INTEGRATIVE PREDICTIVE MODELING OF METASTASIS IN MELANOMA CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYŞEGÜL KUTLAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

FEBRUARY 2022

Approval of the thesis:

INTEGRATIVE PREDICTIVE MODELING OF METASTASIS IN MELANOMA CANCER

Submitted by AYŞEGÜL KUTLAY in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Head of Dept., **Health Informatics** _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Supervisor, **Health Informatics Dept., METU** _____

Examining Committee Members:

Prof. Dr. Tolga CAN
Computer Engineering Dept., METU _____

Assoc Prof. Dr. Yeşim AYDIN SON
Health Informatics, METU _____

Asst. Prof. Dr. Aybar Can ACAR
Health Informatics, METU _____

Assoc. Prof. Dr. Özlen KONU
Molecular Biology and Genetics Dept., İ.D. Bilkent University _____

Prof. Dr. Hasan OĞUL
Computer Engineering Dept., Çankaya University _____

Date: **08.02.2022**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name AYŞEGÜL KUTLAY

Signature : _____

ABSTRACT

INTEGRATIVE PREDICTIVE MODELING OF METASTASIS IN MELANOMA CANCER

KUTLAY, Ayşegül

Ph.D., Department of Health Informatics

Supervisor: Assist. Prof. Dr. Yeşim AYDIN SON

February 2022, 121 pages

This study focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation. We developed multiple machine learning models to distinguish the metastasis by integrating miRNA, mRNA, and DNA methylation markers. We used the TCGA melanoma dataset to differentiate metastatic melanoma samples by assessing a set of predictive models. An iterative combination of differentially expressed miRNA, mRNA, and methylation signatures is used as candidate markers to reveal each new biomarker category's impact. In each iteration, the performances of the combined models are calculated. The choice of feature selection method and under and oversampling approaches are analyzed during all comparisons. Selected biomarkers of the highest performing models are further analyzed for the biological interpretation of functional enrichment. MiRNA biomarkers can identify metastatic melanoma with an 81% F-score in the initial model. The addition of mRNA markers upon miRNA increased F-score to 92 %. In the final integrated model, the inclusion of the methylation data resulted in a similar F-score of 92% but produced a stable model with low variance across multiple trials. Our results support the role of miRNA regulation in metastatic melanoma as miRNA markers models metastasis outcomes with high accuracy. Moreover, the integrated evaluation of miRNA with mRNA and Methylation biomarkers increases the model's accuracy. It populates selected biomarkers on the metastasis-associated pathways of melanoma, such as "Osteoclast," "Rap1 Signaling" "and "Chemokine Signaling" Pathways.

Keywords: Machine Learning, Metastatic Molecular Signatures, miRNA, mRNA, DNA Methylation

ÖZ

MELANOM KANSERİNDE METASTAZIN TAHMİNE DAYALI BÜTELEŞTİRİCİ MODELLENMESİ

KUTLAY, Ayşegül

Doktora Sağlık Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Yeşim AYDIN SON

Şubat 2022, 121 sayfa

Bu çalışma, miRNA, mRNA ve DNA metilasyonunun birleştirilmiş etkisini araştırarak melanomun metastatik moleküler imzalarını izlemek için genetik biyobelirteçlerin düzenleyici etkisini belirlemeye amaçlamaktadır. MiRNA, mRNA ve DNA metilasyon belirteçlerini entegre ederek metastazı ayırt etmek için çoklu makine öğrenme modelleri geliştirdik. Bir dizi tahminleme modeli değerlendirerek metastatik melanom örneklerini ayırt etmek için TCGA melanom veri setini kullandık. Her yeni biyobelirteç kategorisinin etkisini ortaya çıkarmak için aday belirteçler olarak ayırıcı şekilde ifade edilen miRNA, mRNA ve metilasyon özniteliklerini yinelemeli bir kombinasyonu uygulanmıştır. Her yinelemede, birleştirilmiş modellerin performansları hesaplandı. Tüm karşılaştırmalar sırasında, öznitelik seçim yönteminin seçimi ve alt ve üst örnekleme yaklaşımları analiz edilmiştir. En yüksek performans gösteren modellerin seçilmiş biyobelirteçleri, fonksiyonel zenginleştirme kümelerinin analizi biyolojik yorumu için ayrıca irdelenmiştir. İlk modelde, miRNA biyobelirteçleri metastatik melanomu %81 F-skoru ile tanımlayabilir. miRNA üzerine mRNA markörlerinin eklenmesi F-skorunu %92'ye yükseltti. Nihai entegre modelde, metilasyon verilerinin eklenmesi, %92'lik benzer bir F-skoruyla ulaşıldı, ancak birden fazla denemede düşük varyanslı daha kararlı bir model üretildi. Sonuçlarımız, miRNA belirteçleri metastaz sonuçlarını yüksek doğrulukla modellediğinden, metastatik melanomda miRNA düzenlemesinin rolünü desteklemektedir. Ayrıca miRNA'nın mRNA ve Metilasyon biyobelirteçleri ile entegre değerlendirmesi, modelin gücünü artırmaktadır. Modelde, seçilmiş olan belirteçler "Osteoclast", "Rap1 Signaling" "ve "Chemokine Signaling" gibi melanomun metastazla ilişkili patikalarda yoğunlaşmaktadır.

Anahtar Sözcükler: Makine Öğrenimi, Metastatik Moleküler İmzalar, miRNA, mRNA, DNA Metilasyonu

To My Family

ACKNOWLEDGMENTS

First of all, I would like to thank **Assoc. Prof Yeşim AYDIN SON**. I am grateful to her for her fantastic support, guidance and sharing her knowledge during my thesis study. I am glad that you accepted me as a student and have continued to believe in me throughout the years.

Very special thanks to thesis progress committee members **Prof. Dr. Tolga CAN** and **Asst. Prof. Dr. Aybar Can ACAR**. Your suggestions and comprehensive feedback have been really helpful to me.

To my friends … It was important to strike a balance with the outside life while going deeper into the work. Therefore, I cannot emphasize the importance of your being by my side and the strength you have given me while all this is going on. **Hicran BÜYÜKGÖZ, Birce GÜNERİ, Duygu KARA** and of course **Nesrin KÜÇÜK**. Thank God I have you all.

Dear **Cookie**, I hope we have many more years to spend together. Your coming into my life has made this process so much more enjoyable. Maybe you cannot understand what is written here but I believe you can feel it.

Finally, my parents… I am thankful to **Nurten KUTLAY, Selim KUTLAY**, and my brother **Baki KUTLAY** for their support during my whole education life. I am grateful for always being there for me and for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1-NN	1-Nearest Neighbor
3-NN	3-Nearest Neighbor
ABABOOST	Adaptive Boosting
ADTree	Alternating Decision Tree
API	Application Programming Interface
CCLE	Cancer Cell Line Encyclopedia
CGC	Cancer Genomics Cloud
CDC	Center of Disease Control and Prevention
CCP	Compound Covariate Predictor
CT	Computerized Tomography
DCNN	Deep Convolutional Neural Network
DNA	Deoxyribose Nucleic Acid
FPKM	Fragments Per Kilobase of Transcript Per Million Mapped Reads
FPKM-UQ	Fragments Per Kilobase of Transcript per Million Mapped Reads Upper Quartile
RGS1	G-protein Signaling 1
HNSCC	Head and Neck Squamous Cell Carcinoma
LDA	Linear Discriminant Analysis
TUG1	Long Non-Coding RNA Taurine-Upregulated Gene 1
mRNA	Messenger RNA
miRNA	Micro RNA
miRNA-Seq	miRNA Sequence
mRNA-Seq	mRNA Sequence
NC	Nearest Centroid
PCA	Principle Component Analysis
RNA	Ribonucleic Acid
SR-BI	Scavenger Receptor Class B type 1
SKCM	Skin Melanoma
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over-sampling Technique
TCGA	The Cancer Genome Atlas

CHAPTER 1

1. INTRODUCTION

The discovery of the gene regulation mechanism of the microRNAs (miRNA) is one of the critical signs of progress in cancer biology. MicroRNAs, are untranslated sequences, transcript from DNA but do not code into any protein products. So far, about 1400 miRNA (Jansson and Lund 2012) have been discovered in the human genome. The most remarkable function of miRNAs is their capability to suppress almost one-third of human genes. Since the initial discovery of the relationship between cancer and miRNA signatures, many studies have shown that miRNA has a critical role in regulating genes and, thus, is crucial in tumorigenesis.

Today, many techniques for the early detection and diagnosis of tumors are available. Still, when invasive procedures are required for diagnosis or treatment, it is essential to know the metastatic potential of the tumor to estimate the risks benefits of the procedure. Also, in the later stages of tumor development, any information about the metastatic status of the late-stage tumors is required for deciding between therapy choices. So having a tool to predict the metastatic potential may help to decide on a better therapeutic pathway.

In this study, we focused on the “identification and classification of the key miRNA expression profiles for predicting tumor metastasis”. We assessed the predictive potential of miRNA by comparing the findings with the combination of various biomarkers.

So, predictive models based on expression patterns of miRNA, mRNA, and DNA methylation markers were developed to monitor the presence and progression of the metastatic changes in tumors. For this purpose, analysis of miRNA, mRNA expression, and methylation beta values and their impact on metastatic outcomes is conducted. Various technique and machine learning models were applied, and their performance was compared.

As a result, this predictive model, which is based on genetic profiles as biomarkers, is utilized to generate knowledge for metastasis and to develop a personalized medicine approaches in cancer treatment.

1.1. Background of Study

The microRNA profiling in cancer has been used for 12 years (Di Leva and Croce 2013). After discovering the connection between cancer signature and micro-RNA profiling, many studies have been conducted. However, studies until now are mainly focused on the correlation or significance between genes and malignancy of cells. Interrogation metastatic progression of cancer is the main novelty of our proposal.

Two studies with a more generalized scope are pointed out when the literature is reviewed. The first is the study conducted to classify the common cancer types (Volinia et al. 2006). The other study, called "systems genetic analysis to metastatic interrogation progression of cancer," was performed by Faraji and his coworkers (Faraji et al. 2014). Volinia and colloquies (Volinia et al. 2006) did not focus on metastatic cancers. In contrast, Faraji and his coworkers conducted the study by focusing on metastatic breast cancers by working on mice miRNAs. Even though metastasis is the systemic outcome of most cancer diseases that are the leading cause of cancer-related mortality, there is not enough study on the interrogation of the metastatic progression of different cancer types.

This study is distinguished from the other studies in the literature on three main points. First, this study will provide a different perspective for cancer studies by proving a generalized model for metastatic cancer signatures on humans by including various genetic biomarkers miRNAs, mRNA, and DNA methylation. Second, the studies focus on diagnostic purposes, but cancer treatments also need predictive models (tools). This study aims to provide a probabilistic model that could be used as a guide for monitoring of presence and progression of the metastatic changes in tumors. Finally, this study is dissociated from other studies on the literature that the causality of the metastatic signatures will be revealed. A prototype for a predictive model based on the signatures revealed is proposed.

1.2. Aim of Study

The main goals of this study are :

1- To identify the miRNA signatures involved in the regulation of metastatic disease.

2- Comparison of different genetic biomarkers and investigate their contribution to the metastatic outcome.

3- Reveal out the impact of miRNAs among other markers (mRNAs and DNA methylation) and causality relationship for metastatic disease.

4- Use founded relationships of regulators to generate a predictive model that manages metastatic tumor progress within cancers.

The predictive model created during the study is the study's primary outcome.

1.3. Outline of Study

This thesis study is organized into five chapters which are described as follows.

Chapter 1: MicroRNAs' contribution to a cancer diagnosis is discussed as background information. Then the novelty and purposes of the study are listed.

Chapter 2: This chapter contains the study's literature review, which is introduced in three sub-sections. The first is the review of melanoma and metastasis. The second is the list of the potential biomarkers of melanoma metastasis. Finally, previous studies on metastasis are discussed.

Chapter 3: In this chapter methodology of the study is presented. In the initial part of the chapter, the TCGA data set for skin melanoma is given, and attributes are examined. Then the overall method for preprocessing, model training, tuning, and test are discussed in detail. The iterative approach that the study is formed is described in this section.

Chapter 4: Findings of the methods are provided in detail. Comprehensive results for each iteration are presented in this chapter.

Chapter 5: In this final chapter, the discussion and conclusion of the study are explained. Limitations and future works are also introduced.

CHAPTER 2

2. LITERATURE REVIEW

In this chapter, melanoma, metastasis, and biomarkers of metastasis are presented. Also, previous studies on metastasis prediction are discussed.

2.1. Melanoma And Metastasis

Melanoma, cancer with a rapid increase in incidence and high mortality, is a malignant tumor of skin pigmentation cells with a high mortality disease. Melanoma can develop anywhere on the body, but the most observed in areas exposed to the sun, such as the back, legs, arms, and face. With nearly 300,000 cases, melanoma is one of the most common cancer types worldwide (World Cancer Research Fund n.d.).

According to CDC statistics, yearly 85.000 new cases are reported in the USA, where 8.000 people die annually (United States Cancer Statistics n.d.). On the other hand, melanoma cancer incidence reaches 140.00 annual cases in the European Union. It is considered one of the fastest-rising types of cancer, albeit with hotspots in Europe being the Scandinavian countries, Switzerland, and Austria (American Cancer Society 2016). In addition, 16.000 new melanoma cases are reported in the UK, which corresponds to 4% of all cancer types, and it has a rising incidence rate of 135% over thirty years (Cancer Research UK n.d.).

Both distant and regional metastasis is possible in melanomas. The most common metastases sites in melanoma cases are bone, brain, liver, lung, and skin. The presence of skin metastasis may be the first outward sign of lymphatic or hematogenous spreading. So, in melanoma, the prognosis is a critical concern rather than diagnosis. Detecting at least suspicious cases via visual examination or short screening is possible. Early diagnosis leads to high cure rates, but there is still no effective treatment in later stages, where metastasis is observed frequently (Damsky et al., 2011).

In normal tissues, the balance between cell growth and death is essential. This balance can be disrupted as a result of either "uncontrolled cell growth" or "loss of apoptosis ability (Programmed cell death)" (Ma and Weinberg 2008; Oppenheimer 2006), which leads to tumorigenesis. In general, tumorigenesis (abnormal cell) can be malignant or benign. While benign tumors do not spread, a malignant tumor spreads to the other tissues.

Before a malignant tumor develops, the initial conversion of a normal cell into a primary tumor cell occurs (Oppenheimer 2006). This primary tumor may stay stable in this originated tissue (benign) or spread to the other parts of the body (malignant) by invasion or metastasis (Oppenheimer 1983, 2006; Willis and Pp 1953). Invasion, tumor expansion, can be defined as the direct migration of cancer cells into neighboring tissues. On the other hand, metastasis is the spread of tumor cells to areas not directly neighboring the primary tumor (Oppenheimer 2006). In metastasis, five main stages are observed (Leong et al. 2006; Oppenheimer 2006; Willis and Pp 1953):

- Cells from the primary tumor are detached
- Tumor cells, penetration (invasion) of these cells migrate into lymph vessels or blood vessels and disseminate the cells or cell clusters to distant areas.
- Tumor cells lodge in blood vessels of distant organs.
- Invasion of tumor cells through the vessel walls and into the tissue of secondary sites takes place.
- The secondary tumors grow at the secondary sites.

Besides the cellular basis described above, carcinogenesis also has a molecular foundation (Shalaby et al. 2014; Shen, Stass, and Jiang 2013; Tonini, Rossi, and Claudio 2003). It is caused by alterations or mutations in the genetic code (loss of DNA, gain of DNA, changes in nucleotides, or epigenetic effects). Such mutations alter crucial cancer-related pathways. Both "research on abnormalities of cancer-related genes occurring in preneoplastic and neoplastic lesions" and "recent research such as defining signal transduction pathways in cells cycle and the genetic control of the cell cycle" helps to reveal the molecular basis of carcinogenesis. Understanding the molecular basis of carcinogenesis has important implications in the prevention, diagnosis, and treatment of cancer and its metastasis (Harris 1991)

2.2. Signatures of Metastasis

Understanding the molecular basis of carcinogenesis is essential in preventing, diagnosing, and treating cancer and its metastasis (Harris 1991).

Many different markers have been proposed to describe the molecular foundation of metastasis. DNA methylation, gene expression profiles, and microRNAs are frequent biomarkers for predicting metastasis for most cancer types. The initial studies on metastasis biomarkers and predictive models were published in 2004. These initial studies were performed using gene expression profiles of the primary tumor collected by DNA microarray. Until now, many researchers studied the same topic with different datasets. Meanwhile, collective studies performed by three or more previous datasets were also published. After 2015 studies on miRNA expression levels and methylation data occur in the literature.

Although predictive machine learning models for melanoma metastasis are limited, many studies propose predictive biomarkers for different metastatic cancers. While most studies target specific markers, such as microRNA or protein expression, recent studies (De Souza et al., 2017) investigate the integrated usage of miRNA and mRNA signatures. For example, binary logistic regression, which uses mir- 331 and miR-195 as markers, can distinguish metastasis and local breast cancer (Sensitivity = 0.95, Specificity= 0.76) (McAnena et al. 2019). A study conducted by Souza et al. (De Souza et al. 2017) developed an integrated model using expression levels of 27 miRNA and 81 targets mRNA to classify prostate cancer patients from controls with 67% sensitivity and 75% specificity.

The following sections will provide details of previous studies in literature, which utilize different data types as a biomarker.

2.2.1. Micro RNA

MicroRNAs are non-coding RNAs (transcripts from DNA but do not code any protein products). The microRNAs regulate these cancer-related pathways, such as proliferation, cell cycle control, apoptosis, differentiation, migration, and metabolism (Chowdhury et al., 2012; Jansson and Lund 2012; Stahlhut and Slack 2013). So, it is not surprising that these molecules take a role in carcinogenesis. MicroRNAs have a key role as suppressors or promoters of carcinogenesis or metastasis by controlling their target mRNA, which causes the pathogenic activity of cells (Shalaby et al., 2014). For these reasons, in time, microRNAs became the main focus in cancer biology and were proven as crucial

components of normal and pathologic states of cells (Hayes, Peruzzi, and Lawler 2014; Stahlhut and Slack 2013). Their major role is the regulation of genes (He, Xu, and Goldkorn 2011). They are also regulatory actors on carcinogenesis. It is proven that in about 68% of chronic lymphocytic leukemia cases, microRNA genes miR15 and miR16 are either deleted or down-regulated (Calin et al. 2002) (Calin et a.,2002). After this initial finding, many microRNAs are shown as dysregulated or upregulated in different malignant cases (Calin et al., 2002; Chowdhury et al., 2012; Hayes, Peruzzi, and Lawler, 2014; Di Leva and Croce 2013; Lim et al. 2015). So, It is a fact that microRNAs contribute to several different aspects of carcinogenesis (Shalaby et al., 2014).

Shalaby and colloquies conducted one of the initial metastatic early prediction studies based on microRNA expression levels (Shalaby et al., 2014). In the study, the expression levels of the studied miRNAs are analyzed (with Mann-Whitney U test and Kaplan-Meier plots approach) for metastatic characteristics of primary tumor of the renal cell. As a result of the study, miR-155, miR-210, miR-106a, miR-106b, miR-200, and miR-141 are found as differentially expressed over metastatic and non-metastatic tumor cells.

Zhou and colloquies (Zhou et al. 2014) search for the potential of using miR-105 as a prognostic marker for metastases. They have conducted research on miR-105 in their mouse models and observed high circulating miR-105 at premetastatic and metastases stages. Then they used patient data to analyze serum from patients with stage II and III breast cancer.

Zhang and colloquies (L. Zhang et al. 2015) proposed a microRNA-based prediction model which predicts risk and hazard ratio for metastasis primary hepatocellular cancer. For this purpose, they have used five statistically independent factors (vascular invasion, Barcelona Clinic Liver Cancer stage, miR-145, miR-31, and miR-92a). The model sensitivity and specificity were 69.6 and 80.2 %, respectively.

Goossens-Beumer and colloquies (Goossens-Beumer et al. 2015) also proposed a microRNA-based classifier for prediction metastasis in colon cancer. In this study, the combination of miR25-3p and miR339-5p expression levels in tumor cells was founded as an independent prognostic factor for the occurrence of distant metastasis in TNM stage II–III colon cancer with a stable microsatellite phenotype. According to another study (Wu et al., 2015), CD44, MMP7, and β-catenin expression were positively correlated, AI1/CD82 expression showed a negative correlation with distant metastasis colorectal cancer.

In the other study, Wang et al. (R. Wang, Chen, and Shu 2015) developed a model that predicts distance metastatic of primary lung cancer by using Micro RNA expressions on a nude mouse. As a result, 17 microRNAs are founded as up-regulated, and seven are founded as a down-regulated expression between the non-small cell lung cancer metastatic and the non-metastatic cancers.

2.2.2. mRNA And Gene Expression Profiles

Most of the studies in the literature used gene expression profiles of primary tumors obtained by microarray as predictive markers. The distributions of these studies are given below.

In terms of predictive biomarkers of metastasis, the study conducted by Kan et al. (Kan et al. 2004) is one of the initial researches. They have developed a predictive model for "Lymph Node Metastasis" using artificial neural networks. The primary site of the metastasis was the "esophagus." According to their result, the model predicts the metastasis with %77 accuracy by using "gene expression profiles of primary tumor obtain by DNA microarray."

Another study is conducted on "Lung Adenocarcinomas" cancer (Xi et al. 2005) with metastasis of lymph nodes. The prediction model is constructed upon "gene expression profiles of primary tumor obtain by microarray." Analysis of gene expression profiles from primary tumors may predict lymph nodes well but frequently misclassifies negative patients as positive. Classification accuracy is again 94.1% in the metastasis-positive cases but only 21.2% in the metastasis-negative cases.

Moriya et al. (Moriya et al. 2009) also introduce another study on primary lung tumors with Lymph Node metastasis using gene expression profiles. Their prediction model has yielded 71.4% accuracy for forecasting lymph node metastasis with independent test cases.

Besides, SVM (support vector machine) classifier, which uses gene expression profiling with microarray, predicts metastasis with 78% accuracy for breast cancer (Burton et al., 2012).

Bidus et al. proposed to use gene expression profiling of the primary tumors in patients with endometrioid endometrial cancers seems promising for identifying genes associated

with lymph node metastasis (Bidus et al. 2006). As a result of the study, TOB2, CDC2, MAD2L, ZIC2 probes on the microarray are found as differentially expressed between patients with and without lymph node metastasis.

Wang and colloquies proposed a model to predict distance metastasis on breast cancer (Y Wang et al. 2005). Their model predicts distance metastasis with 93% sensitivity and 48% specificity. In another study on Gene expression profiling with microarray, a classifier SVM (support vector machine) classifier is modeled to predict the primary tumor metastasis on Breast Cancer (Thomassen et al. 2007). Through SVM modeling, 24 metastases and 23 non-metastases were classified correctly using 60 samples with 78 % accuracy (Thomassen, 2007). In a later study, metastatic characteristics of breast cancer are modeled by using pathological and histological findings of lymph node biopsy by using ADTree as a prediction model (Takada et al., 2012). The multiple survival screening algorithms predict metastasis with an accuracy of %77 (Li et al., 2010). The predictive accuracy was %87 only for the low-risk group patients.

Dehnavi and colloquies (Dehnavi et al. 2013) also proposed a hybrid model to predict metastasis in breast cancer by using a combination of six data set which includes Lin et al. (Li et al. 2010), Dataset and Wang et al. (Y Wang et al. 2005) data sets. First of all, they generated a rough-set theory-based gene selection method. Afterward, this method was applied to six available data sets on breast cancer to select the most informative genes. This selected gene set is evaluated for prognostic signatures of breast cancer. From the combined gene pool, 18 genes were selected for meta-signature. Their model reached a 71% accuracy level for all risk groups. Radwan and colloquies (Radwan et al. 2013), on the other hand, proposed to use the blood mammaglobin expression level as a marker for the diagnosis and prediction of breast cancer. During the 34 months of follow-up, five mammaglobin-positive patients showed metastatic lesions, and none of the mammaglobin-negative patients developed metastasis.

Computerized tomography(CT) and mRNA expression profiling were combined via statistical analysis (Chang et al., 2008) to predict the lymph node metastasis of primary lung cancer tumors. This method increases accuracy from 55%(CT) to 86% (CT and mRNA).

A statistical model (by using ANOVA and hierarchical Clustering) is proposed by Rickman and colloquies (Rickman et al. 2008) to predict head and neck squamous cell carcinoma (HNSCC) metastasis. Using the expressing mRNA levels (with microarray), the model predicts future metastasis with an accuracy of %77.

In the study of "cancer metastasis networks," done on a large set of patient data, the prediction of progression patterns is generated as a system network for primary tumors and the sides of metastasis (Chen et al., 2009). By using these networks (which are constructed by hierarchical clustering), they have tried to predict the primary site of tumor after a sequence of metastasis multinomial logistic regression with an overall accuracy of 51-(Prostate 84%, colon 80%, lung and bronchus 69%, ovary 64%, larynx 61%, and female breast 56%).

Roessler (Roessler et al. 2010) has generated a risk classifier tool to predict the outcome of hepatocellular tumors by using gene expression levels combined with serum AFP levels or BCLC staging. In this study, six prediction algorithms (Support Vector Machines (SVM), Nearest Centroid (NC), 3-Nearest Neighbor (3-NN), 1-Nearest Neighbor (1-NN), Linear Discriminant Analysis (LDA), or Compound Covariate Predictor (CCP) were used as a prediction model. Among all, CCP had a sensitivity of 76 % and a specificity of 60.3 % on cases from the "Liver Cancer Institute" case. They also tested the model on another case set from "Laboratory of Experimental Carcinogenesis." The model predicts the risk with a sensitivity of 83.9 % - specificity of 64.9 %.

Watanabe and colloquies (Watanabe et al. 2010) proposed a model to predict liver metastasis with primary colorectal tumor by using Gene-expression profiles of samples of DNA microarray with k-nearest- neighbor method (KNN) and 10-fold cross-validation. The model predicts metastasis with 86,2 % Accuracy.

Zemmour and friends (Zemmour et al. 2015) developed three different models (Elastic net, LASSO, and CoxBoost) to predict Early Breast Cancer Metastasis by using DNA micro Array Data. In the study, they have used a publicly available dataset. Then they validate the results on two other datasets. They predict metastasis with 66 % accuracy on one of the datasets and 59% accuracy on the other.

Several prognostic gene expression signatures have been proposed as significant for colorectal cancer. Ramaswamy et al. (Ramaswamy et al. 2003) studied patients with lung, breast, prostate, colorectal, uterus, ovary" on expression levels of oligonucleotide microarray and found five genes as significant for prognostic purposes. In addition, Wang (2004) (Yixin Wang et al. 2004), Barrier (2005) (Barrier et al. 2005, Yamasaki (2007) (Yamasaki et al. 2007) is also studied patients with Stage I-III colorectal Cancer and found 23 genes, 47 genes, 119 genes as significant respectively. Yoshida (Yoshida et al. 2010), Cavalieri (Cavalieri et al. 2007), and lin (Lin et al. 2007) also studied patients with all Stages(I-II-III-IV) of colorectal cancer. Nevertheless, they have proposed different genes with prognostic significance. Moreover, overlap in the gene expression signatures is little.

The low consistency between the different studies may be in part attributed to methodological and technical variances.

2.2.3. DNA Methylation

DNA methylation is a chemical process that adds methyl groups to DNA. This process modifies the functionality of the DNA itself. DNA methylation is an important regulator and plays a crucial role in normal development. It is essential for genomic imprinting, X-chromosome inactivation, repression of repetitive elements, and aging. Besides all these, DNA methylation is also found as associated with many types of cancer (F. F. Zhang et al., 2011). Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms (Craig et al., 2011). Typically, there is hypermethylation of tumor suppressor genes and hypomethylation of oncogenes (Gonzalo 2010).

Melchers and colloquies (Melchers et al. 2015), used DNA methylation as marker for metastasis. According to the result of their study, five out of 28 methylation markers (OCLN, CDKN2A, MGMT, MLH1, and DAPK1) were frequently differentially methylated in patients with oral and oropharyngeal squamous cell carcinoma.

2.2.4. Other Markers

Several other studies propose different markers besides mRNA and miRNA expression profiles and DNA methylation biomarkers.

