8-2021

Significant Gene Array Analysis and Cluster-Based Machine Learning for Disease Class Prediction

Myrine A. Barreiro-Arevalo
The University of Texas Rio Grande Valley

SIGNIFICANT GENE ARRAY ANALYSIS AND CLUSTER-BASED MACHINE LEARNING FOR DISEASE CLASS PREDICTION

A Thesis

by

MYRINE A. BARREIRO-AREVALO

Submitted to the Graduate College of
The University of Texas Rio Grande Valley
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2021

Major Subject: Applied Statistics and Data Science

SIGNIFICANT GENE ARRAY ANALYSIS AND CLUSTER-BASED MACHINE LEARNING FOR DISEASE CLASS PREDICTION

A Thesis
by
MYRINE A. BARREIRO-AREVALO

COMMITTEE MEMBERS

Dr. Hansapani Rodrigo
Chair of Committee

Dr. George Yanev
Committee Member

Dr. Tamer Oraby
Committee Member

Dr. Zhijun Qiao
Committee Member

August 2021

ABSTRACT

Barreiro-Arevalo, Myrine A., Significant Gene Array Analysis and Cluster-Based Machine Learning for Disease Class Prediction. Master of Science (MS), August, 2021, 90 pp., 45 tables, 19 figures, 76 references.

Gene expression analysis has been a major interest to biostatisticians for many decades. Such studies are necessary for the understanding of disease risk assessment and prediction, for a creation of better treatment plans, to lessen symptoms, and perhaps find cures. In this study, we have investigated how to incorporate clusters of genes based on prior biological knowledge into machine learning models for effective gene expression data analysis and to uncover differentially expressed (DE) genes for different disease pathologies. Gene expression datasets for multiple pathologies have been used to test model evaluation metrics and will be obtained using the Affymetrix U133A platform (GPL96).

Significant Analysis of Microarrays (SAM) had been used to identify potential disease biomarkers, followed by the predictive models: (a) random forest, (b) random forest with Gene eXpression Network Analysis (GXNA), (c) RF++, (d) LASSO, and (e) Bayesian Neural Networks. Deferentially expressed genes within the clusters of co-expressed networks, have been successfully identified where they may be used as potential biomarkers within their particular disease pathology. Moreover, we were able to utilize the Automatic Relevancy Determination prior to identify the relatively important genes with Bayesian neural networks effectively.

DEDICATION

To my parents, Margarita Barreiro and Juan J. Arevalo, whom supported me throughout my journey to higher knowledge and sacrificed so much for my well-being.

And to the rest of my family and friends, for their countless encouraging words. Thank you so much for believing in me and for cheering me on. I couldn't have done it without you.

ACKNOWLEDGMENTS

I would like to sincerely express my gratitude to my advisor, Dr. Hansapani Rodrigo, for providing me with continuous support and motivating me throughout my graduate research. I couldn't have had a better advisor than her.

My appreciation goes out to Dr. George Yanev, Dr. Tamer Oraby, and Dr. Zhijun Qiao, for serving on my thesis committee.

And I also would like to acknowledge those who helped along the way: Rebecca Bernal, Armando Garces, Viannay Cantu, Jonathan Arsola, Juan Hernandez, Scarlett Basurto, and Alejandra Baez.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION TO GENE EXPRESSION

Gene expression is the process by which the DNA information encoded in a gene is used to direct the assembly of a functional gene product, protein/coding or non-coding RNA, and ultimately affect a phenotypic manifestation, as the final effect by process of genetic transcriptions and translations (Figure 1.1) [1]. Here, transcription refers to the DNA being copied into an RNA molecule, and translation refers to the synthesis of organic compounds known as amino acids from the RNA coding.

In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype [2], better seen as an observable trait in an organism. The genetic information stored in DNA represents the genotype, whereas the phenotype results from the "interpretation" of that information from the coding RNA. Gene expression analysis determines the patterns of expression within a given genomic sequence in a specific cell or for specific phenotypes. Phenotypes such as diseases will be the primary interest in this study.

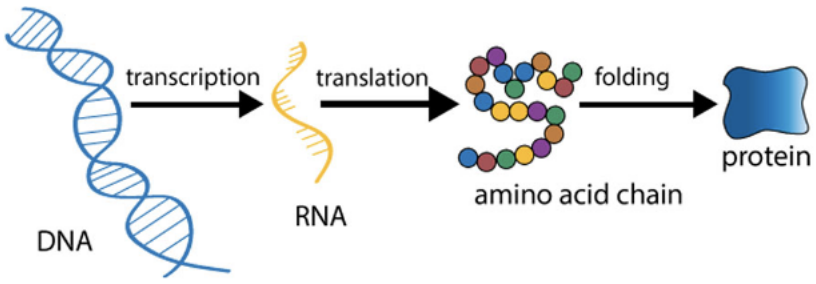


Figure 1.1: Visual representation of gene expression.

1.1 Types of RNA

Ribonucleic acids (RNA) are the building blocks of our genomic sequence and essential to the study of gene expression and gene expression analysis. There are two main types of RNA, these include (a) Coding RNA and (b) Non-coding RNA. We will focus on coding RNA, specifically what is known as messenger RNA or mRNA. Messenger RNA (mRNA) carries genetic code from DNA in a form that can be recognized by proteins. In the cell, mRNA arrives from the splicing of the RNA after its been transcribed from the DNA inside in the nucleus. It is then exported into the cytoplasm after splicing of the non-coding sequences of the RNA before being translated into the amino acid chain necessary for protein formation (Figure 1.2). There are about 23,000 mRNAs encoded in human genome [3]. mRNA is utilized in a technique that will be explored further known as DNA microarray analysis, wherein the mRNA will be used to extract expression values of a genomic data from subjects presenting with known diseases.

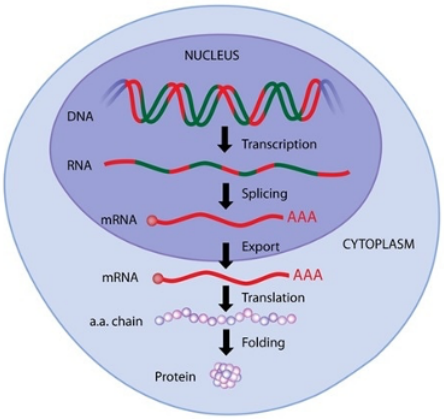


Figure 1.2: A cell showing the process of DNA to protein formation.

1.2 Microarray Technology

DNA microarray technology (also called DNA chip technology) has had a specific significant breakthrough in the field of molecular biology because of its capability of handling thousands of gene expression data simultaneously [4]. A microarray is a hybridization of a sample of mRNA to a

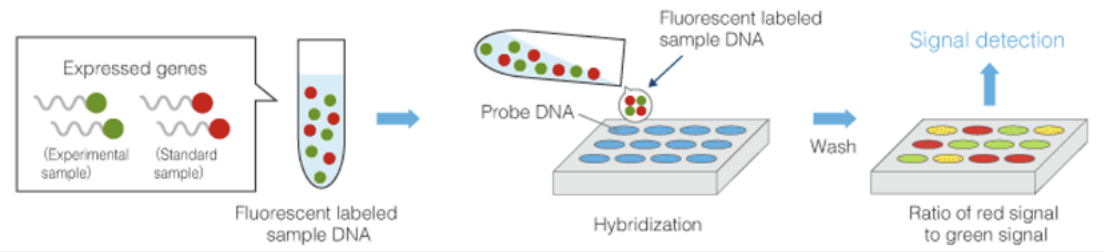


Figure 1.3: Visual representation of a microarray.

set of oligonucleotide probes on an agar plate, which is then run under a fluorescent microscope to measure expression levels of a gene (Figure 1.3).

The discrepancies in microarray results are a consequence of differences in microarray measures, such as accuracy (i.e. the degree of conformity of the measured quantity to its actual (true) value), sensitivity (i.e. the concentration range of target molecules in which accurate measurements can be made), and specificity (i.e. the ability of a probe to provide a signal that is influenced only by the presence of the target molecule) [4].

Affymetrix pioneered the field of microarrays analysis with oligonucleotides as probes, thereby expanding the study of gene expression in various types of organisms [5]. The datasets used in this research study have come from the Affymetrix Human Genome U133A platform (GPL96).

1.3 Literature Review

Several previous studies have explored using machine learning techniques on gene expression data to improve diagnoses in healthcare and clinical applications [6]. In fact, machine learning methods such as k-nearest neighbor (KNN), support vector machines (SVM) [7, 8], and even random forests [9] have been widely used to explain relationships between genes and disease states. These methods have high interpretability, they suffer from low accuracy due to their inability to associate gene-to-gene interactions and co-functionality of genes in their models. Artificial intelligence (AI) methods within the machine learning field such as convolutional neural networks (CNN) have been used previously to generate images of gene expression have been developed with robust imaging

algorithm and have improved accuracy of disease state classification but in contrast suffer from low interpretability due to their complex nature [10]. Bayesian networks for gene expression data has been performed numerous times as discussed in de Campos et al. [11] and have yielded good accuracy results in classification of disease state, but Bayesian Neural Networks (BNNs) are not among common methods for studying gene data.

As per our knowledge, no previous studies have incorporated gene interaction networks such as those identified by the GXNA into the random forest models as a hybrid approach to analyzing gene expression data as we will be doing in this study. It is generally well known that genes do not act by themselves, but rather in closely linked sets of genes or pathways that are all necessary to perform a function, known to us as "clusters"; if one gene is missing from the cluster, the intended phenotypic function cannot be performed. However, while the idea that co-expression is touched on, nothing has been introduced as of yet to determine the most important genes in these classifications. Artificial neural networks (ANNs) have been used widely for building cancer prediction models from microarray data, described thouroughly by Daoud et al. [12]. In turn, we will be introducing the concept of using the Automatic Relevance Determination prior for gene expression data in order to identify the most important genes according to the α value within a ANN model trained under Bayesian inference. These two novel methods have the possibility to bring better explainability to AI machine learning within the medical and scientific community.

1.4 Research Objectives

Identification of co-expressed, or functionally related pathways, from an organism's genome has been a challenging research problem for the scientific community, one that we will be investigating in this study. More specifically, the research objectives of this study are as follows:
(1) Use of Significant Analysis of Microarrays to determine the differentially expressed genes among different pathologies (breast cancer, lung cancer, Parkinson's disease).
(2) Identification of co-expressed gene clusters using Gene eXpression Network Analysis.

(3) Use of gene clusters with machine learning methods for improving pathophysiology.

(4) Incorporation of Automatic Determination prior in analyzing gene expression data with Bayesian Neural Networks.

(5) Analyzing impact on the predictive power of machine learning methods on gene expression data with different dimension reduction techniques (use of gene clusters identified through GXNA and LASSO).

In Chapter 2, we will investigate methods of clustering genes based on their interactions within their genetic pathways to be used in our relevant machine learning models for gene expression analysis and its effects on accuracy of disease state classification. Chapters 3 and 4 will reveal the results of objectives (1)-(5).

CHAPTER II

METHODOLOGY

In this chapter, we present the details of our four datasets used in the gene expression analysis and discuss the technicalities and relevant ideas of the predictive models used.

2.1 Subjects

Multiple gene expression datasets will be used to test model accuracy and were obtained from the Gene Expression Omnibus (GEO) [13], a public access database maintained by the National Center for Biotechnology Information (NCBI):

GSE 2034: this dataset contains the gene expression data of 230 subjects of which 144 subjects were lymph-node negative relapse free patients and 86 subjects were lymph-node negative patients that developed a distant metastasis (breast cancer). All patients were female. Of the 86 subjects identified with breast cancer, 10 subjects further metastasized to the brain and spinal cord through the blood or lymph system.

GSE 2990: this dataset contains the gene expression data of 149 subjects of which 93 subjects were lymph-node negative relapse free patients and 56 subjects were lymph-node negative patients that developed a distant metastasis (breast cancer). All patients were female. Of the 56 subjects identified with breast cancer, the use of gene expression grade indexing was performed to determine histologic grading of tumors was an effective method to predict recurrence.

GSE 4115: this dataset contains the gene expression data of 197 subjects of which 90 subjects were frequent cigarette smokers with abnormalities found in the bronchial epithelium (lung

cancer) and 97 subjects were frequent cigarette smokers without abnormalities found. Abnormalities were obtained through a clinical bronchoscopy.

GSE 8397: this dataset contains the gene expression data of 47 subjects of which there were 29 post-mortem brain tissue samples from individuals diagnosed with Parkinson's disease and 18 post-mortem brain tissue samples from control individuals.

2.2 Data Processing

DNA microarrays measure gene product levels of thousands of genes by attaching probes on a gene array platform or a collection of gene-specific nucleic acids on a solid surface that have defined "targets" for mRNA within a DNA sample [14]. All gene expression data was first retrieved using a custom chip description file (CDF) package from the current R Bioconductor or retrieved manually from Bioconductor source files. Custom CDFs in Affymetrix GeneChips were based on the best gene clustering and genomic sequence information available at the time of chip design [15]. Affymetrix gene platforms contain a significant portion of cluster where a subset of oligonucleotide probes in a probe set may be assigned to another gene or more than one gene, to remedy this situation the custom CDF allows for removal of extra probes to ease interpretation of data [15].

Probe data containing 25 base pair length oligonucleotides used to match to the RNA targets of specified genes was normalized using *gcrma* normalization techniques from R Bioconductor. *gcrma* adjusts for background intensities in Affymetrix gene data, including optical noise and non-specific binding (NSB) using probe sequence information to estimate probe affinity [16]. In addition to normalization, Affy control probes were removed. A non-specific filtration of genes called *pOverA* [17] was applied to each dataset before any analysis was performed. The function takes a single vector, x, as an argument. When the returned function is evaluated it returns TRUE if the proportion of values in x that are larger than A is at least p [17]. Specifically, genes which had an unlogged normalized intensity of greater than 100 ($A = 100$) in at least 20% of samples ($p = 0.2$) and genes which has a coefficient variance between a minimum of 0.7 and maximum of 10 (genes

with high variation are cut off to limit our focus on genes that do not have high variation compared to their mean). Dataset subjects were partitioned into 70%:30% training and testing subsets for the purpose of model evaluations.

2.2.1 Principal Component Analysis

Filtered log_2 gene expression data was used to fit a linear model with weighted least squares with empirical Bayes moderation of the standard error. This approach is well suited to identify differentially expressed genes which are not normally distributed when the expression values differ between genes. We apply the Benjamin-Hochberg correction for multiple comparison testing. Genes with log_2 fold change $> |1.1|$ and adjusted $p \leq 0.01$ with a Benjamin-Hochberg correction for multiple testing comparisons were identified as differentially expressed (DE) genes. PCA was preformed to identify subjects with similar gene profiles. There were no clearly evident clusters among the No Relapse and Relapse groups, Lung Cancer and No Lung Cancer groups, and Parkinson's and Control groups.

2.3 Significant Analysis of Mircroarrays

Significant Analysis of Microarrays (SAM) [18] identifies statistically significant genes by gene specific *t*-tests and computes a new scoring statistic called the "relative difference", d_i for each gene i, which measures the strength of the relationships between gene expression and the response variable (disease state) [18]. Typically, a *t*-test for disease state response would apply a "pairwise fold change" method, where the measurement units of the classes are different i.e. disease and control state groups. This fold change method attempts to account for uncertainty in the data by identifying genes as significantly changed if an *R*-fold change is observed consistently between paired samples [18]. This is advantageous over a normal *t*-test as it is more robustly dynamic and can be applied to samples that are not independent and or are not normally distributed.

Although robust, fold change methods to no account for very low levels of expression, and for higher levels of expression, smaller changes in gene expression may be real, but these

changes are rejected by fold-change methods [18]. The pairwise fold change method provides modest improvement but remains inferior to SAM [18]. SAM does not depend on normality or homoscedasticity; this ensures a low number of "false discoveries" by use of a modified t-function with at least 1.5-fold. To account for correlation of genes, permutations were used to calculate the q-value. The q-value is the lowest false discovery rate (FDR) at which the gene is called significant; it is similar to the p-value but adapted to the analysis of a large number of genes [19]. The q-value measures how significant the gene is: as $d_i > 0$ increases, the corresponding q-value decreases [19]. We define differentially expressed genes to be the genes with at least 2-fold change within a 10% FDR. The equation for the test statistic calculating relative difference in gene expression is given as

$$d_i = \frac{r_i}{(s_i + s_0)};\ i = 1, \ldots,\ p, \tag{2.1}$$

r_i is score of the average level of expression for gene i between disease and control states, "gene-specific scatter" s_i is the standard deviation of r_i, and s_0 is an "exchangeability factor" [18, 19].

r_i is the linear regression coefficient of gene i on the outcome calculated as

$$r_i = \bar{x_{i2}} - \bar{x_{i1}}, \tag{2.2}$$

and s_i is defined as

$$[(1/n_1 + 1/n_2) \sum_{j \in C_1} (x_{ij} - \bar{x_{i1}})^2 + \sum_{j \in C_2} (x_{ij} - \bar{x_{i2}})^2 / (n_1 + n_2 - 2)]^{1/2} \tag{2.3}$$

where the data is x_{ij} for genes $i = 1, \ldots, p$ and samples $j = 1, \ldots, n$, and the response data is two class, unpaired data: disease and control state [19].

s_0 can be found automatically or can be set manually to lower the FDR [19]. To compare values of d_i across all genes, the distribution of d_i should be independent of the level of gene expression [18]. At low expression levels, variance in d_i can be high because of small values of

s_i [18]. s_0 must be a positive constant added to the denominator of Equation 2.1 to minimize the coefficient of variation [18], $cv(\alpha)$. Let s_α be the α percentile of the s_i values, then $d_i^\alpha = r_i/(s_i + s_\alpha)$ [19].

To find significant changes in gene expression, genes were ranked by magnitude of their d_i values, so that d_1 was the largest relative difference, d_2 was the second largest relative difference, and d_i was the i^{th} largest relative difference [18].

2.4 Random Forest

Random Forest (RF) is a popular supervised machine learning modeling technique often used for classification data. It is a collection of tree structured classifiers containing independent identically distributed random vectors and each tree casts a vote for the most popular class [20].

In this analysis, classes are disease states which we will be referring to as the "target." An RF model breaks down a dataset into homogenous subset while incrementally developing a decision tree. Each decision tree in a RF model is built using a bootstrapped sample of observations and a best split is chosen from a random subset of predictors rather than all of them to reduce correlation between trees. It is well-suited for DNA microarray data as it is efficient in handling high-dimensional data with good predictive performance. RF modeling is capable of identifying the most substantial set of genes which depict significant variation between the disease state subjects. RF is also capable for handling data with a lot of "noise" (in the case of microarrays [21], noise would be considered non-specific binding of probes or NSB as mentioned in Section 2.2) and outliers [20]. This method utilizes the concept of bootstrap aggregation, commonly known as bagging, in which the results of multiple trees is aggregated, and the random vector is generated as counts towards a target [22].

