest were discarded. Even after this, the number of models was still quite high (often in thousands). Now, in contrast to MS-SPCA which chooses the best performing model, in the third and final module, a generic ensemble method is applied on these models to obtain a single final model. The ensemble method proposed in this thesis is essentially an aggregation method called weighted majority voting ensemble techniques. Weighted majority voting ensemble technique outputs the weighted average of the prediction accuracies of all the models obtained from module 2, where higher weights are provided automatically to the better performing models from module 2 and lower weights are performed to the worse performing models from module 2. Therefore, this proposed ensemble technique automatically selects the final model as the average of the better performing models of module 2. Empirical evaluations on DNA methylation dataset from [4], demonstrates that the method proposed in this thesis significantly outperforms prediction accuracies obtained by applying MS-SPCA on the same dataset.

The rest of the thesis is organized as follows. In Chapter 2, previous research work in this direction is discussed. In Chapter 3, the proposed model is described in details. Dataset

description, empirical performance of the proposed method, and implementation issues are presented in Chapter 4. Finally, conclusions are presented in Chapter 5.

CHAPTER 2

BACKGROUND / LITERATURE SURVEY

In this chapter various existing research work in finding the significant CPGs and related DNA microarray data for detecting cancer is discussed. There are three major research work that analyzes DNA methylation data for finding association of DNA methylation pattern to cancer. Among these the MS-SPCA [4] is the most recent and state of the art technique. MS-SPCA is an advanced version of another technique called Supervised Principal Component Analysis (SPCA) [25] and it outperforms another recently proposed method called EVORA [19]. We describe these methods in details below.

2.1 Supervised Principle Component Analysis

In this work, Eric Bair et al [25] explain the challenges of microarray data analysis and how they approached these challenges. The main problem they discuss is the difficulty of finding satisfactory results when the number of features greatly exceed the number of samples; i.e. the curse of dimensionality problem. Here the author assumes that there are two types of cells: good cells and bad cells. In [25], they found that there was considerable overlap between good and bad cells. Survival time of the patients varied depending upon the cell type. If the patient has the good cell type, this signifies that the patient may live longer than the average survival time but if the patient has the bad cell type, this signifies that the patient may live lesser than the average survival time. Genes that lead to the production of good cells are known as survival genes. Using their proposed Supervised Principle Component Analysis (SPCA) algorithm, their idea was to discover the readable patterns responsible for producing the underlying cell types. Using the Supervised Principle Component Analysis (SPCA) algorithm, weights were assigned to genes based on the relationship between a gene and the survival genes using the procedure below. The aim was to predict which gene patterns affected the survival time of the patient.

SPCA relies on principal component analysis (PCA), principal component analysis the standard approach for modeling correlation. In PCA, the first few principal components significantly indicate the correlation between the features and the labels. Because PCA does not always return good results, SPCA finds a subset of significant features and then applies PCA rather than directly applying PCA. In this method, they considered *n* samples of DNA microarray data $X_1, X_2, \dots, X_p$ and p features, where $p \gg n$ and each feature corresponds to the measurement of one particular gene.

After using PCA to find the principal components, Eric Bair et al used regression to predict the case status Y, which was modeled as a quantitative variable. Below is a summary of the steps they followed:

1. Calculate the (univariate) standard regression coefficients for each feature.
2. Consider only those features whose values exceed some threshold in absolute value, where the threshold is calculated by cross-validation.
3. Compute the first few principal components of the reduced data matrix.
4. Use those principal components to predict the outcome in a regression model.

2.2 Model Selection SPCA

In [4] T. Wilhelm et al presented a method called MS-SPCA (Model Selection Supervised principle component analysis). In [4] author presents a model which is used to identify the significant CPGs from DNAm data. This method determines "several models that perform well in the training data and selects specific ones for the prediction of test data" [4]. This method proposed the below steps in order to predict the CPGs from the DNA microarray data.

1. Splitting of the training data into k-fold training and test sets using the Leave One Out method, which considers one out of *n* samples were *n* is the number of samples for testing and the remaining for training.
2. Calculating the p-values for CPGs for all the k training datasets, which consists of all the features of CPGs methylation: average methylation difference, methylation variation difference, and methylation-age correlation.
3. Identifying the best models which perform well in training. This is done via cross-validation. Test models were considered with different weights from all the k different training datasets by combining all the three p-value rank lists [4].
4. Predicting the particular test datasets that perform well in the models.
5. "The model Selection step of MS-SPCA: ranking these models according to the criterion Eval1-EV1dist and using the cumulative risk scores of the first n best ranking models for final prediction (here they used n = 5)"[4].

The method used for identifying significantly associated CPGs is described in detail in the following section.

2.2.1 Significantly Associated CPGs

Significantly associated CPGs were determined using statistical testing. In this testing, T. Wilhelm et al included the CPG data of all the six datasets and all the three components of the CPGs: average methylation difference, methylation variation difference, and methylation-age-correlation. For average methylation difference they used the t-test and the Mann–Whitney U test. For methylation variation difference they used Bartlett's test and Levene's test and methylation-age-correlation. For each test they calculated p-values on all the CPGs, only on hyper CPGs and only on hypo-CPGs. In this tests, they struggled to differentiate between cases

and controls in the three normal datasets but, by applying principal components for the CPGs, they were able to differentiate between cases and controls. This is why they chose to use the supervised PCA approach.

The method for selecting the best models described in detail in the following section.

2.2.2 Selecting the Best Models from Cross-Validation in Training Data

In [4], T. Wilhelm et al combined all the three features of the best ranking CPGs which they had used as their training data and selected the models based on cross-validation prediction accuracies. They considered the models greater that 65% for the three normal datasets, 82% for CIN2+ (a) and 95% for CIN2+ (b). This resulted in greater than 300 models used for training Normal datasets, greater than 700 models for CIN2+ (a) and greater than 1000 models for CIN2+ (b). They found that the models with best predicting accuracy had 80% for normal datasets, 92% for CIN2+ (a) and 100% for CIN2+ (b).

The method for selecting the best predicting models described in detail in the following section.

2.2.3 Best Predicting Models and Final Prediction

In [4] T. Wilhelm et al explains two parameters used in generating all of the models: Eval1 and EV1dist. Eval1 measures how much of the variation in the test data is likely captured by first principal component PC1, and EV1dist is a measure for how well the model obtained from the training data fits the test data. These parameters significantly determine the effectiveness of the model.

2.3 EVORA (Epigenetic Variable Outliers for Risk Prediction Analysis)

In [18] they considered only two features, the level of methylation and methylation variability, in order to identify the risk CPGs. Teschendorff AE et al needed a statistical method which was robust and independent of the scale used in order to “guarantee that classification

thresholds are generalizable for independent datasets" [19]. For this reason, they used COPA (Cancer Outlier Profile Analysis) to identify candidate gene fusions. They proposed identification of the DNAm data matrix outliers for independent datasets.