Yang and colloquies (Yang et al. 2016) developed a predictive statistical model for Lymph Node Metastasis in Endometrial Cancer using Serum CA125 combined with immunohistochemical markers PR and Ki67. Their model predicts the lymph node metastasis sensitivity and specificity of the model were 84.6% and 67.4%, respectively. Son et al. (Son et al. 2015) also studied on prediction of primary Endometrial cancer. In the study, they proposed using “serum CA-125” as a biomarker for early prediction of metastasis.

In one of the recent studies, Schell et al. (2016) (Schell et al. 2016) developed a prognostic signature score with a propensity to detect non-EMT(epithelial-to-mesenchy- mal transition) features. The study has proposed a new composite gene expression signature as prognostic score (DPC1.EMT).

Lim & Chung (2014) (Lim et al. 2015) proposed serum ENA78/CXCL5, SDF-1/CXCL12, and their combinations as biomarkers to predict the presence and distant metastasis of primary gastric cancer. Combination of serum ENA78/CXCL5, SDF-1/CXCL12, and CEA achieved 92.8% specificity at 75.0% sensitivity to predict distant metastasis of gastric cancer.

2.3. Predictive Models for Metastasis of Melanoma

Unlike other cancers, there are limited studies on modeling melanoma metastasis. Recently serum levels of the cytokines IL-4, GM-CSF, DCD, and the Breslow thickness were proposed as a marker to predict melanoma metastasis, where a linear regression achieved the best balance accuracy (83%) in the test set (Mancuso et al. 2020). A deep convolutional neural network (DCNN) study to predict BAP1 mutation also identified decisive prognostic factors for predicting metastatic risk via whole slide images with an area under curve 0.90 (H. Zhang et al. 2020). Additionally, Mir-205-5p is found as a significant biomarker for metastatic melanoma by Valentine (Valentini et al. 2019). Also, Wei et al. (Wei et al. 2019) indicate TRIM44 -tripartite motif-containing protein-44, regulated by miR-26b-5p, is identified as amplified on melanoma tissues. The same study reports miR-26-5p as downregulated on melanoma. The study conducted by Kinslechner et al. (Kinslechner et al. 2019) shows that the scavenger receptor class B type 1 (SR-BI) protein expression contributes to metastatic melanoma. Wang et al. (Yanqian Wang et al. 2019) proposed long non-coding RNA TUG1 as a prognostic biomarker of metastatic melanoma. Besides, they have also indicated miR-29c-3p, which is the target for G-protein signaling 1 (RGS1), suppresses the expression of Long non-coding RNA taurine-upregulated gene 1 (TUG1).

2.4. Summary

Overall, transcriptional regulation is one of the critical mechanisms underlying cancer development. Even though mRNA, microRNA, and DNA methylation mechanisms critically impact metastatic outcomes, there are no comprehensive data mining models that combine all aspects of transcriptional regulation for metastasis prediction. This study focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation. We used differentially expressed miRNA, mRNA, and methylation signatures on the TCGA melanoma dataset to distinguish metastatic melanoma samples by assessing a set of predictive models. The highest

performing model is selected, and its biomarkers are further analyzed for the biological interpretation of functional enrichment and to determine regulatory networks.

CHAPTER 3

3. METHODOLOGY

3.1. Dataset Collection

In the study, opened data for Skin Melanoma (SKCM) (The Cancer Genome Atlas Network 2015) of TCGA(The Cancer Genome Atlas) database is used, which is a part of the TCGA dataset served on the Cancer Genomics Cloud (CGC). The Cancer Genomics Cloud (CGC) (Institute 2020) hosts a large genomic dataset and provides tools for searching and analyzing genomic data, serving as a computational environment on the cloud. The data browser tool provided by CGC is used to search on TCGA cases and Cancer Cell Line Encyclopedia (CCLE) Cell Lines. On TCGA, melanoma data set with 470 cases composed of 352 Metastatic and 97 primary tumor samples used during this study, with three experimental strategies in the data set, namely miRNA Expression, mRNA Expression, and methylation.

3.1.1. TCGA Skin Melanoma Data (SKCM)

We have collected the melanoma data for miRNA sequencing, RNA sequencing, and methylation array. For 470 different cases with primary and metastatic melanoma, tissue samples are compared to distinguish the metastatic melanoma from the primary tumor. We finalized the predictive model input preprocessing by applying data cleaning, normalization, and scaling preprocessing steps for the remaining 449 cases (Figure 1). 470 distinct cases and 11.265 opened files have been found by using three filters:

1. Primary Site (Skin)
2. Project (TCGA-SKCM)
3. Experimental Strategy (miRNA-Seq; Methylation array; RNA-Seq)
4. File Access (Open)

We generated a subset of cases, which contains all data for “miRNA sequences,” “Methylation array,” and “RNA sequences.” In the current interface of GDC Data Portal, the following search query provides the data files in the repository:

cases.primary_site in ["skin"] and cases.project.program.name in ["TCGA"] and cases.project.project_id in ["TCGA-SKCM"] and files.access in ["open"] and files.experimental_strategy in ["Methylation Array","RNA-Seq","miRNA-Seq"]

TCGA provides various attributes for "miRNA sequences," "Methylation array," and "RNA sequences." For miRNA, we used "miRNA Expression Quantification," which is miRNA expressions provided as a table that associates miRNA IDs with reading count and a normalized count in reads-per-million-miRNA-mapped. Raw Read Counts, the number of reads aligned to each gene, calculated by the HT-Seq algorithm, is used for mRNA. Ensemble Gene Id represents this data and the number of reads aligned mRNA. For methylation analysis, TCGA provides Beta-values, which approximates the percentage of methylation of the gene (Figure 1).

Table 1: TCGA provides separate files for each data type.

SUPPLEMENTARY DATA	1. Clinical 2. Biospecimen
MIRNA	3. Isoform Expression Quantification 4. miRNA Expression Quantification
MRNA	5. Gene Expression Quantification (HT-SEQ) 6. Gene Expression Quantification (FPKM) 7. Gene Expression Quantification (FPMK-UQ)
GENOTYPING	8. Copy Number Segment 9. Masked Copy Number Segment 10. Gene Expression Quantification 11. Masked Somatic Mutation
METHYLATION	12. Methylation Beta Value

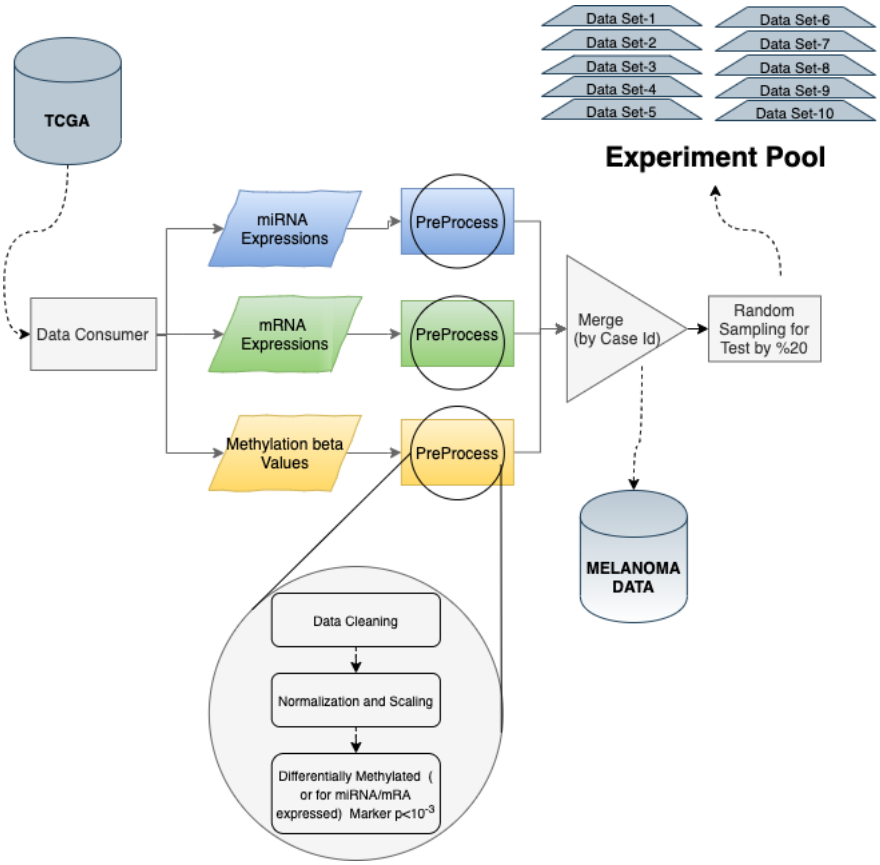

Figure 1: Experimental Pool Generation

Each method is evaluated using a sample experimental pool under the same circumstances. miRNA, mRNA, Methylation Data consumed through TCGA were processed separately and merged to generate the whole melanoma marker dataset. Then, through random splinting, ten individual sample datasets are constructed. Each random split is saved by applying both under-sampling and oversampling (SMOTE) techniques.

TGCA provides two supplementary data, namely clinical data and biospecimen data. **Clinical data** provides clinical values such as gender, race, ethnicity, year of birth, year of death, diagnosis and treatment, family history,

Biospecimen provides detailed data on the samples. On TCGA, each sample is represented by a barcode number. A TCGA barcode is composed of a collection of identifiers. The following image provides those identifiers.

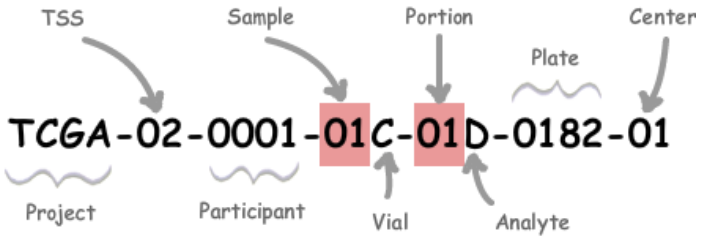

Figure 2: TCGA Sample Barcode

Biospecimen files are used to identify tumor types. Tumor types range from 01 - 09, normal types from 10 - 19, and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes. (Pls see codes table (Sample Type Codes | NCI Genomic Data Commons n.d.) *https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes)*

3.1.1.1.MiRNA Expressions

Each expression value is provided with more than one type of data. There were two types of data "miRNA Expression Quantification" and "Isoform Expression Quantification."

miRNA Expression Quantification: miRNA expressions are provided as a table that associates miRNA IDs with reading count and a normalized count in reads-per-million-miRNA-mapped (Bioinformatics Pipeline: miRNA Analysis - GDC Docs n.d.).

miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-mapped
hsa-let-7a-1	41777	13051.800750	N
hsa-let-7a-2	41593	12994.316217	N
hsa-let-7a-3	41904	13091.477574	N
hsa-let-7b	28180	8803.881206	N
hsa-let-7c	1609	502.677248	N
hsa-let-7d	1771	553.288631	N
hsa-let-7e	17535	5478.213518	N
hsa-let-7f-1	13039	4073.591449	N
hsa-let-7f-2	13289	4151.695434	N
hsa-let-7g	2137	667.632865	N
hsa-let-7i	1086	339.283711	N
hsa-mir-100	50195	15681.718138	N
hsa-mir-101-1	4498	1405.246901	N
hsa-mir-101-2	4482	1400.248246	N
hsa-mir-103a-1	70128	21909.105081	Y
hsa-mir-103a-2	70328	21971.588270	Y
hsa-mir-103b-1	0	0.000000	N
hsa-mir-103b-2	0	0.000000	N
hsa-mir-105-1	499	155.895554	N
hsa-mir-105-2	450	140.587173	N
hsa-mir-106a	41	12.809054	Y
hsa-mir-106b	4219	1318.082853	N
hsa-mir-107	284	88.726127	Y
hsa-mir-10a	50441	15758.572459	N
hsa-mir-10b	256207	80043.150890	N
hsa-mir-1-1	5	1.562080	N
hsa-mir-1178	0	0.000000	N
hsa-mir-1179	3	0.937248	N
hsa-mir-1180	107	33.428506	N
hsa-mir-1181	0	0.000000	N
hsa-mir-1182	0	0.000000	N
hsa-mir-1183	0	0.000000	N

Figure 3: Sample Data File For miRNA Expression Quantifications

Isoform Expression Quantification: this data contains a table with the same information as the miRNA Expression Quantification files with the addition of isoform information such as the coordinates of the isoform and the type of region it constitutes within the full miRNA transcript (Bioinformatics Pipeline: miRNA Analysis - GDC Docs n.d.)

miRNA_ID	isoform_coords	read_count	reads_per_million_miRNA_mapped	cross-mapped	miRNA_region
hsa-let-7a-2	hg38:chr11:122146522-122146545:-	3	0.937248	N	mature,MIMAT0010195
hsa-let-7a-2	hg38:chr11:122146523-122146545:-	49	15.308381	N	mature,MIMAT0010195
hsa-let-7a-2	hg38:chr11:122146524-122146545:-	11	3.436575	N	mature,MIMAT0010195
hsa-let-7a-2	hg38:chr11:122146524-122146548:-	1	0.312416	N	mature,MIMAT0010195
hsa-let-7a-2	hg38:chr11:122146528-122146545:-	1	0.312416	N	mature,MIMAT0010195
hsa-let-7a-2	hg38:chr11:122146566-122146590:-	10	3.124159	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146567-122146589:-	1	0.312416	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146567-122146590:-	749	233.999539	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146568-122146586:-	3	0.937248	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146568-122146587:-	10	3.124159	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146568-122146589:-	4	1.249664	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146568-122146590:-	29138	9103.175677	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146568-122146591:-	10	3.124159	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146569-122146586:-	1	0.312416	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146569-122146587:-	1	0.312416	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146569-122146589:-	1	0.312416	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146569-122146590:-	8207	2563.997624	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146569-122146591:-	8	2.499328	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146570-122146590:-	3268	1020.975294	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146570-122146591:-	5	1.562080	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146571-122146590:-	102	31.866426	N	mature,MIMAT0000062
hsa-let-7a-2	hg38:chr11:122146590-122146610:-	2	0.624832	N	precursor
hsa-let-7a-2	hg38:chr11:122146590-122146611:-	1	0.312416	N	precursor
hsa-let-7a-2	hg38:chr11:122146590-122146612:-	7	2.186912	N	precursor

Figure 4:Sample Data File Isoform Quantifications

3.1.1.2.MRNA Expressions

For mRNA, there were three types of data, namely, “Raw Read Count,” “FPKM,” and “FPKM-OU.”

Raw Read Counts: The number of reads aligned to each gene, calculated by the HT-Seq algorithm. Ensembl Gene Id represents data and the number of reads aligned (Bioinformatics Pipeline: mRNA Analysis - GDC Docs n.d.).

Fragments Per Kilobase of transcript per Million mapped reads (FPKM): FPKM, which is an expression level normalization method (Bioinformatics Pipeline: mRNA Analysis - GDC Docs n.d.)., is formulated as follows :

$$\mathbf{FPKM = [RM_g * 10^9] / [RM_t * L]}$$

- **RMg**: The number of reads mapped to the gene
- **RMt:** The total number of readings mapped to protein-coding sequences in the alignment
- **L**: The length of the gene in base pairs

Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-UQ): This data type contains a modified version of the FPKM formula, where the 75th percentile read count is used (Bioinformatics Pipeline: mRNA Analysis - GDC Docs n.d.). The formulation FPKM-UQ value is as follows:

FPKM = [RMg * 109] / [RM75 * L] where

- ***RMg:*** The number of reads mapped to the gene
- ***RM75:*** The number of reads mapped to the 75th percentile gene in the alignment.
- ***L:*** The length of the gene in base pairs

3.1.1.3.DNA Methylation

TGCA methylation data is provided as beta values, representing the ratio between the methylated array intensity and total array intensity. Beta values take 0 for the lower level of methylation and 1 for the highest methylation level. Methylation data is represented by the association of array props with CpG sites in the dataset. Apart from the CpG site, Gene symbol, gene type, Transcript Id CGI coordinates are also available in the data file(Bioinformatics Pipeline: Methylation Liftover Pipeline - GDC Docs n.d.) (Figure 5)

Composite E	Beta_value	Chromosome	Start	End	Gene_Symbol	Gene_Type	Transcript_ID	Position_to_	CGI_Coordin	Feature_Type
cg00000029	0.57591297	chr16	53434200	53434201	RBL2;RBL2;RBL2	protein_codi	ENST00000262133.9	-221;-1420;2	CGI:chr16:53	N_Shore
cg00000108	NA	chr3	37417715	37417716	C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C	lincRNA;lincF	ENST00000328376.8	18552;18552	CGI:chr3:374	.
cg00000109	NA	chr3	172198247	172198248	FNDC3B;FNDC3B;FNDC3B;FNDC3B;FNDC3B	protein_codi	ENST00000336824.7	157692;1586	CGI:chr3:172	.
cg00000165	0.8558218	chr1	90729117	90729118	.	.	.	.	CGI:chr1:907	S_Shore
cg00000236	0.9068647	chr8	42405776	42405777	VDAC3	protein_codi	ENST00000022615.7	13872	CGI:chr8:424	.
cg00000289	0.68336249	chr14	68874422	68874423	ACTN1;ACTN1;ACTN1;ACTN1	protein_codi	ENST00000193403.9	105019;1048	CGI:chr14:68	N_Shore
cg00000292	0.82333157	chr16	28878779	28878780	ATP2A1;ATP2A1;ATP2A1;ATP2A1;ATP2A1	protein_codi	ENST00000357084.6	373;290;-127	CGI:chr16:28	N_Shore
cg00000321	0.75368807	chr8	41310283	41310284	SFRP1	protein_codi	ENST00000220772.6	-785	CGI:chr8:413	S_Shore
cg00000363	0.75984787	chr1	230425047	230425048	PGBD5	protein_codi	ENST00000391860.4	683	CGI:chr1:230	N_Shore
cg00000622	0.01559547	chr15	22838620	22838621	NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIF	protein_codi	ENST00000337451.6	-41;-39;-68;-	CGI:chr15:22	Island
cg00000658	0.85205338	chr9	137103472	137103473	MAN1B1;MAN1B1;MAN1B1;MAN1B1;MAN	protein_codi	ENST00000371589.7	16544;14290	CGI:chr9:137	Island
cg00000714	0.10458747	chr19	54191827	54191828	TSEN34;TSEN34;TSEN34;TSEN34;TSEN34;T	protein_codi	ENST00000302937.7	1558;1095;5	CGI:chr19:54	S_Shore
cg00000721	0.95147603	chr6	25282551	25282552	LRRC16A;LRRC16A	protein_codi	ENST00000329474.6	3122;3472	CGI:chr6:252	S_Shelf
cg00000734	0.04680678	chr3	129183534	129183535	CNBP;CNBP;CNBP;CNBP;CNBP;CNBP;CNBP;(	protein_codi	ENST00000422453.5	389;377;366;	CGI:chr3:129	Island
cg00000769	0.0227429	chr12	123601930	123601931	DDX55;DDX55;DDX55;DDX55;DDX55;DDX55;I	protein_codi	ENST00000238146.7	-170;-201;-2	CGI:chr12:12	Island
cg00000807	NA	chr2	23690544	23690545	KLHL29;KLHL29;KLHL29	protein_codi	ENST00000288548.5	48151;1758;-	CGI:chr2:236	Island
cg00000884	NA	chr4	153688705	153688706	.	.	.	.	CGI:chr4:153	S_Shelf
cg00000905	0.04919231	chr15	59493107	59493108	FAM81A	protein_codi	ENST00000288228.8	54933	CGI:chr15:59	.
cg00000924	0.20986946	chr11	2699233	2699234	KCNQ1;KCNQ1;KCNQ1OT1	protein_codi	ENST00000155840.8	254241;2377	CGI:chr11:26	Island
cg00000948	0.91480638	chr8	48978050	48978051	.	.	.	.	CGI:chr8:489	Island
cg00000957	0.91984804	chr1	5877193	5877194	NPHP4;NPHP4;NPHP4;NPHP4;NPHP4;NPHP4	protein_codi	ENST00000378156.7	115279;1152	CGI:chr1:587	Island
cg00001099	NA	chr8	86069324	86069325	ATP6V0D2;PSKH2;PSKH2;PSKH2	protein_codi	ENST00000521564.1	82000;374;1	CGI:chr8:860	N_Shore
cg00001245	0.03864679	chr3	15065203	15065204	MRPS25;MRPS25;MRPS25;MRPS25;MRPS2	protein_codi	ENST00000253686.5	133;109;115;	CGI:chr3:150	Island
cg00001249	0.5150283	chr14	59923068	59923069	LRRC9;LRRC9;LRRC9;LRRC9	protein_codi	ENST00000254271.7	3354;3354;3	CGI:chr14:59	S_Shelf
cg00001261	0.21301308	chr16	3413964	3413965	LA16c-306E5.2	protein_codi	ENST00000575785.1	48851	CGI:chr16:34	.
cg00001269	NA	chr20	50342467	50342468	.	.	.	.	CGI:chr20:50	N_Shore
cg00001349	0.65105112	chr1	166989202	166989203	MAEL;MAEL;MAEL;MAEL;MAEL	protein_codi	ENST00000367870.5	112;92;-87;-	CGI:chr1:166	Island
cg00001364	0.92386923	chr1	213997033	213997034	PROX1;PROX1;PROX1;PROX1;PROX1	protein_codi	ENST00000261454.7	8531;9089;8	CGI:chr1:213	

Figure 5:Sample Data File for DNA Methylation

3.1.2. Data Collection Process

Data is downloaded from Cancer Genome Atlas (The Cancer Genome Atlas Program - National Cancer Institute n.d.) Genomic Data Commons data portal (GDC Data Portal n.d.), by implementing Application Programming Interface (API) (GDC Application Programming Interface (API) | NCI Genomic Data Commons n.d.).A new application has been implemented for this purpose (Downloading Files - GDC Docs n.d.).

The implementation has been designed to read each case, including manifest file, clinical data, biospecimen supplementary, miRNA expressions, mRNA expression (including all data files). A sample data directory for Case "**TCGA-GN-A265**" is represented in Figure-6.

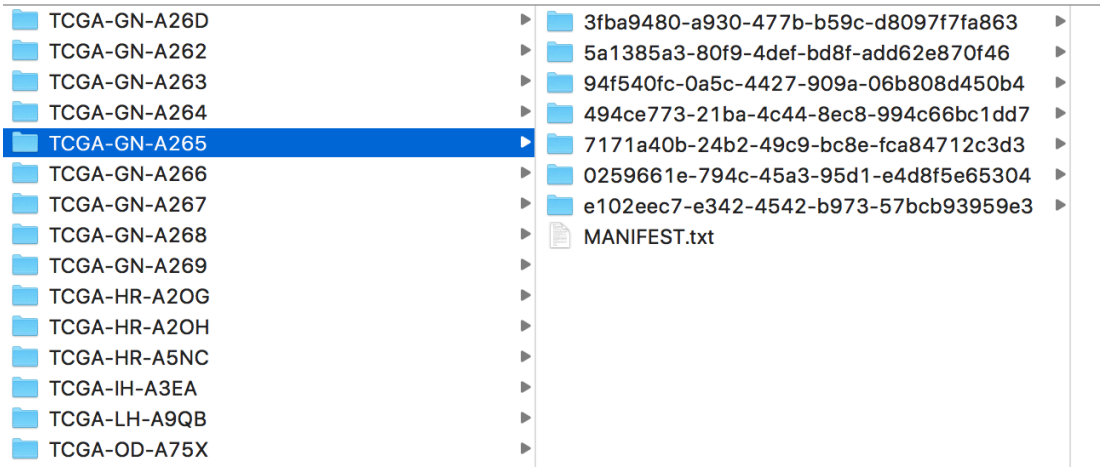

Figure 6: Case Directory After downloading

After the data is downloaded via API, the files on the data directory list for CASES are read by file IO operations and to parse data. Case clinical data and sample information are also read by XML parser and combined with expression values. 11.265 files were processed during this process, with an algorithm implemented for this purpose.

Once the data parser of expression values and combination of bio-specimen and clinical data is collected, the initial data cleaning has been started.

1- miRNA IDs and RNA Ensemble IDs, containing all null/zero values, have been removed.

2- Three separate databases have been created for
   a. Cases for miRNA expressions (452 Sample)
   b. Cases for mRNA Expression (472 Sample)
   c. Cases for DNA Methylation (483 Sample)
3- By intersection, all three database cases containing miRNA expressions, mRNA expressions, and DNA methylation are merged.

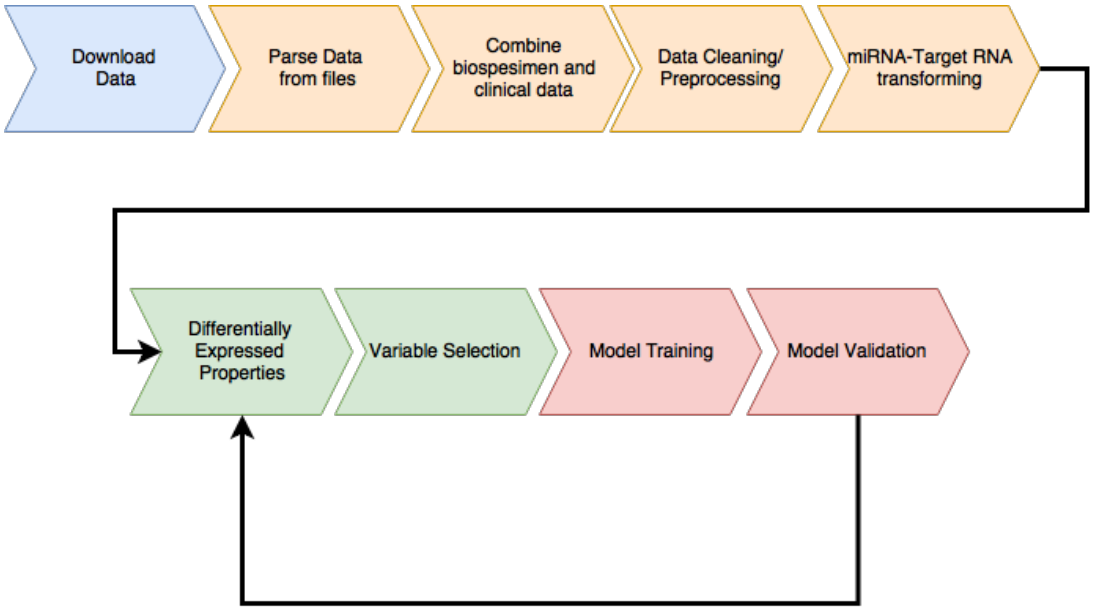

Figure 7: Summary of Analysis Steps

Besides, a new database for miRNA, its target mRNA, has been prepared to be used during variable selection for Differentially Expressed Patterns of miRNA and target mRNA. All differentially expressed variables (miRNAs & mRNAs) are analyzed by comparing their target mRNA or miRNA expression levels.

In order to identify matching, identify miRNA expressions and RNA expressions of Data 4 database has been used

1- TCGA miRNA ID Data List: (List for miRNA IDs has been exported from TCGA Data portal)
2- miRTARBASE (mirTAR Base Database has been downloaded to find matching RNA Name) (H. Y. Huang et al. 2020).
3- BioMart database for ensemble has been used to get ensemble Gene Id from RNA Name. For this purpose, the biomart R package has been used (BiomaRt Bioc R package n.d.)
4- TCGA mRNA Ensemble ID List: (List for mRNA ensemble ID has been exported from TCGA Data portal)

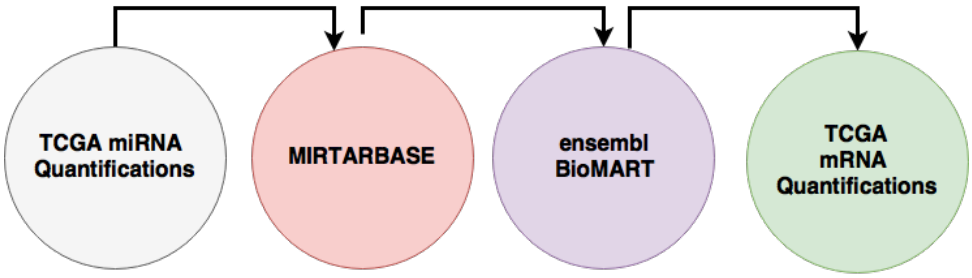

Figure 8: miRNA And Target RNA identification cycle

This mapping database is extended during methylation biomarkers analysis to cover their miRNA expressions with target genes' DNA methylation beta-values. For this purpose, the same method is used, and the target gene of the miRNA is defined by checking the gene symbol of the DNA methylation data values. All these mapping databases were used to analyze the pattern of selected miRNA with respect to target mRNA and Gene.