RF models have a lower likelihood of overfitting as more trees are created in large datasets, instead the generalized error is limited to a certain value due to nature of bootstrapping. The bootstrapped of a specified size is drawn with replacement from the original dataset [20]. Within

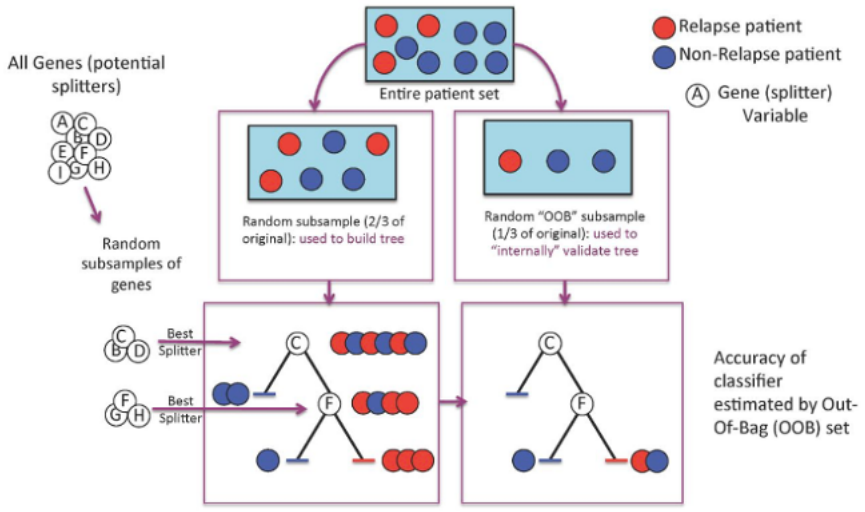


Figure 2.1: Visual representation of RF 1 decision tree algorithm for dataset GSE 2034 [image source: [24]].

the training set, 1/3 of patients are left out of the model or bag, to calculate an "out-of-bag" (OOB) error rate for the validity of the model. The OOB generalized error rate is used as a method of measuring prediction error rate of decision trees for a sub-sample set used for training. For instance in GSE 2034, OOB subjects are assigned a predictor classification based on where they end up after going through the "forest" and the OOB measures accuracy of their classification into 1: No Relapse or 2: Relapse. The goal of this is to find the decision tree within the RF model with the lowest OOB score, or most accurate prediction [23]. RF will hereafter be referred to as RF 1. A sample algorithm is shown in Appendix B.

2.4.1 Gene eXpression Network Analysis

Here, the terms "networks," "pathways," and "clusters" is used interchangeably. Substantial gene identification by one-at-a-time models like random forest is not the most appropriate technique as many genes are co-expressed and function cooperatively. Typically, gene expression analysis experiments will compare two or more phenotypes or disease states with many subject replicates [25]. Each subject replicate measures expression data for a large number of genes. Standard

analyses first start with filtering and normalizing gene data with computations following for each gene to compare the expression levels between the different phenotypes or disease states [25].

First, all four datasets were processed using the R packages *bioconductor* [26] and *limma* [27]. An M-value is computed using vsn-transformation [28] and normalization, and a t-statistic is computed using these M-values before being used as inputs for GXNA. To reduce number of false positives, multiple testing is performed to filter out genes that do not show enough variability against a 0.5 standard deviation for M-values as a threshold [25].

GXNA [25] computes a score that measures to what extent a gene or set of genes is differentially expressed. Consider a set $S = \{g_1, \ldots, g_k\}$ of genes $i = 1, \ldots, k$. We can average the t-statistic of its individual genes to derive a "new" scoring function known as ΣT [25],

$$f_1(S) = \frac{1}{k} \sum_{i=1}^{k} T_{g_i}. \tag{2.4}$$

Often genes can be up-regulated and down-regulated, so the absolute values of the t-statistic is taken [25]. The distribution is dependent on the size of set S. A nonparametric normalization method using the permutations of the phenotypes is used to estimate a null distribution of the score [25]. The disadvantage of this method is it is not reliable if the number of phenotypes is small [25]. As an alternative, a parametric assumption is made by normalizing all sets of size k by sampling from random sets of k genes resulting in a function

$$f_2(S) = \frac{1}{\sqrt{k}} \left(\sum_{i=1}^{k} |T_{g_i}| - k\mu \right), \tag{2.5}$$

where μ is the mean of $|T|$ over all genes [29]. The assumptions here are that the normalizations only need to depend on the size of S and the individual gene scores are independent (which is unrealistic) [25]. Instead, we normalize by sampling among connected sets of k genes, given as

$$f_3(S) = \frac{1}{\sqrt{\sigma_k}} \left(\sum_{i=1}^{k} |T_{g_i}| - \mu_k \right), \tag{2.6}$$

where μ_k and σ_k are the mean and standard deviation score for randomly connect sets of k genes, respectively, and σ_k is not needed to be proportional to $\sqrt{k}$ [25]. Scores based on single gene t-statistics ignore the coexpression of genes, so the sum of expression levels for the genes in S within a microarray is computed before the t-statistic is found to derive a scoring function known as $T\Sigma$.

Let X_{ij} be the expression level (normalized M-value) for gene i on array j [25]. The group expression is given as

$$S_j = \sum_{i=1}^{k} X_{g_i j}. \tag{2.7}$$

The score of the group S as the t-statistic of those values is calculated as

$$f_4(S) = (\mu_{i1} - \mu_{i0}) / \sqrt{\sigma_{i1}^2/n_1 + \sigma_{i0}^2/n_0}, \tag{2.8}$$

where the mean and standard deviation μ_{i1}, σ_{i1} are for the set S_j where j is a disease state, and μ_{i0}, σ_{i0} are for S_j where j is a control [25].

To allow for both up-regulated and down-regulated genes in the same pathway the signs are included as

$$S_j = \sum_{i=1}^{k} \varepsilon_i X_{g_i j}, \tag{2.9}$$

where ε_i is -1 if gene i is under expressed in the disease state and +1 otherwise [25]. This yields a more sensitive scoring function, but may produce more false positives [25]. This method takes into account probe correlations across arrays and is less likely to need normalization [25].

2.4.2 Significance levels

Sets of interacting genes or pathways are searched for based on target classification using permutation-based significance levels in addition to adjustment with multiple testing when sampling a large sample set to discourage the discovery of high scoring clusters that have no biological significance [25]. The standard measures used in multiple testing problems are family-wise error rate (FWER) and false discovery rate (FDR). In this instance, FWER is used because it has better protective power over false positives despite being more difficult in use to find significant genes [25]. The nonparametric techniques described in Section 2.4.1 were applied for standard microarray data in the two-phenotype case [25]. In the two-phenotype using dataset GSE 2034 as an example (eg. relapse state and no-relapse state), the indices are permutated so that some relapse states are relabeled as no-relapse and vice versa. The analysis is repeated enough for each permutation and if enough are available, this gives a reliable estimation of the null distribution of the networks scores and allow for the computation of adjusted p-values that control the FWER [25]. According to Nacu et al., the use of $T\Sigma$ algorithms will identify groups of genes with p-values very close to the conventional threshold of 5%, reflecting underlying biological significance rather than random chance [25].

2.4.3 RF 2 Method

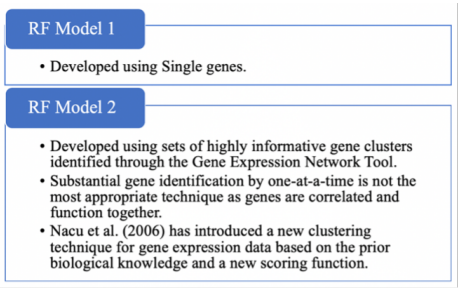


Figure 2.2: Overview of RF 1 and RF 2.

In a standard model such as the random forest model or BNN, genes are normally sorted in increasing order of significance based on measures tested in their respective models and then used to form biological hypotheses or used for experimental validation [25]. This strategy is limited, as single gene analysis does not account for co-expression of genes. For this reason we will revisit the random forest model but now with the intention of using the found gene pathways from GXNA as inputs for the model (Figure 2.2). This approach will hereafter be referred to as RF 2.

2.5 RF++

RF++ is a novel Generalized Random Forest-based classifier for cluster-correlated data like genomic data. It can classify cluster-correlated non-independent data in a statistically valid fashion. RF++ also identifies important variables in datasets with large numbers of variables and few subjects using permutation based proportioning. Subject-level bootstrapping, an alternative sampling method that obviates the need to average or otherwise reduce each set of replicates to a single independent sample [30], in RF++ assures that approximately 63% of all replicates will end up in the in-bag training set for a particular tree, thus the tree will be unaware of the rest of the independent samples in the dataset (OOB samples) and the error estimate will be unbiased.

2.6 LASSO

LASSO Penalized Regression (LPR) is a multivariate logistic regression model that can be used to generate a set of genes for the best performing model. In LASSO regression, the coefficients of some less contributive variables are forced to be zero (ie. they are "shrunken") by introducing a penalty constraint L1-norm to the regression for having too many variables in the initial model and only the most significant variables are kept in the final model [31]. In the instance of breast cancer relapse data, only the most significant genes for predicting the targets, Relapse and No Relapse, are found and kept in the model.

The constant λ, known as the tuning variable, will be defined to adjust the amount of the coefficient shrinkage. The best λ for the data will be defined as the λ that minimize the cross-

validation prediction error rate [31], but can be selected according to cross-validation or to achieve a specific number of non-zero coefficients [32]. LASSO was fit using the *glmnet* R package with λ set to the value of λ corresponding to the minimum mean cross-validation error.

Gene set selection by LASSO Penalized Regression (SLPR) has been previously introduced by Frost [32], where SLPR is performed by mapping of multiset gene set testing to penalized multiple linear regression. Gene set testing, or pathway analysis, is an important bioinformatics technique that lets researchers step back from the level of individual genomic variables and explore associations for biologically meaningful groups of genes [32]. By focusing the analysis on a smaller number of functional gene sets, this approach can substantially improve statistical power, biological interpretation and replication relative to an analysis focused on individual genomic variables [32].

LPR produces simpler and more interpretable models that incorporate only a reduced set of the inputs [33]. With this in mind, LASSO will be analyzed for its efficiency as a dimension reduction technique.

2.7 Artificial Neural Networks

In the last decade, interest in artificial intelligence and machine learning, has boomed in the computer science and psychological science fields, and one such method to show promising results is neural networks. Neural networks closely imitate the millions of processing nodes of the human brain [34] and had evolved ever since to measure just about anything that has a defined probability space, and it is most useful for describing relationships in classification modelling. Throughout this chapter we will be referencing to a Bayesian framework first introduced by David. J. C. MacKay in 1992 [35] for quantitative and practical applications of neural networks [36]. We begin by designing a model to classify the output of a variable, such as the one used in our analysis which will be predicting the state of a disease based on the genomic data of a subject. A neural network model first contains a set of adjustable parameters whose values are determined with the help of data, or input variables [36]. These parameters can be written simply as

$$\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x^i}; \boldsymbol{w}) \tag{2.10}$$

for the i^{th} patient where the output of a model from a set of input variables $\boldsymbol{x}^i = \boldsymbol{x_1}, \ldots, \boldsymbol{x_d}$ are set to an output variable $\boldsymbol{y}$ representing the class state. In the example of cancer relapse, the outcome of the classification in terms of variable y may take the values of 1 if the subject relapses from the disease and 0 if they do not. The parameters comprising $\boldsymbol{w}$ are called the weights of the neural network. The advantage of neural networks over simpler classification models like ones previously discussed is that they offer non-linear modeling from several input variables to the output variable, even for multiple output variables. In the case of neural networks, the process the model takes of determining the values of these parameters is called "learning" and the process where the output of the class state known is called "supervised learning [37, 38]."

2.7.1 Multi-Layer Perceptron and Activation

Also known as an MLP, a multi-layer perceptron is a popular feed-forward model of an artificial neural networks in which the nonlinear function of several input variable vectors $\boldsymbol{x} = \boldsymbol{x^1}, \boldsymbol{x^2}, , \boldsymbol{x^N}$ and their respective target variable vectors $\boldsymbol{D} = \boldsymbol{t^1}, \boldsymbol{t^2}, , \boldsymbol{t^N}$, where N is the number of variables [39]. An MLP is formed by a series of "nodes" much like neurons in the brain that are organized in layers. The structure of a neural network is formed by an "input" layer, one or more "hidden" layers, and the "output" layer (refer to Figure 2.3). The number of nodes in a layer and the number of layers depends on how the network is implemented. In an MLP, each node in a layer relates to each node in the next layer through a weight, $\boldsymbol{w}$. The value of the weight $\boldsymbol{w}$ indicates the strength of the connection between the i^{th} node in a layer and the j^{th} node in the next one. As an input enters a node, it gets multiplied by a weight value and the resulting output is either observed or propagated (passed) to the next layer in the neural network without the formation of loops among layers between a node, hence the name "feed forward." By training an MLP, minimizing the difference between the actual and network class predictions by adjusting the weights (including

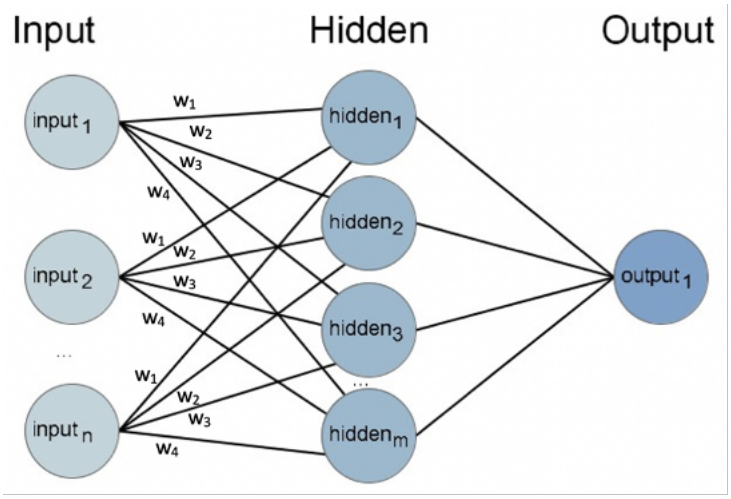


Figure 2.3: A multi-layer perceptron neural network.

biases) using some optimization algorithms is the most important goal. A well trained MLP is capable of making reasonable predictions on unseen data, which is known as generalization.

The role of biases in a neural network allows us to shift the activation function to the left or right of network, which is critical for successful learning during the training phase. In the case of a one-to-one input $\boldsymbol{x}$ to output network with $\boldsymbol{w_0}$ that has no bias, the output of the network is computed by multiplying the input by the weight $\boldsymbol{xw_0}$ and passing the result through an logistic sigmoid activation function given as

$$\boldsymbol{y}(\boldsymbol{x^i};\boldsymbol{w}) = \frac{1}{(1+e(-a))}. \tag{2.11}$$

Weight and biases are randomized before the training process and then adjusted to the desired values for an optimal output. Our models in this study will be making use of the Bayesian evidence procedure [39].

2.7.2 Bayesian Inference and Decision

A Bayesian neural network (BNN) is an artificial neural network trained under the influence of the Bayesian inference architecture which helps to overcome the issues associated with regular

MLP [39]. These Bayesian networks are a type of probabilistic model made of nodes and specified layers which can capture uncertainty in the weight distribution. BNN models can be prepared from data, then used for inference to estimate the probabilities for events.

Suppose in the event of breast cancer relapse a person who developed T1 cancer with one to three positive lymph-nodes has a 20% cumulative risk of distant relapse in 20 years since treatment, cumulative risk being the "total risk that something will happen over time." To classify a new patient with consideration of minimizing the probability of misclassification [37], we collect a large number of patients and classify them first into two classes, relapse (C_r) and no-relapse (C_{nr}). In Bayesian inference, the $\boldsymbol{y}(\boldsymbol{x^i};\boldsymbol{w})$ can be interpreted as the probability of belonging to class C_k given the input vector x. Prior probabilities $P(C_k)$ of a patient belong to each of the classes is introduced as C_k, where k can be either r or nr. We determine that the prior probability that a subject who meets the criteria of one to three positive lymph-nodes thus has $P(C_r)$=0.20 and $P(C_{nr})$=0.80.

If we were introduced to a new subject meeting the same criteria, then we could infer that the most probable class to assign her would be the no-relapse class given that $P(C_{nr}) > P(C_r)$ and therefore lowering the probability of misclassification despite the possibility that they are a part of the relapse class [37].

As given in Equation 2.12, Bayes' Theorem allows the posterior probability to be expressed in terms of the prior probability $P(C_k)$, along with the class-conditional probability $P(x|C_k)$ of x for class C_k. The denominator of the fraction in Bayes' Theorem is the normalization factor and ensures the posterior probability to make optimal decisions regarding the classification of new data. Assigning a new patient to the class having the largest posterior probability minimizes the probability of misclassification of that patient [37],

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}. \tag{2.12}$$

The normalization factor can be expressed in terms of the prior probabilities and class-

conditional probabilities [37]. Any new patient must be assigned to one of the two classes C_k, where k is r or nr,

$$P(C_r|x) + P(C_{nr}|x) = 1. \tag{2.13}$$

Substituting Equation 2.12 of Bayes' Theorem into 2.13, we get

$$P(x) = P(x|C_r)P(C_r) + P(x|C_{nr})P(C_{nr}). \tag{2.14}$$

2.7.3 Bayes' Theorem, prior and posterior probability distributions

Given input variable data $\boldsymbol{x}$ and parameter $\boldsymbol{w}$, a simple Bayesian analysis starts with a prior probability distribution (the prior) $p(\boldsymbol{w})$ which expresses known parameters of the dataset before it is analyzed and a likelihood (measure of belief of an event occurring) function for the data $p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x})$ to compute a posterior probability $p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x})p(\boldsymbol{w})$. By Bayes' Theorem to invert conditional probabilities, the posterior density of data $\boldsymbol{x}$ and parameter $\boldsymbol{w}$ is

$$p(\boldsymbol{w}|\boldsymbol{D},\boldsymbol{x}) = \frac{p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x})p(\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{D}|\boldsymbol{x})}, \tag{2.15}$$

where $p(\boldsymbol{D}|\boldsymbol{x})$ (the evidence) [39] is a normalization in the parameter space

$$p(\boldsymbol{D}|\boldsymbol{x}) = \int p(\boldsymbol{D}|\boldsymbol{w'},\boldsymbol{x})p(\boldsymbol{w'}|\boldsymbol{x})d\boldsymbol{w'}. \tag{2.16}$$

Computing the posterior distribution is known as the "inference problem." Once the posterior is found, the inference is made by integrating over the above distribution. To make a prediction of a new input $\boldsymbol{x}^*$, a marginal prediction distribution is calculated as

$$p(\boldsymbol{y}|\boldsymbol{x}^*,\boldsymbol{D}) = p(\boldsymbol{y}|\boldsymbol{x}^*,\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{D},\boldsymbol{x})d\boldsymbol{w}, \tag{2.17}$$

but unfortunately not all of these are analytically tractable [39] due to high dimensionality [12]. As such, approximations of posteriors (evidence procedures) are a large reason why Bayesian inference in neural networks is used. Network training (minimizing the difference between the actual and network predictions) can be done in two ways, using conventional maximum likelihood and Bayesian approaches. With a maximum likelihood approach, a single set of most likely values for the weights are generated whereas with Bayesian, a probability distribution for weights is obtained to represent the uncertainty in the weight estimation [39]. Bayesian inference is important because it takes into account parameter uncertainty (incomplete knowledge of inputs in the model) so that "overfitting" is not a problem unlike in a regular MLP [37, 40]; Overfitting is an inherited problem in the maximum likelihood estimation approach which leads to poor generalization. It regularizes parameters during the training process of the neural network, and parameter uncertainty is accounted in network predictions [39].