These are the steps followed in EVORA, where the authors assume that samples are labeled with binary phenotype with N as normal state and T as transformed phenotype.

1. They did COPA transformation on the whole DNAm data matrix without considering any phenotype information.
2. They did 10-fold cross-validation making sure that each fold consisted of an equal number of normal state and transformed state phenotype in both training and testing samples.
3. From the original β-valued data matrix. Where β is the normalized methylation values of the CPGs [18]. They identified candidate risk CPGs using Bartlett's test in the training set. The author performed FDR (false discovery rate) in order to identify that almost all the selected risk CPGs are crossed the appropriate threshold.
4. Identified risk CPGs were later transformed from COPA thresholds into EVORA binary format. If the COPA score of the risk CPG was greater than the threshold it was represented as 1. Otherwise it was represented as 0.
5. In the test set for each sample, they calculated the fraction of the sampled CPG which were labeled with 1. The calculated fraction was the risk index of that sample which was dependent on COPA threshold including top ranked CPGs.
6. They repeated all the steps from 3-5 in turn, with each sample serving as the test sample for a single run and then they assigned the risk score.

7. From all the COPA thresholds they calculated the area under curve (AUC) from the resulting risk scores for different numbers of the top ranked risk CPGs.
8. Finally they found the COPA threshold and all the risk CPGs that optimized the AUC over the internal cross-validation.
9. After identifying the optimized parameters they obtained the risk scores of the independent samples that satisfied the optimized COPA threshold.

CHAPTER 3

PROPOSED METHOD

In this chapter a novel framework proposed in this thesis for cancer prediction accuracies in later stages of cancer using DNA methylation data is presented in details. In section 3.1 we present the overview of the proposed modular framework. In section 3.2 we present the Dimension reduction via statistical tests, 3.3 we discuss about Randomly Generating the Datasets from Tests, 3.4 we discuss about Generic Classifier: SUPPORT VECTOR MACHINE, 3.5 we discuss about Passing the datasets to SVM classifier with different kernels, and in 3.6 we discuss about Ensemble method - Majority voting and weighted majority voting, details of each modules are described in details.

3.1 Overview of Method

The proposed framework is divided into three modules. Figure (1) below shows the overview of the framework.

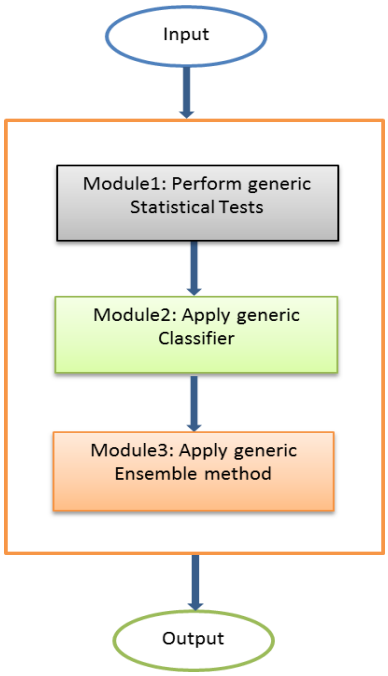

Figure 1. Flowchart depicting the process used for analyzing.

Module 1: Dimension reduction

The main goal of this module is to reduce the number of features, as it is difficult to classify the samples when the number of features are much larger than the number of samples. In this module various statistical tests have been implemented and the top 1000 features considered that are significant according to those tests. For this thesis, the Cox test, t-test, cosine test and univariate regression test have been used. In the general case any other tests can be considered. More details about the statistical tests used in this thesis are discussed in section 3.2.

Module 2: Generic classifier

The main goal of this module is to classify 1000 reduced datasets obtained after applying the dimension reduction technique as described above. Each of the reduced datasets consists of 1000 attributes, as compared with the original datasets, which consist of 27,578 attributes. These reduced datasets will be subsampled into smaller datasets which need to be classified. There are many classifiers available, a few of which are like Decision tree [41], KNN [42], SVM [43], Naive Bayes [44] and many more. In this thesis, SVM classifier was chosen for classifying the samples as it performed much better than the other classifiers considered. For the classifier, 50 features were randomly chosen from the features obtained from the statistical test and the same features from the test set for testing the classifier. The classifier was trained with training samples and the model obtained. The test dataset was then provided to obtain the models. Advantages of using SVM are that it avoids over-fitting, is a powerful tool for two class classification, offers good generalization, and allows for use of a kernel function which transforms low dimensional data to high dimensional data. SVM provides a unique solution since the optimality curve is convex. SVM can also handle high dimensional data. More details about SVM is discussed in topic 3.4

Module 3: Ensemble method

The main goal of this module is to automatically combine large number of good performing models generated in module 2 to obtain the final classifier with improved accuracy. Any ensemble method can be applied like Boosting Trees [45], Bagging Trees [46], and performance weighting [21]. In this thesis, the weighted majority voting algorithm was applied. In this algorithm all the prediction accuracies were considered and a positive weight given for each of the prediction accuracies. This positive weight effectively decreases the contribution of bad classifiers and increases the contribution of the good classifiers. More details about weighted majority voting are discussed in topic 3.6.

Figure (2) gives the detailed explanation of each module.

3.2 Dimension reduction via statistical tests

Dimension reduction was performed using various statistical tests, including cox test, t-test, cosine test and univariate standard regression coefficient test. The following hypothesis were assumed in the cox test and t-test.

1. Null Hypothesis: The relationship between both groups are same.
2. Alternative Null Hypothesis: The relationship between the both groups are different.

P-values: A p-value is a function in hypothesis testing which defines how extreme the distribution of the observations is [13]. The following p-values were considered as significant in the standard statistical analysis testing.

I. If p <= 0.05 indicates that the relationship between two distributions are the different. In this case, the null hypothesis rejected.