3.2. Method

The data analysis is started with data preprocess and variable selection. miRNA expression is used for the initial cycle of the spiral analysis method. Then, 11.265 separate files that contain miRNA mRNA expressions for each case are downloaded from TCGA with a manifest file that contains metadata for the specific case. The manifest file is used to read and combine case files to generate a data pool. The final data pool contains 472 observations with 60.492 properties for mRNA, 450 observations with 1904 properties for miRNA, and 483 observations with 34014 variables for methylation. We only chose the cases which have all three experiments, namely miRNA, mRNA, and Methylation.

The sample type property is used for the class variable, which is a categorical variable with four levels, namely: "Primary Tumor," "Solid Tissue Normal," "Metastatic," and "Additional Metastatic." "Solid Tissue Normal." Only the samples with "Primary Tumor" and "Metastatic" are selected for further analysis.

There were variables for miRNA and mRNA expressions with a constant (1 or 0) value for all samples. These attributes have been removed from the dataset. The remaining samples are subject to a significance test concerning class variables: log normalization & Z-score normalization used for relevant markers. Markers are scaled 0-1 range. T-TEST

has been used as a significance test (P-value is defined as 0.001). As a result of the test, 425 miRNA, 2061 mRNA, and 8698 Methylation variables were significantly expressed between two groups ("Primary Tumor" and "Metastatic").

For a detailed analysis of the results, all possible miRNA patterns and their target mRNA and gene methylation are calculated. Then, depending on the evaluation of the significance level, different patterns are defined.

Random selection is applied for each class with a 20 % ratio to separate unseen data for testing during the analysis. We repeated this randomization process to create ten different splits, which are used as a separate trial. By generating more than one split, we aim to decrease the bias due to random splitting and test the repeatability. So, as an experiment environment, we created an experiment pool constructed by ten random partitions for the test set and training set generated by applying both under-sampling and oversampling (SMOTE) (Fernández et al. 2018) techniques for addressing class imbalance issues. So, 80% of the data is used for training and validation (Figure 9). In each trial, both dimensional reduction and feature selection techniques were applied separately to solve the curse of dimensionality problem for both undersampling and oversampling methodology, and different machine learning techniques were evaluated with 10-fold cross-validation. Final models are tested against the unseen data separated at the beginning. All these processes were repeated ten times for each data set in the experimental pool. Finally, the mean values of prediction parameters are calculated for the results reported in this study.

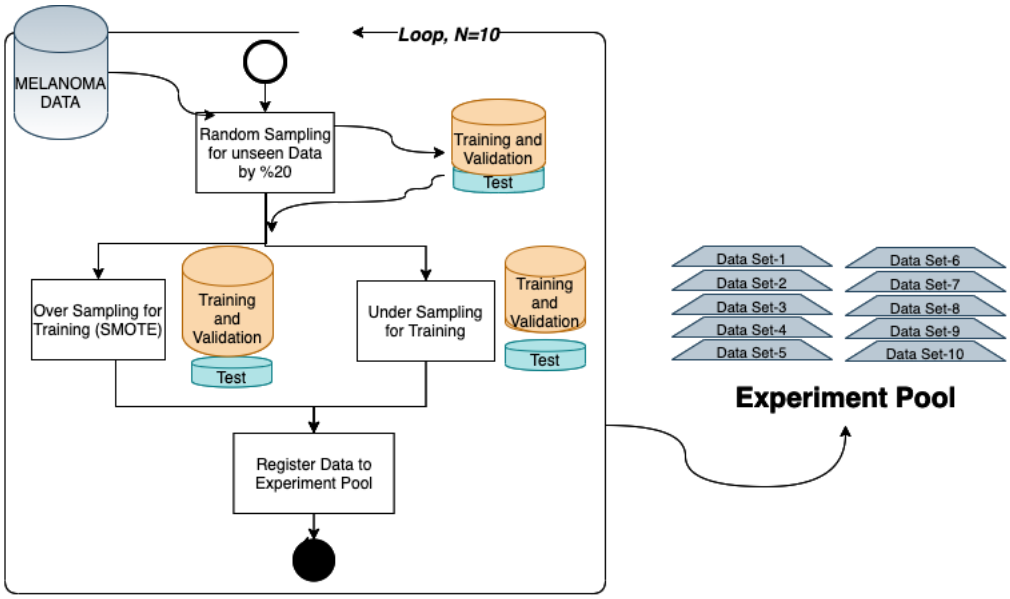

Figure 9: Training Validation and Unseen Test Data Generation

In each trial, for this purpose following steps are followed for both undersampling and oversampling. 1)The significant variables listed in the given category are selected from the dataset. 2) A set selected randomly from whole data with an 80 % ratio of each class is kept for unseen data. 3) Apply technique to solve the curse of dimensionality problem(For dimensional Reduction, Principal component analysis is applied and for Oversampling runs SMOTE algorithm is used with K=3) 4) Steps 1 to 3 is repeated for each data split in experiment pool.

Each test/training subsets listed in the experiment pool were trained and tested for different models by adding miRNA expressions, mRNA expressions, and methylation beta values iteratively. Besides, to address the curse of dimensionality, we tried both dimensional reduction and feature selection techniques. Seven methods, namely SVM with linear, radial, polynomial kernels, neural network, random forest, Adaboost, and Naive Bayes, have been applied to generate and test a predictive model (Figure 3). Neural Networks and Support Vector Machines are frequent models that have been applied to similar classification models. However, as we searched the literature, we did not see any research which applied bagging, boosting, or probabilistic methods. So, we choose at least one representative of various classification algorithm categories, namely Artificial Neural Networks, Bagging Methods, Boosting methods, and probabilistic models one or more. Apart from Support Vector Machines and Neural Networks, we included Adaptive Boosting, an ensemble method that composes a robust classifier from various weak classifiers, and Random Forest, which relies on bagging techniques to increase classification performance more than the single decision trees (see Figure 10). Apart from all these, Naïve Bayes also chooses an alternative since it is a fundamental model based on probabilistic techniques. Mean F-score and Mean P-value are evaluated as

performance indicators for validation and test classifications of data sets. Box Plot distribution of classification scores is investigated for each data set in the experimental pool. The best model for each category is made by comparing mean F scores and mean P-values. If these results are the same two or more best model candidates, we have reviewed the box plot of significance and sensitivity distributes.

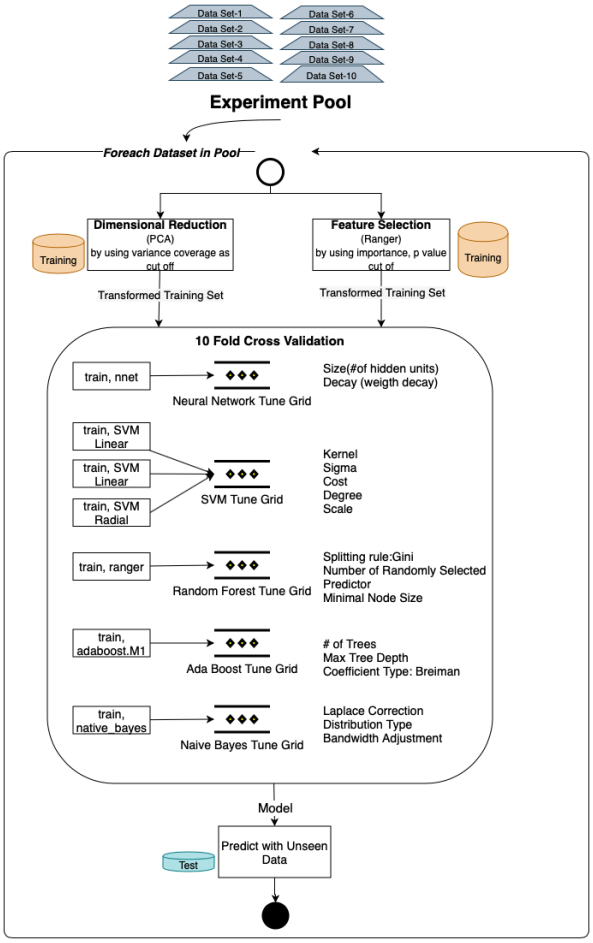

Figure 10: Model Training And Testing Process

Experiment flow initiated by applying alternative dimensionality solutions, namely PCA and Feature Selection. Through each experiment flow, models are trained with seven (SVM with linear, radial, polynomial kernels, neural network, random forest, AdaBoost, and Naive Bayes) machine learning algorithms and tested with the same unseen data. Overall flow repeated for each data-subsets in the experiment pool.

This thesis follows the following coding mechanism to map the alternative scenarios of class imbalance and dimensionality solution techniques for each category. This annotation is used as the naming convention of the given result set in the following sections :

- a1: miRNA biomarkers modeled with feature selection and undersampling
- b1: miRNA biomarkers modeled with feature selection and SMOTE

- c1: miRNA biomarkers modeled with PCA and undersampling
- d1: miRNA biomarkers modeled with PCA and SMOTE
- a2: miRNA and mRNA biomarkers modeled with feature selection and undersampling
- b2: miRNA and mRNA biomarkers modeled with feature selection and SMOTE
- c2: miRNA and mRNA biomarkers modeled with PCA and Undersamplining
- d2: miRNA and mRNA biomarkers modeled with PCA and SMOTE
- a3: miRNA, mRNA, and methylation biomarkers modeled with feature selection and undersampling
- b3: miRNA, mRNA, and methylation biomarkers modeled with feature selection and SMOTE
- c3: miRNA, mRNA, and methylation biomarkers modeled with PCA and undersampling
- d3: miRNA, mRNA, and methylation biomarkers modeled with PCA and SMOTE
- a4: miRNA, mRNA, and methylation biomarkers modeled with feature selection and undersampling
- b4: miRNA, mRNA, and methylation biomarkers modeled with feature selection and SMOTE
- c4: miRNA, mRNA, and methylation biomarkers modeled with PCA and undersampling
- d4: miRNA, mRNA, and hypo methylation biomarkers modeled with PCA and SMOTE
- a5: miRNA, mRNA, and hypo methylation biomarkers modeled with feature selection and undersampling
- b5: miRNA, mRNA, and hyper methylation biomarkers modeled with feature selection and SMOTE
- c5: miRNA, mRNA, and hyper methylation biomarkers modeled with PCA and undersampling
- d5: miRNA, mRNA, and hyper methylation biomarkers modeled with PCA and SMOTE
- a6: mRNA, and methylation biomarkers modeled with feature selection and undersampling
- b6: mRNA, and methylation biomarkers modeled with feature selection and SMOTE
- c6: mRNA, and methylation biomarkers modeled with PCA and undersampling
- d6: mRNA, and methylation biomarkers modeled with PCA and SMOTE
- a7: mRNA biomarkers modeled with feature selection and undersampling

- b7: mRNA biomarkers modeled with feature selection and SMOTE
- c7: mRNA biomarkers modeled with PCA and undersampling
- d7: mRNA biomarkers modeled with PCA and SMOTE
- a8: methylation biomarkers modeled with feature selection and undersampling
- b8: methylation biomarkers modeled with feature selection and SMOTE
- c8: methylation biomarkers modeled with PCA and undersampling
- d8: methylation biomarkers modeled with PCA and SMOTE
- a9: miRNA and methylation biomarkers modeled with feature selection and undersampling
- b9: miRNA and methylation biomarkers modeled with feature selection and SMOTE
- c9: miRNA and methylation biomarkers modeled with PCA and undersampling
- d9: miRNA and methylation biomarkers modeled with PCA and SMOTE

All preprocessing, training, validation, and test with R studio use various R packages.

- Neural Network (package:nnet) (Ripley 2021)(Ripley and Venables 2021)
- Adaboost (package : adabag) (Alfaro, Gáamez, and García 2013) (Alfaro, Gamez, and Garcia 2018)
- Random Forest (package: ranger) (Wright, Wager, and Probst 2021; Wright and Ziegler 2017)
- Naïve Bayes (package : naivebayes) (Majka and Michal Majka 2020)
- Support Vector Machine (package : kernlab) (Karatzoglou et al. 2004; Karatzoglou, Smola, and Hornik 2016)
- Smote (smotefamily) (Siriseriwan 2019; Wacharasak Siriseriwan 2019)

We followed a systematic cross-comparison technique during the collection and evaluation of the results. First, we collected the prediction scores for different classification models to find the best algorithm. Evaluation of the successors within each feature category identified the winner. Finally, model progress and contributions of adding new feature categories are assessed based on these results collected. The illustration of this process is summarized in Figure 11.

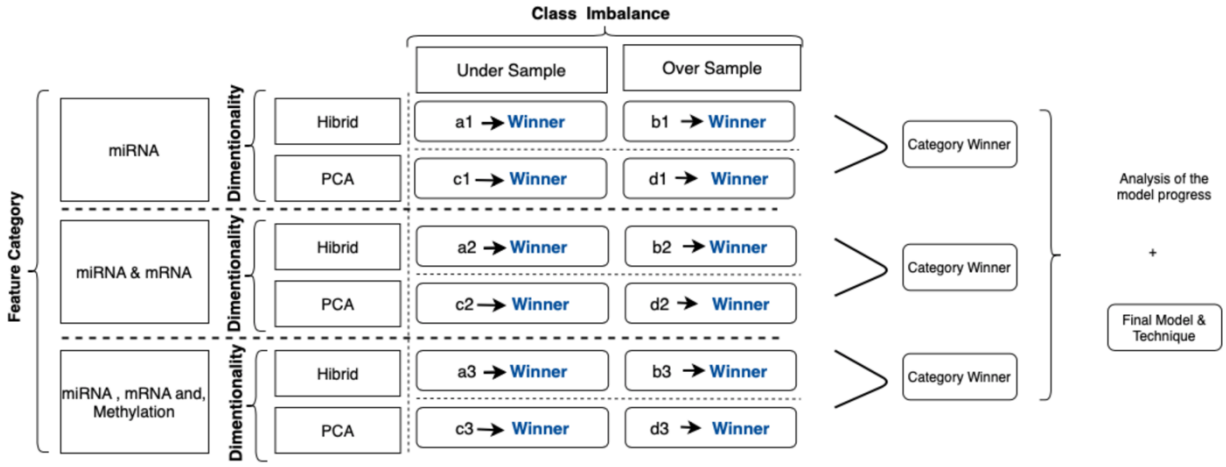

Figure 11: Illustration For Category Based Analysis With Techniques Applied

Each evaluation criterion is represented with a code. For example, (a1) represents the predictive models by using miRNA signatures with the hybrid method -that is, random forest to calculate feature importance undersampling for class imbalance solution. Similarly, d3 represents the outcomes of models applied to predict metastasis using significant miRNA & mRNA, methylation biomarkers using PCA as a dimensional solution, and SMOTE as a class imbalance solution.

The experiment is repeated for each subset in the data pool to find the prediction scores, and mean values were calculated. We have assessed the predictive algorithm using seven different machine-learning models, including representatives of various classification algorithm categories, namely artificial neural networks, bagging methods, boosting methods, and probabilistic models. We applied ten-fold cross-validation for each subset and calculated mean F-score; mean P-value to evaluate each category's best model. If the results are the same for two or more model candidates, we have reviewed the box plot of significance and sensitivity distributes to choose the one with low variance.

$$FScore = 2 * \frac{(Sensitivity * Predictivity)}{Sensitivity + Predictivity} \quad (1)$$

As a final step, we performed functional and pathway enrichment analysis using DAVID (Dennis et al., 2003; D. W. Huang et al., 2007). KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers are compared for sets of "miRNA," "miRNA and mRNA," and "miRNA, mRNA, and methylation" to better understand contributing factors behind the higher precision and consistency after including methylation data to the models.

CHAPTER 4

4. RESULTS

In this study, we have evaluated the potential genetic biomarkers of melanoma metastasis. In addition, we developed multiple predictive models to predict the metastatic outcome by integrating miRNA, mRNA, and DNA methylation markers by using the TCGA melanoma dataset. This study's experimental strategy is composed of a Multi-cycled evaluation, each of which targets different feature categories. In each cycle, different techniques to solve using dimensionality and class imbalance problem solutions are evaluated. Figure 12 summarizes the results of all evaluation techniques for each cycle.

4.1. Classification with miRNA

At the first step of the initial cycle, we have implemented a predictive model (a1) with a microRNA biomarkers model using feature selection through importance (hybrid model) and class imbalance solution through under-sampling. The predictive model with adaptive boosting (AdaBoost) demonstrates the best results among all trials with the highest F-score and accuracy. Besides, the variance of the results for the different datasets in the experiment was also low compared to other models. Similarly, the random forest has the second-best results among all trials (F-score 80%). In the second scenario (b1), when we replace the class imbalance solution with smote, random forest demonstrates similar results with an F-score of 79 %. In parallel, adaptive boosting (AdaBoost) presents a comparable performance (F-score 80%) to the random forest model with a slightly higher score. In the third trial (c1), we have used under-sampling and dimensional reduction with PCA. According to our results, adaptive boosting (AdaBoost) showed better scores (F-score 80 %), but for this time, SVM with the linear kernel (F-score 78%) was better than random forest (F-score 72%), demonstrating the second-best results. Finally, we applied SMOTE to address the class imbalance issues (d1). The results were similar to the first trial; adaptive boosting showed the best results (F-score: 80%), the random forest also had the better results (F-score: 79%) compared with other models (Figure 13).

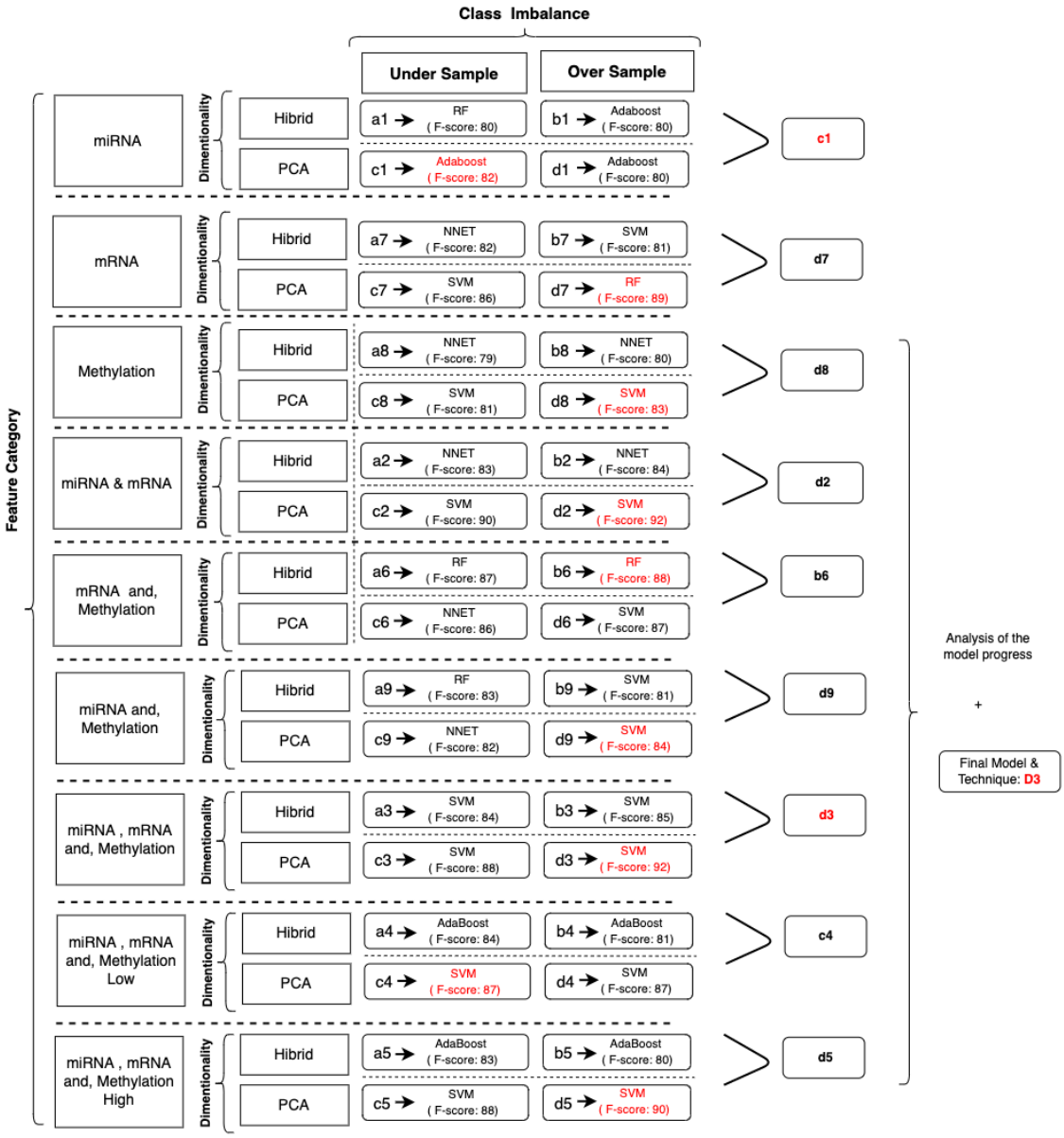

Figure 12: Illustration For Results Of Category Based Analysis With Techniques Applied To Solve Significant Issues

As a result of the evaluation process, c1 is selected as the successor model for miRNA markers. When two markers, miRNA and mRNA, are combined, the winner is identified as d2. In the final cycle, the merge of all biomarkers resulted in d3 as the successor. Among all d3 was the winner to predict the metastatic outcome

As a result of the initial cycle, microRNA biomarkers predict the primary tumor's metastatic outcome with an F-score of almost 80%. In predictive models, all workflows showed similar classification accuracy by using miRNA markers. We selected (c1) the adaptive boosting with the PCA and undersampling, resulting in the highest F-score. Both

random forest and adaptive boosting (AdaBoost) demonstrated better results in each workflow (Figure 13).

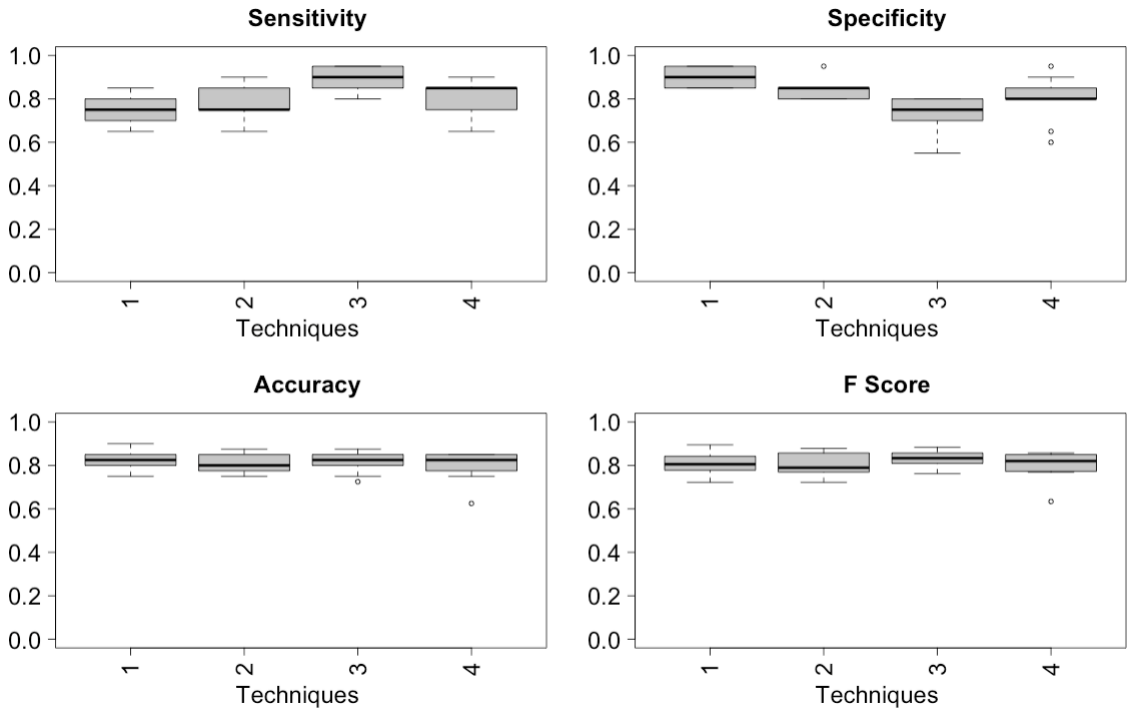

Figure 13: Model Comparison Of Techniques Used For miRNA Biomarkers
(1:a1, 2:b1, 3:c1, 4:d1): Category 1, which uses a hybrid model of feature selection and Adaboost classifier(c1), has the best results among all scenarios.

4.2. Classification with mRNA

Only mRNA markers are selected as biomarkers for the classification model in this step. By following the same methodology for dimensional reduction and class imbalance solution techniques, random forest algorithm(d7) with principal component analysis and Smote (as oversampling) technique classification achieved the highest accuracy with 88% mean F score and P-Value of 1.18×10^{-05}. Undersampling with PCA also showed similar performance with the AdaBoost algorithm(c7) (F score 86%, P-Value 1.44×10^{-05}), but feature selection techniques were behind these two trials. Neural networks technique(a7), by applying feature selection as dimensional reduction and undersampling as class imbalance solution, accuracy is observed as 82% Mean F Score and $9.09 \times 10^{-04.}$ When we replaced the class imbalance solution with smote, SVM (with the linear kernel) is listed

as the best model, but almost no change is observed in prediction accuracy (F Score 81 %, P-Value 4.88 x 10^{-04}).

At the end of the cycle, we saw that model using mRNA markers winner models had F-scores ranging between 82% and 88%. The prediction scores using the feature selections technique were not as good as PCA for undersampling and oversampling techniques. Principle component analysis produced better results for both oversampling and undersampling. Since F-score for (d7), Random Forest using PCA and SMOTE, has the highest scores, it is selected. (See Figure 14).

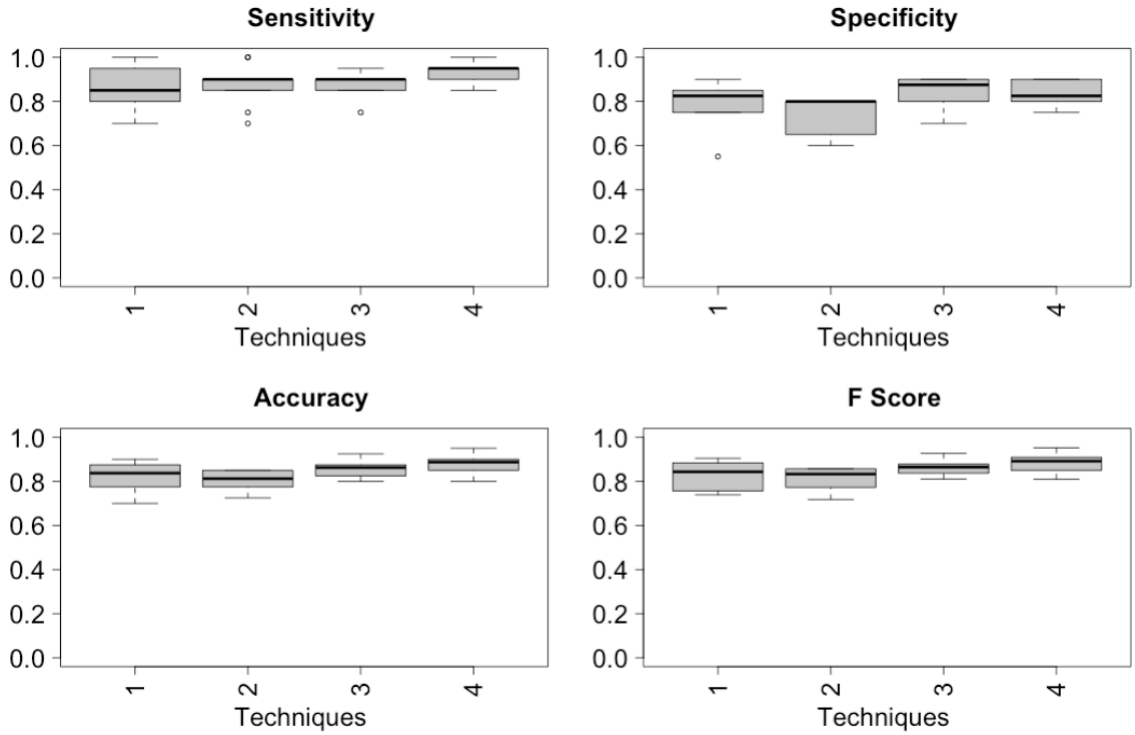

Figure 14: Model Comparison Of Techniques Used For mRNA Biomarkers
(1:a7, 2:b7, 3:c7, 4:d7): Category 4, which uses (d7) PCA and Smote techniques and Random Forest classifier, has the best results among all scenarios.