In the event of classification modeling, there are two stages of classifying data. The first stage is inference where input variable data is used to determine the values of posterior distribution probabilities, the second stage is decision making in which the posterior distribution probabilities are used to make decisions such as classifying an output in a class state [37].

2.7.4 Evidence Procedure

Once a prior has been constructed for a neural network and placed in the data structure, the maximum likelihood can be determined by the model for a two-class disease state problem as

$$p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x}) = \prod_{n=1}^{N} \boldsymbol{y}(\boldsymbol{x}^n;\boldsymbol{w})^{\boldsymbol{t}^n}(1-\boldsymbol{y}(\boldsymbol{x}^n;\boldsymbol{w}))^{1-\boldsymbol{t}^n} \tag{2.18}$$

and the error function can be determined as

$$E_D(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x}) = -\sum_{n=1}^{N} \boldsymbol{t}^n ln\boldsymbol{y}(\boldsymbol{w}^n;\boldsymbol{w}) + (1+\boldsymbol{t}^n) ln(1-\boldsymbol{y}(\boldsymbol{x}^n;\boldsymbol{w})). \tag{2.19}$$

Here, $p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x})$ represents the maximum likelihood function of the Bernoulli variable. We take the negative log likelihood function to attain the corresponding error function, $E_D(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x})$. The final output of these functions gives us the logistic sigmoid activation function

$$f(a) = \frac{1}{(1+e(-a))} \tag{2.20}$$

the same as in Equation 2.11, thus implying $f = y(\boldsymbol{x^i};\boldsymbol{w})$.

If the assumption that weight posterior is around the most probable weight, $\boldsymbol{w}_{MP}$, occurs, the output function $y(a)$ is not linear [39]. Then the prediction is no longer the most probable output $\boldsymbol{y}(\boldsymbol{x},\boldsymbol{w}_{MP})$. Under this assumption, the result of a can be found in the logistic sigmoid function in the place of y. a has a Gaussian distribution with mean a_{MP} given through forward propagation of $\boldsymbol{x}$ through the network with weights $\boldsymbol{w}_{MP}$ and a variance of

$$s^2(\boldsymbol{x}) = g^T A^{-1} g, \tag{2.21}$$

where g is the gradient, or the numeric calculation of the parameters in the network to reduce misclassification, of a with respect to the weights $\boldsymbol{w}$ at their most probable $\boldsymbol{w}_{MP}$ [39]. The MacKay approximation of the predictions on the input vector can be rewritten as

$$P(C_{nr}|\boldsymbol{x},\boldsymbol{D}) \approx f(k(s)a_{MP}) \tag{2.22}$$

where

$$k(s) = \left(1 + \frac{\pi s^2}{8}\right)^{-\frac{1}{2}}. \tag{2.23}$$

Decision boundaries to minimize the probability of misclassification is usually set at $P(C_{nr}) = 0.5$. Considering the MacKay approximation equation, the network predicts 0.5 if and only if $a_{MP} = 0$ [39]. For this decision boundary, the predictions made by the most probable output

$\boldsymbol{y}(\boldsymbol{x};\boldsymbol{w}_{MP})$ and the marginalized prediction from the effect of marginalizing the network output are the same. Equation 2.23 shows because $0 < k < 1$, we can assume

$$|f(k(s)a_{MP})| \leq |f(a_{MP})| = |\boldsymbol{y}(\boldsymbol{x};\boldsymbol{w}_{MP})|, \tag{2.24}$$

although the inequality means $a_{MP} \neq 0$ [39].

The evidence procedure used for a multi-layer perceptron model is an algorithm for finding the optimal values of the hyperparameters and weight of each input variable. It is based on the Bayesian regularization technique in attempts to reduce network "overfitting," or the occurrence of a model being unable to make predictions on data it was not trained on, such as its ability to classify targets in a training dataset but not the testing dataset. Predictions can be made by taking the integral over the posterior weight distribution [39].

2.7.5 Prior and Posterior Distribution of the Weights

First, we choose the prior probability distribution for the weights in a neural network. The neural network favors small values for the weight parameter thus introducing the Gaussian prior distribution with a mean of

$$P(\boldsymbol{w}) = \frac{1}{Z_{\boldsymbol{w}}(\alpha)} e\left(-\frac{\alpha}{2}||\boldsymbol{w}||^2\right), \tag{2.25}$$

where α is the inverse variance of the distribution and

$$Z_{\boldsymbol{w}}(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{\boldsymbol{w}}{2}} \tag{2.26}$$

is the normalization constant that does not depend of the weight [10]. α is thus known as a hyperparameter of the network because it is a parameter of the prior distribution, in implementation it contains a vector of hyperparameters of weights and biases for each singular input variable [39].

The error term $E_p(\boldsymbol{w})$ is given in the form of

$$E_p(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}, \tag{2.27}$$

and usually penalizes large weights to generate a better generalization of the data.

By taking the negative log likelihood of the posterior distribution for weights and adding the error term according to the Bayes' Theorem we get the following equation

$$P(\boldsymbol{w}|\boldsymbol{D},\boldsymbol{x}) = \frac{1}{Z_s}e\left[-\left(ln\ p(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{x}) + \alpha E_p\boldsymbol{w}\right)\right] = \frac{1}{Z_s}e\left[-S(\boldsymbol{w})\right] \tag{2.28}$$

where Z_s is the normalization constant for the posterior [39]. This is known as $S(\boldsymbol{w})$, the regularized cost function. This follows with the posterior distribution of the weights $\boldsymbol{w}$ is given in Equation 2.28 over the total cost function $S(\boldsymbol{w})$ and A is the Hessian matrix of $S(\boldsymbol{w})$ approximated at $\boldsymbol{w}_{MP}$, or the most probable weight vector found by optimization of $S(\boldsymbol{w})$ following the second order Taylor series expansion of $S(\boldsymbol{w})$ in

$$S(\boldsymbol{w}) \approx S(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{MP})^T A(\boldsymbol{w}-\boldsymbol{w}_{MP}). \tag{2.29}$$

where $Z_S{}^*$ is the normalization constant.

Assuming this, the output function is thus "locally" linear as suggested by MacKay [39], that is it is sufficiently small for a tangent line to closely approximate the function over an interval of weights with a distribution of

$$p(a|\boldsymbol{x}^*,\boldsymbol{D}) = \int p(a|\boldsymbol{x},\boldsymbol{w})\,p(\boldsymbol{w}|\boldsymbol{D})\ d\boldsymbol{w}, \tag{2.30}$$

we can obtain the network predictions on new inputs of $\boldsymbol{x}^*$ using the above posterior distribution.

2.7.6 Gaussian Priors

A Gaussian weight prior will capture the significance of small weights by showing the relationship between hyperparameters and the different layers of a multi-layer perceptron [39]. In implementation, we can set the parameters as:

1. aw1 is the hyperparameter for the input layer weights containing the input values for each input variable
2. ab1 is the hyperparameter for the input layer biases
3. aw2 is the hyperparameter for the second-layer weights
4. ab2 is the hyperparameter for the second-layer biases

and so on [39].

In a model with a consistent prior, the regularization parameter is inconsistent with linear relationships of the input and output patterns, so the optimal values for the output layer bias weights are the unconditional means of the corresponding output variables. For this reason, we normalize the target data to zero mean so that the prior does not have too much effect on the target [39]. The implementation of a zero mean Gaussian prior will distinguish between two types of hyperparameters: (1) a single hyperparameter α for all the weights in the network, and (2) separate hyperparameters for different groups of weights known as the ARD prior.

2.7.7 Automatic Relevance Determination

In Bayesian inference modeling, a separate hyperparameter for each input variable representing the inverse variance of the prior distribution of the weights associated with a certain input variable is considered as optimal when obtained using the evidence procedure [41]. Weights associated to "irrelevant" input variables are set to small values and thus known as an ARD prior. Automatic relevance determination (ARD) is used to decide the relative importance of input variable.

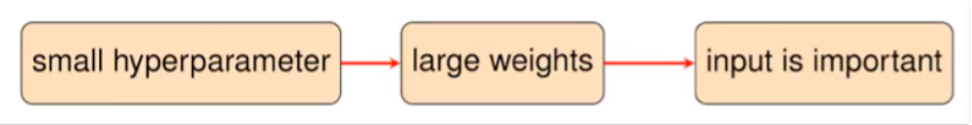


Figure 2.4: Automatic relevance determination prior.

We will add a separate regularization factor/hyperparameter for each gene/input variable. This hyperparameter will represent the inverse variance of the prior distribution of the weights coming from that input. It is normally distributed, thus assuming a Gaussian prior for each class we can define $E_{\boldsymbol{W}(c)} = \sum_{i \in c} \boldsymbol{w}_i^2/2$ so the ARD prior is calculated as

$$\frac{1}{\prod_c Z_{\boldsymbol{W}(c)}} e(-\sum_c \alpha_c E_{\boldsymbol{W}(c)}. \tag{2.31}$$

The evidence framework can be used to optimism all the regularization constants simultaneously by finding their most probable value, i.e., the maximum over α_c of the evidence [42]. During training of the neural network model we can modify the hyperparameter using the evidence procedure to find their optimal values [41]. Since hyperparameter represent the inverse variance of the weights, a small hyperparameter value means that large weights are allowed and we may conclude that the corresponding input variable is important (Figure 2.4) by so-called "fine tuning" of α.

A large hyperparameter value pushes the weight value to nearly zero, thus meaning the opposite of an important variable [41]. During model implementation, the function for finding α with ARD prior returns a data structure for the priors in the form of matrix where each of the first columns contains ones in the positions for the corresponding weight of an input variable. A sample algorithm is shown in Appendix B.

2.8 DAVID

The DAVID tool was released in 2003 as one of the pioneering works of high-throughput functional annotation bioinformatics system. Since then, a series of novel bioinformatics algorithms

have been continually developed and reported in peer-reviewed papers. DAVID is a web-based online bioinformatics resource that aims to provide tools for the functional interpretation of large lists of genes/proteins. Information regarding the ontology of genes identified in this study were obtained from DAVID.

CHAPTER III

GENE EXPRESSION ANALYSIS ON BREAST CANCER

Breast cancer accounts for 15% of total cancer deaths and is the second leading cause of cancer death in women [43]. It is estimated that 1 in 8 women will develop breast cancer in the U.S. in her lifetime [43], of these women, approximately 1 in 39 will die from breast cancer. Additionally, there are rare cases of breast cancer developing in men and children. In 2020 alone, an estimated 276,480 new cases of metastatic breast cancer were diagnosed in women and 2,620 new cases were diagnosed in men [43]. Despite a 90% 5-year relative survival rate for breast cancer, early detection in patients is crucial to reducing the deadly threat to their life significantly. In this chapter we present our findings for breast cancer pathology related to objectives (1)-(5). GSE 2034 and GSE 2990 contain data related to breast cancer pathologies.

Prior to running machine learning methods for GSE 2034 and GSE 2990, the datasets were split into a training subset (70%) and a testing subset (30%). The training set was used to "train" the model parameters and the testing set was used to evaluate model metrics. The specifications of the dataset partitions created after *pOverA* filtration are as follows:

Table 3.1: Summary of genes in breast cancer pathology datasets.

	No. Genes after *pOverA*	Subjects in training dataset	Subjects in testing dataset
GSE 2034	775	230	189
GSE 2990	1038	142	34

3.1 Results of SAM

SAM analysis was performed to identify the differentially expressed genes for breast cancer pathology in datasets GSE 2034 and GSE 2990. Tables 3.1 and 3.2 present the results of the SAM analysis for GSE 2034 and GSE 2990, respectively.

Table 3.2: GSE 2034 up-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
TRAF5	Tumor necrosis factor receptor binding	2.2679	2.2667	0.00

Table 3.3: GSE 2990 down-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
KCNA3	Voltage-gated potassium channel activity	-1.7836	0.4780	23.2315
GSTT1	Glutathione transferase/peroxidase activity	-1.7475	0.4629	23.2315
CBR1	Carbonyl reductase (NADPH) activity	-1.6621	0.4507	23.2315

GSE 2034 had 1 up-regulated gene identified as statistically significant (q-value<0.05), TRAF5 (TNF Receptor Associated Factor 5). This protein encoding gene is primarily responsible for tumor cell necrosis. Cell necrosis is the early death of cells in the body usually due to significantly low blood flow to these cells. It may produce an inflammatory response, which may participate in tumor regression during cancer therapy. TRAF5 is not prognostic in breast cancer, but given its up-regulated state, we can assume that it is increased in subjects who had relapsed with malignant breast cancer against a control of those who have not.

The malignant subjects within GSE 2990 had 3 down-regulated genes identified: KCNA3, GSTT1, and CBR1. All of these genes were insignificant based on their q-value but are still relevant to the pathology.

Voltage-gated potassium (Kv) channels are widely expressed in the plasma membranes of numerous cells such as epithelial cells, or cells that line the outer surfaces of organs and blood vessels throughout the body [44]. Recently, it has been demonstrated that Kv channels are associated with the proliferation of several types of cancer cells [44]. Specifically, Kv1.3 (the channel related to the KCNA3 (Potassium Voltage-Gated Channel Modifier Subfamily A Member 3) gene) seems to be involved in cancer cell proliferation and apoptosis. The expression level of Kv1.3 has been evaluated in each stage of breast cancer using mRNA isolated from breast cancer patients and has been found to be a potential biomarker for breast cancer [45].

Recently, GSTT1 (Glutathione S-Transferase Theta 1) has been confirmed as a potentially cancer susceptible gene [46]. It's role is in the detoxification of toxic, potentially carcinogenic compounds located within the body that may contribute to gene mutations that lead to various types of cancers.

Decreased CBR1 (Carbonyl Reductase 1) expression is associated with poor prognosis in ovarian cancer [47]. Often times, women who inherit the gene mutation for breast cancer are also at risk for ovarian cancer. Previous studies have shown that genetic inhibition of the CBR1 enzyme encoded by the CBR1 gene improved the anticancer effects of certain cancer therapies in breast cancer patients [48].

GSE 2034 and GSE 2990 shared no similar genes identified using SAM despite being filtered to have the same genes present in their dataset prior to analysis.

3.2 Results of RF 1

When creating the decision trees in the random forest, the following parameters have been fine-tuned: the number of trees and the number of nodes per tree. After training, Out-of-Bag error rate for GSE 2034 and GSE 2990 were found to be 32.61% and 44.37%, respectively. The top 25 genes that were identified within GSE 2034 and GSE 2990 according to the mean decrease in accuracy are presented in Figures 3.1 and 3.2. The higher the mean decrease associated with a gene,

the more important it is in correctly predicting the output.

Table 3.4: Results of RF 1 for breast cancer pathologies.

	Accuracy	Sensitivity	Specificity	AUC
GSE 2034	0.6739 ± 0.0012	0.7229 ± 0.0019	0.2326 ± 0.0111	0.4372 ± 0.0056
GSE 2990	0.5563 ± 0.0049	0.6333 ± 0.0090	0.4231 ± 0.0084	0.5962 ± 0.0035

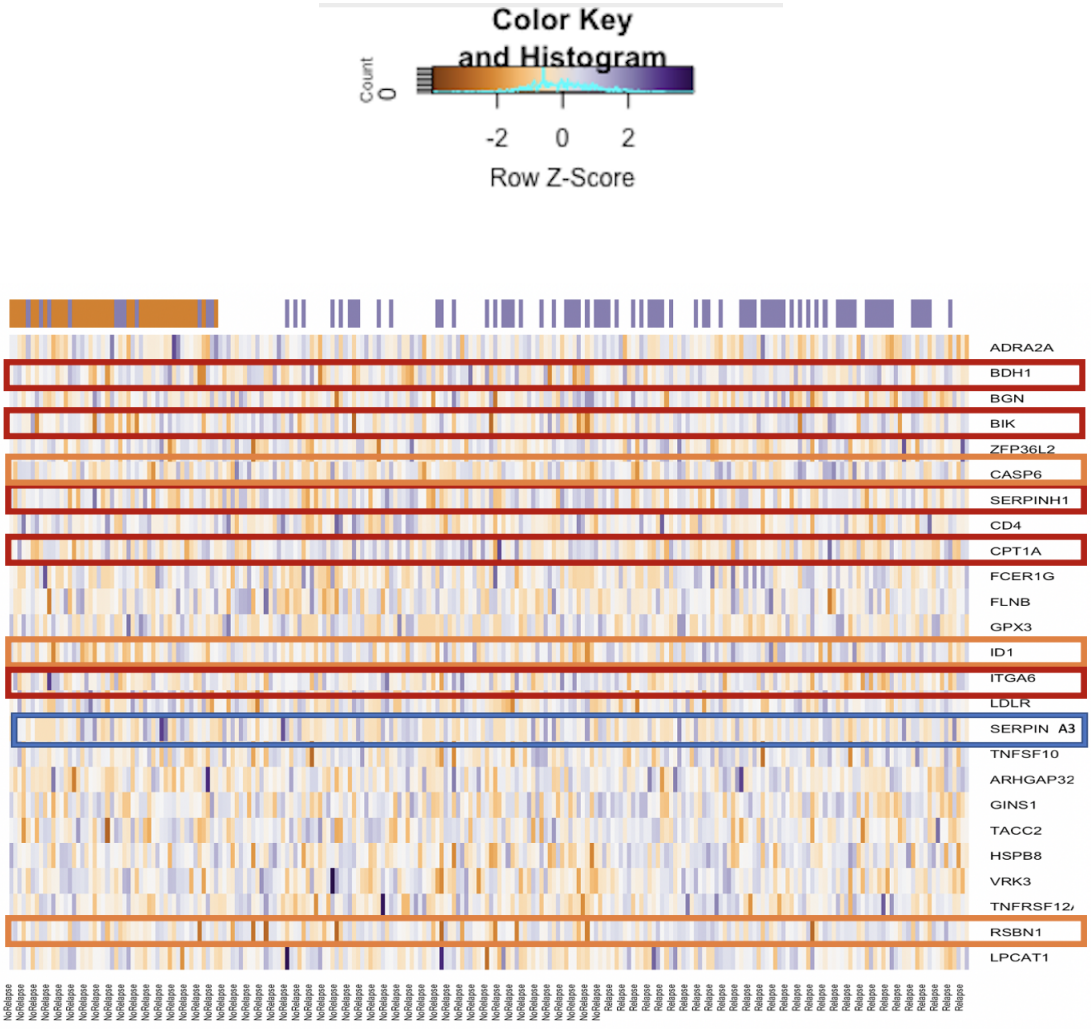


Figure 3.1: RF 1 method (single genes as inputs to RF) heatmap for GSE 2034.