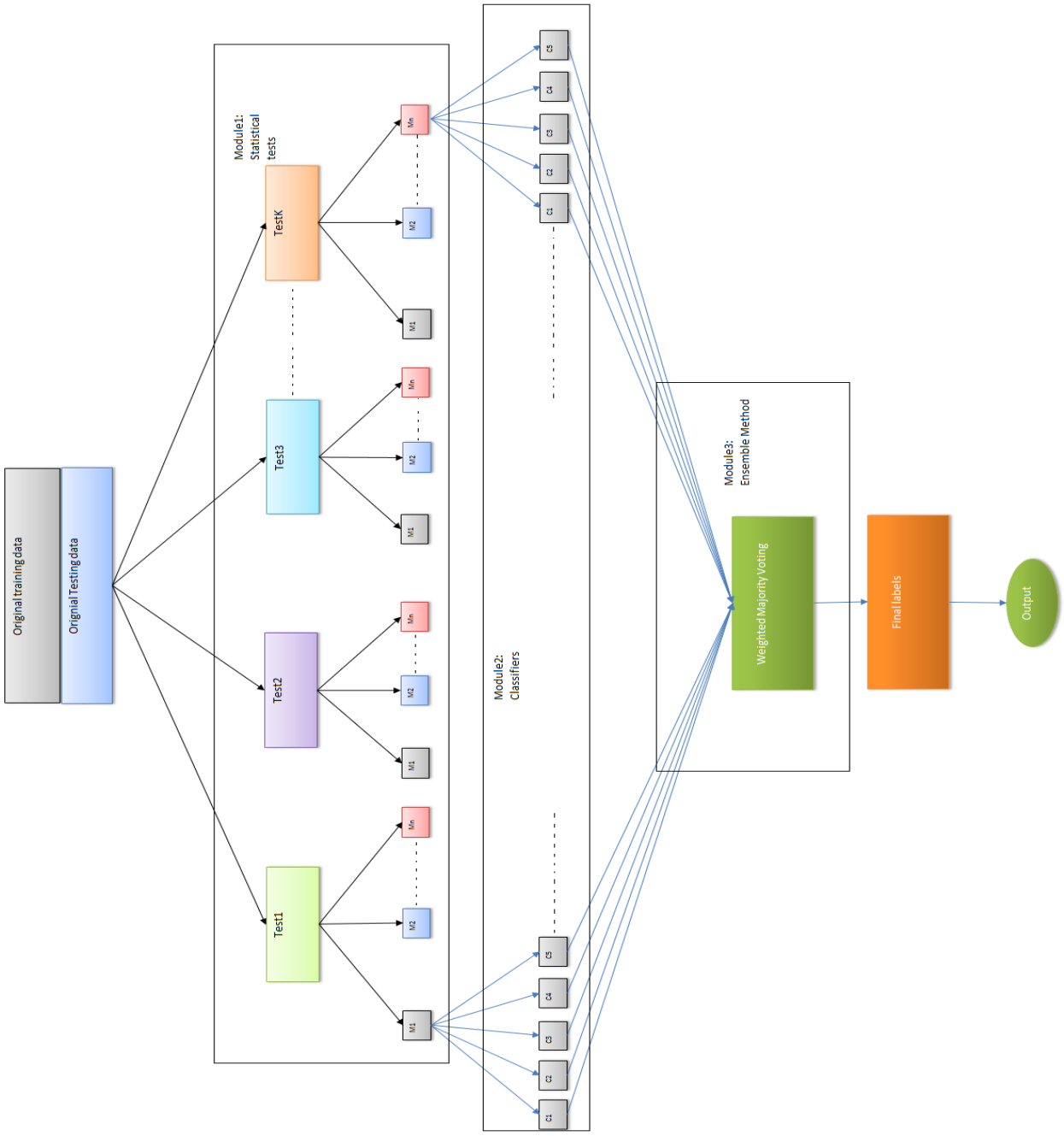

Figure 2. Detailed alogorithm implementation of Novel Framework.

II. If $p > 0.05$ indicates that the relationship between two distributions are the same. In this case, the null hypothesis accepted.

In this thesis the p-values were sorted in increasing order and the top 1000 values considered.

Details of various statistical tests are discussed next.

3.2.1 Wilcoxon Rank Sum Test (cox test)

The Wilcoxon Rank Sum test is a non-parametric statistical test for two groups when samples are independent [31]. Suppose two samples are X and Y independent from one another with different sample size. Then it is possible to apply the Rank Sum test in order to obtain the relation between the two groups. Depending upon the p-values generated by the test, it is possible to decide whether to accept or reject the null hypothesis. In this thesis, this test was used to identify the relationship between cases and controls of the given datasets. The dataset was divided into two samples, one case and the other control, and p-values generated for each CPG. The p-values were sorted in increasing order and the top 1000 CPGs with the highest p-values considered. This was performed in Matlab.

The following Matlab function was used to generate p-values: $\mathrm{p} = \mathrm{ranksum}(\mathrm{x}, \mathrm{y})$

Rank Sum Test

$$U = W - n_x(n_x + 1)/2 \quad (1)$$

3.2.2 T-Test

The two sample t-test is a parametric statistical test used to find the significant difference between two independent samples [32]. Its implementation is similar to what was specified for the cox test. If two samples X and Y are independent from one another with different sample sizes, the two sample t-test must be used in order to obtain the p-values for each of the CPGs between the two groups. It accepts the null hypothesis only if the distributions of the samples have equal means. It rejects the null hypothesis if the samples have unequal means. In this thesis, this test is used to identify the relation between cases and controls of the given datasets. The dataset was divided into two samples, one case and the other control, and p-values generated for each CPG. The p-values were sorted in increasing order and the top 1000 CPGs with the highest p-values considered. This was performed in Matlab.

The following Matlab function was used to generate p-values: $\mathrm{p} = \mathrm{ttest2}(\mathrm{x}, \mathrm{y})$

3.2.3 Cosine Correlation

Cosine similarity is a statistical test which is used to define the relationship between two vectors by using their inner product space and the angle between them. This test measures the difference in orientation between two vectors and returns a value between [0, 1]. Vectors which are closer to 1 are highly correlated. In this test it considers X as the samples and Y as the labels for each CPG in the samples, finds the cosine similarity using the corresponding labels to obtain the outcome. All the CPGs were sorted in decreasing order by outcome and the top 1000 CPGs chosen which defines that those CPGs are highly oriented and they are very close to the labels. This was performed in Matlab.

The following Matlab function was used to generate p-values:

$$score = (dot(x, y))/(\|x\| * \|y\|) \quad (2)$$

3.2.4 Standard Regression Coefficients

In this thesis, scores were calculated for each CPG with its corresponding label using the Supervised Principal Component Analysis (SPCA) method. In Eric Bair et al [25] explain that by calculating the univariate standard regression coefficients of the features they can identify the features which are highly correlated to the labels. These coefficients measure the univariate effect of each feature separately on labels. The dataset considered in this paper is represented as X as (N x p) matrix where N is number of samples, and p is the number of features associated with it and Y is the corresponding labels. In this method, they calculated the score for each feature correlating to the label and selected those features that were above the certain threshold.

In this thesis, scores were calculated for each CPG with their corresponding label and sorted in descending order. The top 1000 CPGs were selected which were highly correlated to the labels. This was performed in Matlab.

The following Matlab function was used to generate scores:

$$s = (x' * y)/\sqrt{x' * x} \quad (3)$$

3.3 Randomly Generating the Datasets from Statistical Tests

Each dataset contained 27,578 features. Using the cox test, the top 1000 features/CPGs were chosen, yielding the training dataset. The dimensionality of the other datasets were reduced to include only the indices of those features chosen, yielding the testing datasets. Next, n subsets of 50 features each were chosen. This final *n* different sets of datasets served as the final training and testing datasets for the sample.