4.3. Classification with Methylation

In this cycle, we filter down methylation biomarkers. Like previous cycles, we applied a combination of each class imbalance and dimensionality solution techniques. We decided on Support Vector Machine since model significance demonstrated improvement in our results. Firstly, all Neural network predicts metastasis with an F-score of 79% using under-sampling and feature selection through importance techniques(a8). SVM with all radial,

polynomial and linear kernel produced similar results in this technique (F score 72%, 74%, and 72% relatively). Random Forest predicts with similar F-scores (75%). AdaBoost, on the other hand, showed %74 accuracies in F-Score. The best prediction accuracy was observed with Neural Network with feature selection and undersampling (F-Score 80%). In the second trial, we have replaced the class imbalance solution technique with SMOTE. Prediction accuracies were quite similar to the previous. Both SVM with linear kernel and the polynomial kernel have similar results with 77% and 78% F-scores. Prediction accuracy of random forest decreased to %72 while AdaBoost produced 67% in F score. In the third trial, under-sampling and dimensional reduction with PCA are applied. SVM with radial kernel was the best model (F-score; 81%). All models showed similar performance in this cycle (Neural Network 79%, SVM -Linear Kernel- 80 %, SVM-Polynomial Kernel- %79, Random Forest %80, Adaboost %79). Finally, when we applied SMOTE instead of under-sampling(d8), SVM with Polynomial kernel demonstrated slightly higher scores (F-score 83%). SVM with Linear and radial kernels had an F-score of 82%. Both Neural networks and AdaBoost had similar results with an F Score % of 78. Similarly, random forest achieved 80% of the mean F score (Figure 15).

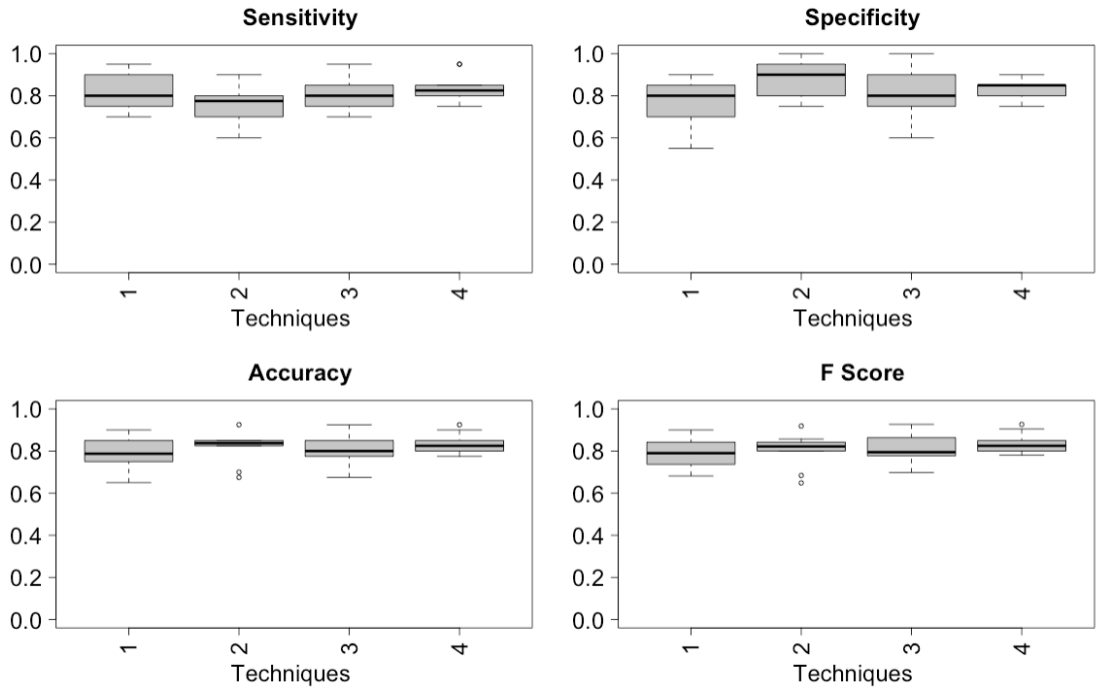

Figure 15: Model Comparison of Techniques Used For methylation

Biomarkers (1:a8, 2:b8, 3:c8, 4:d8): The model listed in d8, is selected as the successor model for the second cycle

As a result, although the prediction accuracies have similar results for all trials, SVM with polynomial kernel bu using SMOTE and principal component analysis(d8), we observed the best accuracy in mean F Score and Low variance in all prediction variables (see Figure 15)

4.4. Classification with miRNA and mRNA

In this cycle, we utilized both miRNA and mRNA as biomarkers. Like the previous cycles, we first used (a2) feature selection through importance (hybrid model) and class imbalance solution through undersampling. When we compared the predictive models, results were quite similar by varying F scores between 81% to %83. However, random forest produces the best results of mean F-score (83%); mean P-value (8.26 x 10-05). SVM with a polynomial kernel was the second-best model to predict the metastatic outcome with the same F-score but with a lower P-value (9.34 x 10-05). As a second trial(b2), we have replaced the class imbalance solution with SMOTE. The results for each model, which vary from 80% to 84% for F-score, were quite similar. The neural network showed the best F-score (84%) and P-values (2.41 x 10-05). SVM with linear and polynomial kernel also had the same F-score (84%), and the neural network showed higher significance. Adaptive boosting and random forest demonstrate better results for the miRNA-mRNA cycle predict the metastatic outcome with equal mean F-scores of 81%. In the third trial (c2), undersampling for class imbalance and dimensional reduction with PCA are applied. SVM with the linear kernel was the best model with the highest F-score (90%). The neural network was the second-best model to predict metastasis with F-score (89%). Nevertheless, this time, adaptive boosting (F-score: 82%) and random forest (F-score 75%) are left behind. As the final trial(d2), we have applied SMOTE and dimensional reduction with PCA(d2). Neural network and SVM with linear kernel produced the best results compared to the rest with F-scores 91% and 92%, respectively. On the other hand, adaptive boosting and random forest showed high variance across different trials (Figure 16).

At the end of the second cycle, we saw that models using MiRNA and mRNA markers as winner models with F-scores ranging between 83% and 92%. The prediction scores for both boosting and bagging techniques were not as good as in the first cycle. Since F-score for (d2), SVM using PCA and SMOTE, has the highest scores, it is selected.

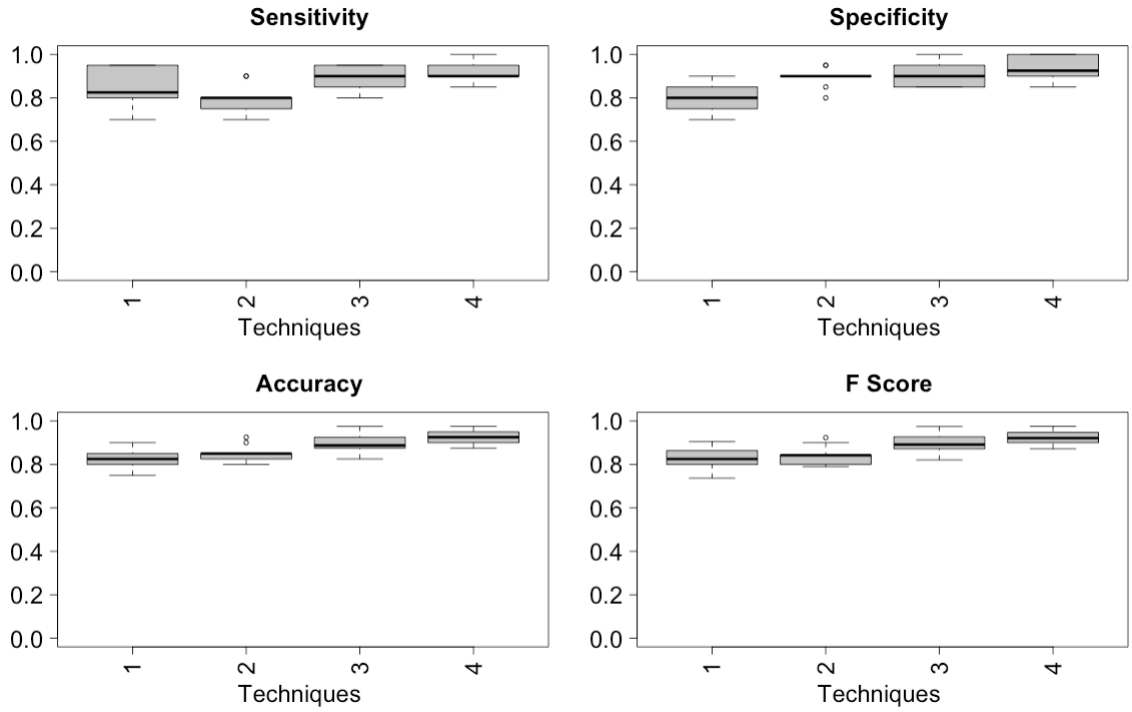

Figure 16: Model Comparison of Techniques Used For miRNA And mRNA Biomarkers
(1:a2, 2:b2, 3.c3, 4:d2): Model listed in 4, which applies d2, is selected as the successor model for the second cycle

4.5. Classification with mRNA and Methylation

The combination of mRNA and Methylation biomarkers is examined during our analysis. First, we tried feature selection as dimensional reduction and undersampling as class imbalance solution(a6) prediction accuracies for neural network, SVM with linear and polynomial kernels observed as 85 % Mean F Score. On the other hand, the SVM radial kernel presents similar but a bit lower results with 84 % F-Score.

While AdaBoost algorithm resulted with %79 accuracies, random forest present the best accuracy with %87 of F score and P-value of 1.35 x 10^{-05}. Secondly, when we replace class imbalance with SMOTE (b6), prediction accuracies result in similar F- scores. While random forest demonstrates the best accuracy in 88 % F-Score, neural network, SVM with linear, radial, and polynomial kernels had F-Scores, had F-Values of 85 %, 86 %, 86%, and 85% relatively.

In the third trial, we tried PCA as a feature selection method using undersampling(c6); the neural networks were the best model with a %86 F- Score. The other algorithms demonstrated quite similar results varying between %81 and %84 in F-score. The worst results were observed in SVM with the polynomial kernel (F score 67 %).

Finally, using PCA with SMOTE(d6), SVM with linear kernel listed with the best F-Score (87%). Neural Network and Adaboots present comparable results with 86% and 85% mean F Scores. Random forest resulted in %80 Mean F Score. Like the previous trial SVM, the polynomial kernel was the worst model (F Score 67 %) (Figure 17).

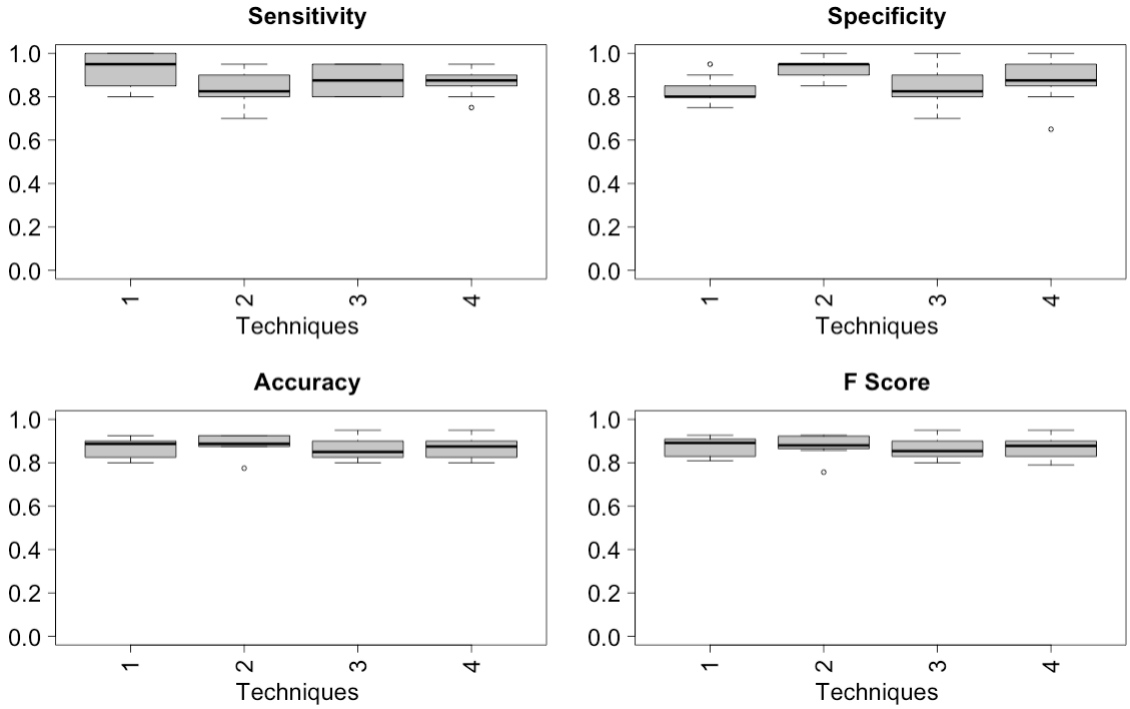

Figure 17: Model Comparison of Techniques Used For mRNA And Methylation
Biomarkers (1:a6, 2:b6, 3:c6, 4:d6): The model listed as 2, which applies b6, is selected as the successor model for the second cycle

4.6. Classification with miRNA and Methylation

We combined miRNA with Methylation markers in the final step of 2-grouped biomarkers. In all previous trials, we started with feature selection as dimensional reduction and undersampling as class imbalance solution(a9). Random Forest

demonstrated the highest result in this initial step with an F value of %83. Adaboost was the second-best model in this trial (F score-80%). On the other hand, SVM resulted in similar F Scores for all kernels with ~77%. Finally, the Neural network showed a 78 % F-score.

When we change the class imbalance method with smote(b9), apart from the random forest, SVM (with the linear kernel) reached the same highest F score (81 %). Neural network and Adaboost were relatively close with 78 %, and 77% mean F scores.

For the third trial, we applied PCA with the undersampling technique(c9), Neural network, SVM with Radial and polynomial kernel demonstrated the highest prediction values with 82% in F Score. Random Forest and AdaBoost, on the other hand, were listed with lower scores with 76% and 78% F scores, relatively.

Finally, we applied SMOTE for the class imbalance solution with PCA(d9). Results were also similar in this run; SVM with radial kernel was the best model with an 84% F-score. Neural networks also produced high results with an 83% F -score. Random Forest and AdaBoost models have the same results in F-scores (%80) (Figure 18).

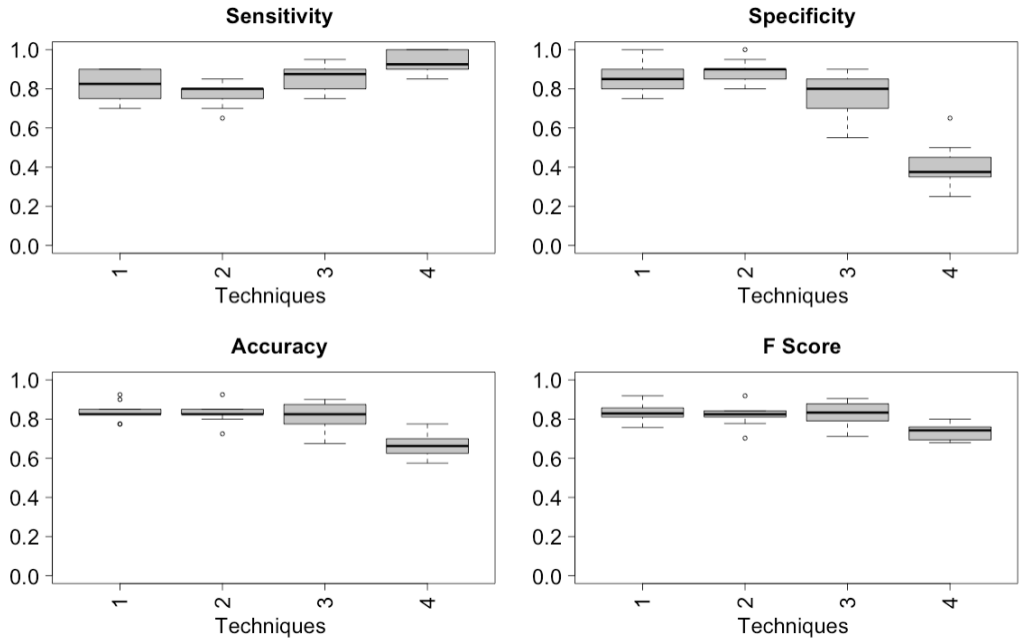

Figure 18: Model Comparison of Techniques Used For miRNA And Methylation Biomarkers
(1: a9, 2:b9, 3.c9, 4.d9): Model listed in 5, which applies d9, is selected as the successor model for the second cycle. Indeed, the results were similar to a9. Also, the variance in the prediction variable was relatively small for a9.

4.7. Classification with miRNA, mRNA, and Methylation

We combined all miRNA, mRNA, and methylation biomarkers in the third cycle. Similar to previous cycles, we applied a combination of each class imbalance and dimensionality solution techniques. We decided on neural networks since model significance demonstrated improvement in our results.

Firstly, all Neural networks, SVM with linear and polynomial kernel predicts metastasis with an F-score of 83% by using under-sampling and feature selection through importance techniques(a3). Both SVM with radial kernel and random forest predict with similar F-scores (83%). So, the prediction model results were close to each other for this trial. However, the lowest variance across different trials was observed with SVM (linear kernel).

In the second trial (b3), we have replaced the class imbalance solution technique with SMOTE. Both SVM with linear kernel and the polynomial kernel were the two best-performing models with 84% and 85% F-scores.

In the third trial(b4), sampling and dimensional reduction with PCA are applied. SVM was the best model regardless of the selected kernel (F-score; 88%).

Finally, when we applied SMOTE instead of under-sampling(d3), SVM with linear kernel demonstrated slightly higher scores (F-score 92%). In contrast, SVM with polynomial kernel and Neural network had 91% and 90% F-score. The best predictive model was SVM, trained using dimensional reduction with PCA and SMOTE (d3). Like the second cycle, both SVM and Neural Network models resulted in better results in all trials. In addition, both under-sampling and oversampling techniques produced similar results (Figure 19).

4.8. Classification with miRNA, mRNA and Hypo Methylation

After we analyzed the triple model, we were also curious about changes for hypo and hyper methylated genes. First, we combined miRNA and mRNA markers with Hypo Methylated genes.

In the initial trial, we started with feature selection and undersampling techniques (a4). Adaboots was the successor model in this step with 84% accuracy. The random forest also produced similar results with 82 % accuracy. On the other hand, SVM achieved a 78% F

score with Linear Kernel. On the other hand, Neural Network showed the worst results with a 67% F score.

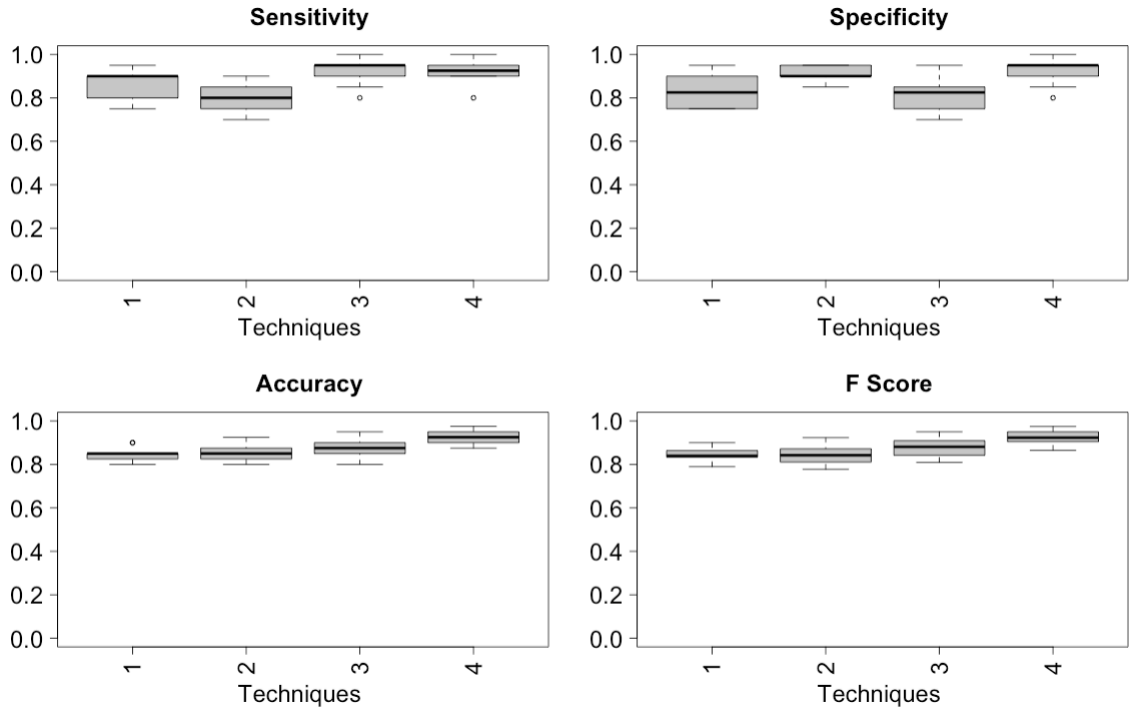

Figure 19: Model Comparison Of Techniques Used For miRNA, mRNA, and Methylation Biomarkers *(1:a3, 2:b3, 3:c3, 4:d3). The model listed in 4, which applies d3, is selected as the successor model for the final cycle*

Secondly, we tried smote ad class imbalance solution and combined it with the feature selection(hybrid) technique(b4). SVM with the linear kernel is the winner among all models with 81% accuracy. While AdaBoost presented similar results (F score, 79%), neural network and Random forest demonstrated the worst results with 77% accuracies.

Next, we switched to PCA for dimensional reduction and combined it with undersampling (c4). Similar to the previous trial, SVN with Liner kernel was the successor model (F Score 87%). The results for neural networks were also quite similar, with an 86% F Score. Random forest and AdaBoost were lower with 82% and 80% F scores.

Finally, we tried smote as a class imbalance solution. As we observed, linear SVM was the winner model again (F score %87). SVM with polynomial kernel and random forest were the second-best model with a %86 F Score. Neural network and AdaBoost shared similar scores with 84% and 83% F scores.

As a result, when PCA is selected, the undersampling technique results were higher for all models. In general, SVM demonstrated high scores in all trials (Figure 20).

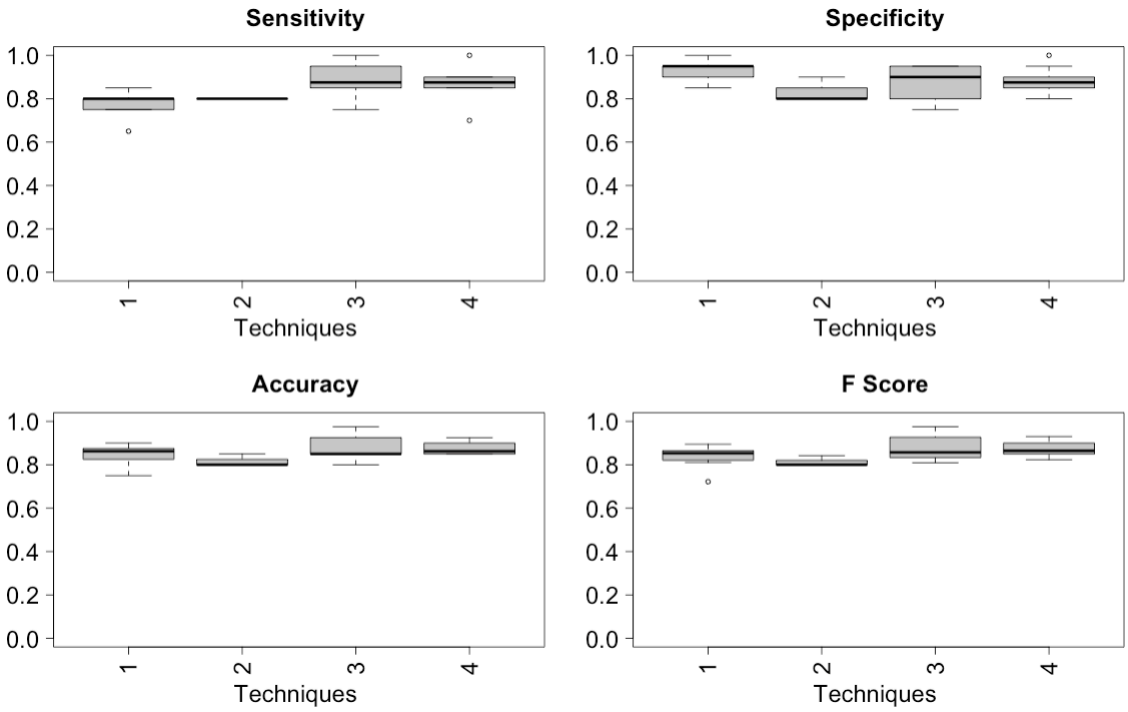

Figure 20: Model Comparison Of Techniques Used For miRNA, mRNA, and Hypo Methylation Biomarkers
(1:a4, 2:b4, 3:c4, 4:d4) The model listed in 3, which applies c4, is selected as the successor model for the final cycle.

4.9. Classification with miRNA, mRNA and Hyper Methylation

We filtered down hypermethylation biomarkers for the next cycle and ran the same analysis. The results were better when compared with hypomethylation, but they did not exceed the score of the triple model (listed in d3).

First of all, when we select feature selection as dimensional reduction and undersampling(a5) for class imbalance solution, AdaBoost had the best F score with 83 %. Random forest was the second-best predictive model with an 82% F score. On the other hand, SVM reached a 78% F score with Linear Kernel. Finally, the neural network showed the worst F score with 67%.

Secondly, we tried Smote instead of undersampling(b5). Results were similar; AdaBoost was the winning model again (F score 80 %). SVM also reached an F score of 80 % accuracy with linear kernel. Random forest listed behind with an F score of 75 %.

Next, we switched the dimensional reduction method with PCA and used undersampling as a class imbalance solution(b6). SVM with polynomial and radial kernels, both listed as successor model with F score 88%. SVM radial kernel and neural networks were also similar regarding 86% and 85% in F score. Random Forest and AdaBoost had relatively similar results with 82% and 81% F scores.

Finally, we used SMOTE as a class imbalance solution(d5) and achieved 90% accuracy with SVM linear and polynomial kernels. Neural network listed just behind these two and produced 88% of F score. Random forest and AdaBoost demonstrated relatively 85 % and 82% F scores (Figure 21).