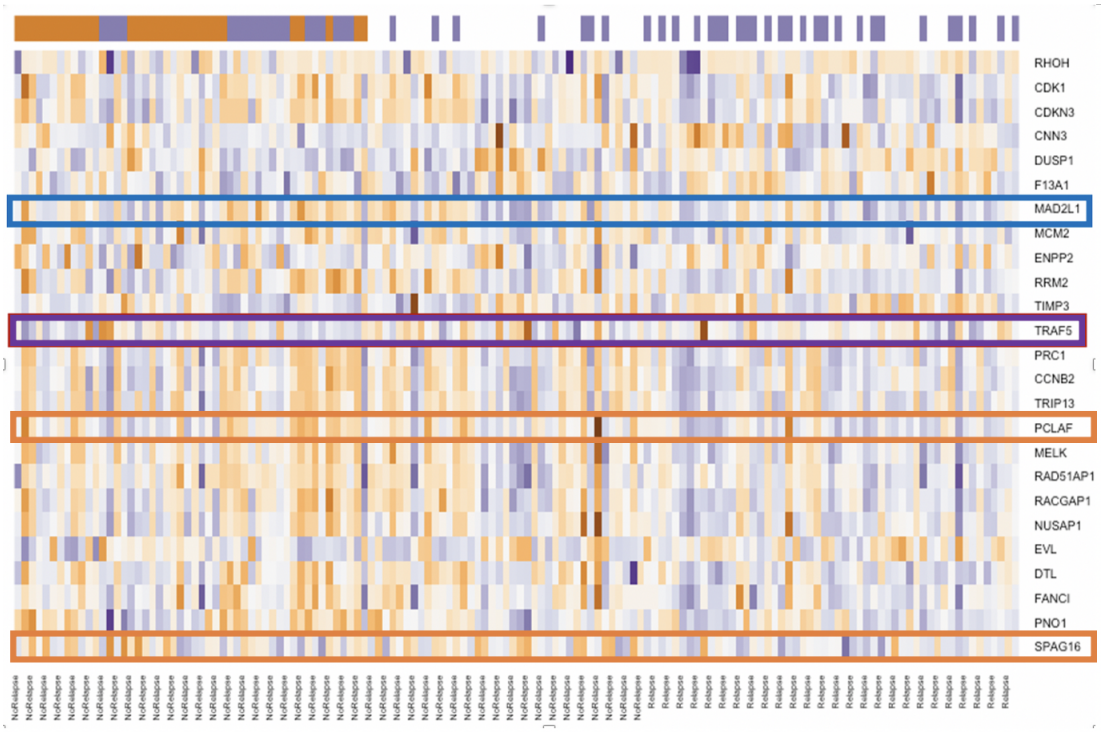


Figure 3.2: RF 1 method (single genes as inputs to RF) heatmap for GSE 2990.

Although the gene identified as up-regulated by the SAM analysis for GSE 2034 (TRAF5) did not present in these top 25 genes in the heatmap for RF 1, it was present amongst the top 25 genes identified in GSE 2990. This is noteworthy given that both datasets share common pathology since TRAF5 is related to tumor necrosis in multiple cancers types. Included in the maps are highlighted genes that will be identified using other methods in the continuation of the analysis on GSE 2034 and GSE 2990.

3.3 Results of RF 2 (Random Forest with GXNA Clusters)

The top 25 most influential clusters with at most 15 genes in each cluster; clusters with more than one gene were used as input to RF Model 2. Out-of-Bag error rate for GSE 2034 and GSE 2990 were 42.33% and 31.99% respectively. The top 8 genes in GSE 2034 and top 25 genes in GSE 2990 were identified according to the mean decrease in accuracy and given in Figures 3.3 and 3.4.

Table 3.5: Results of RF 2 for breast cancer pathologies.

	Accuracy	Sensitivity	Specificity	AUC
GSE 2034	0.5767 ± 0.0017	0.7000 ± 0.0033	0.2931 ± 0.0042	0.6954 ± 0.0004
GSE 2990	0.6801 ± 0.0050	0.8109 ± 0.0071	0.3345 ± 0.0182	0.7777 ± 0.0049

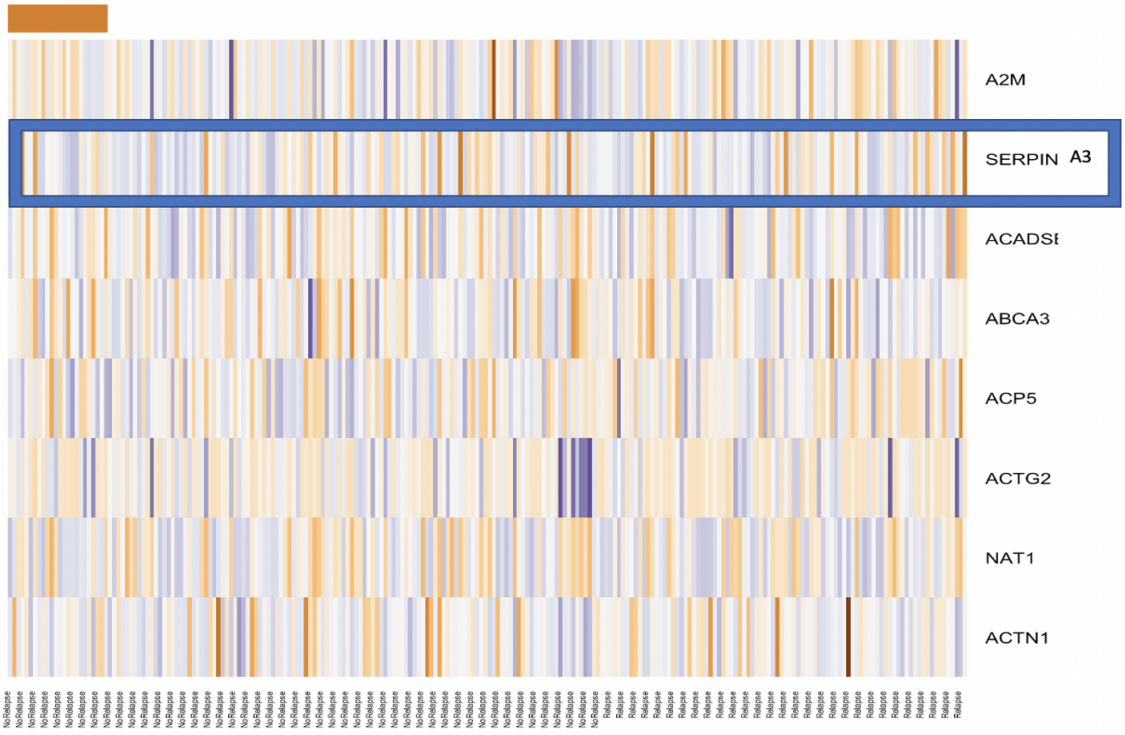


Figure 3.3: RF 2 method (clustered genes as inputs to RF) heatmap for GSE 2034.

SERPINA3 was present in both the RF 1 and RF 2 methods of GSE 2034, showing that it could be a relatively important gene for breast cancer.

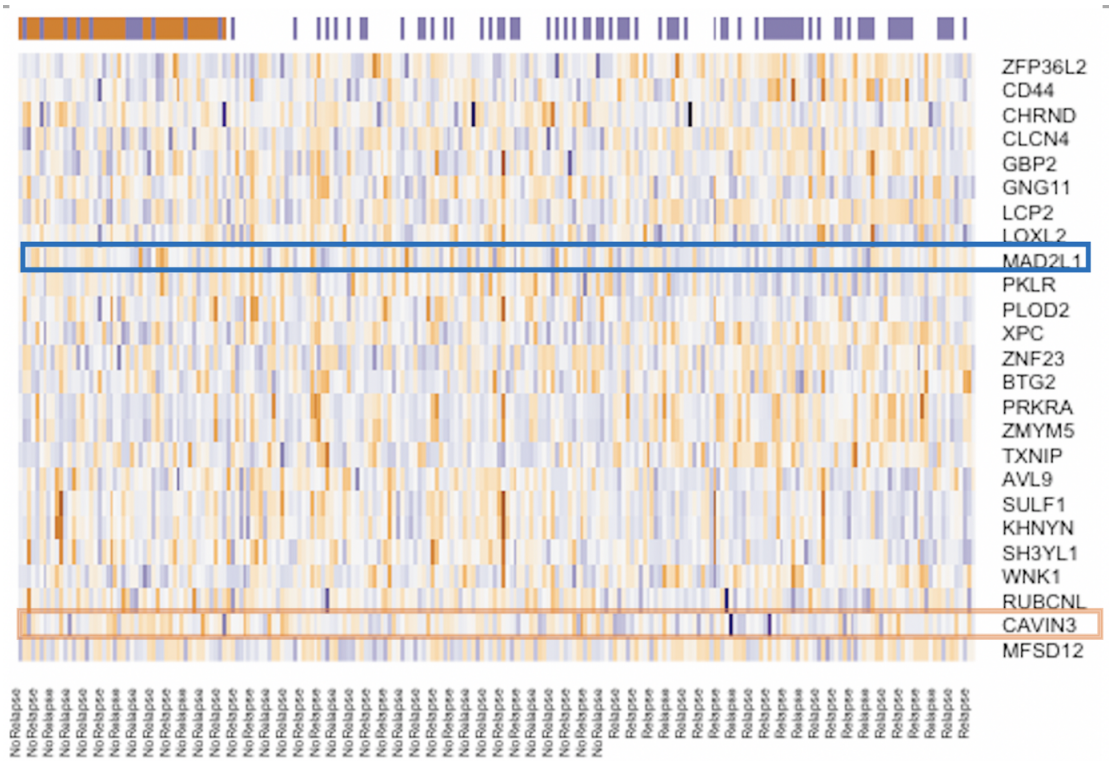


Figure 3.4: RF 2 method (clustered genes as inputs to RF) heatmap for GSE 2990.

MAD2L1 was present in both the RF 1 and RF 2 methods for GSE 2990, showing that it could be another relatively important gene for breast cancer.

Table 3.6: Change (Δ) in OOB for RF1/RF2 in breast cancer pathologies.

	Δ OOB error rate
GSE 2034	+9.72
GSE 2990	-12.38

3.4 Results of RF++

Our RF++ yielded sample-level error rate of 50.79% and 38.24% on predictions of the subject disease class in for GSE 2034 and GSE 2990 testing datasets, respectively. RF++ with subject-level bootstrapping (SLB) was performed following both RF methods to analyze its use as a

dimension reduction technique. The top 20 genes identified as important with permutation-based proportioning are presented in Tables 3.8 and 3.9.

Table 3.7: Results of RF++ for breast cancer pathologies.

	Accuracy	Sensitivity	Specificity	AUC
GSE 2034	0.6402 ± 0.0190	0.4000 ± 0.0096	0.7206 ± 0.0087	0.5709 ± 0.0263
GSE 2990	0.6176 ± 0.0322	0.9091 ± 0.0425	0.0833 ± 0.0081	0.5000 ± 0.0378

Table 3.8: Variable Importance in the GSE 2034 with RF++.

Gene Name	Score
ID1*	0.0009
IL6ST	0.0008
CPT1A*	0.0007
TYMS	0.0005
TANK	0.0005
ARL4C	0.0004
SPTLC2	0.0004
RSBN1*	0.0004
TMEM176A	0.0004
TPP2	0.0004
RUNX3	0.0004
CTSD	0.0004
NDUFA6	0.0003
SERPINA3*	0.0003
HSPB8*	0.0003
TACC2*	0.0003
PNMA8A	0.0003
EFNA3	0.0003
FAH	0.0003
SATB1	0.0003

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

Table 3.9: Variable Importance in the GSE 2990 with RF++.

Gene Name	Score
RRM2	0.0023
SPAG16*	0.0016
RACGAP1	0.0010
CDK1*	0.0009
NELL2	0.0009
RAD51AP1*	0.0009
TIMP3	0.0008
MMP7	0.0007
PCLAF*	0.0006
CNN3	0.0006
FANCI*	0.0005
DTL*	0.0005
TRAF5*	0.0005
ENPP2*	0.0004
NBN	0.0004
GINS1	0.0004
CBR1*	0.0004
CXCL13	0.0004
MELK*	0.0004
NAAA	0.0004

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

3.5 Results of LASSO

Our LASSO yielded accuracies of 63.49% and 70.59% for GSE 2034 and GSE 2990, respectively (Table 3.10). Figures A.1 and A.2 in Appendix A display the cross-validation error according to the log of λ for GSE 2034 and GSE 2990. The optimal λ was found by minimizing the cross-validation prediction error, which will give the most accurate model.

Table 3.10: Results of LASSO for breast cancer pathologies.

	Accuracy	Sensitivity	Specificity	AUC
GSE 2034	0.6349 ± 0.0020	0.9947 ± 0.0021	0.8125 ± 0.0058	0.5794 ± 0.0004
GSE 2990	0.7059 ± 0.0037	0.6250 ± 0.0261	0.7308 ± 0.1409	0.6023 ± 0.0115

Genes found to be to be significant as regression coefficients in the GSE 2034 and GSE 2990 LASSO are available in Tables 3.11 and 3.12, respectively:

Table 3.11: Regression coefficients in the GSE 2034 with LASSO.

Gene Name	Coeff. Value
AQP1	-0.0391
CASP6*	0.0117
ID1*	0.1869
RARA	0.03868
TRAF5*	0.0123
SPTLC2	0.0228
TVP23B	0.1137
RSBN1*	0.0331
CAVIN3*	-0.0391

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

Table 3.12: Regression coefficients in the GSE 2990 with LASSO.

Gene Name	Coeff. Value
KRT15	0.0035
NELL2*	0.0040
PCLAF*	0.1571
SPAG16*	0.0473

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

3.6 Results of Bayesian Neural Network

Using the Bayesian approach, we trained two types of networks with different weight regularization techniques. The first network was a standard network. The second type of the network was trained using Bayesian evidence procedure with ARD prior.

Each network was trained with 10-fold cross validation and 3 hidden nodes and ran with 5 random initializations to obtain an average on all validation measures.

The best network in terms of the highest accuracy and specificity was found to be the network trained using Bayesian evidence procedure along with the ARD prior for both breast cancer datasets. The results are shown in Tables 3.13 and 3.14. As can be seen, use of evidence procedure along with the ARD prior has led to better sensitivities and higher AUC values indicating higher discrimination power.

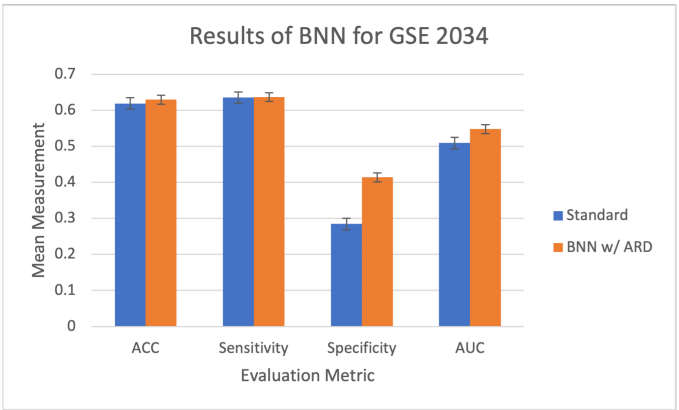


Table 3.13: Results of BNN for GSE 2034.

	Accuracy	Sensitivity	Specificity	AUC
Standard	0.61891 ± 0.016	0.6357 ± 0.0054	0.2846 ± 0.1894	0.5093 ± 0.0488
BNN w/ ARD	0.62958 ± 0.012	0.6365 ± 0.0028	0.4139 ± 0.0906	0.5478 ± 0.0196

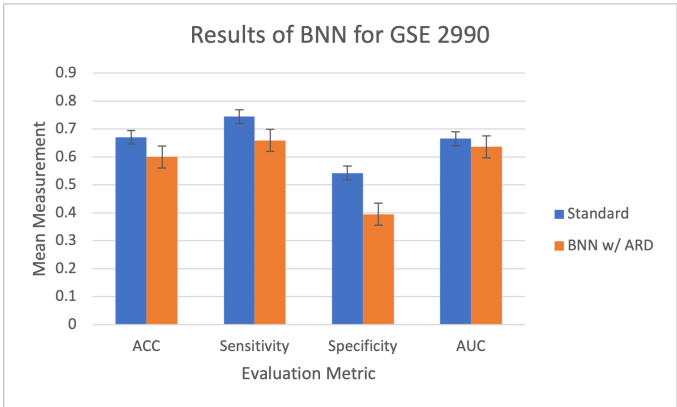


Table 3.14: Results of BNN for GSE 2990.

	Accuracy	Sensitivity	Specificity	AUC
Standard	0.6706 ± 0.0246	0.7442 ± 0.0248	0.5426 ± 0.0296	0.6659 ± 0.0118
BNN w/ ARD	0.5998 ± 0.0395	0.6591 ± 0.0248	0.3952 ± 0.0852	0.6364 ± 0.0347

For the purposes of our study, the following definitions of validation measures apply. Accuracy is the ratio of correctly predicted disease class of subjects, it is measured on a partitioned testing set. Sensitivity is the ability of our model to correctly identify patients with the disease. Specificity is the ability of our model to correctly identify subjects without the disease.

3.6.1 Relative Importance of Genes Based on ARD prior

Final mean α values for the ANN trained under Bayesian evidence with the ARD prior for GSE 2034 was 0.009, and 0.294 for GSE 2990. First 25 genes identified as relatively important by the ARD prior for GSE 2034 and GSE 2990 are given in Tables 3.15 and 3.16, respectively. Here, the smaller the variable importance value the more significant.