For example if *n* = 100, 100 different training datasets were generated by selecting 50 features randomly from the top 1000 features. Each dataset consisted of 50 unique features and the test dataset retained the features of the training dataset. The above process was repeated for all the tests: t-test, cosine similarity and standard regression coefficients.

3.4 Generic Classifier: SUPPORT VECTOR MACHINE

In this thesis SVM [43] is used as a generic classifier. A support vector machine is a supervised machine learning algorithm is used to analyze labeled data in classification and regression problems. It provides the optimum hyperplane which is used to categorize the new samples. In other words, it is a hyperplane based classifier formulated explicitly on the maximum margin principle. It finds not just an arbitrary hyperplane separating the data but one having the maximum margin on the training data. The main goal of the SVM is to find the values of w and b

which provides the maximum margin, where w refers to weight and b refers to bias. Figure (3) shows different nonoptimal hyper planes.

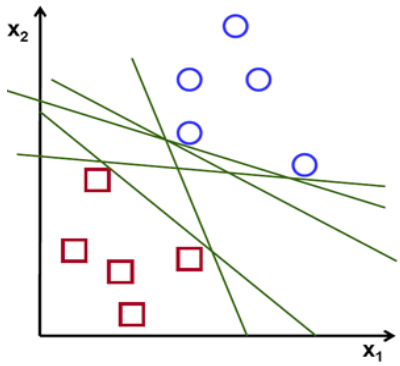

Figure (3). Example of different hyperplanes.
[http://docs.opencv.org/2.4/_images/separating-lines.png]

The optimal hyper plane is computed by Equation (2)

$$y = sign(w^T * x + b) \quad (4)$$

Where *w* is weight, *b* is the bias term and *x* is the training examples

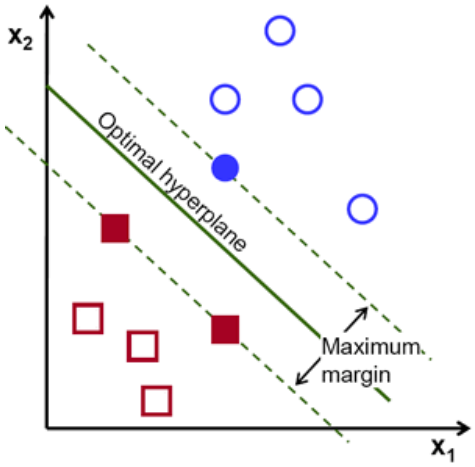

Figure 4. Example of optimal hyperplane.
[http://docs.opencv.org/2.4/_images/optimal-hyperplane.png]

Primary Lagrangian problem with constraints

$$f(w,b,\alpha) = min_{w,b}\left(L_P \equiv \frac{\|w\|^2}{2} - \sum_{i=1}^{n}\alpha_i y_i(x_i \cdot w + b) + \sum_{i=1}^{n}\alpha_i\right), \alpha_i \geq 0, \frac{\partial L_P}{\partial \alpha_i} = 0 \quad (5)$$

In this thesis, SVM is used to find the optimal hyperplane that classifies the data which is in 50 dimensional space and finds the optimal 49 dimensional plane which is used to classify the samples as represented by weights and biases. Here different kernels are used to generate different hyperplanes in order to obtain the maximum margin (linear, quadratic, polynomial, RBF, mlp). Each of them is discussed below.

3.4.1 Kernels in SVM

Kernels are used to map the training data onto a kernel space [14], making linear data work in nonlinear settings. The author of the book Support Vector and Kernel Machines, N.Cristianini, explains that kernels use the information of the "inner product between the data items" [20]. If the kernel of the algorithm is given then there is no need to specify the features of the data used.

Inner product of two vectors

$$\langle \bar{x}, \bar{y} \rangle = \sum_i x_i Z_i \quad (6)$$

To illustrate the concept, consider the binary classification problem. Each sample consists of a single feature represented as x as there is no linear separable plane exits for classification.

Figure (5). Example of 1- dimensional representation

Now consider the new representation of the dataset which is represented as two features x and x^2. Figure (6) represents a kernel mapping of this representation with a hyperplane that linearly separates the two classes.

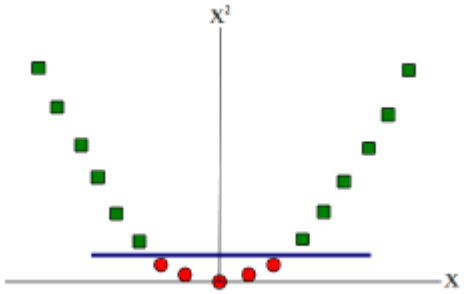

Figure (6). Example of multi-dimensional representation

3.4.1.1 Linear Kernel

A linear kernel is a dot product operation [14] which is used to find the linear relationship between data points and labels (case, control). Each feature is compared with the labels. If the feature correlates with the label it will be classified as 1. Otherwise, it will be 0.

3.4.1.2 Quadratic Kernel

This kernel is the special case of a polynomial kernel where degree of polynomial is 2. This is the standard choice in order to avoid the over-fitting of the datasets. This is because, as the number of degree increases probability of over-fitting increases.

Quadratic Kernel

$$k(x, z) = (x^T * z)^2 \quad (7)$$

3.4.1.3 Polynomial Kernel

This kernel is used when the models are non-linear in nature. A polynomial kernel not only provides a way to compare the given features with the labels but also allows a different combination of features to be compared with labels. This allows the classifier to increase the probability of identifying the pattern between features and labels.

For degree-d polynomials, the polynomial kernel is defined as

Polynomial Kernel

$$K(x, y) = (x^T y)^d \quad (8)$$

Where x and y are in the input samples, i.e. vectors of features computed from training samples

3.4.1.4 RBF Kernel

This kernel is used to find the distance from the origin, typically Euclidean distance. Finding the distance from the origin is called the radial function. For the vector x and the center y the radial basis function depends upon the distance between x and y [21]

RBF Kernel

$$K(x, y) = \exp(-gamma * |x - y|^2) \quad (9)$$

Here, gamma is a hyperparameter (also called the kernel bandwidth) and value of the gamma is calculated by cross validation. Regularization parameter lambda is used on the validation set to find the suitable gamma. For simplicity in most cases lambda is set to 0.0001

3.4.1.5 MLP kernel

This is one of the oldest kernels, but it is still used. It works based on "error driven learning"[15].This means that it iteratively runs on the training dataset and, whenever it finds incorrect classification, it improves the model accuracy by updating the model parameters. Where k and θ are constants.