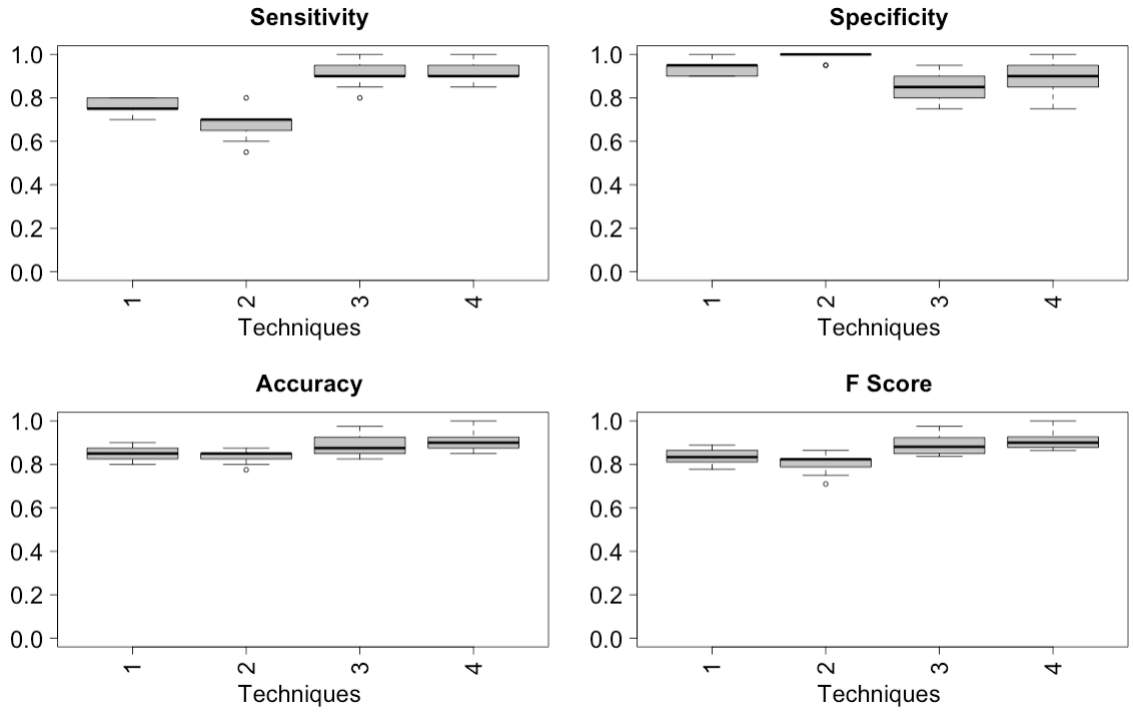

Figure 21: Model Comparison Of Techniques Used For miRNA, mRNA, and Hyper Methylation Biomarkers
(1:a5, 2:b5, 3:c5, 4:d5). The model listed in 4, which applies d5, is selected as the successor model for the final cycle

4.10. Comparison for All Findings

This study examines the predictive model progress when new biomarkers sets are added upon micro-RNAs. We started the analysis by selecting miRNA and adding mRNA upon it in the next cycle. We included DNA methylation upon these two and re-analysis the model progress as the next step.

Once these three cycles are completed, we also check the other combinations. Predictive power changes are also investigated to select hypomethylated and hypermethylated genes. Then we continue with the analysis of various other combinations of biomarkers.

As a result of all evaluations (see supplementary material), we came up with successors for each biomarker category (Table 2, Figure 22).

First of all, random forest with (a1) feature selection and undersampling achieved best results for miRNA markers (F-score=81 %, sensitivity= 75 %, Specify = 90 %, accuracy= 82 % $p = 1.7 \times 10^{-4}$). In addition, SVM (d2) with PCA and SMOTE was the most successful technique for combination of miRNA and mRNA markers (F-score=92 %, sensitivity= 92 %, Specify = 93,5 %, accuracy= 93% $p = 1.0 \times 10^{-7}$). Finally, by using all miRNA, mRNA and methylation markers (d3), SVM reached the same results with the previous one with higher consistency across different trials (F-score=92 %, sensitivity= 92 %, Specify = 93 %, accuracy= 92 % $p = 1.05 \times 10^{-7}$) (Figure 22).

For mRNA markers predictive model achieve 88% for F-score (sensitivity= 93 %, Specify = 83 %, accuracy= 82 % $p = 1.18 \times 10^{-5}$) by using Adaptive boosting model via undersampling and PCA methods. Methylation biomarkers on the other hand shows 81% of F-score (sensitivity= 82 %, Specify = 81 %, accuracy= 81 % $p = 2.1 \times 10^{-5}$) by using Support vector machine algorithm.

The combination of miRNA with methylation resulted in the same F-Score (82 %). Similarly, grouping mRNA with Methylation demonstrates the same F-score with the mRNA model (88 %).

Finally, selecting hyper and hypomethylated genes for the triple model and combining them with miRNA and mRNA markers generated two predictive models with an F-score of 90% and 87%, respectively.

Table 2: Summary For Iterative Progress On Model Precision Scores:
The miRNA model applied by feature selection through importance (hybrid model) and class imbalance solution through under-sampling is the method to be applied for prediction. For both the "miRNA- mRNA" and "miRNA-mRNA-Methylation" triple model, principal component analysis for dimensionality and SMOTE for Class imbalance solution was the best method to increase predictive power and stability of the model

Biomarker Group	Method	Sen.	Spe.	Accuracy	F Value	Kappa	P-Value
miRNA	PCA, Undersample Adaboost	90%	73%	81.25%	82.71%	62.50%	4.58E-04
mRNA	PCA, SMOTE Random Forest	93%	83%	88.00%	88.57%	76.00%	1.18E-05
methylation	PCA, SMOTE SVM (radial Kernel)	82%	81%	81.00%	81.16%	62.00%	2.10E-03
miRNA and mRNA	PCA, SMOTE SVM (Linear Kernel)	92%	94%	92.50%	92.43%	85.00%	1.00E-07
mRNA and methylation	Feature Selection, SMOTE Random Forest	84%	94%	88.75%	88.00%	77.50%	3.43E-05
miRNA -methylation	PCA, SMOTE SVM (Radial Kernel)	93%	72%	82.25%	84.11%	64.50%	9.03E-04
miRNA - mRNA and methylation	**PCA, SMOTE SVM (Linear Kernel)**	**92%**	**93%**	**92.50%**	**92.47%**	**85.00%**	**9.92E-08**
miRNA - mRNA and methylation L	PCA, SMOTE SVM (Linear Kernel)	88%	87%	87.50%	87.53%	75.00%	1.29E-05
miRNA - mRNA and methylation H	PCA, SMOTE SVM (Linear Kernel)	92%	90%	90.75%	90.90%	81.50%	5.87E-07

As a result of evaluations, the combination of miRNA, mRNA, and methylation markers (d3), SVM, by using Smote and PCA at method selected as successor model.

In third model, ten miRNA biomarkers, namely *hsa-mir-142, hsa-mir-29c, hsa-mir-3124, hsa-mir-3130, hsa-mir-326, hsa-mir-331, hsa-mir-4419b, hsa-mir-4444, hsa-mir-4474, hsa-mir-4491, hsa-mir-4523, hsa-mir-625 and hsa-mir-766* are found as upregulated and 1 miRNA ,hsa-mir-203a, found as down-regulated. Hence, 11 miRNA markers have been used as a biomarker in our successor model to predict metastasis. In addition, 163 methylation and 1770 mRNA markers are selected in the final triple biomarker model. All miRNA biomarkers and their targets miRNA and Methylation information in their target genes are presented in the Appendixes (Table S2 and S3).

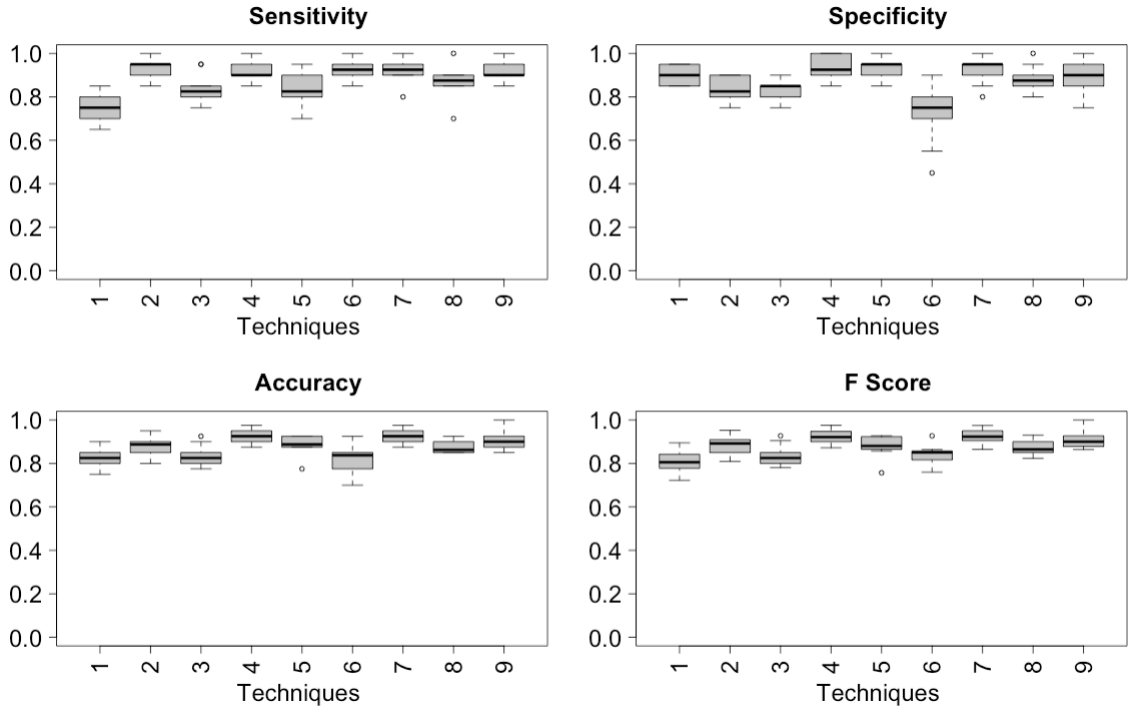

Figure 22: Comparison of Best Models for Each Biomarker Sets

1) The Performance of the predictive model by using miRNA, 2) The Performance of the predictive model by using mRNA, 3) The Performance of the predictive model by using Methylation, 4) The Performance of the predictive model by using miRNA and mRNA markers, 5) The Performance of the predictive model by using mRNA and methylation markers 6) The Performance of the predictive model by using miRNA and methylation markers 7) The Performance of the predictive model by using miRNA, mRNA and Methylation markers 8) The Performance of the predictive model by using miRNA, mRNA, and Hypo Methylation marker 9) The Performance of the predictive model by using miRNA, mRNA and Hyper Methylation markers.

4.11. Pathway Analysis

Evaluation of the overall results at the functional level is completed with an enrichment analysis. We used DAVID ((Dennis et al. 2003; D. W. Huang et al. 2007))Tools for biological interpretation of selected features used in selected "miRNA and mRNA" classification and "miRNA, mRNA, and Methylation" classification.

Based on the functional enrichment analysis, KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers are compared for "MiRNA and mRNA" with "miRNA, mRNA, and Methylation" to examine the reason higher precision and

consistency of addition of methylation. In the model with methylation markers, the significance of the Osteoclast, Rap1 Signaling Pathway, and Chemokine Signaling pathways increased (Figure 10). Osteoclast Differentiation also appealed within the top 15 pathways when all three biomarker categories were combined. In addition, Rap1 Signaling Pathway and Chemokine Signaling were listed in the top 3 among the most significant pathways (Table 3).

Table 3: Comparison of Top 15 Pathways of Different Biomarkers Sets.
P values of Osteoclast, Rap 1 Signaling Pathway, and Chemokine Signaling Pathways gradually increased after adding a new biomarker set. In addition, Rap1 Signaling Pathway and Chemokine Signaling were listed among the top three pathways with increasing significance with Osteoclast Differentiation. Other pathways with increasing significance, such as Cytokine-cytokine receptor interaction and Ras signaling pathway also observed.

	P-value		
	miRNA	miRNA-mRNA	miRNA-mRNA- Methylation
6.3.2.-		1.60×10^{-02}	
cAMP signaling pathway	7.40×10^{-04}		
Chemokine signaling pathway (*)		2.40×10^{-07}	$1.60\ 10^{-10}$
Cytokine-cytokine receptor interaction			$1.90\ 10^{-04}$
Endocytosis	4.30×10^{-04}	6.30×10^{-11}	$1.40\ 10^{-04}$
Focal adhesion	6.10×10^{-09}	$1.40\ 10^{-06}$	$2.80\ 10^{-07}$
Hepatitis B	4.40×10^{-09}		
HTLV-I infection	5.10×10^{-07}	3.60×10^{-13}	2.00×10^{-09}
MAPK signaling pathway	1.60×10^{-03}	6.70×10^{-11}	4.90×10^{-04}
Osteoclast differentiation (*)			2.90×10^{-14}
Pathways in cancer	4.60×10^{-13}	3.10×10^{-16}	1.20×10^{-12}
PI3K-Akt signaling pathway	7.00×10^{-05}	8.00×10^{-06}	1.90×10^{-05}
Proteoglycans in cancer	2.00×10^{-10}	5.70×10^{-10}	2.70×10^{-08}
R-HSA-212436	3.40×10^{-05}	6.00×10^{-03}	
R-HSA-983168	3.60×10^{-03}	3.80×10^{-05}	7.30×10^{-03}
Rap1 signaling pathway(*)	4.80×10^{-07}	4.30×10^{-06}	3.70×10^{-10}
Ras signaling pathway	3.80×10^{-06}	1.50×10^{-07}	3.00×10^{-08}
Regulation of actin cytoskeleton	1.80×10^{-03}	1.70×10^{-05}	1.50×10^{-04}
Viral carcinogenesis	9.70×10^{-06}	5.10×10^{-04}	3.80×10^{-04}

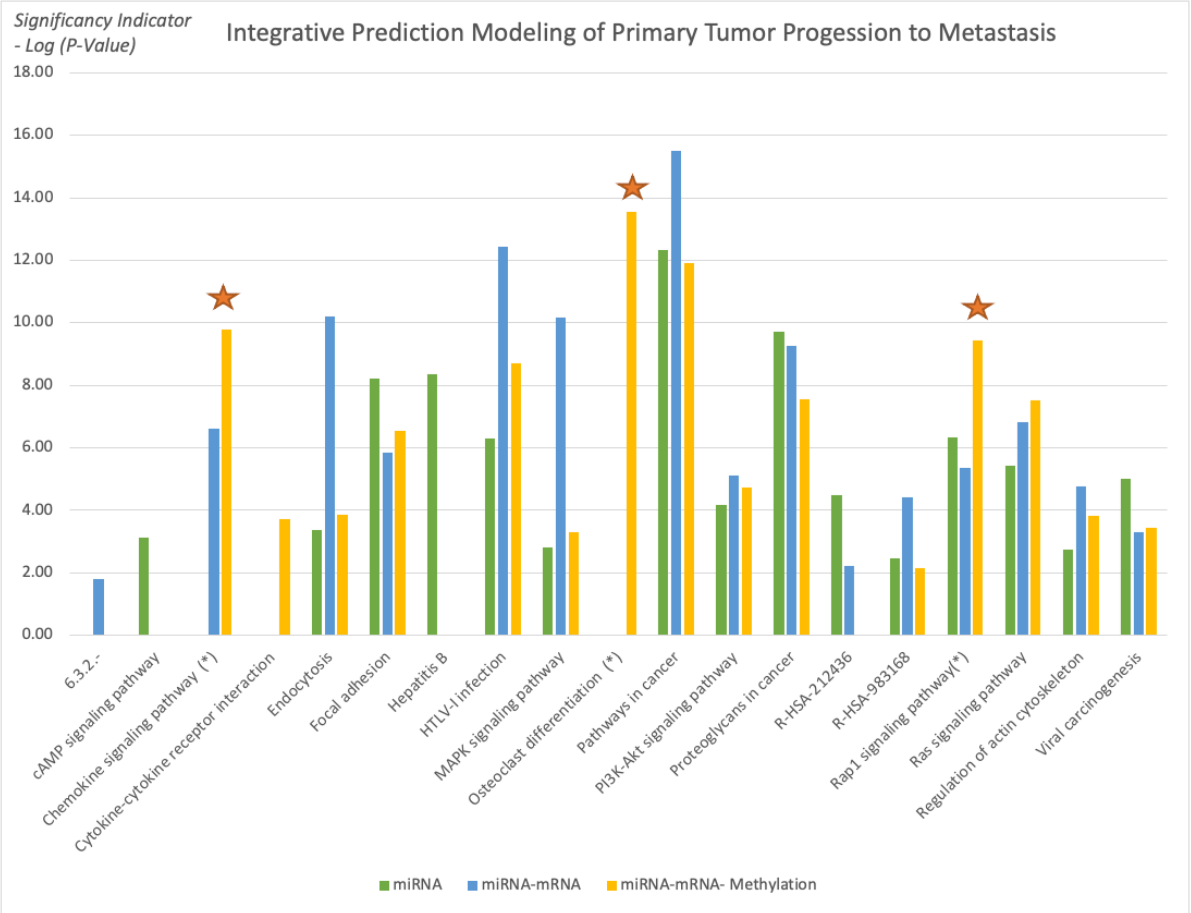

Figure 23: Significant Pathways Functionally Enriched in All Three Feature Sets
*As the new biomarker set **is** added, the significance of the pathways is evaluated. Osteoclast, Rap1 Signaling Pathway, and Chemokine Signaling Pathways' showed a significant increase in the third model.*

CHAPTER 5

5. CONCLUSIONS AND DISCUSSION

5.1. Discussion

Melanoma can be distinguished with visual assessment or through a short screening. Although there is an opportunity for the cure when detected early, treatment is challenging in later stages. Likewise, metastasis is an undesired outcome in such cases, and differential diagnosis is crucial for the treatment decision. So, the opportunity to diagnose metastatic melanomas in earlier stages may support therapeutic decisions, advise for more frequent and in-depth screening, and provide a higher chance of curing or preventing further metastatic progress.

This study shows that miRNA plays an essential role in the metastatic progression of primary melanoma and predicts metastasis outcomes with high accuracy. miRNA biomarkers anticipated metastatic results with an F-score of 82 %. Expansion of mRNA markers upon miRNA reached an F-score of 92 %. The ultimate model, which includes DNA methylation, results in a comparative F-score of 92 % but delivers a steady model with low variation over different trials. Moreover, the integrated evaluation of miRNA with mRNA and methylation biomarkers increases the model's predictive power. Another remarkable finding in this study is that boosting and bagging models' performance was better for miRNA signatures. However, we got higher prediction scores for neural networks and support vector machine classifiers when we added new mRNA and DNA methylation.

So far, different markers have been proposed to describe the molecular foundation of metastasis. DNA methylation, gene expression profiles, and microRNAs are used as markers of metastasis. In the literature, the predictive models for cancer metastasis were revealed by 2004. These initial studies were performed using gene expression profiles of the primary tumor collected by DNA microarray. After 2015 studies on miRNA expression levels and methylation data were published in the literature. Most of the studies of miRNA focused on identifying differentially expressed miRNAs for different cancer types based on statistical analysis.On the other hand, many more studies were reported for mRNA, where machine learning models apart from the statistical analysis were utilized. The prediction accuracy of metastasis observed to be between %50 and %86 percent.

For the methylation, there was one remarkable study. They identified 28 methylation markers for metastasis of primary colorectal cancer. On the other hand, several studies proposed different miRNA or proteins as significant for metastasis for melanoma. For example, miR-26b-5p (Wei et al. 2019), Mir-205-5p (Valentini et al. 2019), miR-26-5p (Wei et al. 2019), scavenger receptor class B type 1 (SR-BI) protein, and miR-29c-3p (Kinslechner et al. 2019) were found as a significant marker for melanoma metastasis. In addition, using serum levels of the cytokines IL-4, GM-CSF, DCD, and the Breslow thickness linear regression predicts metastatic outcome with a %83 accuracy (Mancuso et al. 2020). Also, prediction BAP1 mutation is used to identify metastatic outcomes by using whole slide images and a deep convolutional neural network (AUC: %90).

As can be seen, few studies try to predict metastatic outcomes of melanoma. Our results were compatible with them. Most of the studies in the literature are based on diagnostic purposes, so we focused on metastatic progress. There was not enough research that built up a generalized perspective and investigated the contribution of different genetic markers. On the other hand, we focused on this goal and examined the contribution of miRNA, mRNA, and Methylation to the metastatic outcome of primary melanoma cancer. Genetic material provides various biomarkers, which can be utilized either for diagnostic or predictive purposes. However, it is impossible to examine each marker, so we need to find an effective smallest biomarkers set that provide high precision. Our study is focused on this idea as bases and tried to evaluate the effectiveness of the proposed miRNA by comparing their predictive contribution with other possible biomarkers. From these points of view, our study provides valuable input to the field.

In machine learning studies, undersampling techniques are also used to deal with class imbalance issues. So, we performed oversampling and undersampling methods and evaluated their outcomes. The SMOTE, a synthetic minority oversampling method based on the k-nearest-neighbors, has been tested with different k values between 3 to 6, and the final k is chosen as 3. We used the 1:2 ratio for oversampling of the minority class. Under the given circumstances, we generated similar results for undersampling and oversampling. Overall, our results present satisfactory evidence that the synthetic minority oversampling technique can also be applicable for prediction studies for genomics data.

Although the results are compatible with SMOTE, this technique created a shortcoming for our study; that is, synthetic data generation may cause overfitting of the machine learning model. So evaluation of our predictive model in new data sets would be valuable to eliminate this issue. Unfortunately, our data size was limited and skewed. In order to

continue extended research on metastasis prediction, new integrative datasets that hold different genetic markers should be collected. We can observe many more samples for metastastic tumors in many datasets, but primary tumor samples are restricted. This also prevents the diversity of the machine learning used for prediction. Therefore, collecting more samples for primary tumors will also contribute to developing more effective predictors.

As our model is based on the differences between primary and metastatic melanomas, the markers identified here can be used for differential diagnosis. We believe it will become possible to predict melanomas with metastatic potential (prediction of prognosis). In those cases, several actions can be taken in the clinic, such as intensive scanning for metastasis or frequent follow-ups of patients. In the future, patients with higher risk can be offered prevention from metastasis with gene therapies based on emerging technologies like miRNA therapies or gene editing.

Our study is initiated with an iterative approach to include more biomarkers set upon miRNA signatures. In the initial run, we included mRNA, and then we included methylation biomarkers. The triple model resulted in the highest predictive value. After investigating biomarkers with a recursive approach, we also inquire predictive accuracy of different combinations of biomarker sets. All biomarkers generated relatively high performant models with above %80 F values for all single biomarker sets. However, the mRNA model was more potent than both MiRNA and Methylation predictive models. The sensitivity of both mRNA and miRNA was quite similar, but variance specificity was higher in the miRNA model for different trials. The addition of methylation markers upon miRNA and mRNA does not improve the predictive accuracy of miRNA and mRNA.

Nevertheless, the combination of miRNA and mRNA exceeds the predictive values of models. On the other hand, the addition of methylation reduces the variance of model scores among different trials. Finally, the selection of hypo-methylated and hyper-methylated genes reduces the predictive scores.

During the study, we have identified 128 miRNA for model c1 (miRNA model), 18 mi-RNA in model d2 (mRNA and miRNA model), and ten miRNA for model d3 (miRNA, mRNA, and Methylation). There were only a few studies that we can identify on modeling melanoma metastasis in the literature. Valentine et al. (Valentini et al. 2019) found that mir-205, which is also listed in our attribute list for model c1 (miRNA model), is significant in distinguishing metastatic melanoma. Similarly, Wang et al. (Yanqian Wang et al. 2019) mentioned that miR-29c is a suppressor of non-coding RNA taurine-upregulated gene 1 (TUG1), which is identified as a prognostic marker of metastatic

melanoma. Our results also support this finding since miR-29c is one of the markers that we included in all models c1, d2, and d3.

Moreover, Mancuso et al. (Mancuso et al. 2020) used serum levels of the cytokines IL-4, GM-CSF, and DCD to model linear regression, which achieves 80% accuracy. Another study by Zhang et al. approaches the issue from a different perspective and tries to predict BAP 1 mutation for predicting metastatic risk via whole slide images. In their study, AUC is reported as 0.90. As it can be seen, our results, which aim to distinguish metastatic melanoma from the primary tumor, are compatible with other similar studies in the literature.

This study focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated effect of miRNA, mRNA, and DNA methylation. We used the TCGA melanoma dataset to predict metastatic melanoma samples by assessing a set of predictive models. Throughout the study, differentially expressed miRNA, mRNA, and methylation signatures are used as biomarkers. The highest performing models' selected biomarkers are further analyzed for the biological interpretation of functional enrichment and determining regulatory networks. So we focused on gradually including new feature sets. We have performed functional enrichment analysis to reveal our evaluation pattern for including a new biomarkers set. The functional enrichment of KEGG, Reactome, EC Number, and Biocarta Pathways of selected biomarkers and compare sets for "miRNA," "miRNA and mRNA," and "miRNA, mRNA, and Methylation" we tried to search for the reason behind the higher precision and consistency achieved after addition of methylation.

Osteoclast, **Rap1 Signaling Pathway,** and **Chemokine Signaling Pathways** significantly increased and listed the top 15 pathways when all three biomarker sets were used for modeling. So combined model populates selected biomarkers on the metastasis-associated pathways of melanoma.

Osteoclasts are multinucleated cells responsible for bone resorption. Molecular pathways involved in osteoclast proliferation, differentiation, and survival are essential players of bone metastasis. **Osteoclast Differentiation** is a systemic pathway that controls bone renovation. Since the main metastasis sites for melanoma cancer include bone, liver, lung, and skin/muscle [54], functional enrichment of osteoclast-related pathways within top-level pathways is a supporting finding for our study design.

Ras-associated protein-1 (Rap1) is an essential regulator for basic cell functions such as cellular migration and polarization. This pathway has critical role in tumor metastasis, so

such an increase in the significance level is also critical for the metastatic outcome (Y. L. Zhang et al. 2017).

Chemokines are involved in controlling the migration of cells during normal processes of tissue maintenance or development. The chemokine-receptor system plays critical roles in various physiological processes, including immune homeostasis, inflammatory responses, and cancer progression. Chemokines have essential roles in tumor progression, involved in the growth of many cancers and metastasis (Sarvaiya et al., 2013).

Since the initial discovery of the relationship between cancer and miRNA signatures, many studies have shown that miRNA has a critical role in regulating genes and, thus, has a critical role in tumorigenesis. Today, many techniques for the early detection and diagnosis of tumors are available. Still, when invasive procedures are required for diagnosis or treatment, it is vital to know the tumor's metastatic potential to estimate the risks vs. benefits of the procedure. Also, in the later stages of tumor development, any information about the metastatic status of the late-stage tumors is required for deciding between therapy choices. Hence, the miRNA reported in this study can be candidates for therapeutic targets of melanoma metastasis.

5.2. Limitations

One limitation of the study was the data imbalance and small sample size. We validated and tested our models in restricted data size since we could not access additional data sets on GEO or CGC, combining all three markers at the time of the study. We utilized oversampling techniques and ran the overall process multiple times to reduce the bias to address this limitation. Additionally, we were able to compare various machine learning models as they were appropriate for the data size in the study. However, we realize that deep learning methods would be competitive with these techniques. Therefore, repetition of the study with a balanced or more extensive data set in the future can further validate the biomarkers reported here.

5.3. Future Research

Comparing significant pathways of the biomarker’s groups can give information about the featured pathways for Melanoma Metastasis. In this study, we only focus on pathway analysis of specific groups. In future studies, pathways analysis of mRNA, Methylation, miRNA- Methylation, and mRNA- Methylation can be done, and the results can be compared with those included in this study.

In addition, the final predictive model of this study can be extended to conduct an ablation study, where input modalities would be removed, to see how much predictive accuracy can be extracted from individual data modalities.

5.4. Conclusion

Since this initial discovery of the gene regulation mechanism of the microRNAs, many studies have been contacted to reveal out their impact. In this study we tried to investigate contriutoon of miRNAs for metastatic outcome of the primary melanoma cancer.

Untill now, there were many studies on impact molecular biomarkers and gene regulation mechanism of the microRNAs on various cancer types. However, there was not enough studies to investigate metastic progress from genereralized perspective and to investigate the contribution of different genetic markers.

We consternate on molecular foundation of metastasis by combining all miRNA and mRNA and Metthlaton as possible markers, for idenifing metastasis of melanoma which is a cancer with a rapid increase in incidence and high mortality. In Melanoma, metastasis is frequent and deadly. Therefore predicting possiple outcome of the melanoma in early stages may help to make better terapatic decisions.

We used TCGA melanoma dataset by combining miRNA and mRNA expressions with Methylation Beta Values.We developed Predictive models by using combinations of all these biomarker groups. Different techniques for dimensional reduction and class imbalance solutions are applied. We trained the model by using various machine learning algorithms, and compare their performance. Our goal was comparing different biomarkers and investigating their contribution to metastasis. By this way we want to identify miRNA signatures and reveal out their impact.

According to our results combining miRNAs with mRNA and Methylation improves models predictive accuracy and precision. However, miRNAs alone proposed in this study, predicts the melanoma metastasis with high accuracy as well. Therefore, miRNAs proposed in this study can be effective smallest set to predicts the metatastic outcome.

REFERENCES

Alfaro, Esteban, Matías Gáamez, and Noelia García. 2013. "Adabag: An R Package for Classification with Boosting and Bagging." *Journal of Statistical Software* 54(2): 1–35. https://www.jstatsoft.org/index.php/jss/article/view/v054i02/adabag_An_R_Package_for_Classification_with_Boosting_and_Bagging.pdf (June 23, 2021).