Table 3.15: Genes identified as relatively important by the ARD prior for GSE 2034.

Gene Name	Var. Imp.
BDH1*	972.4011
TVP23B*	1407.4657
AQP1*	1469.6753
PPIF	3351.2864
LTBP1	3767.2501
ITGA6*	3925.2669
RSBN1*	4030.3748
ID1*	4108.9872
TRIL	5034.1607
TASOR2	5646.7008
DKK3	7518.3705
SPTLC2*	8721.3112
EXOSC4	10205.2575
PLAU	10925.9211
SERPINH1*	10945.2185
RARA*	11075.3203
FHL2	11103.9455
SERPINA3*	11551.8215
CPT1A*	11665.9103
EFNB2	13318.8532
RBPMS	14089.5247
PYGB	15278.2917
MZF1	15852.7832
RBBP8	16348.1513
BIK*	16557.7346

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

Table 3.16: Genes identified as relatively important by the ARD prior for GSE 2990.

Gene Name	Var. Imp.
DHRS2	175.8987
RBP1	217.8083
GSTT1*	222.4862
PEG3	231.6678
PNMA8A	233.5499
ANPEP	239.6462
NELL2*	240.7688
POMZP3	252.9264
CALML5	267.3526
KCNK1	267.8039
MET	271.3244
SERPINA5	274.8154
MMP3	287.4289
PEG10	316.2320
MATN2	317.1113
ACOX2	318.4414
SLC24A3	319.7252
CA2	321.3489
KCNE4	325.0765
TMPRSS3	332.0550
CSTA	344.5998
STC1	344.6604
EEF1A2	346.3719
RTN1	349.0472
S100P	354.6985

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

3.7 Potentially Relevant Genes for Breast Cancer

Table 3.17: Genes identified in GSE 2034.

Gene Name	Models
TRAF5	SAM, RF 1, & LASSO
SERPINA3	RF 1, RF 2, RF++, & BNN
ID1	RF 1, RF++, LASSO & BNN
RSBN1	RF 1, RF++, LASSO & BNN

In GSE 2034, the genes found to be relatively important were: TRAF5, SERPINA3, ID1, and RSBN1. Relatively important genes were genes that were identified in three or more models or significantly by SAM and found in the following models.

TRAF5 (TNF Receptor Associated Factor 5); This protein encoding gene is primarily responsible for tumor cell necrosis. Cell necrosis is the early death of cells in the body usually due to significantly low blood flow to these cells. It may produce an inflammatory response, which may participate in tumor regression during cancer therapy TRAF5 was found as significant in GSE 2990 RF 1 method; this could be an indication that it is statistically significant as a differentially expressed gene for malignant breast cancer across both datasets. TRAF5 has not yet been identified to manifest in direct connection to malignant breast cancer. It has been identified with multiple myeloma, which is a cancer that develops in the bone marrow, the spongy tissue found in the center of most bones. If this cancer presents near the upper chest region it could metastasize into the breast.

SERPINA3 (Serpin Peptidase Inhibitor, Clade A (Alpha-1 Antiproteinase, Antitrypsin), Member 3) was present and down-regulated in relapsed patients in both the Random Forest and Random Forest with GXNA gene clusters model. SERPINA3 has been found significant in Malignant Fibrous Histiocytoma, a type of cancer that usually forms in the soft tissue. Recent studies

have indicated that SERPINA3 is a potential biomarker associated with tumor progression, which connoted that SERPINA3 is related to malignant phenotypes in cancer [49]. While SERPINA3 was found to be significant in our analysis, it has not yet been identified to manifest a direct connection to malignant breast cancer. However, we believe it should be further investigated.

ID1 (Inhibitor Of Differentiation 1) has been shown to play an important role in cell differentiation, tumor angiogenesis, cell invasion, and metastasis. Despite the data establishing ID1 as a critical factor for lung metastasis in breast cancer, the pathways and molecular mechanisms of ID1 functions in metastasis remains to be defined until recent studies suggested its role in promoting breast cancer metastasis [50].

RSBN1 (Round Spermatid Basic Protein 1) has been discovered recently as a potential HIF target, hypoxia inducible factor, for breast cancer. Hypoxia is a characteristic of breast tumours indicating poor prognosis over time [51].

Table 3.18: Genes identified in GSE 2990.

Gene Name	Models
SPAG16	RF 1, RF++, & LASSO
PCLAF	RF 1, RF++, & LASSO

In GSE 2990 the genes found to be relatively important were: SPAG16 and PCLAF. Relatively important genes were genes that were identified in three or more models or significantly by SAM and found in the following models.

SPAG16 (Sperm Associated Antigen 16) has not been found prognostic to breast cancer but interacting protein SPAG6 has been confirmed by analysis of tumor and normal tissue microarrays to be up-regulated in lung and breast cancer [52].

PCLAF (PCNA Clamp Associated Factor), also known as KIAA0101, has been seen along with SERPINA3 to be associated with oestrogen regulation and whose expression is stimulated in

a variety of oestrogen-sensitive systems [53]. Oestrogens have a major role in the regulation and function of many developmental processes in multiple target organs [54]. Additionally abnormal, excessive or prolonged stimulation by oestrogen can result in malfunction and be associated with diseases such as breast cancer [54].

CHAPTER IV

GENE EXPRESSION ANALYSIS ON LUNG CANCER

Lung cancer is the third most common cancer in the United States, proceeded by skin cancer and breast cancer (among women) [55]. More people in the US die from lung cancer than any other type of cancer. Cigarette smoking is the number one risk factor for lung cancer [55]. In the US, cigarette smoking is linked to about 80%-90% of all lung cancer deaths [54]. Using other tobacco products such as cigars or pipes also increases the risk for lung cancer [55]. In this chapter we present our findings for lung cancer pathogology related to objectives (1)-(5).

Prior to running machine learning methods for GSE 4115, the datasets were split into a training subset (70%) and a testing subset (30%). The training set was used to "train" the model parameters and the testing set was used to evaluate model metrics. The specifications of the dataset partitions created after *pOverA* filtration are as follows:

Table 4.1: Summary of genes in lung cancer pathology dataset.

	No. Genes after *pOverA*	Subjects in training dataset	Subjects in testing dataset
GSE 4115	816	129	56

4.1 Results of SAM

SAM analysis was performed to identify the differentially expressed genes for lung cancer pathology in dataset GSE 4115. Tables 4.2 and 4.3 present the results of the SAM analysis for the top 10 up- and down-regulated genes, respectively.

Table 4.2: GSE 4115 up-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
PITPNA	Phospholipid transporter activity	1.7490	1.3135	0.00
MFN2	Nucleotide binding	1.7222	1.3368	0.00
SRPRA	Nucleotide binding	1.7171	1.3162	0.00
FAM129A	Apoptosis, migration and proliferation	1.6743	1.3065	0.00
CIRBP	Nucleic acid binding	1.4901	1.3811	0.00
ITPR3	Inositol hexakisphosphate binding	1.4459	1.4002	0.00
PTPRF	Phosphoprotein phosphatase activity	1.3323	1.4497	8.4844
RPS10	Structural constituent of ribosome	1.3075	1.3638	8.4844
PAFAH1B1	Microtubule binding	1.2899	1.3405	8.4844
PTPRC	Phosphoprotein phosphatase activity	1.2141	1.4409	12.7266

Table 4.3: GSE 4115 down-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
TMED2	Protein binding	-1.4962	0.7453	12.7266
HSBP1	Transcription corepressor activity	-1.4608	0.7181	12.7266

GSE 4115 had 10 up-regulated genes and 2 down-regulated genes identified in total. Of the up-regulated genes found statistically significant (q-value<0.05), the genes MFN2, FAM129A, and ITPR3 have been found previously relevant to lung cancer.

MFN2 (Mitofusin 2), is a nucleotide binding gene and was previously associated with hypertension. However in studies, MFN2 expression has been found significantly higher in lung adenocarcinoma tissues [56] as opposed to control tissue samples. Given its up-regulated state, we can assume it presents higher in the patients that have been diagnosed with lung cancer.

FAM129A (Family with sequence similarity 129, member A) inhibits apoptosis, or cell death, and promotes migration and proliferation in human cancers. Studies have show that FAM129A may promote tumor proliferation and invasion of non-small cell lung carcinoma (NSCLC) in human lymph nodes [57].

High ITPR3 (Inositol 1,4,5-Trisphosphate Receptor Type 3) expression levels in tumors may be associated with a better survival of NSCLC patients [58]. Because of mutations of this gene are rare, high expression may predict survival of non-small cell lung carcinoma (NSCLC) according to Wu et al. [58].

Of the down-regulated genes found in GSE 4115, HSBP1 has been found previously relevant to lung cancer albeit not statistically significant.

HSBP1 (Heat Shock Factor Binding Protein 1) expression may have distinct prognostic values in non-small cell lung carcinoma (NSCLC) patients according to Huang et al. [59]. Low expression is correlated to better survival rates.

4.2 Results of RF 1

When creating the decision trees in the random forest, the following parameters have been fine-tuned: the number of trees and the number of nodes per tree. After training, Out-of-Bag error rate for GSE 4115 was 27.91%. The top 25 genes that were identified within GSE 4115 according to the mean decrease in accuracy are presented in Figure 4.1. The higher the mean decrease associated with a gene, the more important it is in correctly predicting the output.

Table 4.4: Results of RF 1 for lung cancer pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 4115	0.7209 ± 0.0114	0.7111 ± 0.0049	0.6034 ± 0.0084	0.6954 ± 0.0144

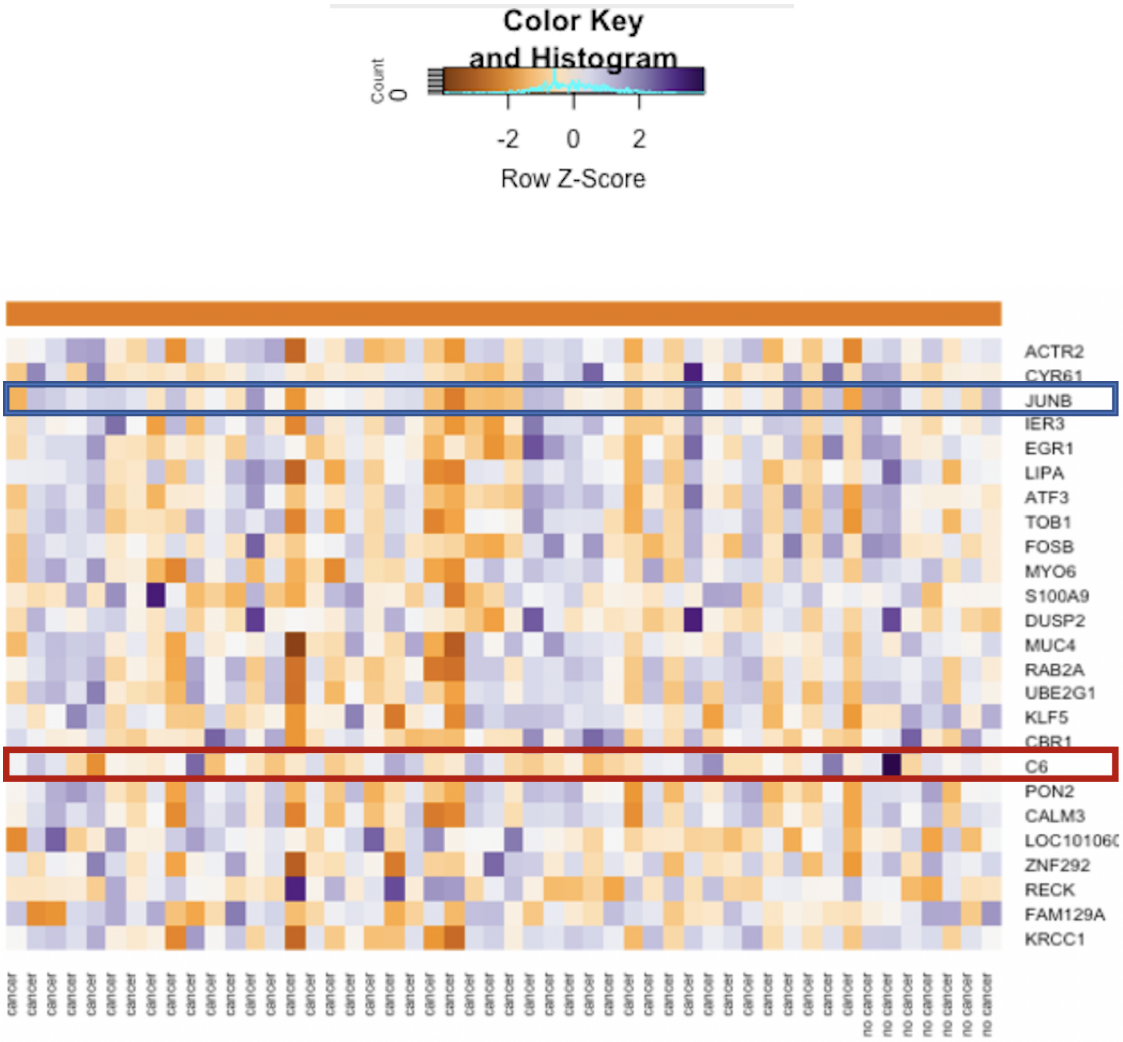


Figure 4.1: RF 1 method (single genes as inputs to RF) heatmap for GSE 4115.

4.3 Results of RF 2 (Random Forest with GXNA Clusters)

The top 25 most influential clusters with at most 15 genes in each cluster; clusters with more than one gene were used as input to the RF model. Out-of-Bag error rate for GSE 4115 was 22.19%. The top 13 genes in GSE 4115 were identified according to the mean decrease in accuracy and given in Figure 4.2.

Table 4.5: Results of RF 2 for lung cancer pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 4115	0.7781 ± 0.0062	0.6689 ± 0.0088	0.4207 ± 0.0179	0.6702 ± 0.0081

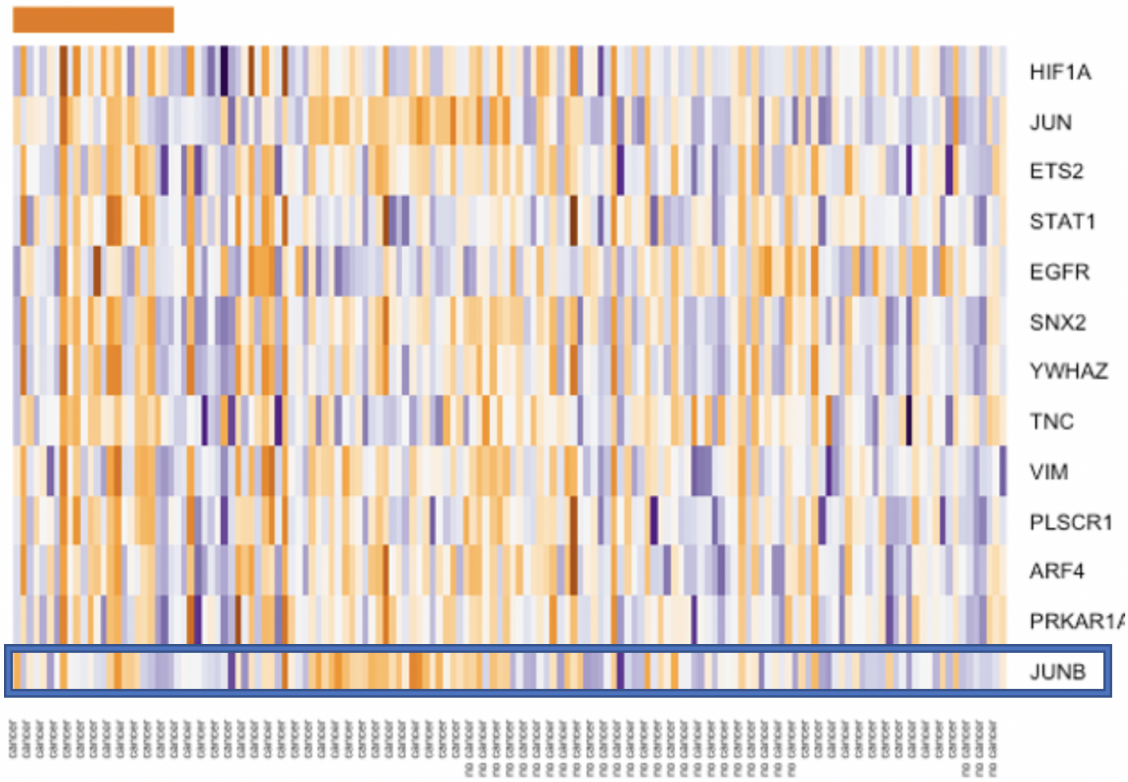


Figure 4.2: RF 2 method (clustered genes as inputs to RF) heatmap for GSE 4115.

JUNB gene was present and down-regulated in lung cancer patients in both the RF 1 and RF 2 methods, showing that it could be a relatively important gene for lung cancer.

Table 4.6: Change (Δ) in OOB for RF1/RF2 in lung cancer pathology.

	Δ OOB error rate
GSE 4115	-5.72

4.4 Results of RF++

Our RF++ yielded sample-level error rate of 23.21% on predictions of the subject disease class in for GSE 4115 testing dataset. RF++ with subject-level bootstrapping (SLB) was performed following both RF methods to analyze its use as a dimension reduction technique. The top 20 genes identified as important with permutation-based proportioning are presented in Table 4.8.

Table 4.7: Results of RF++ for lung cancer pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 4115	0.7679 ± 0.0317	0.6000 ± 0.0331	0.9032 ± 0.0780	0.7802 ± 0.0673

Table 4.8: Variable Importance in the GSE 4115 with RF++.

Gene Name	Score
CYR61*	0.0059
LOC100509457	0.0030
CYR61*	0.0030
FOS	0.0023
FOSB*	0.0020
STAT1*	0.0019
KLF5*	0.0015
DUSP6*	0.0014
AF010144	0.0014
TMEM47	0.0013
LOC101060275*	0.0012
ZNF160*	0.0011
INSR	0.0011
TMEM45A	0.0011
SLC35E1	0.0011
CXCR4	0.0010
NRGN	0.0010
CBR1*	0.0010
IER3	0.0010
CST6	0.0009

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

Several of the top genes identified as relatively important in RF++ were also identified in the top genes identified of the previous two RF methods.

4.5 Results of LASSO

Our LASSO yielded accuracy of 64.1% for GSE 4115 (Table 4.9). Figure A.3 in Appendix A displays the cross-validation error according to the log of λ for GSE 4115. The optimal λ was

found by minimizing the cross-validation prediction error, which will give the most optimal model.