MLP Kernel

$$y = tanh(kx_i^t x + \theta) \quad (10)$$

3.5 Passing the datasets to SVM classifier with different kernels

In this thesis, after obtaining the training datasets and corresponding test datasets from the statistical test. SVM classifier was fed with training datasets and the test datasets to generate

the models using different kernels. This function is repeated for all the statistical tests (t-test, cosine test, univariate regression coefficients).

The Matlab function used to generate the model and to test the model with the corresponding test datasets is as follows.

SVMStruct = svmtrain (data_new {1, e}, data758c,'kernel_function','linear');

Svmlinear ( ) = svmclassify (SVMStruct, data_new {6, e});

In this code snippet SVMStruct represents the SVM model generated by training a particular dataset. The test datasets are passed to the model in order to obtain the labels. Svmlinear consists a vector of labels for each dataset. This method is repeated for *n* number of times in order to obtain the different labels for *n* different datasets. This method is repeated for different kernels to obtain *n* different labels from each kernel then repeated the entire process for all the statistical tests i.e. t-test, Cosine similarity and univariate regression coefficients.

3.6 Ensemble method - Majority voting and weighted majority voting

The main goal of implementing the Ensemble method is to automatically generate an average classifier by giving less importance to poor performing classifiers, and more importance to top performing classifiers. Here in this thesis the weighted majority voting algorithm is chosen as the ensemble method. The models were obtained from module 2 and, using the accuracies of the models and the amplification term, weights were calculated using equation 9. These weights were assigned to all the final models to obtain the AUC curve.

Calculating weights

$$w = \frac{e^{\alpha a_1}}{e^{\alpha a_1} + \cdots + e^{\alpha a_n}} \tag{11}$$

Where α is the constant amplification term, w is the weight and a, is the accuracy of each classifier.

Below gives the sample example of the weighted majority algorithm. Let us consider the accuracies of three classifiers .20, .65, .90 with amplification $\alpha = 1,5,10$

Case1: classifier with accuracy = .20 and $\alpha = 1,5,10$

$$w = \frac{e^{1*.20}}{e^{1*.20}+e^{1*.65}+e^{1*.90}} \qquad w = .22$$

$$w = \frac{e^{5*.20}}{e^{5*.20}+e^{5*.65}+e^{5*.90}} \qquad w = .023$$

$$w = \frac{e^{10*.20}}{e^{10*.20}+e^{10*.65}+e^{10*.90}} \qquad w = .0008$$

Case 2: classifier with accuracy = .65 and $\alpha = 1,5,10$

$$w = \frac{e^{1*.65}}{e^{1*.20}+e^{1*.65}+e^{1*.90}} \qquad w = .34$$

$$w = \frac{e^{5*.65}}{e^{5*.20}+e^{5*.65}+e^{5*.90}} \qquad w = .22$$

$$w = \frac{e^{10*.65}}{e^{10*.20}+e^{10*.65}+e^{10*.90}} \qquad w = .075$$

Case 3: classifier with accuracy = .90 and $\alpha = 1,5,10$

$$w = \frac{e^{1*.90}}{e^{1*.20}+e^{1*.65}+e^{1*.90}} \qquad w = .44$$

$$w = \frac{e^{5*.90}}{e^{5*.20}+e^{5*.65}+e^{5*.90}} \qquad w = .76$$

$$w = \frac{e^{10*.90}}{e^{10*.20}+e^{10*.65}+e^{10*.90}} \qquad w = .92$$

From the above results it can be explained that as the amplification term increases, the weights of top performing classifier increases whereas the weights of the poor performing classifier decreases. In this thesis, the master algorithm is implemented as follow:

Training Phase: 1

D_{train} = Training Dataset
y_{train}= Training Labels
x = Number of statistical tests
y = Number of sample indexes
z = Number of machine learning algorithms
b = Sub sample count
$sample_{ij}$= 2D array whose *(i, j)* entry ($sample_{ij}$) corresponds to top 50 randomly chosen features as per the p-values of the statistical test ***i***
Start:

Input: D_{train}, y_{train}
Output: θ

for ***i*** from 1 to x (where x = 4 in this thesis)
    Let ***S*** be the top $\boldsymbol{m_i}$ features corresponds to the top $\boldsymbol{m_i}$ p-values of statistical test ***i*** (here m_i = 1000)
    for ***j*** from 1 to b (b =100)
        $sample_{ij} = rand(50)$ (Random sample of 50 features from m_i)
        Let $D_{subset,ij}$ be the reduced subsample having features from $sample_{ij}$
        for ***k*** from 1 to z (where z = 5: SVM with 5 different kernels)
            θ is the 3D array whose *(i, j, k)* entry θ_{ijk} corresponds to the parameters of an ML algorithm k trained using the $D_{subset,ij}$ and y_{train}
        end
    end
end
Stop

Training Phase: 2

D_2^{CV}= Cross-validation training sample
y_2^{CV}= Cross-validation labels
α = Amplification term
w = Weights
θ= Parameters of machine learning algorithm
x = Number of statistical tests
y = Number of sample indexes
z = Number of machine learning algorithms
b = Sub sample count
Start:

Input: $D_2^{CV}, \theta, \alpha, y_2^{CV}, sample$
Output: w

Let $D^{CV}_{subsample\_2,ij}$ be the reduced subsample having features from $sample$
for ***i*** from 1 to x (where x = 4 in this thesis)

for ***j*** from 1 to *b* (*b* =100)

for ***k*** from 1 to *z* (where *z* = 5: SVM with 5 different kernels)

Let a_{ijk} be the accuracy after applying ML algorithm *k* with parameters θ_{ijk} on $D^{CV}_{subsample\_2,ij}$ and y_2^{CV}

end

end

end

for all *i, j, k* as above normalize w_{ijk} as shown in the below equation

$$w_{ijk} = \frac{e^{\alpha a_{ijk}}}{\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z} e^{\alpha a_{ijk}}}$$

Stop

Testing Phase:

w = Weights

θ= Parameters of machine learning algorithm

D_2^{Test}= Final samples for testing set

c = Random sample from D_2^{Test}

Start:

Input:$w,\ \theta, sample, c$

Output: Label predicted for *c*

variable $sum = 0$

for ***i*** from 1 to *x* (where *x* = 4 in this thesis)

for ***j*** from 1 to *b* (*b* =100)

for ***k*** from 1 to *z* (where *z* = 5: SVM with 5 different kernels)

Let $pred$ be the prediction of ML algorithm *k* with parameters θ_{ijk} run on *c* using features from $sample$

$sum = sum + w_{ijk} * pred$

end

end

end

$$final\_labels = \begin{cases} 1, & if\ sum \geq 0.5 \\ 0, & otherwise \end{cases}$$

return $final\_labels$

Stop

In the field of bioinformatics we often encounter a situation with a small number of samples and a large number of features. This makes it hard for the predictive algorithms to understand these features given the small number of samples. The main advantage of using this

complex framework is that it has great flexibility in each module for adding or removing various algorithms. In module 1, we can add or remove various statistical tests according to the different datasets. In module 2, we can add or remove various machine learning algorithms and can also use different combinations of machine learning algorithms in order to attain better accuracy. Finally, in module 3, it is possible to use various ensemble methods. This makes the framework dynamic and universal, making it useful for the field of bioinformatics.