Alfaro, Esteban, Matias Gamez, and Noelia Garcia. 2018. "CRAN - Package Adabag." *CRAN R Project*. https://cran.r-project.org/web/packages/adabag/index.html (June 23, 2021).

American Cancer Society. 2016. *European Commission Melanoma Skin Cancer. Atlanta*. http://ec.europa.eu/eurostat/statistics-explained/index.php/Causes_of_death_statistics#Further_Eurostat_information (June 3, 2021).

Barrier, Alain et al. 2005. "Gene Expression Profiling of Nonneoplastic Mucosa May Predict Clinical Outcome of Colon Cancer Patients." *Diseases of the Colon and Rectum* 48(12): 2238–48. https://journals.lww.com/dcrjournal/Fulltext/2005/48120/Gene_Expression_Profiling_of_Nonneoplastic_Mucosa.10.aspx (January 15, 2022).

Bidus, Michael A. et al. 2006. "Prediction of Lymph Node Metastasis in Patients with Endometrioid Endometrial Cancer Using Expression Microarray." *Clinical Cancer Research*.

"Bioinformatics Pipeline: Methylation Liftover Pipeline - GDC Docs." https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Methylation_LO_Pipeline/ (January 17, 2022).

"Bioinformatics Pipeline: MiRNA Analysis - GDC Docs." https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/ (January 17, 2022).

"Bioinformatics Pipeline: MRNA Analysis - GDC Docs." https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/ (January 17, 2022).

"BiomaRt Bioc R Package." http://www.ensembl.org/info/data/biomart/biomart_r_package.html (January 17,

2022).

Burton, Mark, Mads Thomassen, Qihua Tan, and Torben A. Kruse. 2012. “Prediction of Breast Cancer Metastasis by Gene Expression Profiles: A Comparison of Metagenes and Single Genes.” *Cancer Informatics*.

Calin, George Adrian et al. 2002. “Frequent Deletions and Down-Regulation of Micro-RNA Genes MiR15 and MiR16 at 13q14 in Chronic Lymphocytic Leukemia.” *Proceedings of the National Academy of Sciences of the United States of America* 99(24): 15524–29. https://pubmed.ncbi.nlm.nih.gov/12434020/ (January 17, 2022).

Cancer Research UK. “Melanoma Skin Cancer Incidence Statistics.” https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer#:~:text=Melanoma skin cancer incidence,new cancer cases (2017). (June 3, 2021).

Carter, R. L. 1974. “The Spread of Tumours in the Human Body.” *Journal of Clinical Pathology* 27(5): 432–33.

Cavalieri, Duccio et al. 2007. “Analysis of Gene Expression Profiles Reveals Novel Correlations with the Clinical Course of Colorectal Cancer.” *Oncology research* 16(11): 535–48. https://pubmed.ncbi.nlm.nih.gov/18306933/ (January 15, 2022).

Chang, Jee Won et al. 2008. “Prediction of Lymph Node Metastasis Using the Combined Criteria of Helical CT and MRNA Expression Profiling for Non-Small Cell Lung Cancer.” *Lung Cancer* 60(2): 264–70. http://linkinghub.elsevier.com/retrieve/pii/S0169500207005971.

Chen, Ll et al. 2009. “Cancer Metastasis Networks and the Prediction of Progression Patterns.” *British Journal of Cancer* 101: 749–58.

Chowdhury, Ashis, Kalidindi K. Raju, Swathi Kalurupalle, and Sundaresan Tharun. 2012. “Both Sm-Domain and C-Terminal Extension of Lsm1 Are Important for the RNA-Binding Activity of the Lsm1–7–Pat1 Complex.” *RNA* 18(5): 936. /pmc/articles/PMC3334702/ (January 17, 2022).

Craig, Editor Jeffrey M et al. 2011. “Epigenetics : A Reference Manual | Book Epigenetics : A Reference Manual | Book.” *DNA Sequence*: 2–5.

Damsky, William E., Jr., Lara E. Rosenbaum, and Marcus Bosenberg. 2011. “Decoding Melanoma Metastasis.” *Cancers* 3(1): 126. /pmc/articles/PMC3756353/ (July 27, 2021).

Dehnavi, Alireza Mehri, Mohammad Reza Sehhati, Hossein Rabbani, and Alireza Mehridehnavi. 2013. “Hybrid Method for Prediction of Metastasis in Breast Cancer

Patients Using Gene Expression Signals." *J Med Signals Sens* 3(2): 79–86.

Dennis, Glynn et al. 2003. "DAVID: Database for Annotation, Visualization, and Integrated Discovery." *Genome biology* 4(5). https://pubmed.ncbi.nlm.nih.gov/12734009/ (June 22, 2021).

"Downloading Files - GDC Docs." https://docs.gdc.cancer.gov/API/Users_Guide/Downloading_Files/ (January 17, 2022).

Faraji, Farhoud et al. 2014. "An Integrated Systems Genetics Screen Reveals the Transcriptional Structure of Inherited Predisposition to Metastatic Disease." *Genome research* 24(2): 227–40. https://pubmed.ncbi.nlm.nih.gov/24322557/ (January 17, 2022).

Fernández, Alberto, Salvador García, Francisco Herrera, and Nitesh V. Chawla. 2018. "SMOTE: Synthetic Minority over-Sampling Technique: Journal of Artificial Intelligence Research: Vol 16, No 1." *Journal of Artificial Intelligence Research* 61: 863–905. https://dl.acm.org/doi/10.5555/1622407.1622416 (December 26, 2019).

"GDC Application Programming Interface (API) | NCI Genomic Data Commons." https://gdc.cancer.gov/developers/gdc-application-programming-interface-api (January 17, 2022).

"GDC Data Portal." https://portal.gdc.cancer.gov/ (January 17, 2022).

Gonzalo, Susana. 2010. "Epigenetic Alterations in Aging." *Journal of Applied Physiology* 109(2): 586–97. https://www.physiology.org/doi/10.1152/japplphysiol.00238.2010.

Goossens-Beumer, Inès J. et al. 2015. "MicroRNA Classifier and Nomogramfor Metastasis Prediction in Colon Cancer." *Cancer Epidemiology Biomarkers and Prevention*.

Harris, Curtis C. 1991. "Molecular Basis of Multistage Carcinogenesis." *Princess Takamatsu symposia* 22(22): 3–19. http://www.ncbi.nlm.nih.gov/pubmed/1844248.

Hayes, Josie, Pier Paolo Peruzzi, and Sean Lawler. 2014. "MicroRNAs in Cancer: Biomarkers, Functions and Therapy." *Trends in Molecular Medicine* 20(8): 460–69.

He, Kaijie, Tong Xu, and Amir Goldkorn. 2011. "Cancer Cells Cyclically Lose and Regain Drug-Resistant Highly Tumorigenic Features Characteristic of a Cancer Stem-like Phenotype." *Molecular cancer therapeutics* 10(6): 938–48. https://pubmed.ncbi.nlm.nih.gov/21518726/ (January 17, 2022).

Huang, Da Wei et al. 2007. "DAVID Bioinformatics Resources: Expanded Annotation

Database and Novel Algorithms to Better Extract Biology from Large Gene Lists." *Nucleic Acids Research* 35(SUPPL.2).

Huang, Hsi Yuan et al. 2020. "MiRTarBase 2020: Updates to the Experimentally Validated MicroRNA-Target Interaction Database." *Nucleic acids research* 48(D1): D148–54. https://pubmed.ncbi.nlm.nih.gov/31647101/ (January 17, 2022).

Institute, National Cancer. 2020. "The Cancer Genome Atlas Program - National Cancer Institute." *National Institute of Health*. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (June 22, 2021).

Jansson, Martin D., and Anders H. Lund. 2012. "MicroRNA and Cancer." *Molecular Oncology* 6(6): 590–610.

Kan, Takatsugu et al. 2004. "Prediction of Lymph Node Metastasis with Use of Artificial Neural Networks Based on Gene Expression Profiles in Esophageal Squamous Cell Carcinoma." *Annals of Surgical Oncology* 11(12): 1070–78. http://www.springerlink.com/index/10.1245/ASO.2004.03.007.

Karatzoglou, Alexandros, Kurt Hornik, Alex Smola, and Achim Zeileis. 2004. "Kernlab - An S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11(1): 1–20. https://www.jstatsoft.org/index.php/jss/article/view/v011i09/v11i09.pdf (June 23, 2021).

Karatzoglou, Alexandros, Alex Smola, and Kurt Hornik. 2016. "Kernlab: Kernel-Based Machine Learning Lab. R Package Kernlab. Version 0.9-29." https://cran.r-project.org/package=kernlab (June 23, 2021).

Kinslechner, Katharina et al. 2019. "Loss of SR-BI down-Regulates MITF and Suppresses Extracellular Vesicle Release in Human Melanoma." *International Journal of Molecular Sciences* 20(5).

Leong, Stanley P.L. et al. 2006. "Clinical Patterns of Metastasis." *Cancer metastasis reviews* 25(2): 221–32. https://pubmed.ncbi.nlm.nih.gov/16770534/ (January 17, 2022).

Di Leva, Gianpiero, and Carlo M. Croce. 2013. "MiRNA Profiling of Cancer." *Current opinion in genetics & development* 23(1): 3. /pmc/articles/PMC3632255/ (January 17, 2022).

Li, Jie et al. 2010. "Identification of High-Quality Cancer Prognostic Markers and Metastasis Network Modules." *Nature Communications* 1(4): 1–8. http://www.nature.com/doifinder/10.1038/ncomms1033.

Lim, Jong Baeck et al. 2015. "Serum ENA78/CXCL5, SDF-1/CXCL12, and Their

Combinations as Potential Biomarkers for Prediction of the Presence and Distant Metastasis of Primary Gastric Cancer." *Cytokine* 73(1): 16–22. http://www.spandidos-publications.com/10.3892/ijo.2019.4699 (June 19, 2019).

Lin, Yu Hsin et al. 2007. "Multiple Gene Expression Classifiers from Different Array Platforms Predict Poor Prognosis of Colorectal Cancer." *Clinical cancer research : an official journal of the American Association for Cancer Research* 13(2 Pt 1): 498–507. https://pubmed.ncbi.nlm.nih.gov/17255271/ (January 15, 2022).

Ma, Li, and Robert A. Weinberg. 2008. "Micromanagers of Malignancy: Role of MicroRNAs in Regulating Metastasis." *Trends in Genetics* 24(9): 448–56.

Majka, Michal, and Michal Majka. 2020. "CRAN - Package Naivebayes." *CRAN R Project*. https://cran.r-project.org/web/packages/naivebayes/index.html (June 23, 2021).

Mancuso, Filippo et al. 2020. "Serum Markers Improve Current Prediction of Metastasis Development in Early-stage Melanoma Patients: A Machine Learning-based Study." *Molecular Oncology* 14(8): 1705–18.

McAnena, Peter et al. 2019. "Circulating MicroRNAs MiR-331 and MiR-195 Differentiate Local Luminal a from Metastatic Breast Cancer." *BMC Cancer* 19(1): 436. https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5636-y (November 20, 2019).

Melchers, L. J. et al. 2015. "Identification of Methylation Markers for the Prediction of Nodal Metastasis in Oral and Oropharyngeal Squamous Cell Carcinoma." *Epigenetics* 10(9): 850–60.

Moriya, Yasumitsu et al. 2009. "Prediction of Lymph Node Metastasis by Gene Expression Profiling in Patients with Primary Resected Lung Cancer." *Lung Cancer*.

Oppenheimer, Steven B. 1983. "Cancer: A Biological and Clinical Introduction." *Transactions of the American Microscopical Society* 102(2): 182. https://www.jstor.org/stable/3225891?origin=crossref.

———. 2006. "Cellular Basis of Cancer Metastasis: A Review of Fundamentals and New Advances." *Acta Histochemica* 108(5): 327–34.

Radwan, Wafaa M. et al. 2013. "Peripheral Blood Mammaglobin Gene Expression for Diagnosis and Prediction of Metastasis in Breast Cancer Patients." *Asia-Pacific Journal of Clinical Oncology*.

Ramaswamy, Sridhar, Ken N. Ross, Eric S. Lander, and Todd R. Golub. 2003. "A Molecular Signature of Metastasis in Primary Solid Tumors." *Nature genetics* 33(1):

49–54. https://pubmed.ncbi.nlm.nih.gov/12469122/ (January 15, 2022).

Rickman, Ds et al. 2008. “Prediction of Future Metastasis and Molecular Characterization of Head and Neck Squamous-Cell Carcinoma Based on Transcriptome and Genome Analysis by Microarrays.” *Oncogene* 27.

Ripley, Brian. 2021. “Feed-Forward Neural Networks and Multinomial Log-Linear Models [R Package Nnet Version 7.3-16].” https://cran.r-project.org/package=nnet (June 23, 2021).

Ripley, Brian, and William Venables. 2021. “CRAN - Package Nnet.” *CRAN R Project*. https://cran.r-project.org/web/packages/nnet/index.html (June 23, 2021).

Roessler, Stephanie et al. 2010. “A Unique Metastasis Gene Signature Enables Prediction of Tumor Relapse in Early-Stage Hepatocellular Carcinoma Patients.” *Cancer Research*.

“Sample Type Codes | NCI Genomic Data Commons.” https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes (January 17, 2022).

Sarvaiya, Purvaba J. et al. 2013. “Chemokines in Tumor Progression and Metastasis.” *Oncotarget* 4(12): 2171–85. /pmc/articles/PMC3926818/?report=abstract (October 14, 2020).

Schell, Michael J. et al. 2016. “A Composite Gene Expression Signature Optimizes Prediction of Colorectal Cancer Metastasis and Outcome.” *Clinical Cancer Research*.

Shalaby, Tarek, Giulio Fiaschetti, Martin Baumgartner, and Michael A. Grotzer. 2014. “MicroRNA Signatures as Biomarkers and Therapeutic Target for CNS Embryonal Tumors: The Pros and the Cons.” *International journal of molecular sciences* 15(11): 21554–86.

Shen, Jun, Sanford A. Stass, and Feng Jiang. 2013. “MicroRNAs as Potential Biomarkers in Human Solid Tumors.” *Cancer Letters* 329(2): 125–36.

Siriseriwan, Wacharasak. 2019. “A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE [R Package Smotefamily Version 1.3.1].” https://cran.r-project.org/package=smotefamily (June 23, 2021).

Son, Joo-Hyuk et al. 2015. “Prediction of Lymph Node Metastasis in Patients with Apparent Early Endometrial Cancer.” *Obstetrics & Gynecology Science* 58(5): 385. http://synapse.koreamed.org/DOIx.php?id=10.5468/ogs.2015.58.5.385.

De Souza, Marilesia Ferreira et al. 2017. “Circulating MRNAs and MiRNAs as Candidate

Markers for the Diagnosis and Prognosis of Prostate Cancer." *PLoS ONE* 12(9).

Stahlhut, Carlos, and Frank J Slack. 2013. "MicroRNAs and the Cancer Phenotype: Profiling, Signatures and Clinical Implications." *Genome Medicine* 5(12): 111.

Takada, Masahiro et al. 2012. "Prediction of Axillary Lymph Node Metastasis in Primary Breast Cancer Patients Using a Decision Tree-Based Model." *BMC Medical Informatics and Decision Making* 12(1): 54. http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-54.

The Cancer Genome Atlas Network. 2015. "Genomic Classification of Cutaneous Melanoma." *Cell* 161(7): 1681–96. http://dx.doi.org/10.1016/j.cell.2015.05.044 (June 22, 2021).

"The Cancer Genome Atlas Program - National Cancer Institute." https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (January 17, 2022).

Thomassen, Mads et al. 2007. "Comparison of Gene Sets for Expression Profiling: Prediction of Metastasis from Low-Malignant Breast Cancer." *Clinical Cancer Research*.

Tonini, Tiziana, Francesca Rossi, and Pier Paolo Claudio. 2003. "Molecular Basis of Angiogenesis and Cancer." *Oncogene* 22(43): 6549–56.

United States Cancer Statistics. "CDC United States Cancer Statistics Data Visualizations." https://www.cdc.gov/cancer/uscs/dataviz/index.htm (June 3, 2021).

Valentini, Virginia et al. 2019. "MiRNAs as Potential Prognostic Biomarkers for Metastasis in Thin and Thick Primary Cutaneous Melanomas." *Anticancer Research* 39(8): 4085–93. http://ar.iiarjournals.org/lookup/doi/10.21873/anticanres.13566 (November 17, 2019).

Volinia, Stefano et al. 2006. "A MicroRNA Expression Signature of Human Solid Tumors Defines Cancer Gene Targets." *Proceedings of the National Academy of Sciences of the United States of America* 103(7): 2257–61. https://pubmed.ncbi.nlm.nih.gov/16461460/ (January 17, 2022).

Wacharasak Siriseriwan. 2019. "CRAN - Package Smotefamily." *CRAN R Project*. https://cran.r-project.org/web/packages/smotefamily/index.html (June 23, 2021).

Wang, Rong, Xiao-Feng Chen, and Yong-Qian Shu. 2015. "Prediction of Non-Small Cell Lung Cancer Metastasis-Associated MicroRNAs Using Bioinformatics." *Am J Cancer Res* 5(1): 32–51. www.ajcr.us.

Wang, Y et al. 2005. “Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer.” *www.thelancet.com* 365.

Wang, Yanqian et al. 2019. “Long Non-Coding RNA TUG1 Recruits MiR-29c-3p from Its Target Gene RGS1 to Promote Proliferation and Metastasis of Melanoma Cells.” *International Journal of Oncology* 54(4): 1317–26. http://www.spandidos-publications.com/10.3892/ijo.2019.4699 (November 17, 2019).

Wang, Yixin et al. 2004. “Gene Expression Profiles and Molecular Markers to Predict Recurrence of Dukes’ B Colon Cancer.” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 22(9): 1564–71. https://pubmed.ncbi.nlm.nih.gov/15051756/ (January 15, 2022).

Watanabe, Toshiaki et al. 2010. “Prediction of Liver Metastasis after Colorectal Cancer Using Reverse Transcription-Polymerase Chain Reaction Analysis of 10 Genes.” *European Journal of Cancer*.

Wei, Chuan-Yuan Yuan et al. 2019. “TRIM44 Activates the AKT/MTOR Signal Pathway to Induce Melanoma Progression by Stabilizing TLR4.” 38(1): 137. https://jeccr.biomedcentral.com/articles/10.1186/s13046-019-1138-7 (November 17, 2019).

Willis, R A, and F R C P Pp. 1953. “The Spread of Tumours in the Human Body.” *Postgraduate Medical Journal* 29(329): 160. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2500336/ (January 17, 2022).

World Cancer Research Fund. “Skin Cancer | World Cancer Research Fund International.” https://www.wcrf.org/dietandcancer/skin-cancer/ (June 3, 2021).

Wright, Marvin N., Stefan Wager, and Philipp Probst. 2021. “CRAN - Package Ranger.” *CRAN R Project*. https://cran.r-project.org/web/packages/ranger/index.html (June 23, 2021).

Wright, Marvin N., and Andreas Ziegler. 2017. “Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77(1).

Xi, Liqiang et al. 2005. “Prediction of Lymph Node Metastasis by Analysis of Gene Expression Profiles in Primary Lung Adenocarcinomas.” *Clinical cancer research : an official journal of the American Association for Cancer Research* 11(11): 4128–35. http://www.ncbi.nlm.nih.gov/pubmed/15930348.

Yamasaki, Makoto et al. 2007. “The Gene Expression Profile Represents the Molecular Nature of Liver Metastasis in Colorectal Cancer.” *International Journal of Oncology* 30(1): 129–38. http://www.spandidos-

publications.com/10.3892/ijo.30.1.129/abstract (January 15, 2022).

Yang, Bingyi et al. 2016. "Predicting Lymph Node Metastasis in Endometrial Cancer Using Serum CA125 Combined with Immunohistochemical Markers PR and Ki67, and a Comparison with Other Prediction Models." *PLoS ONE*.

Yoshida, Tsuyoshi et al. 2010. "Clinical Omics Analysis of Colorectal Cancer Incorporating Copy Number Aberrations and Gene Expression Data." *Cancer informatics* 9: 147–61. https://pubmed.ncbi.nlm.nih.gov/20706620/ (January 15, 2022).

Zemmour, Christophe et al. 2015. "Prediction of Early Breast Cancer Metastasis from Dna Microarray Data Using High-Dimensional Cox Regression Models." *Cancer Informatics*.

Zhang, Fang Fang et al. 2011. "Physical Activity and Global Genomic DNA Methylation in a Cancer-Free Population." *Epigenetics* 6(3): 293–99.

Zhang, Hongrun et al. 2020. "Piloting a Deep Learning Model for Predicting Nuclear BAP1 Immunohistochemical Expression of Uveal Melanoma from Hematoxylin-and-Eosin Sections." *Translational Vision Science and Technology* 9(2): 1–13.

Zhang, Li et al. 2015. "A MicroRNA-Based Prediction Model for Lymph Node Metastasis in Hepatocellular Carcinoma." *Oncotarget* 7(3). www.impactjournals.com/oncotarget.

Zhang, Yi Lei et al. 2017. "Roles of Rap1 Signaling in Tumor Cell Migration and Invasion." *Cancer Biology and Medicine* 14(1): 90–99. /pmc/articles/PMC5365179/?report=abstract (October 14, 2020).

Zhou, Weiying et al. 2014. "Cancer-Secreted MiR-105 Destroys Vascular Endothelial Barriers to Promote Metastasis." *Cancer cell* 25(4): 501–15. https://pubmed.ncbi.nlm.nih.gov/24735924/ (January 15, 2022).

APPENDICES

APPENDIX A

Prediction Model Results

Table 4: Model Prediction Results for Each Experiment Cycle and Technique in Unseen Test Data

		Model	Sensitivity	Specificity	PValue	Accuracy	F Value
miRNA	**A1**	Neural Network	80%	75%	2.50E-03	77.25%	77.50%
	A1	SVM(Linear)	78%	72%	6.69E-03	74.75%	75.18%
	A1	SVM (Polynomial)	79%	74%	1.45E-03	76.00%	76.38%
	A1	SVM (Radial)	79%	72%	5.91E-03	75.50%	75.52%
	A1	Random Forest	75%	90%	1.70E-04	82.25%	80.79%
	A1(*)	AdaBoost	78%	79%	1.59E-03	78.25%	77.96%*
	A1	Naïve Bayes	56%	89%	8.77E-03	72.50%	66.98%
	B1	Neural Network	72%	80%	3.39E-03	75.75%	74.61%
	B1	SVM(Linear)	72%	77%	3.78E-03	74.50%	73.65%
	B1	SVM (Polynomial)	68%	78%	1.28E-02	73.00%	71.09%
	B1	SVM (Radial)	65%	80%	2.24E-02	72.00%	69.06%
	B1	Random Forest	71%	92%	2.59E-04	81.00%	78.66%
	B1(*)	AdaBoost	78%	84%	3.11E-04	80.75%	79.91%*
	B1	Naïve Bayes	56%	91%	6.81E-03	73.50%	67.53%
	C1	Neural Network	81%	73%	3.25E-03	77.00%	77.87%
	C1	SVM(Linear)	77%	74%	5.21E-03	75.25%	75.53%
	C1	SVM (Polynomial)	81%	67%	5.79E-03	73.50%	75.24%
	C1	SVM (Radial)	79%	75%	1.95E-03	77.00%	77.40%
	C1	Random Forest	85%	80%	2.37E-04	82.25%	82.63%
	C1(**)	AdaBoost	90%	73%	4.58E-04	81.25%	82.71%*
	C1	Naïve Bayes	92%	62%	2.61E-03	76.50%	79.57%
	D1	Neural Network	68%	79%	8.79E-03	73.25%	71.33%
	D1	SVM(Linear)	80%	74%	3.97E-03	76.75%	77.46%
	D1	SVM (Polynomial)	70%	75%	1.35E-02	72.25%	71.46%
	D1	SVM (Radial)	59%	87%	1.24E-02	72.75%	68.15%
	D1	Random Forest	63%	91%	2.38E-03	77.00%	72.97%
	D1(*)	AdaBoost	81%	80%	7.85E-03	80.25%	80.30%*
	D1	Naïve Bayes	70%	80%	2.98E-02	75.00%	73.29%
miRNA and mRNA	**A2(*)**	Neural Network	85%	80%	1.68E-04	82.50%	82.80%*
	A2	SVM(Linear)	87%	77%	1.79E-04	81.75%	82.67%
	A2	SVM (Polynomial)	86%	78%	9.34E-05	82.00%	82.58%
	A2	SVM (Radial)	85%	77%	4.10E-03	81.00%	81.76%
	A2	Random Forest	81%	86%	8.26E-05	83.25%	82.71%
	A2	AdaBoost	86%	76%	2.83E-04	80.75%	81.43%
	A2	Naïve Bayes	59%	92%	6.59E-03	75.25%	70.09%
	B2(*)	Neural Network	80%	90%	2.41E-05	84.75%	83.92%*
	B2	SVM(Linear)	81%	88%	5.67E-05	84.50%	83.80%
	B2	SVM (Polynomial)	78%	94%	2.24E-04	85.50%	83.81%
	B2	SVM (Radial)	86%	73%	8.66E-04	79.00%	80.21%
	B2	Random Forest	72%	93%	1.70E-04	82.00%	79.65%
	B2	AdaBoost	76%	90%	3.73E-04	82.75%	81.18%
	B2	Naïve Bayes	56%	95%	6.47E-03	75.25%	68.43%
	C2	Neural Network	95%	81%	3.52E-05	87.75%	88.69%
	C2(*)	SVM(Linear)	90%	91%	2.40E-06	90.00%	89.93%*
	C2	SVM (Polynomial)	96%	75%	1.24E-04	85.00%	86.57%
	C2	SVM (Radial)	96%	35%	2.08E-01	65.25%	74.46%