Table 4.9: Results of LASSO for lung cancer pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 4115	0.6410 ± 0.0254	0.6410 ± 0.0026	0.3590 ± 0.0268	0.5448 ± 0.0203

Genes found to be to be significant as regression coefficients in the GSE 4115 LASSO are available in Table 4.10.

Table 4.10: Regression coefficients in the GSE 4115 with LASSO.

Gene Name	Coeff. Value
IER3*	-0.1547
LTF	0.0346
CPNE3	-0.0596
MIR4680	-0.5190
NELL2	-0.1753
HBG2	-0.0417
SNCA	-0.1116
LOC100996809	0.1344
TYMP	-0.0596
MX2	0.3723
KRT17	-0.1174
SCGB1A1	0.0665
HBD	-0.0632
CD46	-0.2062
UGT2A2	-0.2307
IGI16	-0.1072
KLF5*	0.3891
CBR1*	0.1209
SPP1	-0.0304
IGFBP3	0.0239
C6	-0.0249
FN1	-0.0598
RECK*	-0.9628
FAM129A*	-0.3795
DNAJC12	-0.1067
SPRR3	-0.0042
PDZK1IP1	-0.3191
PLPPR3	-0.2527
NUCKS1	0.1134

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

4.6 Results of Bayesian Neural Network

Using the Bayesian approach, we trained two types of networks with different weight regularization techniques. The first network was a standard network. The second type of the network was trained using Bayesian evidence procedure with ARD prior. Each network was trained with 10-fold cross validation and 3 hidden nodes and ran with 5 random initializations to obtain an average on all validation measures. The best network in terms of the highest accuracy and specificity was found to be the network trained using Bayesian evidence procedure along with the ARD prior for both breast cancer datasets. The results are shown in Table 4.10. As can be seen, use of evidence procedure along with the ARD prior has led to better sensitivities and higher AUC values indicating higher discrimination power.

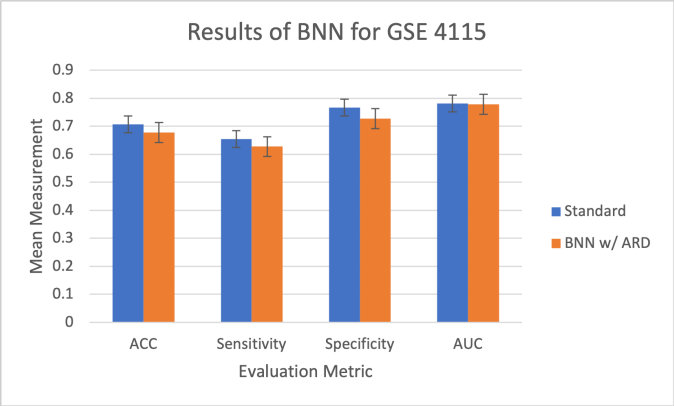


Table 4.11: Results of BNN for GSE 4115.

	Accuracy	Sensitivity	Specificity	AUC
Standard	0.7071 ± 0.0299	0.6546 ± 0.0139	0.7671 ± 0.0636	0.7817 ± 0.0164
BNN w/ ARD	0.6786 ± 0.0357	0.6277 ± 0.0353	0.7279 ± 0.0394	0.7794 ± 0.0194

4.6.1 Relative Importance of Genes Based on ARD prior

First 25 genes identified as relatively important by the ARD prior for GSE 4115 are given in Table 4.12. Final α values for the ANN trained under Bayesian evidence with the ARD prior for GSE 4115 was 0.234.

Table 4.12: Genes identified as relatively important by the ARD prior for GSE 4115.

Gene Name	Relative Var. Importance
C6*	169.7731
AZGP1	384.7405
CFD	440.8326
PSPH	867.9941
EIF5A	1045.9526
ATP8B1	1060.1301
PTN	1335.7198
TMEM47*	1380.3078
HSPA1B	1578.4274
LOC100509457*	1884.2459
ZDHHC11B	1909.9062
LY6D	2053.2058
RARRES1	2218.3027
RARRES2	2252.5971
NTS	2577.9195
NELL2*	2672.1974
APOD	2703.9649
UCHL1	2756.7809
HLA-DQB1	2773.0040
CRISP2	2907.2497
OXTR	2946.5680
NPIPA5	3161.4209
CST6*	3171.6443
G0S2	3172.4094
FOSB*	3211.7444

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

4.7 Potentially Relevant Genes for Lung Cancer

Table 4.13: Genes identified in GSE 4115.

Gene Name	Models
JUNB	RF 1 & RF 2
CST6	RF 1, RF++ & BNN
FOSB	RF1, RF++, & BNN

In GSE 4115, the genes found to be relatively important were: JUNB, C6, and NELL2. Relatively important genes were genes that were identified in three or more models or significantly by SAM and found in the following models.

JUNB (Jun B Proto-Oncogene) was present in both the RF 1 and RF2 methods. JUNB has been found to promote distant metastasis of head and neck squamous cell carcinoma (HNSCC) to other parts of the body [60], perhaps even to the lungs.

CST6 (Cystatin 6) has been found to be an associated gene in lung cancer that presents at significantly different levels between lung cancer subjects who have a history of smoking and those who didn't [61]. Tessema et al. state that it should be further studied as prognostic biomarker for lung adenocarcinoma [61].

FOSB (FosB Proto-Oncogene, AP-1 Transcription Factor Subunit) has a direct network connection to JUNB, along with TRAF5 identified in Chapter 3 as a potentially relevant gene to breast cancer. The methylation status of the FOSB gene in non-small cell lung cancinoma (NSCLC) and its clinical significance in progression was evaluated and been found to be a predictive biomarker for NSCLC prognosis [62]. FOSB may have a tumor suppressor function in the progression of NSCLC [62].

CHAPTER V

GENE EXPRESSION ANALYSIS ON PARKINSON'S DISEASE

Parkinson's disease (PD) is an incurable neurodegenerative brain disease mostly affecting those in old age. Nearly one million people in the U.S. are living with PD [63]. Incidences of Parkinson's disease increase with age, but an estimated 4% of people with PD are diagnosed before age 50 [63]. PD patients can live a fulfilling life with the correct treatment plans but the disease is often debilitating, leading to the main cause of death in those who are diagnosed [63]. Nerve cell damage in the brain causes dopamine levels to drop, leading to the symptoms of PD [63], and estimated 50–80% of people with PD eventually develop dementia.

Prior to running machine learning methods for GSE 8397, the datasets were split into a training subset (70%) and a testing subset (30%). The training set was used to "train" the model parameters and the testing set was used to evaluate model metrics. The specifications of the dataset partitions created after *pOverA* filtration are as follows:

Table 5.1: Summary of genes in PD pathology dataset.

	No. Genes after *pOverA*	Subjects in training dataset	Subjects in testing dataset
GSE 8397	596	38	8

5.1 Results of SAM

SAM analysis was performed to identify the differentially expressed genes for PD pathology in dataset GSE 8397. Tables 5.2 and 5.3 present the results of the SAM analysis for the top 10 up- and down-regulated genes, respectively.

Table 5.2: GSE 8397 up-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
TARDBP	Regulation of gene expression and mRNA processing	3.7008	3.3830	0.00
CCL5	Regulation of chronic inflammatory response	3.6919	4.3560	0.00
ATF4	Transcription by RNA polymerase II	3.6292	4.7593	0.00
CCND2	Long-term memory	3.6201	3.7897	0.00
HSPD1	MyD88-dependent toll-like receptor signaling pathway	3.6163	3.0827	0.00
AHCY	Sulfur amino acid metabolic process	3.5498	5.0944	0.00
PTPN21	Phosphoprotein phosphatase activity	3.4347	3.0086	0.00
CTNNA1	Establishment or maintenance of cell polarity	3.4094	2.9301	0.00
SERP1	Protein binding	3.3569	2.6945	0.00
EIF5B	Nucleotide binding	3.2938	3.4720	0.00

Table 5.3: GSE 8397 down-regulated genes.

Gene Name	Gene Ontology	Score (d)	Fold Change	q-value
C11ORF58	Biological process	-4.3589	0.2728	0.00
ACTR2	Associative learning	-4.0530	0.3563	0.00
DNAJB1	Response to unfolded protein	-4.0465	0.2856	0.00
MARCKSL1	Calmodulin binding	-3.8907	0.3864	0.00
GABARAP	GABA receptor binding	-3.5939	0.3381	0.00
SNORA52	RNA processing	-3.4608	0.4329	0.00
MCL1	Protein homorodimerization/heterodimerization activity	-3.3686	0.3186	0.00
RPL37	Structural constituent of ribosome	-3.2409	0.4865	0.00
HSBP1	Transcription corepressor activity	-3.2297	0.4468	0.00
EPAS1	RNA polymerase II regulatory region sequence-specific DNA binding	-3.0853	0.4988	0.00

GSE 8397 had 65 up-regulated genes and 45 down-regulated genes identified in total. Of the up-regulated genes found statistically significant (q-value<0.05), the genes TARDBP, CCL5, ATF4, HSPD1, AHCY, PTPN21, and CTNNA1 have been found previously relevant to PD.

TARDBP (TAR DNA Binding Protein) is the gene encoding TDP-43 protein, which has been observed across a spectrum of neurodegenerative disorders, including Alzheimer's disease (AD) and PD [64]. Due to its critical role in the pathogenesis of these diseases, we believe it deserves more consideration as a potential biomarker for PD.

CCL5 (C-C Motif Chemokine Ligand 5) is responsible for regulation of chronic inflammatory response, which may reflect a role of systemic inflammation in the neurodegenerative process of PD [65].

ATF4 (Activating Transcription Factor 4) has been found to play a previously undescribed protective role in PD-related neuronal death by maintaining levels of parkin, a recessive autosomal gene in PD. These findings have implications regarding potential disease-modifying strategies for PD [66].

Mutations in HSPD1 (Heat Shock Protein Family D (Hsp60) Member 1) have been associated with autosomal-recessive neurodegenerative disorder [67], like PD.

Dysfunctional AHCY (Adenosylhomocysteinase) activity can result in serious pathological consequences, such as childhood death, AD, PD, age-related diseases, neuroblastoma, and large-artery atherosclerotic stroke [68].

CTNNA1 (Catenin Alpha-1) showed significant change in co-expression levels between disease and control states in Rakshit et al. [69]. It may be a potential biomarkers for therapeutic targets for PD applications developments after further investigation.

Of the down-regulated genes found statistically significant in GSE 8397, the genes DNAJB1, GABARAP, MCL1, and HSBP1 have been found previously relevant to PD.

Although mutations of DNAJB1 (DnaJ Heat Shock Protein Family (Hsp40) Member B1) has been found primarily in the liver tissue, post-mortem PD brain tissues showed immunoreactivity for DNAJB1 in Lewy bodies [70] which are often found in a type of progressive dementia known as Lewy Body Dementia (LBD). Changes to proteins in the brain, such as Lewy bodies, can lead to dementia in Parkinson's patients.

Expression of GABARAP (GABA Type A Receptor-Associated Protein) was able to successfully segregate PD patients from healthy controls in El Haddid et al. [71]. This pilot study suggested that autophagy genes, or genes responsible for cell degredation, expression is dysregulated in PD patients and may open new perspectives for the characterisation of prediction markers.

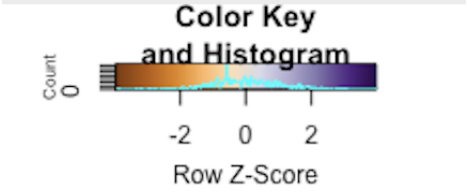
MCL1 (Myeloid Cell Leukemia 1) inhibition may result in apoptosis of neuronal cells and of dopamine neurons. With further investigation, enhancing MCL1 may be a therapeutic strategy to delay apoptosis of dopamine neurons in PD [72].

5.2 Results of RF 1

When creating the decision trees in the random forest, the following parameters have been fine-tuned: the number of trees and the number of nodes per tree. After training, Out-of-Bag error rate for GSE 8397 was 12.82%. The top 25 genes that were identified within GSE 8397 according to the mean decrease in accuracy are presented in Figure 5.1. The higher the mean decrease associated with a gene, the more important it is in correctly predicting the output.

Table 5.4: Results of RF 1 for PD pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 8397	0.8718 ± 0.0087	0.8667 ± 0.0085	0.8750 ± 0.0171	0.8972 ± 0.0387



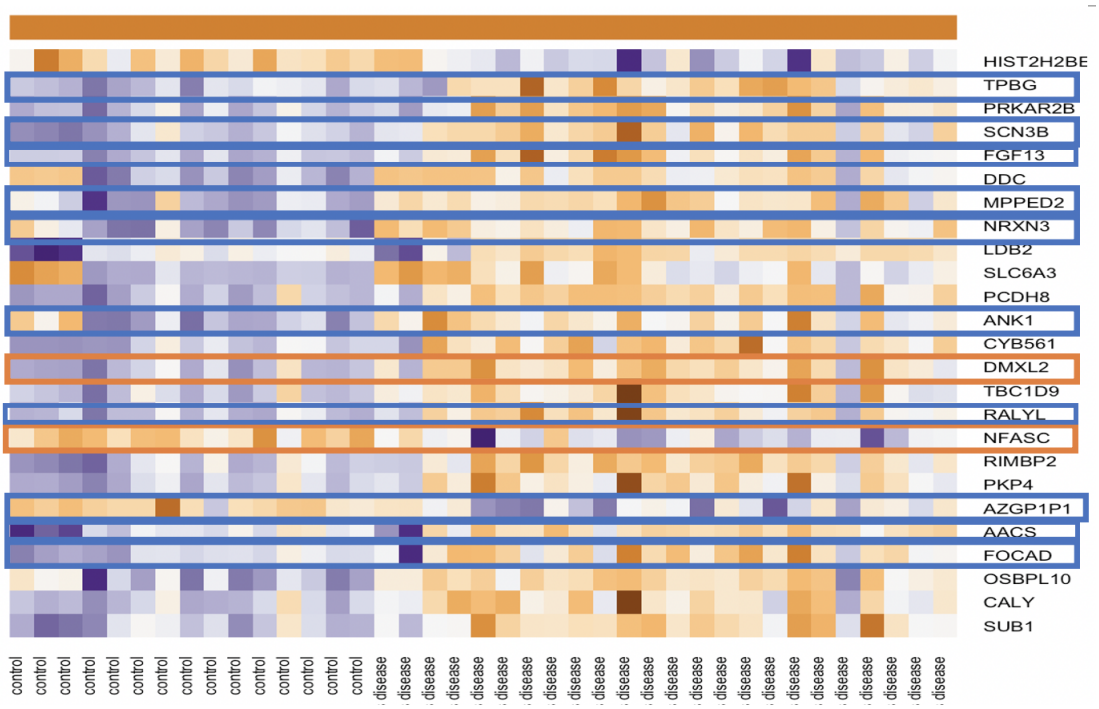


Figure 5.1: RF 1 method (single genes as inputs to RF) heatmap for GSE 8397.

5.3 Results of RF 2 (Random Forest with GXNA Clusters)

The top 25 most influential clusters with at most 15 genes in each cluster; clusters with more than one gene were used as input to the RF model. Out-of-Bag error rate for GSE 8397 was 16.98%. The top 31 genes in GSE 8397 were identified according to the mean decrease in accuracy and given in Figure 5.2.

Table 5.5: Results of RF 2 for PD pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 8397	0.8302 ± 0.0185	0.8273 ± 0.0174	0.5531 ± 0.0244	0.8892 ± 0.0641

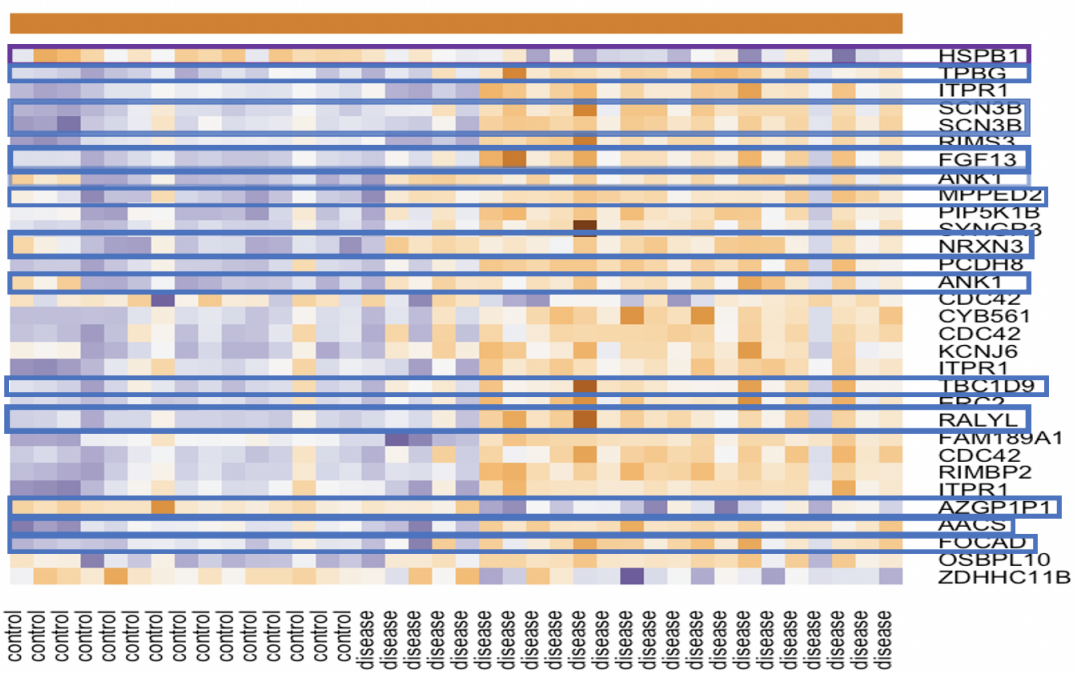


Figure 5.2: RF 2 method (clustered genes as inputs to RF) heatmap for GSE 8397.

HSBP1 gene that was identified with SAM was present and down-regulated in lung cancer patients in the RF 2 method, showing that it could be a relatively important gene for lung cancer.

Genes TPBG, SCN3B, FGF13, MPPED2, NRXN3, ANK1, TBC1D9, RALYL, AZGP1P1, and FOCAD were present in both the RF 1 and RF 2 methods, showing that they could be relatively important genes for Parkinson's disease.

Table 5.6: Change in OOB for RF1/RF2 in PD pathology.

	Δ OOB error rate
GSE 8397	+4.16

5.4 Results of RF++

Our RF++ yielded sample-level error rate of 12.5% on predictions of the subject disease class in for GSE 8397 testing dataset. RF++ with subject-level bootstrapping (SLB) was performed

following both RF methods to analyze its use as a dimension reduction technique. The top 20 genes identified as important with permutation-based proportioning are presented in Table 5.8.