CHAPTER 4

DATASETS AND RESULTS

In this chapter details of empirical evaluation and comparison with various state of the art methods are presented. In section 4.1 we present the datasets used in this framework. In section 4.2 we present the evaluation metric used in this thesis. In section 4.3 we present results of the proposed modular.

4.1 Datasets

In this thesis, four independent datasets were accessed from the paper MS-SPCA which in turn accessed the datasets from the Gene Expression Omnibus repository [16] accession numbers GSE20080, GSE30760 (SuperSeries comprising GSE30758 and GSE30759) and GSE37020. CpG methylation was measured using Illumina's Infinium Human Methylation 27K Beadchips [17]. Here pre-processed datasets from MS-SPCA were used. Each sample consists of 27,578 CPGs/features. Datasets were labeled as Normal, CIN2+ (a), CIN2+ (b) and Cancer, described in Table (1).

Table 1. Datasets

GEO	Name	Cases	Controls
GSE30758	Normal	75	77
GSE30758	Normal HPV+	44	48
GSE30758	Normal HPV-	31	29
GSE20080	CIN2+(a)	18	30
GSE37020	CIN2+(b)	24	24
GSE30759	Cancer	48	15

4.2 Evaluation metric

Area under the ROC curve or Area under curve (AUC) is the most common evaluation metric used for binary classifiers. The AUC curve explains the performance of the binary classifier when the specific threshold of the classifier is varied. The AUC curve is obtained by plotting the true positive rate (TPR), where TPR is a measure of classifying the labels correctly against the false positive rate (FPR), where FPR is the measure of classifying the labels incorrectly. The range of the AUC values is between 0 and 1. If the performance of the classifier is good, the true positive rate increases very quickly making the AUC very close to 1. If the performance of the classifier is bad, the growth of the true positive rate will be very slow and the score of the AUC will be very low [38]. If AUC = 1 then all the labels are classified accurately and if AUC = 0 than all the labels are classified inaccurately.

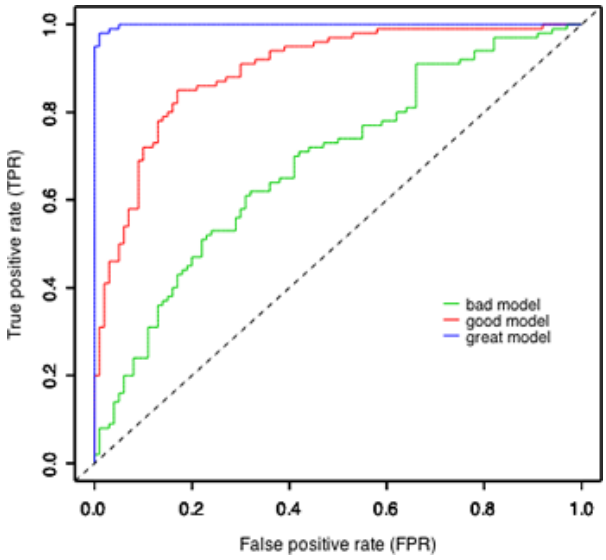

Figure (7). Example of AUC curve
[https://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture22/fig4.png]

4.3 Results

Here a novel modular framework has been developed which considers models generated from statistical tests and combines them to obtain the final prediction. This is in contrast to [4],

which considers several models that perform well in cross-validation and selects the final independent models for predicting the test data.

In this thesis five different datasets were used for generating the models, excluding the Cancer dataset, which was only used in the testing phase. For each dataset, the 1000 most significant features were selected using each of four statistical tests: the cox-test, t-test, cosine test, and univariate regression test. These features corresponded to the top 1000 CPGs in terms of the difference between cases and controls for the CPG. For the t-test and Cox test, the difference was measured in terms of p-values. For the univariate regression test and cosine test, the difference was measured in terms of scores. The purpose of this step was to help the classifier to predict results more accurately than it would if considering all features in the dataset.

For each statistical test, the top 1000 features were selected. For each model, 50 of the top 1000 features were randomly selected. In this thesis 100 models were generated.

LIBSVM was used on each of the other five datasets from [16] and the results were used to generate models. For datasets CIN2+(a) and CIN2+(b), each SVM kernel available in LIBSVM (i.e. linear, polynomial, quadratic, RBF, and MLP) was used to generate a separate model, yielding a total of 500 different sets of models for each hypothesis test. Thus, a total of 2000 sets of models were generated for CIN2+(a) and CIN2+(b). The process was the same for all other datasets; however, for the Normal and Normal HPV+ datasets, the linear SVM model could not find the optimal hyperplane for any hypothesis test, so only 1600 sets of models were generated. For the Normal HPV- dataset, the linear SVM model could only find the optimal hyperplane for the Cox test, so 1700 models were generated.

Weighted majority voting was then used to choose the final set of models. During this process, 10% of the models were used as the validation set and the other 90% as the test set. The models of the validation set were compared with the true labels to obtain the accuracy of each model. The models were then sorted in decreasing order and the weights calculated using the weighted majority voting algorithm. Finally, the model produced using these weights was applied to the test dataset and the AUC scores were obtained from the results.

All of the 21 AUC scores in comparison with those of [4] are shown in the table below, separated by slashes. The first result in each cell belongs to framework and the second belongs to [4].

Table 2. Results of framework compared with [4]. Row indicates the train dataset and the Column indicates the test dataset.

	Normal	Normal HPV+	Normal HPV-	CIN2+(a)	CIN2+(b)	Cancer
Normal				**.90**/.93	**.88**/.81	**1**/1
Normal HPV+			**.69**/.52	**.90**/.93	**.83**/.84	**1**/1
Normal HPV-		**1**/.61		**.71**/.92	**.64**/.64	**.90**/1
CIN2+(a)	**.67**/.60	**.61**/.63	**.80**/.53		**.86**/.83	**.99**/1
CIN2+(b)	**1**/.58	**1**/.62	**.63**/.53	**.80**/.87		**.98**/.98

During the weighted majority voting process, the number of models and the amplification term was varied and those settings yielding the optimal accuracies were selected. The settings that produced the results in the table above are listed below.