	C2	Random Forest	100%	56%	8.45E-04	77.50%	81.63%
	C2	AdaBoost	82%	64%	1.13E-02	72.75%	74.94%
	C2	Naïve Bayes	100%	26%	1.06E-01	62.50%	72.71%
	D2	Neural Network	93%	88%	1.40E-06	90.25%	90.50%
	D2(**)	SVM(Linear)	92%	94%	1.00E-07	92.50%	92.43%*
	D2	SVM (Polynomial)	93%	77%	1.59E-04	84.75%	86.03%
	D2	SVM (Radial)	99%	3%	5.31E-01	50.75%	66.69%
	D2	Random Forest	96%	60%	2.50E-03	77.75%	81.33%
	D2	AdaBoost	80%	74%	9.23E-03	77.00%	77.43%
	D2	Naïve Bayes	100%	9%	3.61E-01	54.25%	68.52%
miRNA-mRNA-Methylation	**A3(*)**	Neural Network	86%	84%	2.41E-05	84.50%	84.60%*
	A3	SVM(Linear)	83%	86%	2.28E-05	84.25%	84.10%
	A3	SVM (Polynomial)	87%	81%	6.41E-05	83.75%	84.24%
	A3	SVM (Radial)	84%	82%	8.86E-04	82.75%	83.05%
	A3	Random Forest	92%	71%	2.81E-04	81.25%	83.06%
	A3	AdaBoost	80%	81%	4.23E-04	80.00%	79.77%
	A3	Naïve Bayes	97%	14%	3.36E-01	55.25%	68.51%
	B3	Neural Network	76%	93%	1.18E-04	84.00%	82.51%
	B3	SVM(Linear)	78%	93%	1.51E-05	85.25%	84.11%
	B3(*)	SVM (Polynomial)	80%	91%	2.17E-05	85.50%	84.55%*
	B3	SVM (Radial)	89%	70%	2.36E-03	79.50%	81.46%
	B3	Random Forest	75%	76%	1.62E-02	75.25%	74.90%
	B3	AdaBoost	74%	86%	2.17E-03	79.75%	78.39%
	B3	Naïve Bayes	89%	21%	3.35E-01	55.00%	66.38%
	C3	Neural Network	93%	79%	4.45E-05	86.00%	87.01%
	C3	SVM(Linear)	86%	88%	1.22E-05	87.00%	86.76%
	C3(*)	SVM (Polynomial)	93%	82%	1.92E-05	87.25%	87.89%*
	C3	SVM (Radial)	98%	71%	8.36E-05	84.00%	86.10%
	C3	Random Forest	96%	69%	1.93E-04	82.00%	84.18%
	C3	AdaBoost	84%	78%	2.90E-04	80.50%	80.94%
	D3	Neural Network	84%	90%	3.89E-06	87.00%	86.53%
	D3(*)	SVM(Linear)	92%	93%	9.92E-08	92.50%	92.47%*
	D3	SVM (Polynomial)	95%	85%	9.45E-07	89.75%	90.34%
	D3	SVM (Radial)	96%	78%	8.39E-04	86.50%	87.94%
	D3	Random Forest	95%	76%	1.38E-05	85.25%	86.47%
	D3	AdaBoost	85%	85%	6.40E-05	84.75%	84.51%
miRNA-mRNA-Hypo Methylation	**A4**	Neural Network	60%	82%	6.26E-03	71.00%	67.43%
	A4	SVM(Linear)	75%	85%	9.11E-05	80.00%	78.95%
	A4	SVM (Polynomial)	60%	85%	4.70E-03	72.50%	68.46%
	A4	SVM (Radial)	72%	82%	5.71E-04	76.75%	75.45%
	A4	Random Forest	73%	97%	1.06E-05	84.50%	82.36%
	A4(*)	AdaBoost	78%	93%	1.16E-04	85.25%	84.07%*
	A4	Naïve Bayes	55%	95%	1.11E-03	75.00%	68.75%
	B4	Neural Network	70%	90%	1.16E-04	79.75%	77.57%
	B4(*)	SVM(Linear)	80%	83%	5.97E-05	81.50%	81.25%*
	B4	SVM (Polynomial)	50%	85%	1.92E-02	67.50%	60.61%
	B4	SVM (Radial)	50%	80%	4.03E-02	65.00%	58.82%
	B4	Random Forest	64%	100%	4.21E-05	81.75%	77.80%
	B4	AdaBoost	70%	98%	3.16E-05	83.50%	80.74%
	B4	Naïve Bayes	60%	100%	9.11E-05	80.00%	75.00%
	C4	Neural Network	93%	78%	8.42E-04	85.25%	86.41%
	C4(*)	SVM(Linear)	88%	87%	1.29E-05	87.50%	87.53%*
	C4	SVM (Polynomial)	91%	78%	8.19E-05	84.50%	85.52%
	C4	SVM (Radial)	94%	70%	3.85E-04	82.00%	83.99%
	C4	Random Forest	95%	65%	4.99E-04	80.00%	82.61%
	C4	AdaBoost	85%	75%	4.97E-04	79.50%	80.36%
	D4	Neural Network	85%	83%	2.06E-05	84.00%	84.11%
	D4(*)	SVM(Linear)	87%	88%	2.19E-06	87.50%	87.35%*
	D4	SVM (Polynomial)	94%	77%	4.80E-05	85.00%	86.23%
	D4	SVM (Radial)	97%	66%	1.07E-03	81.00%	83.80%
	D4	Random Forest	92%	80%	2.30E-05	85.50%	86.31%
	D4	AdaBoost	84%	83%	6.85E-05	83.00%	83.14%
miRNA-mRNA-Hypo Methylation	**A5**	Neural Network	60%	81%	7.86E-03	70.50%	67.06%
	A5	SVM(Linear)	75%	85%	9.11E-05	80.00%	78.95%
	A5	SVM (Polynomial)	61%	82%	6.14E-03	71.50%	68.08%

	A5	SVM (Radial)	71%	84%	4.94E-04	77.00%	75.40%
	A5	Random Forest	73%	97%	3.97E-05	85.00%	82.91%
	A5(*)	AdaBoost	76%	94%	1.69E-05	84.75%	83.29%*
	B5	Neural Network	70%	90%	1.16E-04	79.75%	77.57%
	B5	SVM(Linear)	80%	80%	9.11E-05	80.00%	80.00%
	B5	SVM (Polynomial)	51%	85%	1.81E-02	67.75%	61.02%
	B5	SVM (Radial)	50%	80%	4.03E-02	65.00%	58.82%
	B5	Random Forest	62%	99%	1.38E-04	80.25%	75.83%
	B5(*)	AdaBoost	68%	99%	4.95E-05	83.50%	80.30%*
	C5	Neural Network	91%	78%	5.87E-05	84.25%	85.20%
	C5	SVM(Linear)	88%	89%	6.91E-06	88.25%	88.07%
	C5(*)	SVM (Polynomial)	91%	86%	3.52E-06	88.50%	88.79%*
	C5	SVM (Radial)	97%	73%	3.57E-04	84.75%	86.49%
	C5	Random Forest	97%	62%	4.85E-04	79.25%	82.37%
	C5	AdaBoost	88%	72%	2.33E-03	79.75%	81.18%
	D5	Neural Network	89%	88%	4.80E-06	88.50%	88.60%
	D5(**)	SVM(Linear)	92%	90%	5.87E-07	90.75%	90.90%*
	D5	SVM (Polynomial)	95%	85%	3.09E-06	89.75%	90.39%
	D5	SVM (Radial)	96%	74%	5.00E-05	84.75%	86.42%
	D5	Random Forest	94%	75%	1.52E-04	84.25%	85.72%
	D5	AdaBoost	81%	86%	8.13E-05	83.00%	82.44%
mRNA-Methylation	**A6**	Neural Network	92%	77%	3.34E-04	84.25%	85.45%
	A6	SVM(Linear)	91%	79%	1.24E-04	84.75%	85.63%
	A6	SVM (Polynomial)	93%	75%	1.26E-04	84.00%	85.39%
	A6	SVM (Radial)	99%	63%	1.03E-03	80.75%	83.90%
	A6(*)	Random Forest	92%	83%	1.35E-05	87.25%	87.76%*
	A6	AdaBoost	84%	73%	1.43E-03	78.50%	79.75%
	B6	Neural Network	87%	84%	1.61E-05	85.50%	85.75%
	B6	SVM(Linear)	93%	79%	7.30E-05	85.75%	86.91%
	B6	SVM (Polynomial)	97%	71%	3.91E-04	83.75%	85.87%
	B6	SVM (Radial)	99%	70%	5.67E-05	84.25%	86.32%
	B6(**)	Random Forest	84%	94%	3.43E-05	88.75%	88.00%*
	B6	AdaBoost	81%	86%	3.79E-05	83.25%	82.76%
	C6(*)	Neural Network	88%	85%	2.34E-05	86.00%	86.26%*
	C6	SVM(Linear)	89%	78%	3.70E-04	83.00%	84.04%
	C6	SVM (Polynomial)	100%	5%	4.56E-01	52.25%	67.71%
	C6	SVM (Radial)	85%	77%	2.63E-04	80.75%	81.44%
	C6	Random Forest	85%	84%	2.22E-05	84.50%	84.49%
	C6	AdaBoost	81%	85%	5.85E-05	82.75%	82.29%
	D6	Neural Network	84%	90%	3.89E-06	87.00%	86.53%
	D6(*)	SVM(Linear)	87%	88%	2.06E-05	87.25%	87.29%*
	D6	SVM (Polynomial)	100%	2%	5.13E-01	51.00%	67.12%
	D6	SVM (Radial)	73%	89%	4.90E-04	80.75%	78.61%
	D6	Random Forest	75%	90%	9.17E-05	82.25%	80.71%
	D6	AdaBoost	80%	93%	2.17E-05	86.25%	85.09%
MRNA	**A7(*)**	Neural Network	86%	80%	9.09E-04	82.50%	82.89%*
	A7	SVM(Linear)	90%	68%	6.56E-04	79.00%	81.10%
	A7	SVM (Polynomial)	100%	7%	4.08E-01	53.50%	68.32%
	A7	SVM (Radial)	87%	75%	4.20E-04	80.75%	81.74%
	A7	Random Forest	81%	81%	9.15E-04	80.75%	80.44%
	A7	AdaBoost	77%	86%	8.98E-04	81.00%	79.79%
	B7	Neural Network	80%	85%	1.00E-04	82.00%	81.57%
	B7(*)	SVM(Linear)	88%	74%	4.88E-04	80.75%	81.84%*
	B7	SVM (Polynomial)	100%	2%	5.13E-01	51.00%	67.12%
	B7	SVM (Radial)	78%	81%	4.20E-03	79.25%	78.84%
	B7	Random Forest	76%	88%	4.07E-03	81.75%	79.85%
	B7	AdaBoost	76%	91%	3.61E-04	83.00%	81.22%
	C7	Neural Network	94%	76%	7.48E-05	84.75%	86.07%
	C7	SVM(Linear)	88%	85%	1.44E-05	86.25%	86.49%
	C7(*)	SVM (Polynomial)	91%	81%	2.14E-05	85.75%	86.49%*
	C7	SVM (Radial)	98%	69%	8.60E-04	83.00%	85.36%
	C7	Random Forest	98%	70%	5.37E-05	83.75%	85.84%
	C7	AdaBoost	87%	75%	1.36E-04	80.75%	81.82%
	D7	Neural Network	90%	82%	2.13E-05	86.00%	86.56%
	D7	SVM(Linear)	91%	83%	3.58E-05	87.00%	87.59%

	D7	SVM (Polynomial)	95%	73%	1.41E-04	83.50%	85.20%
	D7	SVM (Radial)	100%	62%	7.75E-04	80.50%	83.78%
	D7(**)	Random Forest	93%	83%	1.18E-05	88.00%	88.57%*
	D7	AdaBoost	78%	85%	8.66E-04	81.25%	80.29%
Methylation	**A8(*)**	Neural Network	81%	77%	5.06E-03	78.75%	79.27%*
	A8	SVM(Linear)	74%	74%	6.07E-03	74.00%	73.66%
	A8	SVM (Polynomial)	85%	57%	3.16E-02	70.75%	74.07%
	A8	SVM (Radial)	71%	78%	1.85E-02	74.25%	72.50%
	A8	Random Forest	75%	77%	3.36E-03	76.00%	75.68%
	A8	AdaBoost	77%	71%	7.14E-03	73.50%	74.37%
	B8(*)	Neural Network	76%	88%	2.76E-03	81.75%	80.48%*
	B8	SVM(Linear)	75%	85%	5.50E-04	80.00%	78.90%
	B8	SVM (Polynomial)	92%	55%	1.71E-02	73.25%	77.55%
	B8	SVM (Radial)	63%	92%	1.33E-03	77.50%	73.24%
	B8	Random Forest	63%	92%	1.65E-03	77.25%	72.99%
	B8	AdaBoost	62%	80%	1.47E-02	71.00%	67.89%
	C8	Neural Network	82%	77%	4.48E-03	79.25%	79.79%
	C8	SVM(Linear)	78%	85%	7.70E-04	81.25%	80.63%
	C8	SVM (Polynomial)	78%	83%	2.89E-03	80.25%	79.72%
	C8(**)	SVM (Radial)	82%	81%	2.10E-03	81.00%	81.16%*
	C8	Random Forest	93%	63%	4.03E-03	77.75%	80.96%
	C8	AdaBoost	85%	70%	1.17E-03	77.25%	78.79%
	D8	Neural Network	76%	84%	8.18E-04	79.75%	78.93%
	D8	SVM(Linear)	82%	83%	4.44E-04	82.25%	82.18%
	D8(*)	SVM (Polynomial)	84%	83%	6.64E-05	83.50%	83.51%*
	D8	SVM (Radial)	88%	75%	5.98E-04	81.00%	82.24%
	D8	Random Forest	80%	82%	4.02E-04	80.75%	80.46%
	D8	AdaBoost	80%	78%	4.56E-04	78.75%	78.75%
miRNA-Methylation	**A9**	Neural Network	78%	80%	3.21E-03	78.75%	78.31%
	A9	SVM(Linear)	79%	77%	8.61E-03	77.50%	77.28%
	A9	SVM (Polynomial)	79%	77%	2.22E-03	77.75%	77.45%
	A9	SVM (Radial)	81%	71%	2.18E-03	75.75%	76.69%
	A9(*)	Random Forest	82%	86%	7.73E-05	83.75%	83.41%*
	A9	AdaBoost	79%	83%	5.16E-04	80.75%	80.22%
	B9	Neural Network	73%	88%	9.30E-04	80.00%	78.32%
	B9(*)	SVM(Linear)	78%	89%	3.40E-04	83.00%	81.98%*
	B9	SVM (Polynomial)	88%	66%	1.65E-03	76.50%	78.58%
	B9	SVM (Radial)	69%	89%	1.43E-03	78.50%	75.50%
	B9	Random Forest	74%	94%	1.33E-04	83.50%	81.56%
	B9	AdaBoost	71%	89%	9.35E-04	79.75%	77.63%
	C9(*)	Neural Network	87%	78%	2.08E-03	82.00%	82.86%*
	C9	SVM(Linear)	79%	84%	2.30E-03	81.25%	80.74%
	C9	SVM (Polynomial)	83%	82%	1.98E-03	82.25%	82.34%
	C9	SVM (Radial)	85%	79%	9.20E-04	81.75%	82.17%
	C9	Random Forest	96%	43%	3.58E-02	69.50%	76.03%
	C9	AdaBoost	84%	70%	4.86E-03	77.00%	78.59%
	D9	Neural Network	79%	90%	6.23E-05	84.25%	83.33%
	D9	SVM(Linear)	77%	88%	2.49E-04	82.25%	81.09%
	D9	SVM (Polynomial)	81%	88%	1.57E-04	84.25%	83.72%
	D9(**)	SVM (Radial)	93%	72%	9.03E-04	82.25%	84.11%*
	D9	Random Forest	88%	71%	1.13E-03	79.25%	80.98%
	D9	AdaBoost	80%	82%	4.13E-04	80.75%	80.48%