Table 5.7: Results of RF++ for PD pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 8397	0.8750 ± 0.0794	0.6667 ± 0.0307	1.00 ± 0.00	0.9167 ± 0.1081

Table 5.8: Variable Importance in the GSE 8397 with RF++ .

Gene Name	Score
TPBG*	0.0081
MPPED2*	0.0077
HSPB1*	0.0073
FGF13*	0.0071
SYNGR3*	0.0059
ANK1*	0.0059
AZGP1P1*	0.0050
RALYL*	0.0041
ITPR1*	0.0040
DMXL2*	0.0036
SCN3B*	0.0035
PCDH8*	0.0033
NRXN3*	0.0030
DRD2	0.0028
ANK1	0.0026
TMEM35A	0.0026
ZNF226	0.0025
OSBPL10	0.0025
GBE1	0.0025
DDC	0.0023

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

As suspected, the top genes identified as relatively important in RF++ very closely resemble the top genes identified in the previous two RF methods.

5.5 Results of LASSO

Our LASSO regression yielded accuracy of 57.14% for GSE 8397. Due to small sample size, predictions on the control state were all misclassified forcing the specificity to zero. Figure

A.4 in Appendix A displays the cross-validation error according to the log of λ for GSE 8397. This λ value will give the most accurate model.

Table 5.9: Results of LASSO for PD pathology.

	Accuracy	Sensitivity	Specificity	AUC
GSE 8397	0.5714 ± 0.0001	0.7500 ± 0.0001	0 ± 0	0.9000 ± 0.0001

Genes found to be to be significant as regression coefficients in the GSE 8397 LASSO are available in Table 5.10.

Table 5.10: Regression coefficients in the GSE 8397 with LASSO.

Gene Name	Coeff. Value
SCN3B*	-0.1757
MPPED2*	-0.2636
NRXN3*	-0.5897
RBM3	-0.0244
HIST1H2BD	0.2568
DMXL2*	-0.2889
NPTX2	0.0219
TRA2A	0.0376
NFASC*	0.1236
AZGP1P1*	0.7435

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

5.6 Results of Bayesian Neural Network

Using the Bayesian approach, we trained two types of networks with different weight regularization techniques. The first network was a standard network. The second type of the network was trained using Bayesian evidence procedure with ARD prior. Each network was trained with 10-fold cross validation and 3 hidden nodes and ran with 5 random initializations to obtain an average on all validation measures. The best network in terms of the highest accuracy and specificity was found to be the network trained using Bayesian evidence procedure along with the ARD prior for both breast cancer datasets. The results are shown in Table 5.11. Again, this method does not have model evaluation values to be presented. Due to small sample size, predictions on the control state were all misclassified, forcing the specificity to zero. Similar to the previous methods presented for the same dataset, this is something to be expected for a relatively small sample size.

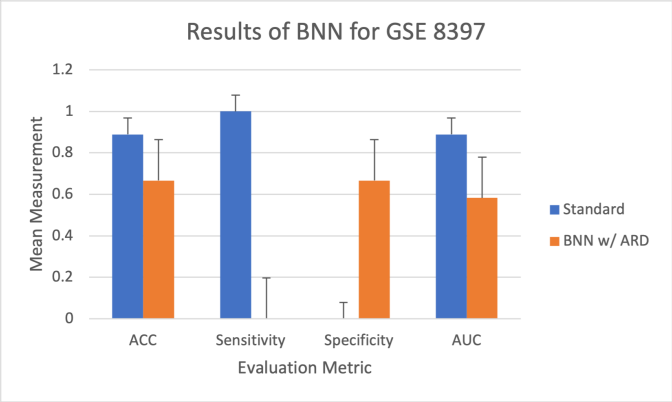


Table 5.11: Results of BNN for GSE 8397.

	Accuracy	Sensitivity	Specificity	AUC
Standard	0.8889 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.8889 ± 0.0785
BNN w/ ARD	0.6667 ± 0.00	0.00 ± 0.00	0.6667 ± 0.00	0.5833 ± 0.1964

5.6.1 Relative Importance of Genes Based on ARD Prior

First 25 genes identified as relatively important by the ARD prior for GSE 8397 are given in Table 5.12. Final α values for the ANN trained under Bayesian evidence with the ARD prior for GSE 8397 was 0.012.

Table 5.12: Genes identified as relatively important by the ARD prior for GSE 8397.

Gene Name	Relative Var. Importance
AZGP1P1*	214.2411
ARHGEF3	1170.4587
ATF4*	1820.5283
ARPP21	4496.3073
PTPRC	4505.9072
SLC18A2	4557.7792
TH	4636.6757
CYR61	5362.7128
F3	6100.9119
DTNA	7051.3135
CERS6	7097.3589
HLA-DRB4	7195.0415
RAB40B	7885.3180
RAB15	7989.3172
RBFOX1	8100.1350
EN1	8160.0887
CA11	8913.4923
RUNDC3A	11668.2999
NPTX2*	12323.6237
NREP	12782.2904
IFI16	18343.0826
ARL4C	18420.4391
CP	18430.9593
ZFPM2	20083.3103
C16orf45	21314.0870

(a) Genes that have been identified in a previous method are indicated by an asterisk (*).

5.7 Potentially Relevant Genes for Parkinson's Disease

Table 5.13: Genes identified overall in GSE 8397.

Gene Name	Models
ATF4	SAM & BNN
HSBP1	SAM, RF++ & RF 2
AZGP1P1	RF 1, RF 2, RF++, LASSO & BNN
SCN3B	RF 1, RF 2, RF++, & LASSO
MPPED2	RF 1, RF 2, RF++, & LASSO
NRXN3	RF1, RF 2, RF++, & LASSO

In GSE 8397, the genes found to be relatively important were: ATF4, HSBP1, AZGP1P1, SCN3B, MPPED2, and NRXN3. Relatively important genes were genes that were identified in three or more models or significantly by SAM and found in the following models.

ATF4 (Activating Transcription Factor 4) has been found to play a previously undescribed protective role in PD-related neuronal death by maintaining levels of parkin, a recessive autosomal gene in PD. These findings have implications regarding potential disease-modifying strategies for PD [66]. ATF4 was identified as statistically significant in SAM and also present in the BNN, indicating that it may be a relatively important gene in PD patients.

HSBP1 (Heat Shock Factor Binding Protein 1) levels are elevated in the cortex of Alzheimer's patients, with higher levels corresponding to increased severity and duration of dementia [73]. HSBP1 was identified as statistically significant in SAM and also present in the RF 2. It's presence in RF 2 means it was found in a high-ranking cluster identified by GXNA.

AZGP1P1 (Alpha-2-Glycoprotein 1, Zinc Pseudogene 1) present in all of the analyzed machine learning methods, strongly suggesting it is relatively important to PD. AZGP1P1 is a

pseudogene, meaning it is primarily nonfunctional but does resemble a functional gene, perhaps one related to PD pathogenesis. It is suggested to be associated with abnormality of refraction, an abnormality in the process of focusing of light by the eye in order to produce a sharp image on the retina. Visual abnormalities can actually be common in PD despite the lack of attention they receive as opposed to other symptoms of the disease. Since we do not know what gene AZGP1P1 resembles in our study, further investigation by experts should be warranted.

SCN3B (Sodium Voltage-Gated Channel Beta Subunit 3) is primarily responsible for the generation and propagation of action potentials in neurons and muscle. SCNB3 was present in both the RF 1 and RF 2 methods as well as the RF++ and LASSO. Although SCNB3 has yet to be found relevant to PD, it has been identified as very significant to syncope and tachycardia which are both symptoms of a PD diagnosis. Syncope refers to a generalized weakness of muscles with loss of postural tone, inability to stand upright, and loss of consciousness. It can be a manifestation of PD and made worse by PD medications. Tachycardia is a rapid heartrate that exceeds the range of the normal resting heartrate for age. It is associated to Wolff-Parkinson-White syndrome, in which patients diagnosed have an extra electrical pathway.

MPPED2 (Metallophosphoesterase Domain Containing 2), also known as C11orf8, is also associated with abnormality of refraction like AZGP1P1. MPPED2 was present in both the RF 1 and RF 2 methods as well as the RF++ and LASSO. It has not been identified as important to PD.

NRXN3 (Neurexin 3) was present in both the RF 1 and RF 2 methods as well as the RF++ and LASSO. NRXN3 has no known associations to PD, but it has been known to carry a mutation in AD patients [74].

5.8 Discussion of Dimension Reduction Techniques/Methodologies

The results of the dimension reduction techniques for the gene expression data we have conducted in the analysis are presented in Table 5.14, each evaluation metric's associated s.d. can be found in their appropriate tables. The dimension reductions techniques were: RF 2, and LASSO.

Table 5.14: Summary of Dimension Reduction Techniques.

	RF 1					RF 2					LASSO				
	No. Genes	ACC	Sens.	Spec.	AUC	No. Genes	ACC	Sens.	Spec.	AUC	No. Genes	ACC	Sens.	Spec.	AUC
GSE 2034	**775**	0.6739	0.7229	0.2326	0.4372	**107**	0.5767	0.7000	0.2931	0.6954	**9**	0.6349	0.8125	0.9947	0.5794
GSE 2990	**1038**	0.5563	0.6333	0.4231	0.5962	**213**	0.6801	0.8109	0.3345	0.7777	**4**	0.7059	0.7308	0.6250	0.6023
GSE 4115	**816**	0.7209	0.7111	0.6034	0.6954	**84**	0.7781	0.6689	0.4207	0.6702	**29**	0.6410	0.3590	0.6410	0.5448
GSE 8397	**596**	0.8718	0.8667	0.8750	0.8972	**165**	0.8302	0.8273	0.5531	0.8892	**10**	0.5741	0.7500	0	0.9000

RF 1 contained the initial number of gene after filtration, RF 2 contained the clusters which could then be consider biologically relevant dimension reduction, and LASSO contained the non-zero coefficients of the LASSO.

CHAPTER VI

CONCLUSIONS, CONTRIBUTIONS, AND FUTURE STUDIES

This study introduces a biological knowledge based machine learning approach to gene expression analysis that is much needed. Moreover, we have investigated the use of ARD prior for possible identification of biomarkers within different disease pathologies. Herewith, we summarize our conclusions and contributions to the five objectives that we intended to achieve in this study.

With SAM, we were able to identify statistically significant genes (Objective (1)) for breast cancer, lung cancer, and Parkinson's disease. Many of the genes found were indeed related to their pathologies as intended.

With the random forest models along with GXNA clusters, we were able to achieve the intended goals for Objectives (2) and (3). We find random forest models to be good at prediction but not at inferences about specific genes. In contrast to that, GXNA based random forest models took into account the correlation between genetic pathways and was an overall better approach as it utilized prior biological knowledge. We would like to extend the suggestion to genomics experts to look closer at the genes that were discovered with GXNA, especially the genes SERPINA3 and MAD2L1 for their roles in breast cancer pathogenesis and the genes HSBP1, SCN3B, MPPED3, and NRXN3 for their roles in Parkinson's disease pathogenesis. The results presented on in this study were made possible with the studious work of Nacu et al., unfortunately GXNA is no longer available for public usage. Future studies include the possible creation of our own clustering method based on biological knowledge to fulfill this missing algorithm.

Our study also confirmed the usage of BNN in improving the accuracy and consistency of

disease state prognosis, achieving Objective (4). In addition, ARD prior provides one sophisticated method to evaluate the relative importance of genes associated to a certain disease pathology which has been investigated for the first time. With further input from experts in these fields, these genes should be investigated for potential disease biomarker identification.

Finally, Objective (5) was achieved with the use of GXNA clusters and LASSO to reduce the number of genes in the model as a form of dimension reduction. In the future, we have planned to examine the impact of negative/positive coefficient values on up-/down-regulated gene associations, as well as explore the use of fractal dimension reduction techniques.

Further identification of DE genes may be made by application of a π-value based approach. A new gene significance score, π-value, is calculated by combining expression biological relevance (such as fold change) and statistical significance (p-value) for better gene ranking [75]. When applied to gene set enrichment analysis (GSEA), π-value has been found to be comparable to p-value and t-statistic based methods, with added protection against false discovery in certain situations [75].

The concept of machine learning has grown over the years to encompass models of high accuracy yet low explainability and interpretability. While models like random forests, linear regressions, and descriptive statistics are easier to explain and interpret to those who are outside of the scope of statistics, they cannot ultimately compare to the useful nature of deep learning models like neural networks [76]. We look forward to expanding the range of models examined on gene expression data to further the field of machine learning and data mining in everyday life.

BIBLIOGRAPHY

[1] National Human Research Institute, “Gene Expression.” [Online]. Available: https://www.genome.gov/genetics-glossary/Gene-Expression

[2] Wikipedia, “Gene expression.” [Online]. Available: https://en.wikipedia.org/wiki/Gene_expression

[3] Future Learn, “Different types of RNAs and their functions.” [Online]. Available: https://www.futurelearn.com/info/courses/translational-research/0/steps/14201

[4] H. Koltai and C. Weingarten-Baror, “Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction,” *Nucleic Acids Research*, vol. 36, issue 7, pp. 2395–2405, Apr. 2008.

[5] National Center for Biotechnology Information, “Microarray.” [Online]. Available: https://www.ncbi.nlm.nih.gov/probe/docs/techmicroarray/

[6] R. Dias and A. Torkamani, “Artificial intelligence in clinical and genomic diagnostics,” *Genome Medicine*, vol. 11, no. 70, Nov. 2019.

[7] J. K. Lee, P. D. Williams and S. Cheon, “Data Mining in Genomics,” *Clinics in Laboratory Medicine*, vol. 28, issue 1, pp. 145-166, Mar. 2008.

[8] C. Devi Arockia Vanitha, D. Devaraj and M. Venkatesulu, “Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection,” *Procedia Computer Science*, vol. 47, pp. 13-21, May 2015.

[9] H. Pang, A. Lin, M. Holford, B. E. Enerson, B. Lu, M. P. Lawton, E. Floyd and H. Zhao, “Pathway analysis using random forests classification and regression,” *Bioinformatics*, vol. 22, issue 16, pp. 2028-2036, Aug. 2006.

[10] M. Mostavi, YC. Chiu, Y. Huang and Y. Chen, “Convolutional neural network models for cancer type prediction based on gene expression,” *BMC Med Genomics*, vol. 13, no. 44, Apr. 2020.

[11] L. M. de Campos, A. Cano, J. G. Castellano and S. Moral, "Bayesian networks classifiers for gene-expression data," *2011 11th International Conference on Intelligent Systems Design and Applications*, pp. 1200-1206, 2011. doi: 10.1109/ISDA.2011.6121822.

[12] M. Daoud and M. Mayo, "A survey of neural network-based cancer prediction models from microarray data," *Artificial Intelligence in Medicine*, vol. 97, pp. 204-214, Jun. 2019.

[13] D. Maglott, J. Ostell, K. D. Pruitt and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 39, issue suppl_1, pp. D52-D57, Jan. 2011.

[14] M. M. Dalkilic, "Gene Expression Arrays," *Encyclopedia of Database Systems*, pp. 52-154, 2009.

[15] Brainarray, "Description of Customized CDF Files." [Online]. Available: http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/

[16] J. Wu and R. Irizarry with contributions from J. MacDonald and J. Gentry, "Background Adjustment Using Sequence Information," Package 'gcrma', version 2.64.0, Aug. 2021. [Online]. Available: https://www.bioconductor.org/packages/release/bioc/manuals/gcrma/man/gcrma.pdf

[17] rdrr.io, "pOver." [Online]. Available: https://rdrr.io/bioc/genefilter/man/pOverA.html

[18] V. G. Tusher, R. Tibshirani and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, issue 9, pp. 5116-5121, Apr. 2001.

[19] G. Chu, M. Seo, J. Li, B. Narasimhan, R. Tibshirani and V. Tusher, "SAM "Significance Analysis of Microarrays" User guide and technical document," Package 'SAM', version 2.20. [Online]. Available: http://statweb.stanford.edu/ tibs/SAM/sam.pdf

[20] L. Breiman, *Random Forests*. Statistics Department, University of California, Berkeley, CA., 2001.

[21] SB. Cho, HH. Won, "Machine learning in DNA microarray analysis for cancer classification," *APBC '03: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, vol. 19, pp. 189-198, Jan. 2003.

[22] L. Breiman, *Out-Of-Bag Estimation*. Statistics Department, University of California, Berkeley, CA., 1996.

[23] O. L. Griffith, F. Pepin, O. M. Enache, L. M. Heiser, E. A. Collisson, P. T. Spellman and J. W. Gray, "A robust prognostic signature for hormone-positive node-negative breast cancer," *Genome Medicine*, vol. 5, no. 92, Oct. 2013.

[24] Biostars Bioinformatics Explained, "Tutorial: Machine Learning For Cancer Classification - Part 1 - Preparing The Data Sets." [Online]. Available: https://www.biostars.org/p/85124/

[25] S. Nacu, R. Critchley-Thorne, P. Lee and S. Holmes, "Gene expression network analysis and applications to immunology," *Bioinformatics*, vol. 23, issue 7, pp. 850-858, Apr. 2007.

[26] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. R80, Sep. 2004.

[27] G. K. Smyth, M. Ritchie, N. Thorne, J. Wetternhall, W. Shi and Y. Hu," limma: Linear Models for Microarray and RNA-Seq Data User's Guide," package 'limma', version 3.13, Jul. 2021. [Online]. Available: https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf

[28] W. Huber, A. von Heydebreck, H. Sultmann, A, Poustka and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol 18, issue suppl_1, pp. S96-S102, Jul. 2002.

[29] T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, issue suppl_1, pp. S233-S240, Jul. 2002.

[30] Y. V. Karpievitch, E. G. Hill, A. P. Leclerc, A. R. Dabney and J. S. Almeida, "An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++," *PLOS ONE*, vol. 4, issue 9, Sep. 2018.

[31] A. Kassambara, "Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net," *STHDA*, Nov. 2018. [Online]. Available: http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/

[32] H. R. Frost and C. I. Amos, "Gene set selection via LASSO penalized regression (SLPR)," *Nucleic Acids Research*, vol. 45, issue 12, Jul. 2017.

[33] A. Kassambara, *Machine Learning Essentials: Practical Guide in R*. Scotts Valley, CA. CreateSpace Independent Publishing Platform, 2018.

[34] MIT News, "Explained: Neural networks," L. Hardesty, Apr. 2017. [Online]. Available: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[35] D. J. C. MacKay, "A Practical Bayesian Framework for Backpropagation Networks," *Neural Computation*, vol. 4, issue 3, pp. 448-472, 1992.