Normal (Training set) Vs CIN2+ (a) (Testing and validation):
Total Models: 1600
Models selected: 1400
Amplification: 0.9
AUC score: 0.90

Normal (Training set) Vs CIN2+ (b) (Testing and validation):
Total Models: 1600
Models selected: 100
Amplification: 0.05
AUC score: 0.88

Normal (Training set) Vs Cancer (Testing and validation):
Total Models: 1600
Models selected: 800
Amplification: 0.05
AUC score: 1

Normal HPV+ (Training set) Vs Normal HPV- (Testing and validation):
Total Models: 1600
Models selected: 100
Amplification: 24
AUC score: 0.69

Normal HPV+ (Training set) Vs CIN2+ (a) (Testing and validation):
Total Models: 1600
Models selected: 1600
Amplification: 1
AUC score: 0.90

Normal HPV+ (Training set) Vs CIN2+ (b) (Testing and validation):
Total Models: 1600
Models selected: 100
Amplification: 0.3
AUC score: 0.83

Normal HPV+ (Training set) Vs Cancer (Testing and validation):
Total Models: 1600
Models selected: 500
Amplification: 1
AUC score: 1.00

Normal HPV- (Training set) Vs Normal HPV+ (Testing and validation):
Total Models: 1700
Models selected: 100
Amplification: 25
AUC score: 1

Normal HPV- (Training set) Vs CIN2+ (a) (Testing and validation):
Total Models: 1700
Models selected: 500
Amplification: 30
AUC score: .71

Normal HPV- (Training set) Vs CIN2+ (b) (Testing and validation):
Total Models: 1700
Models selected: 500
Amplification: 0.05
AUC score: 0.64

Normal HPV- (Training set) Vs Cancer (Testing and validation):
Total Models: 1700
Models selected: 1500
Amplification: 0.05
AUC score: 0.90

CIN2+ (a) (Training set) Vs Normal (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 60
AUC score: 0.67

CIN2+ (a) (Training set) Vs Normal HPV+ (Testing and validation):
Total Models: 2000
Models selected: 500
Amplification: 50
AUC score: 0.61

CIN2+ (a) (Training set) Vs Normal HPV- (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 23
AUC score: 0.80

CIN2+ (a) (Training set) Vs CIN2+ (b) (Testing and validation):
Total Models: 2000
Models selected: 2000
Amplification: 18
AUC score: 0.86

CIN2+ (a) (Training set) Vs Cancer (Testing and validation):
Total Models: 2000
Models selected: 1000
Amplification: 0.05
AUC score: 0.99

CIN2+ (b) (Training set) Vs Normal (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 36
AUC score: 1.00

CIN2+ (b) (Training set) Vs Normal HPV+ (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 17
AUC score: 1.00

CIN2+ (b) (Training set) Vs Normal HPV- (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 18
AUC score: 0.63

CIN2+ (b) (Training set) Vs CIN2+ (a) (Testing and validation):
Total Models: 2000
Models selected: 300
Amplification: 20
AUC score: 0.80

CIN2+ (b) (Training set) Vs Cancer (Testing and validation):
Total Models: 2000
Models selected: 100
Amplification: 0.05
AUC score: 0.98

CHAPTER 5

CONCLUSION

For the majority of the 21 tests performed, the novel framework equaled or outperformed the framework used in [4] as evidenced in Table 2. In four of these tests, framework outperformed [4] by 27% or more. For the tests in which [4] outperformed framework, the difference was no greater than 10% except in the case of Normal HPV- vs CIN2+(a), in which the difference was 21%. This indicates that framework gives more accurate results than the previous method and in some cases a vast improvement is observed.

In addition, proposed framework is simpler than the previous method and more flexible. It is simpler because a maximum of 2000 models are considered for each test in the framework, whereas 35,130 are considered in [4]. It is more flexible because models can be combined to produce a new model and because other classification methods can be used in addition to SVM and their results combined with those of SVM. In contrast, models generated by MS-SPCA cannot be combined and only the univariate regression method can be used.

These results indicate that proposed framework as described in this thesis is an effective method for use in DNA methylation studies. This method may be useful in future studies of this nature as well as making potentially interesting predictions for the data studied herein. Additionally, the framework described in this thesis may be useful for solving other types of feature selection problems in bioinformatics.

5.1 Future Work

It is notable that the CPGs found to significantly affect the disease group of the patient must be examined from a biological perspective before any medical conclusions can be drawn.

This method does not take into account all possible biological relationships between CPGs and cannot singlehandedly determine whether a particular CPG is correlated with a particular stage of cancer.

A potential area of future research is to combine different types of classifiers in cases where SVM alone is not sufficient, such as in the Normal HPV- vs CIN2+ (a) test. This may result in better-performing models due to the higher degree of flexibility. Furthermore, it may prove fruitful to experiment with different statistical tests, as certain statistical tests may be more applicable to particular studies.

REFERENCES

REFERENCES

[1] http://www.news-medical.net/life-sciences/What-is-DNA-Methylation.aspx [Accessed: May 7, 2016].

[2] http://www.whatisepigenetics.com/dna-methylation/ [Accessed: May 7, 2016]

[3] Jin, B., Li, Y., Robertson, K.D., "DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?" Esteller M, ed. Genes & Cancer. 2011;2(6):607-617. doi:10.1177/1947601910393957.

[4] Wilhelm, "Phenotype prediction based on genome-wide DNA methylation data". BMC Bioinformatics 2014 15:193.

[5] Jaenisch, R., Bird, A.,"Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals". (2003) Nature Genetics. 33 Suppl (3s): 245–254. doi:10.1038/ng1089. PMID 12610534.

[6] Bock, C., "Analysing and interpreting DNA methylation data". Nat Rev Genet 2012, 13:705–719.

[7] Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S., "Epigenome-wide association Studies for common human diseases". Nat Rev Genet 2011, 12:529–541.

[8] McKay, J.A., Mathers, J.C., "Diet induced epigenetic changes and their implications for health". Acta Physiol (Oxf) 2011, 202:103–118.

[9] Slomko, H., Heo, H.J., Einstein, F.H., Minireview, "Epigenetics of obesity and diabetes in humans". Endocrinology 2012, 153:1025–1030.

[10] De Carvalho, D.D., Sharma, S., You, J.S., Su, S.F., Taberlay, P.C., Kelly, T.K., Yang, X., Liang, G., Jones, P.A., "DNA methylation screening identifies driver epigenetic events of cancer cell survival". Cancer Cell 2012, 21:655–667.

[11] Feinberg, A.P., Irizarry, R.A., Fradin, D., Aryee, M.J., Murakami, P., Aspelund, T., Eiriksdottir, G., Harris, T.B., Launer, L., Gudnason, V., Fallin, M.D., "Personalized epigenomic signatures that are stable over time and covary with body mass index". Sci Transl Med 2010, 2:49ra67.

[12] Bibikova, M., Fan, J.B., "Genome-wide DNA methylation profiling". Wiley Interdiscip Rev Syst Biol Med 2010, 2:210–223.

[13] Bhattacharya, Bhaskar. Habtzghi, DeSale., "Median of the p value under the alternative hypothesis". (2002) The American Statistician (American Statistical Association) 56 (3): 202–6. doi:10.1198/000313002146.

[14] http://www.mathworks.com/help/stats/svmtrain.html [Accessed: May 7, 2016]

[15] Marvin, Minsky., Seymour, Papert., "Perceptrons: an introduction to computational geometry". Originally published: 1969

[16] Edgar, R., Domrachev, M., Lash, A.E., "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". Nucleic Acids Res 2002,30:207–210.

[17] Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., Gunderson, K.L., "Genome-wide DNA methylation profiling using Infinium assay". Epigenomics 2009, 1:177–200.

[18] Teschendorff, A.E., Widschwendter, M., "Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions". Bioinformatics 2012, 28:1487–1494.

[19] Tomlins S.A., et al. "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer". Science 2005;310:644-648.

[20] http://www.support-vector.net/icml-tutorial.pdf [Accessed: May 7, 2016]

[21] David, Barber., "Bayesian Reasoning and Machine Learning" text book www.cs.ucl.ac.uk/staff/D.Barber/brml .Originally published: 2012

[22] Zhang, F.F.1., Cardarelli, R., Carroll, J., Zhang, S., Fulda, K.G., Gonzalez, K., Vishwanatha, J.K., Morabia, A., Santella, R.M., "Physical activity and global genomic DNA methylation in a cancer-free population". (2011) EPIGENETICS 6 (3): 293-299.doi:10.4161/epi.6.3.14378. PMC 3092677. PMID 21178401.

[23] Craig, J.M., Wong, N.C., "Epigenetics: A Reference Manual Caister". Academic Press. Editor (2011). ISBN 978-1-904455-88-2.

[24] https://en.wiktionary.org/wiki/hypermethylation#English [Accessed: May 7, 2016]

[25] Eric, Bair., Trevor, Hastie., Debashis, Paul., and Robert, Tibshirani., "Prediction by supervised principal components". September 15, 2004

[26] https://en.wiktionary.org/wiki/hypomethylation#English [Accessed: May 7, 2016]

[27] The website of the National Cancer Institute (http://www.cancer.gov) [Accessed: May 7,2016]

[28] Mark, D.M. Leiserson, Hsin-Ta, Wu., Fabio, Vandin., Benjamin J. Raphael
Genome Biol. 2015; 16(1): 160. Published online 2015 August 8. doi: 10.1186/s13059-015-0700-7 PMCID: PMC4531541
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531541/

[29] Jordan M. Thompson., Quy H. Nguyen, Manpreet, Singh., and Olga v. Razorenova., "Approaches to Identifying Synthetic Lethal Interactions in Cancer". Molecular Biology and Biochemistry Department, University of California Irvine, Irvine, California PMCID: PMC4445436 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445436/.

[30] L.J. Engle, C.L. Simpson and J.E. Landers., "Using high-throughput SNP technologies to study cancer Oncogene (2006)". 25, 1594–1601. doi:10.1038/sj.onc.1209368 http://www.nature.com/onc/journal/v25/n11/full/1209368a.html

[31] http://www.mathworks.com/help/stats/ranksum.html [Accessed: July 12, 2016]

[32] http://www.mathworks.com/help/stats/ttest2.html [Accessed: July 12, 2016]

[33] Lior, Rokach.,"Ensemble-based classifiers". http://www.ise.bgu.ac.il/faculty/liorr/AI.pdf, Artif Intell Rev (2010) 33:1–39, DOI 10.1007/s10462-009-9124-7

[34] Harvey, Lodish., Arnold, Berk., Paul, Matsudaira., Chris, A. Kaiser "Molecular Cell Biology (5th ed.)" (2004). New York: W.H. Freeman and Company. ISBN 0-7167-4366-3.

[35] https://ghr.nlm.nih.gov/primer/genomicresearch/snp [Accessed: August 8, 2016]

[36] http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf [Accessed: August 8, 2016]

[37] Ma, Xiao-Jun et al, Cancer Cell, Volume 5 , Issue 6 , 607 – 616, "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen". 2004 June, PMID:15193263, DOI:10.1016/j.ccr.2004.05.015.

[38] http://mlwiki.org/index.php/ROC_Analysis#AUC:_Area_Under_ROC_Curve [Accessed: August 10, 2016]

[39] Gavin J. Gordon, Roderick V. Jensen, Li-Li Hsiao, Steven R. Gullans, Joshua E. Blumenstock, Sridhar Ramaswamy, William G. Richards, David J. Sugarbaker and Raphael Bueno.,"Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", Cancer Res September 1 2002 (62) (17) 4963-4967.

[40] Adi L. Tarca, Roberto, Romero., Sorin Draghici Am J Obstet Gynecol .,"Analysis of microarray experiments of gene expression profiling". Author manuscript; available in PMC 2008 Jun 22.Published in final edited form as: Am J Obstet Gynecol. 2006 Aug; 195(2): 373–388. doi: 10.1016/j.ajog.2006.07.001 PMCID: PMC2435252.

[41] http://hunch.net/~coms-4771/quinlan.pdf [Accessed: September 27, 2016]

[42] Altman, N. S.,"An introduction to kernel and nearest-neighbor nonparametric regression".(1992). *The American Statistician*. **46** (3): 175–185.doi:10.1080/00031305.1992.10475879.

[43] Cortes, C., Vapnik, V., "Support-vector networks".(1995) *Machine Learning*. **20** (3): 273–297. doi:10.1007/BF00994018.

[44] Russell, Stuart., Norvig, Peter., "Artificial Intelligence: A Modern Approach (2nd ed.)". (2003) [1995]. Prentice Hall. ISBN 978-0137903955.

[45] http://statistics.berkeley.edu/sites/default/files/tech-reports/486.pdf [Accessed: September 27, 2016]

[46] Breiman, Leo.,"Bagging predictors". (1996) Machine Learning. 24 (2): 123–140. doi:10.1007/BF00058655. CiteSeerX: 10.1.1.32.9399.

[47] Talluri and Shete. "Evaluating Methods for Modeling Epistasis Networks with Application to Head and Neck Cancer". Cancer Informatics 2015:14(S2) 17–23 doi:10.4137/CIN.S17289.

[48] Adan, Niu., Shuanglin, Zhang., and Qiuying, Sha.,"A Novel Method to Detect Gene–Gene Interactions in Structured Populations: MDR-SP", Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA, doi: 10.1111/j.1469-1809.2011.00681.x.

[49] Xiang, Wan., Can, Yang., Qiang, Yang., Hong, Xue., Nelson, L.S. Tang4., and Weichuan, Yu1., "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study", doi:10.1186/1471-2105-10-13.