[36] D. J. C. MacKay, "Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, issue 3, pp. 469-505, 1995.

[37] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[38] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[39] I. T. Nabney, *NETLAB: Algorithms for Pattern Recognition*. New York: Springer, 2004.

[40] D. J. C. MacKay, "Bayesian Methods for Neural Networks: Theory and Applications," Neural Computing Research Group, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, U.K. 1995.

[41] H. S. Rodrigo, "Bayesian Artificial Neural Networks in Health and Cybersecurity," D. S. Dissertation, University of South Florida, Tampa, FL, Scholar Commons, Jul. 2017.

[42] D. J. C. MacKay, "Bayesian non-linear modeling for the prediction competition," Maximum Entropy and Bayesian Methods, pp. 221-234, 1996.

[43] National Breast Cancer Foundation, Inc., "2020 Breast Cancer Statistics." [Online]. Available: https://www.nationalbreastcancer.org/wp-content/uploads/2020-Breast-Cancer -Stats.pdf

[44] A. W. Hubball., B. Lang, M. A. Souza, O. D. Curran, J. E. Martin and C. H. Knowles, "Voltage-gated potassium channel (kv1) autoantibodies in patients with chagasic gut dysmotility and distribution of kv1 channels in human enteric neuromusculature (autoantibodies in GI dysmotility)," *Neurogastroenterology & Motility*, vol. 24, issue 8, May 2012.

[45] S. H. Jang, K. S. Kang, P. D. Ryu and S. Y. Lee, "Kv1.3 voltage-gated K(+) channel subunit as a potential diagnostic marker and therapeutic target for breast cancer," *BMB reports Korean Society for Biochemistry and Molecular Biology*, vol. 42, issue 8, pp. 535-539, Aug. 2009.

[46] L. F. Miao, X. H. Ye and X. F. He, "Individual and combined effects of GSTM1, GSTT1, and GSTP1 polymorphisms on breast cancer risk: A meta-analysis and re-analysis of systematic meta-analyses," *PLOS ONE*, vol. 15, issue 3, Mar. 2020.

[47] Y. Osawa, Y. Yokoyama, T. Shigeto, M. Futagami and H. Mizunuma, "Decreased expression of carbonyl reductase 1 promotes ovarian cancer growth and proliferation," *International Journal of Oncology*, vol. 46, pp. 1252-1258, Dec. 2014.

[48] A. Jo, T. G. Choi, Y. H. Jo, K. R. Jyothi, M. N. Nguyen, J. H. Kim, S. Lim, M. Shahid, S. Akter, S. Lee, K. H. Lee, W. Kim, H. Cho, J. Lee, K. M. Shokat, K. S. Yoon, I. Kang, J. Ha and S. S. Kim, "Inhibition of Carbonyl Reductase 1 Safely Improves the Efficacy of Doxorubicin in Breast Cancer Treatment," *Antioxidants & Redox Signaling*, vol. 26, no. 2, pp. 73-80, Jan. 2017.

[49] Y. Zhang, J. Tian, C. Qu, Y. Peng, J. Lei, K. Li, B. Zong, L. Sun and S. Liu, "Overexpression of SERPINA3 promotes tumor invasion and migration, epithelial-mesenchymal-transition in triple-negative breast cancer cells," *Breast Cancer*, vol. 28, pp. 859-873, Feb. 2021.

[50] K. Gumireddy, A. Li, A. V. Kossenkov, K. Q. Cai, Q. Liu, J. Yan, H. Xu, L. Showe, L. Zhang and Q. Huang, "ID1 promotes breast cancer metastasis by S100A9 regulation," *Molecular Cancer Research*, vol. 12, issue 9, pp. 1334-1343, Sep. 2014.

[51] B. Abu-Jamous, F. M. Buffa, A. L. Harris and A. K. Nandi, "In vitro downregulated hypoxia transcriptome is associated with poor prognosis in breast cancer," *Molecular Cancer*, vol. 16, no. 105, Jun. 2017.

[52] K. Silina, P. Zayakin, Z. Kalnina, L. Ivanova, I. Meistere, E. Endzelins, A. Abols, A. Stengrevics, M. Leja, K. Ducena, V. Kozirovskis and A. Linē, "Sperm-associated Antigens as Targets for Cancer Immunotherapy: Expression Pattern and Humoral Immune Response in Cancer Patients," *Journal of Immunotherapy*, vol. 34, issue 1, pp. 28-44, Jan. 2011.

[53] W. R. Miller and A. Larionov, "Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase inhibitor, letrozole," *Breast Cancer Research*, vol. 12, no. R52, Jul. 2010.

[54] W. R. Miller and A. Larionov, "Molecular Effects of Oestrogen Inhibition in Breast Cancer," *Molecular and Cellular Endocrinology*, vol. 340, no. 2, pp.127, Jul. 2012.

[55] Center of Disease Control and Prevention, "What Are the Risk Factors for Lung Cancer?" [Online]. Available: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

[56] Y. Lou, Y. Zhang, J. Xu, P. Gu, W. Zhang, X. Zhang, H. Zhong, L. Jiang, and B. Han, "MFN2 might be a risk factor for lung adenocarcinoma," *Journal of Clinical Oncology*, vol. 35, issue 15_suppl, 2017.

[57] N. Zhang, X. M. Zhou, F. F. Yang, Q. Zhang, Y. Miao and G. Hou, "FAM129A promotes invasion and proliferation by activating FAK signaling pathway in non-small cell lung cancer," . *International Journal of Clinical and Experimental Pathology*, vol. 12, issue 3, pp. 893-900, 2019.

[58] Y. Wu, Z. Liu, D. Tang, H. Liu, S. Luo, T. E. Stinchcombe, C. Glass, L. Su, L. Lin, D. C. Christiani, Q. Wang, and Q. Wei, "Potentially functional variants of HBEGF and ITPR3 in GnRH signaling pathway genes predict survival of non-small cell lung cancer patients," *Translational Research : The Journal of Laboratory and Clinical Medicine*, vol. 233, pp. 92-103, Jul. 2021.

[59] Z. C. Huang, H. Li, Z. Q. Sun, J. Zheng, R. K. Zhao, J. Chen, S. G. Sun and C. J. Wu, "Distinct prognostic roles of HSPB1 expression in non-small cell lung cancer," *Neoplasma*, vol. 65, no. 1, pp. 161-166, 2018.

[60] H. Hyakusoku, D. Sano, H. Takahashi, T. Hatano, Y. Isono, S. Shimada, Y. Ito, J. N. Myers and N. Oridate, "JunB promotes cell invasion, migration and distant metastasis of head and neck squamous cell carcinoma," *Journal of Experimental & Clinical Cancer Research*, vol. 35, no. 6, Jan. 2016.

[61] M. Tessema, C. M. Yingling, Y. Liu, C. S. Tellez, L. Van Neste, S. S. Baylin and S. A. Belinsky, "Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers," *Carcinogenesis: Integrative Cancer Research*, vol. 35, issue 6, pp. 1248–1257, Jun. 2014.

[62] D. S. Kim, W. K. Lee and J. Y. Park, "Association of FOSB exon 4 unmethylation with poor prognosis in patients with late stage non small cell lung cancer," *Spandidos Publications: Oncology Reports*, vol. 43, no. 2, pp. 655–661, Dec. 2019.

[63] MayoClinic, "Parkinson's disease." [Online] Available: https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055

[64] S. Rayaprolu, S. Fujioka, S. Traynor, A. I. Soto-Ortolaza, L. Petrucelli, D. W. Dickson, R. Rademakers, K. B. Boylan, N. R. Graff-Radford, R. J. Uitti, Z. K. Wszolek and O. A. Ross, "TARDBP mutations in Parkinson's disease," *Parkinsonism & Related Disorders*, vol. 19, issue 3, pp. 312–315, Mar. 2013.

[65] P. Tang, L. Chong, X. Li, Y. Liu, P. Liu, C. Hou and R. Li, "Correlation between serum RANTES levels and the severity of Parkinson's disease," *Oxidative Medicine and Cellular Longevity*, vol. 2014, no. 208408, Dec. 2014.

[66] X. Sun, J. Liu, J. F. Crary, C. Malagelada, D. Sulzer, L. A. Greene and O. A. Levy, O. A, "ATF4 Protects Against Neuronal Death in Cellular Parkinson's Disease Models by Maintaining Levels of Parkin," *Journal of Neuroscience*, vol. 33, issue 6, pp. 2398–2407, Feb. 2013.

[67] J. J. Hansen, A. Durr, I. Cournu-Rebeix, C. Georgopoulos, D. Ang, M. N. Nielsen, C. S. Davoine, A. Brice, B. Fontaine, N. Gregersen, and P. Bross, "Hereditary Spastic Paraplegia SPG13 is Associated with a Mutation in the Gene Encoding the Mitochondrial Chaperonin Hsp60," *The American Journal of Human Genetics*, vol. 70, issue 5, pp, 1328–1332, May 2002.

[68] L. Zhao, X. Chen, S. Zhou, Z. Lin, X. Yu and Y. Huang, "DNA methylation of AHCY may increase the risk of ischemic stroke," *Bosnian Journal of Basic Medical Sciences*, vol. 20, no. 4, pp. 471–476, Nov. 2020.

[69] H. Rakshit, N. Rathi and D. Roy, "Construction and Analysis of the Protein-Protein Interaction Networks Based on Gene Expression Profiles of Parkinson's Disease," PLOS ONE, vol. 9, issue 8, Aug. 2014.

[70] Zarouchlioti, D. A. Parfitt, W. Li, L. M. Gittings and M. E. Cheetham, "DNAJ Proteins in neurodegeneration: essential and protective factors," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 373, no. 1738, Dec. 2017.

[71] S. El Haddad, A. Serrano, F. Moal, T. Normand, C. Robin, S. Charpentier, A. Valery, F. Brulé-Morabito, P. Auzou, L. Mollet, C. Ozsancak and A. Legrand, "Disturbed expression of

autophagy genes in blood of Parkinson's disease patients," *Gene*, vol. 738, no. 144454, May 2020.

[72] E. J. Robinson, S. Aguiar, M. P. Smidt and L. P. van der Heide, "MCL1 as a therapeutic target in Parkinson's Disease?" *Trends in Molecular Medicine*, vol. 25, issue 12, pp. 1056–1065, Nov. 2019.

[73] S. E. Brownell, R. A. Becker and L. Steinman, "The protective and THERAPEUTIC function of small heat shock proteins in neurological diseases," *Frontiers in Immunology*, vol. 3, issue 74, May 2012.

[74] J. J. Zheng, W. X. Li, J. Q. Liu, Y. C. Guo, Q. Wang, G. H. Li, S. X. Dai and J. F. Huang, "Low expression of aging-related NRXN3 is associated with Alzheimer disease: A systematic review and meta-analysis," *Medicine*, vol. 97, issue 28, no. e11343, Jul. 2018.

[75] Y. Xiao, T. H. Hsiao, U. Suresh, H. I. Chen, X. Wu, S. E. Wolf and Y. Chen, "A novel significance score for gene selection and ranking," *Bioinformatics*, vol. 30, issue 6, pp. 801–807, Feb. 2012.

[76] F. Amato, A. Lopez, E. M. Pena-Mendez, P. Vanhara, A. Hampl and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine*, Jan. 2013.

APPENDIX A

APPENDIX A

ADDITIONAL FIGURES

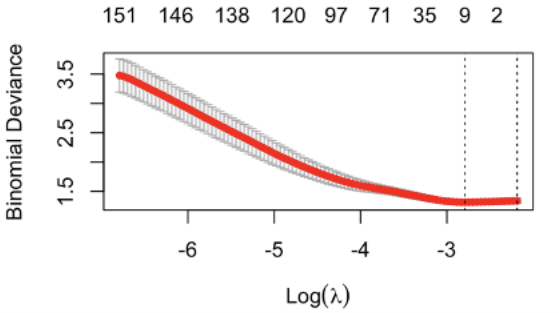


Figure A.1: LASSO lambda for GSE 2034.

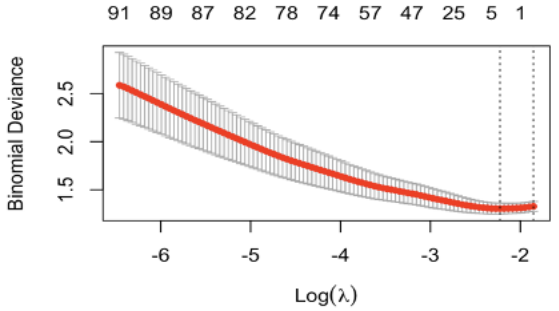


Figure A.2: LASSO lambda for GSE 2990.

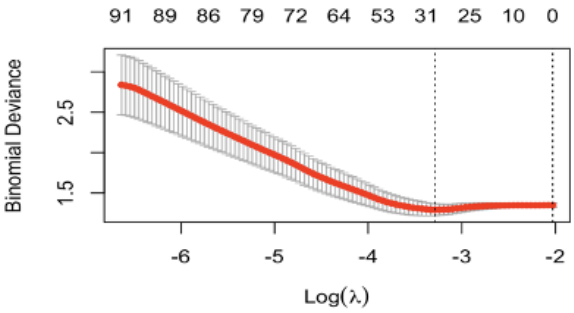


Figure A.3: LASSO lambda for GSE 4115.

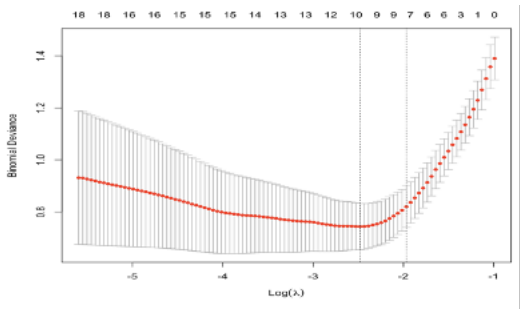


Figure A.4: LASSO lambda for GSE 8397.

APPENDIX B

APPENDIX B

MATLAB COMPUTATIONS

```
%First load the two csv files called "TrainDummy" and "TestDummy" and then define the training and test set
trainx = trainmatlab(:,2:100); %data
traint = trainmatlab(:,1); %target
testx = testmatlab(:,2:100);
testt = testmatlab(:,1);

%for reproducable results, fix the random number generator seeds
rand('state',100);

%Cross Validation
indices = crossvalind('Kfold',traint,3);

trainkfold_x=cell(3,1);
trainkfold_t=cell(3,1);
testkfold_x=cell(3,1);
testkfold_t=cell(3,1);

for i = 1:3
   test = (indices == i); train = ~test;
   trainkfold_x{i} = trainx(train,:);trainkfold_t{i} = traint(train,:);
   testkfold_x{i}=trainx(test,:);testkfold_t{i}=traint(test,:);
end

%create a training error martix
train_error=zeros(10,3);
%creating a testing error matrix
test_error=zeros(10,3);

network = cell(10,3);
network_valid=cell(10,3);

% define network. this specification is for a network with logistic
% (sigmoidal) output neurons; this choice automatically selects the
% cross-entropy error function.
nin=100;
nout=1;
%alpha = 0.05;
alpha = 0.01;
%nhidden = 32;
```

```
    for i=1:10
        nhidden = i*4 ;
        for j = 1 : 3

%net = mlp(nin, nhidden, nout, 'logistic');
%to use a weight decay of size alpha, define the network as
net = mlp(nin, nhidden, nout, 'logistic', alpha);

% options for training: display error values, and set number of iterations
% to 100.
options = zeros(1,18);
options(1) = 1;
options(2)=1e-3;
options(3)=1e-6;
options(14) = 250;

[net]=netopt(net, options,trainkfold_x{j},trainkfold_t{j},'scg');
err=mlperr(net,trainkfold_x{j},trainkfold_t{j});
           train_error(i,j)=err;
      [nettrain]=netopt(net,options,testkfold_x{j},testkfold_t{j},'scg');

  test_error(i,j)=mlperr(nettrain,testkfold_x{j},testkfold_t{j});
  network{i,j}=net;

       network_valid{i,j}=nettrain;
      % [y,z]= mlpfwd(net,tes{j});
      % pred{i,j}=y;
       end;
    end;

z = 1;
a = 1;
%network_valid{i,j}
[testy,z,a] = mlpfwd(net, testx);
hist(testy)
% calculate the 2x2 confusion matrix...
conf = confmat(testy,testt);
% train the network using the scaled conjugate gradient algorithm
```

```
net = netopt(net, options, trainx, traint,'scg');
err=mlperr(net,trainx,traint);

% evaluate performance on the test set
% calculate network outputs for all test data points
[testy,z,a] = mlpfwd(net, testx);
%histogram of the output
hist(testy)
% calculate the 2x2 confusion matrix...
conf = confmat(testy,testt);
% ...and display it
f2 = figure;
set(f2, 'Name', 'Confusion matrix for test set');
plotmat(conf,'b','k',12);

% for comparison, do the same for the training set
trainy = mlpfwd(net, trainx);
% calculate the 2x2 confusion matrix...
conf2 = confmat(trainy,traint);
% ...and display it
f3 = figure;
set(f3, 'Name', 'Confusion matrix for training set');
plotmat(conf2,'b','k',12);

%ROC Curve for All Test Data
LAll = logical(testt);
[XAll,YAll,TAll,AUCAll]= perfcurve(LAll,testy,true);

f3 = figure;
set(f3, 'Name', 'Receiver Operating Characteristic Curve for All Test Data');
AreaAll=plot(XAll,YAll);
fname1 = sprintf('AUC %.4f', AUCAll);
%fname2 = sprintf('ROC for classification by Neural Network with Neural Network with Evidence Procedure ARD for All Test Data%d'
xlabel('False Positive Rate'); ylabel('True Positive Rate');legend(fname1);
%title(fname2)
title('ROC for classification for the All Test Data using the Standard Neural Network')

save fold_alpha01t1.mat
```

BIOGRAPHICAL SKETCH

Myrine A. Barreiro-Arevalo is a South Texas native who has lived in the area for over 24 years. She graduated from the University of Texas Rio Grande Valley with a Bachelor of Science in Biology with a concentration in Biological Sciences and a minor in Statistics with Magna Cum Laude honors before receiving her Master of Science in Applied Statistics and Data Science in August 2021, being the first ever to graduate with the degree from UTRGV. A Presidential Graduate Research Assistantship recipient, she taught undergraduate mathematics courses during her graduate studies and participated in an FDA sponsored internship at the National Center for Toxicology Research which focused on using deep learning techniques on gene sequencing data. Her future goals are to relate biological sciences with her statistical knowledge and pursuit a Doctorate degree in Interdisciplinary Studies with concentration in Biostatistics.

Myrine can be reached via email at maarevalo56@gmail.com.