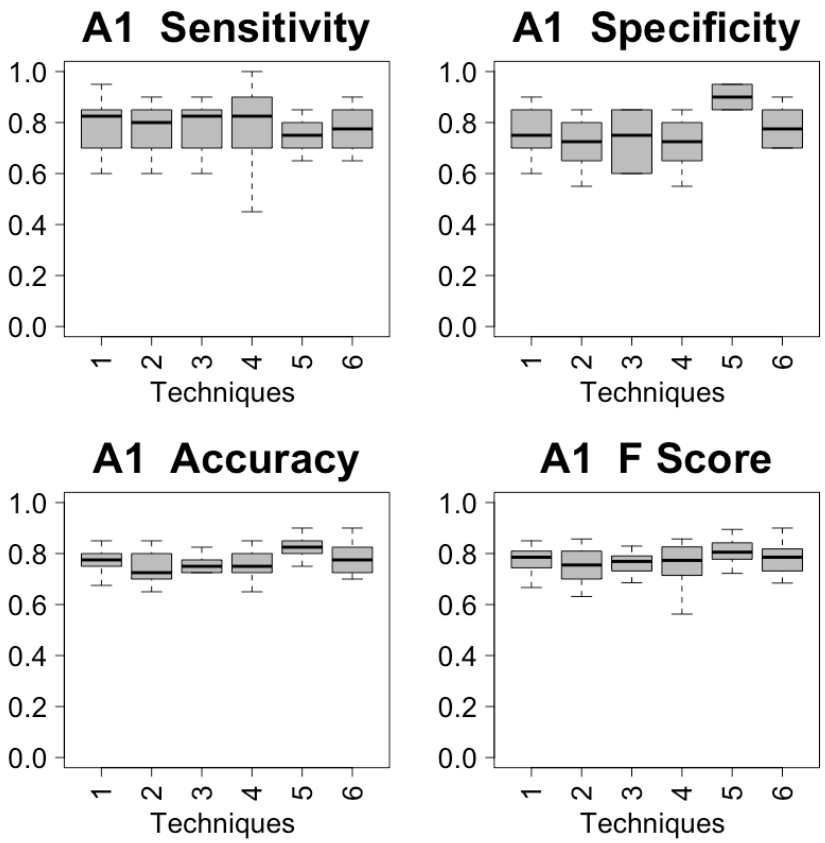

Figure 24: miRNA Models Variance of Predictive Models for A

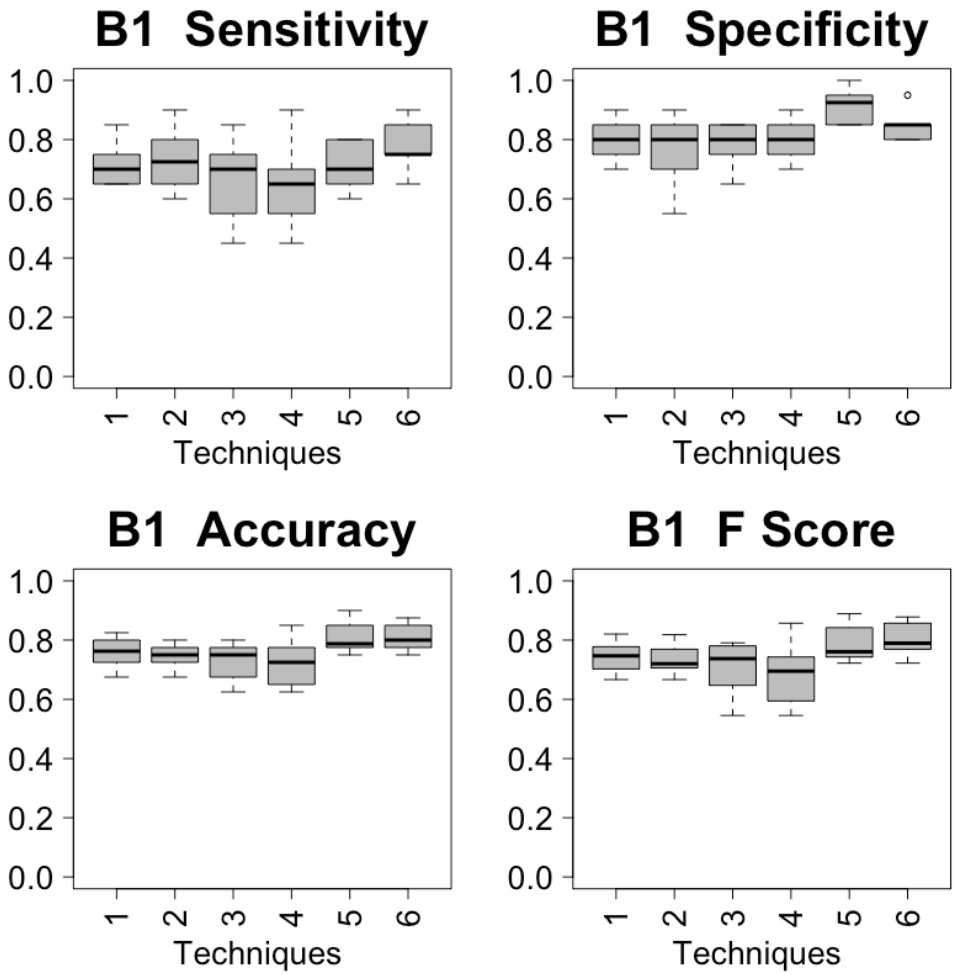

Figure 25: miRNA Models Variance of Predictive Models for B1

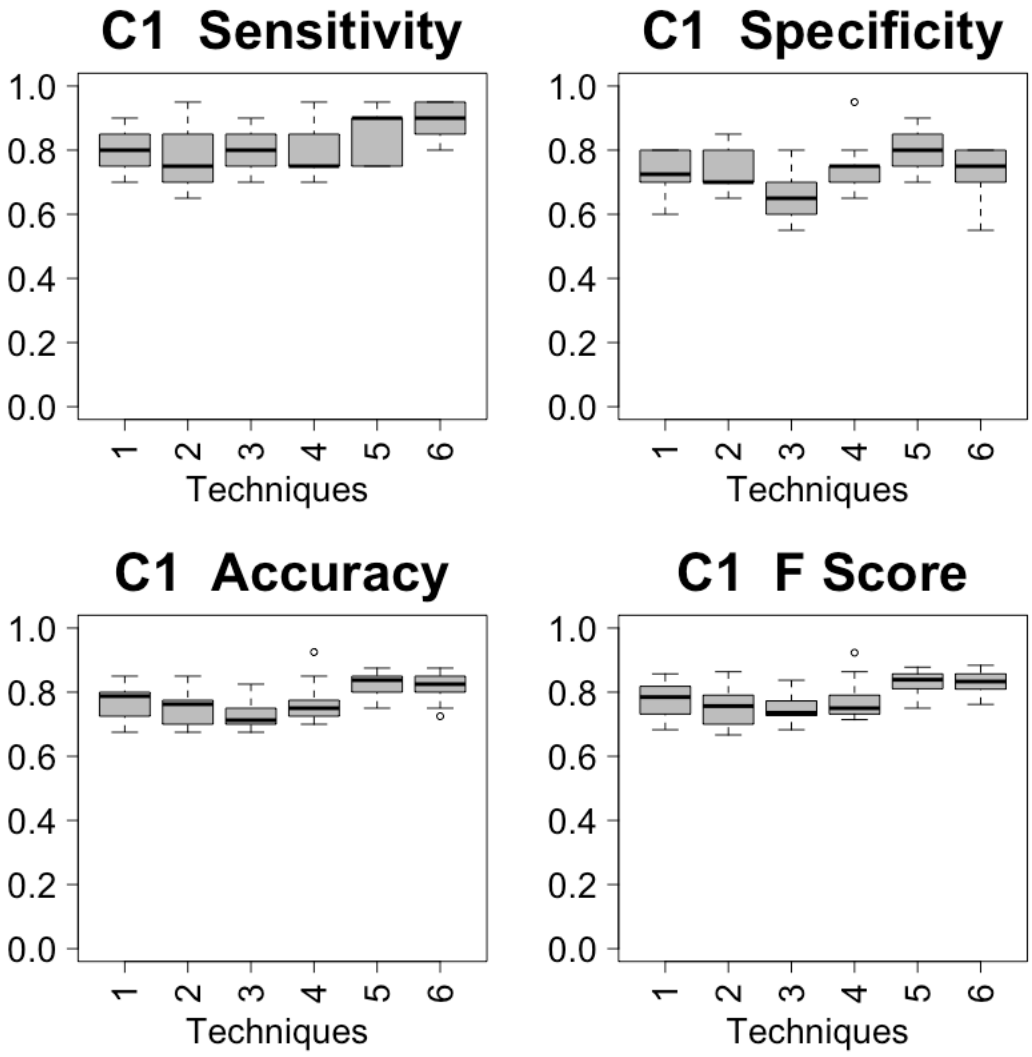

Figure 26: miRNA Models Variance of Predictive Models for C1

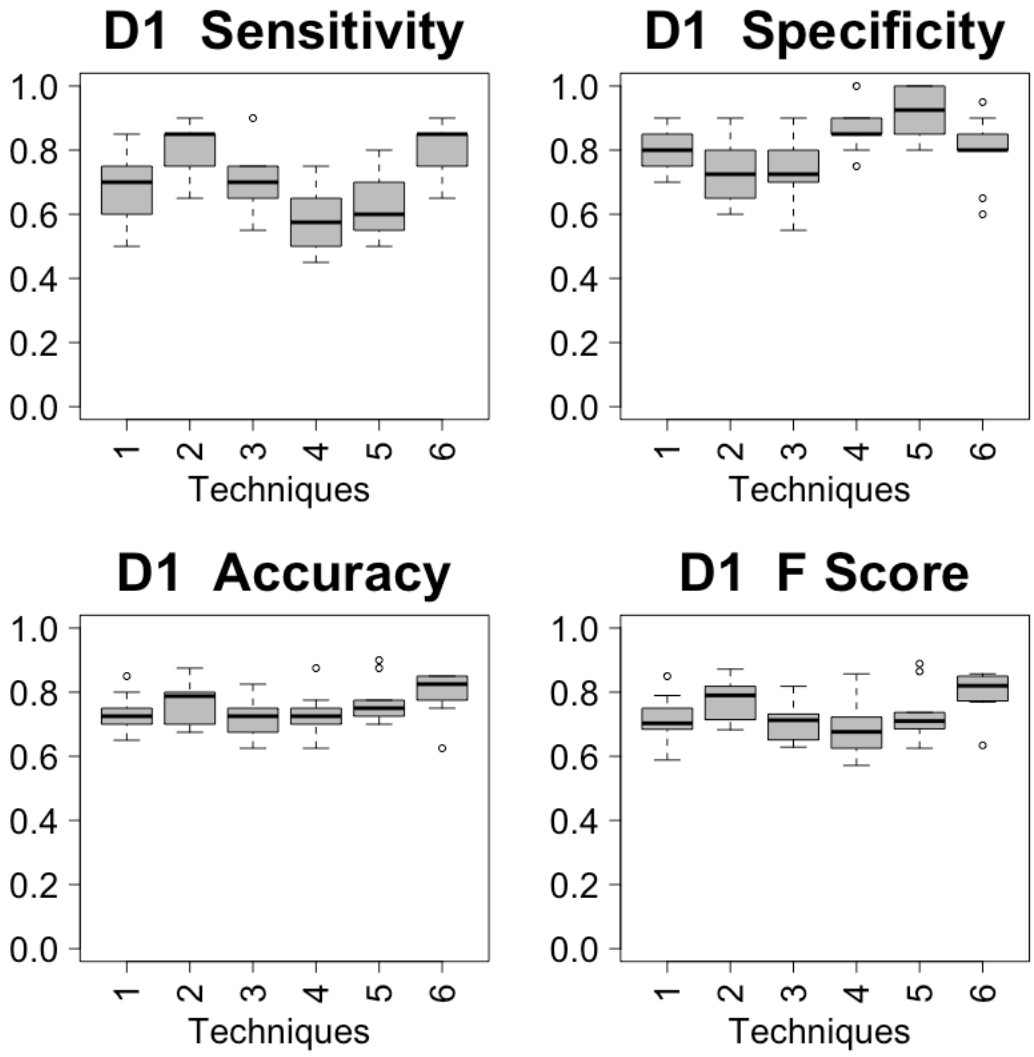

Figure 27: miRNA Models Variance of Predictive Models for D1

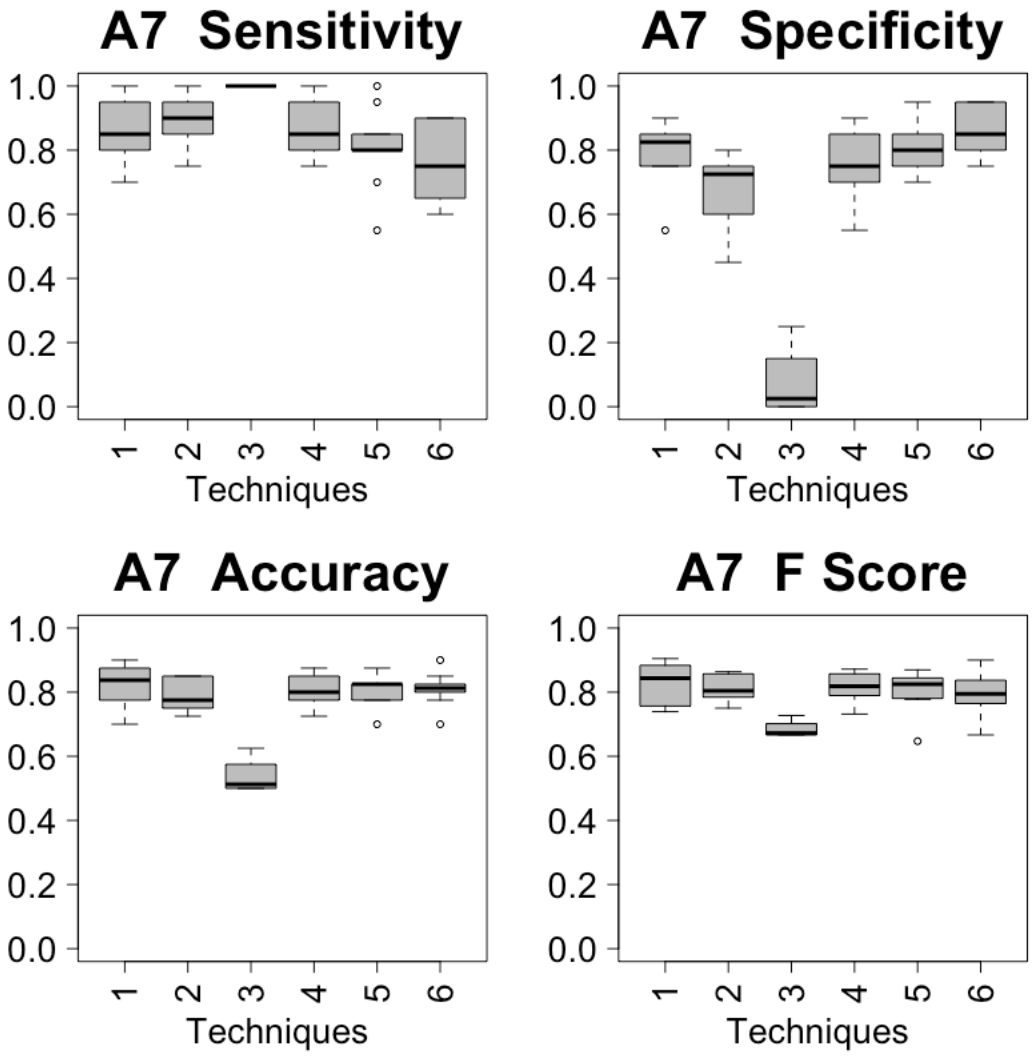

Figure 28: mRNA Models Variance of Predictive Models for A7

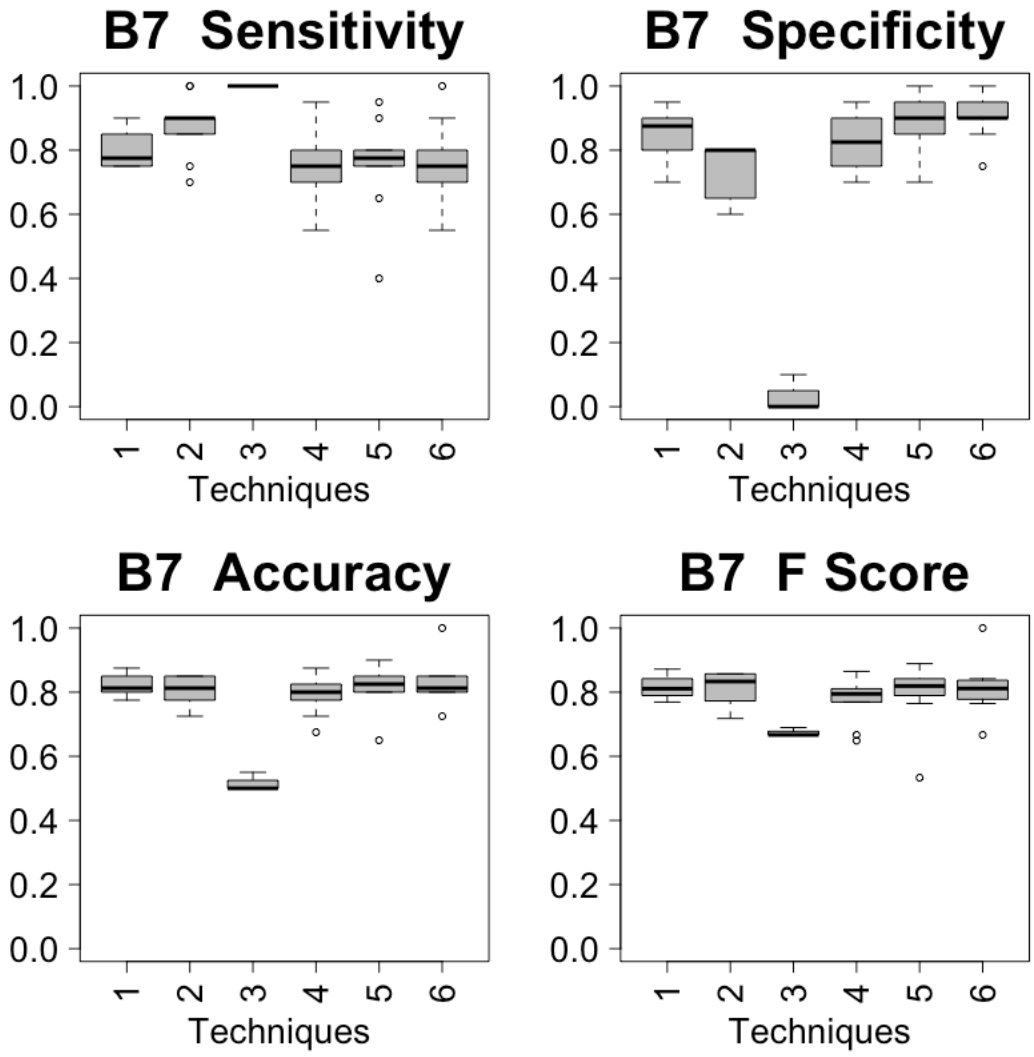

Figure 29: mRNA Models Variance of Predictive Models for B7

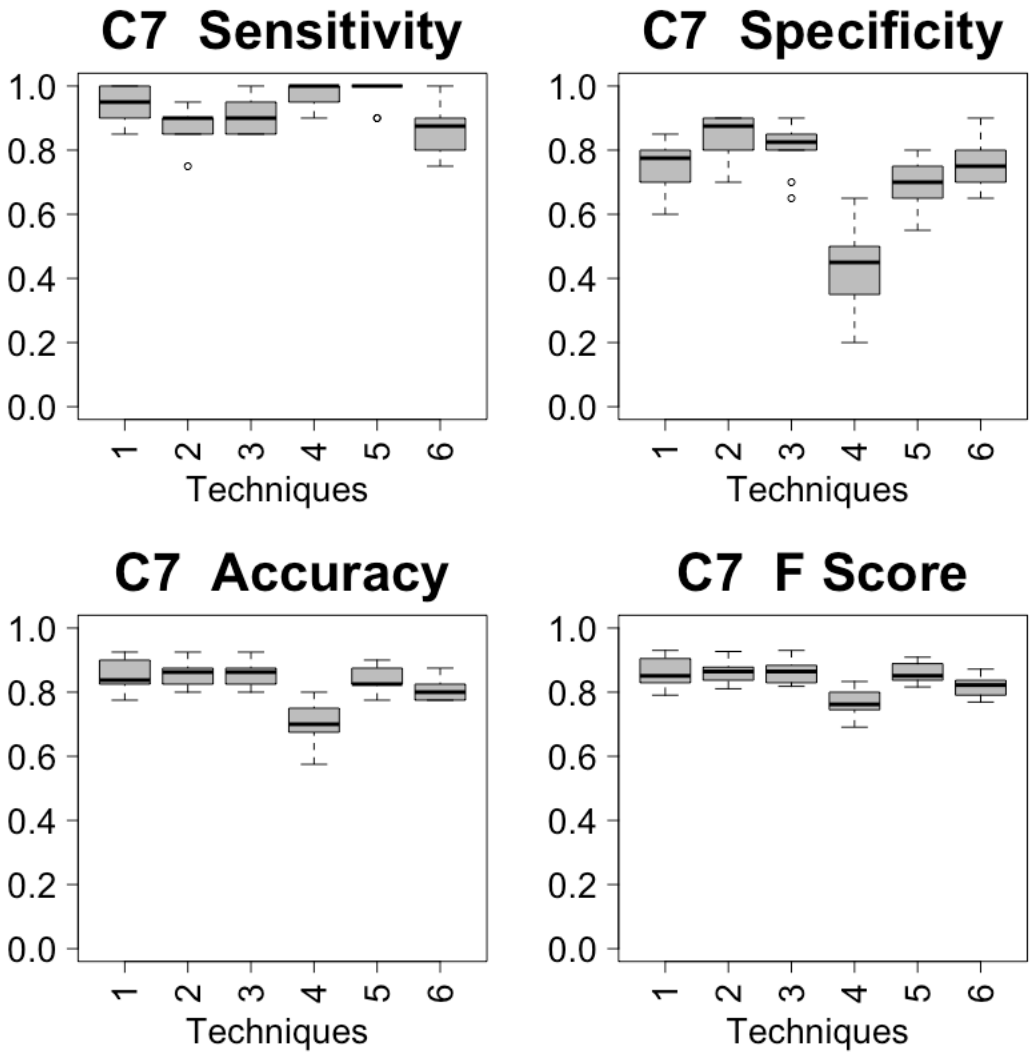

Figure 30: mRNA Models Variance of Predictive Models for C7

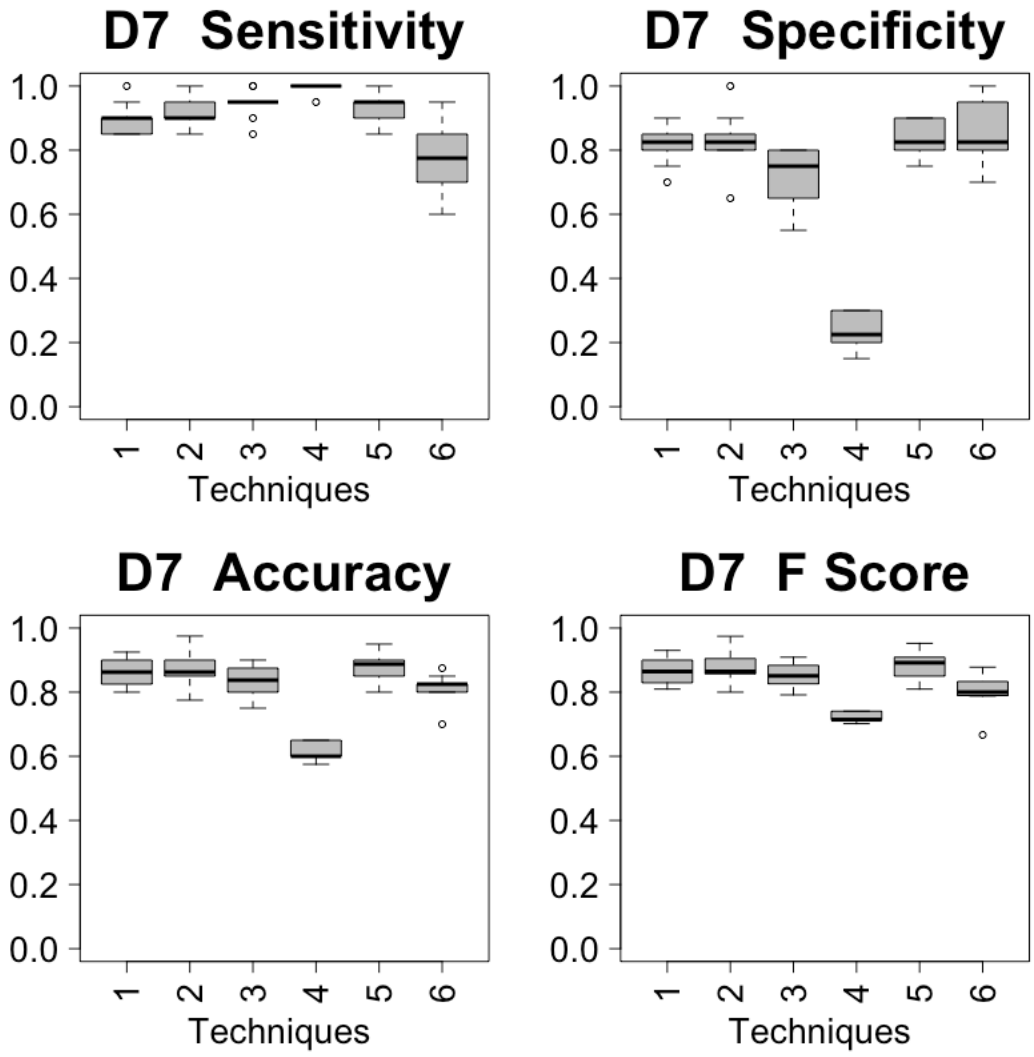

Figure 31: mRNA Models Variance of Predictive Models for D7

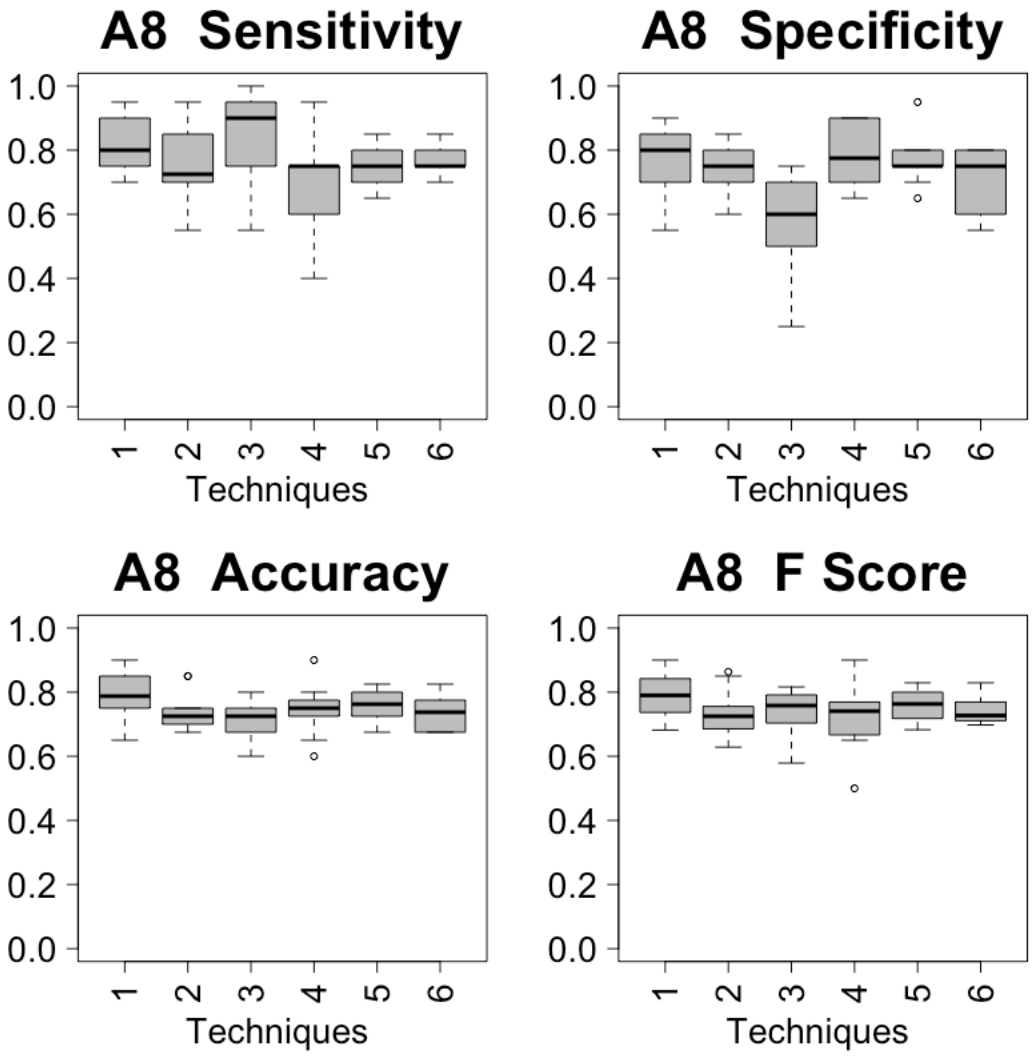

Figure 32: Methylation Models Variance of Predictive Models for A8

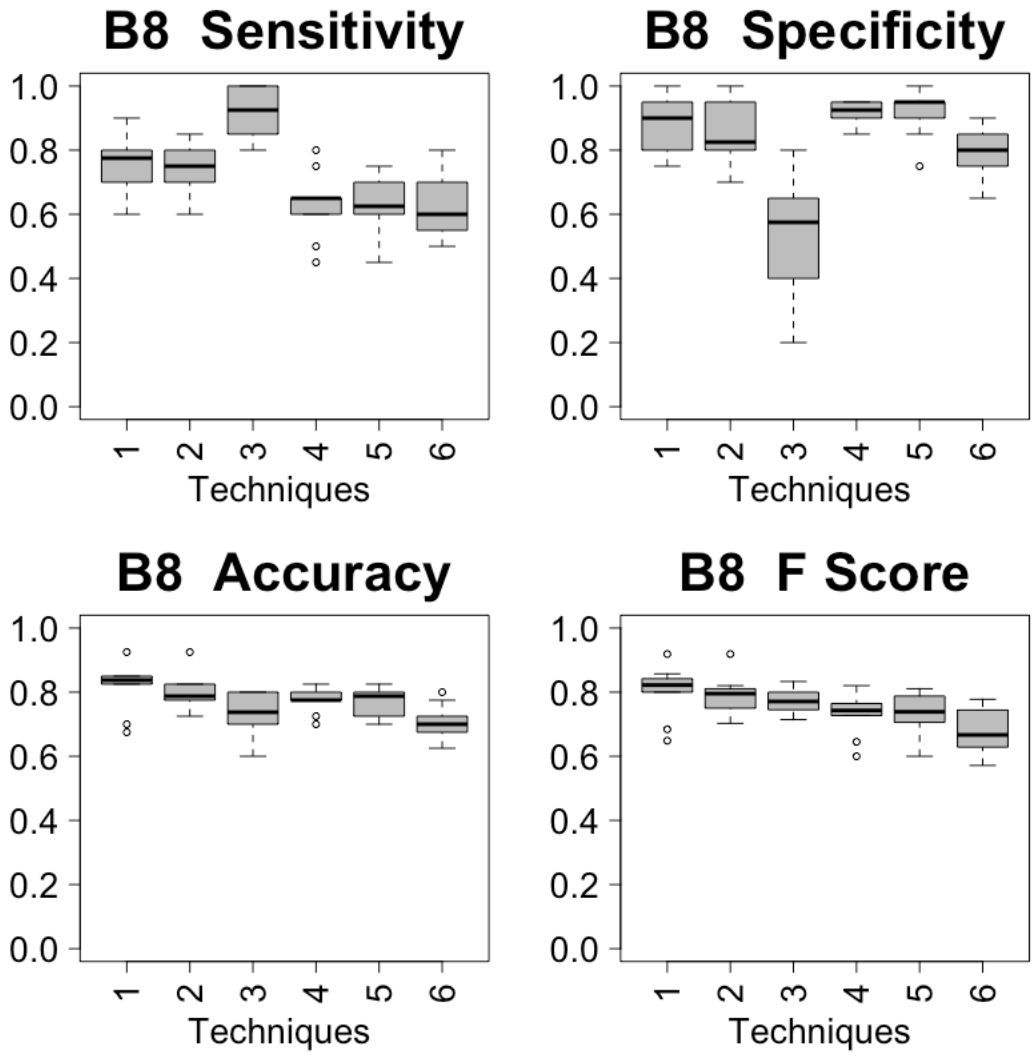

Figure 33: Methylation Models Variance of Predictive Models for B8

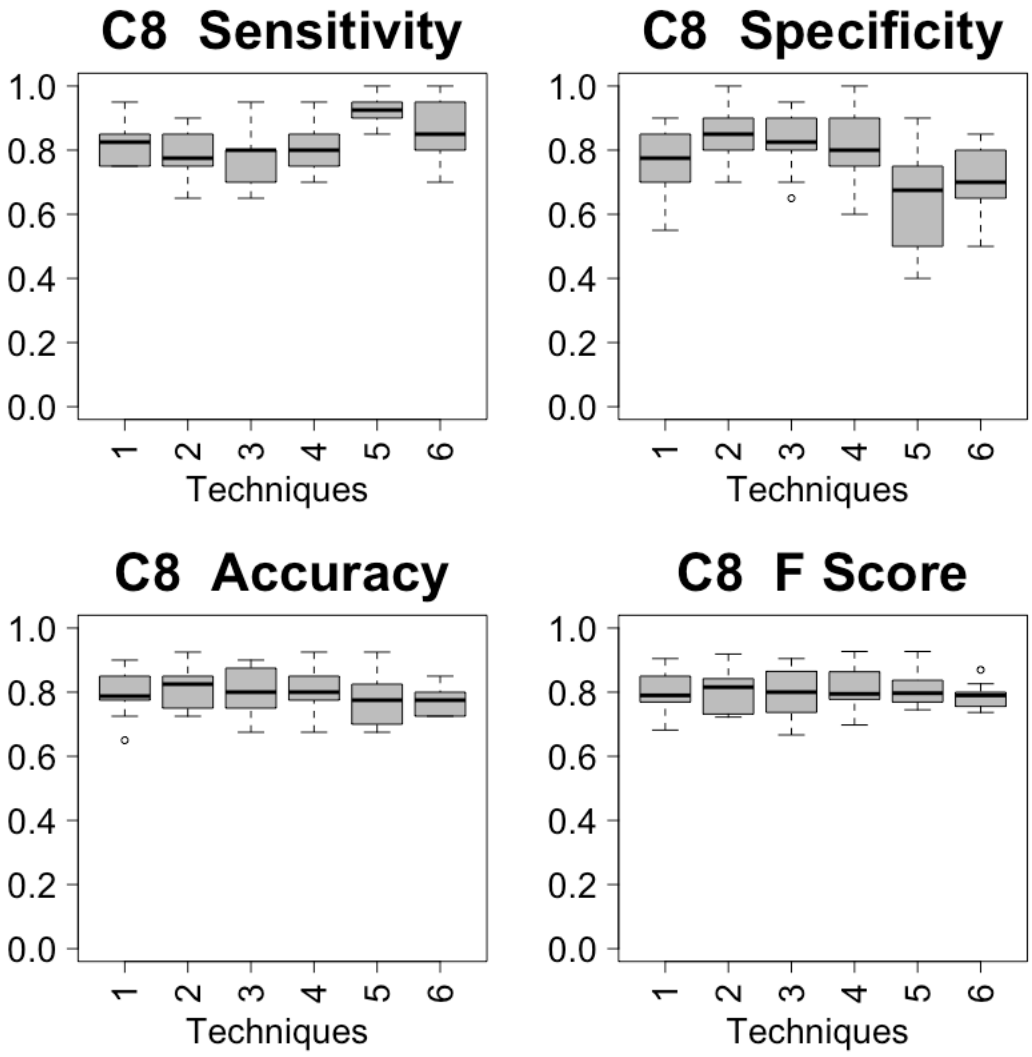

Figure 34: Methylation Models Variance of Predictive Models for C8

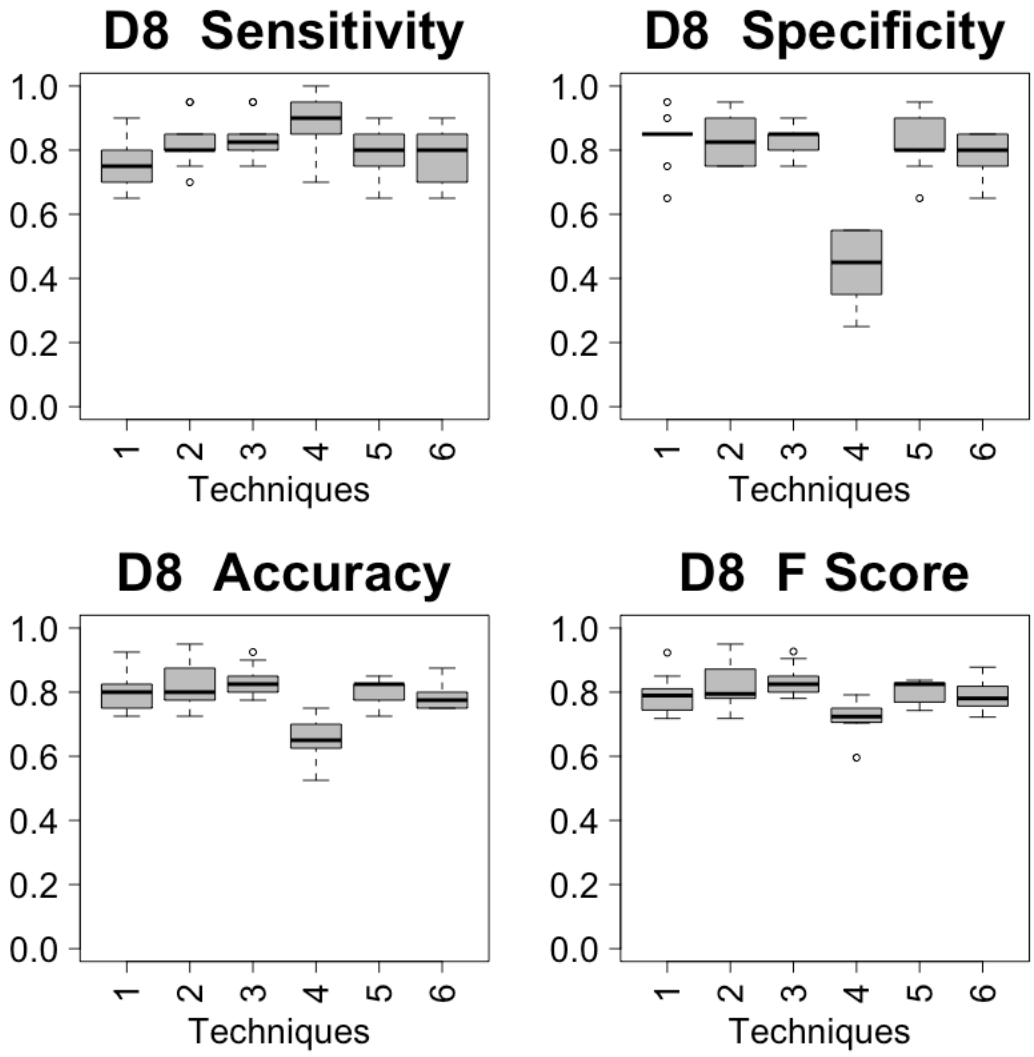

Figure 35: Methylation Model Variances of Predictive Models of D8

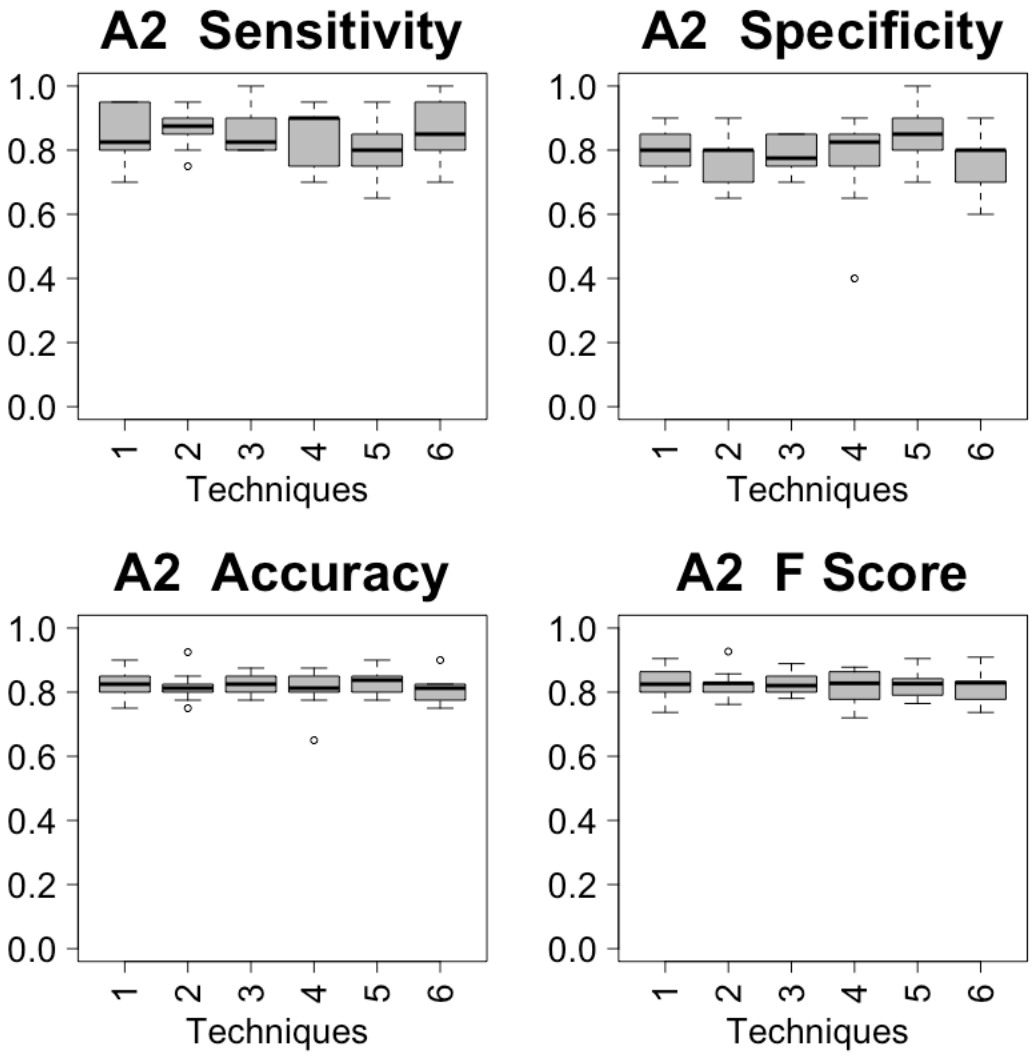

Figure 36: miRNA and mRNA Model Variances of Predictive Models of A2

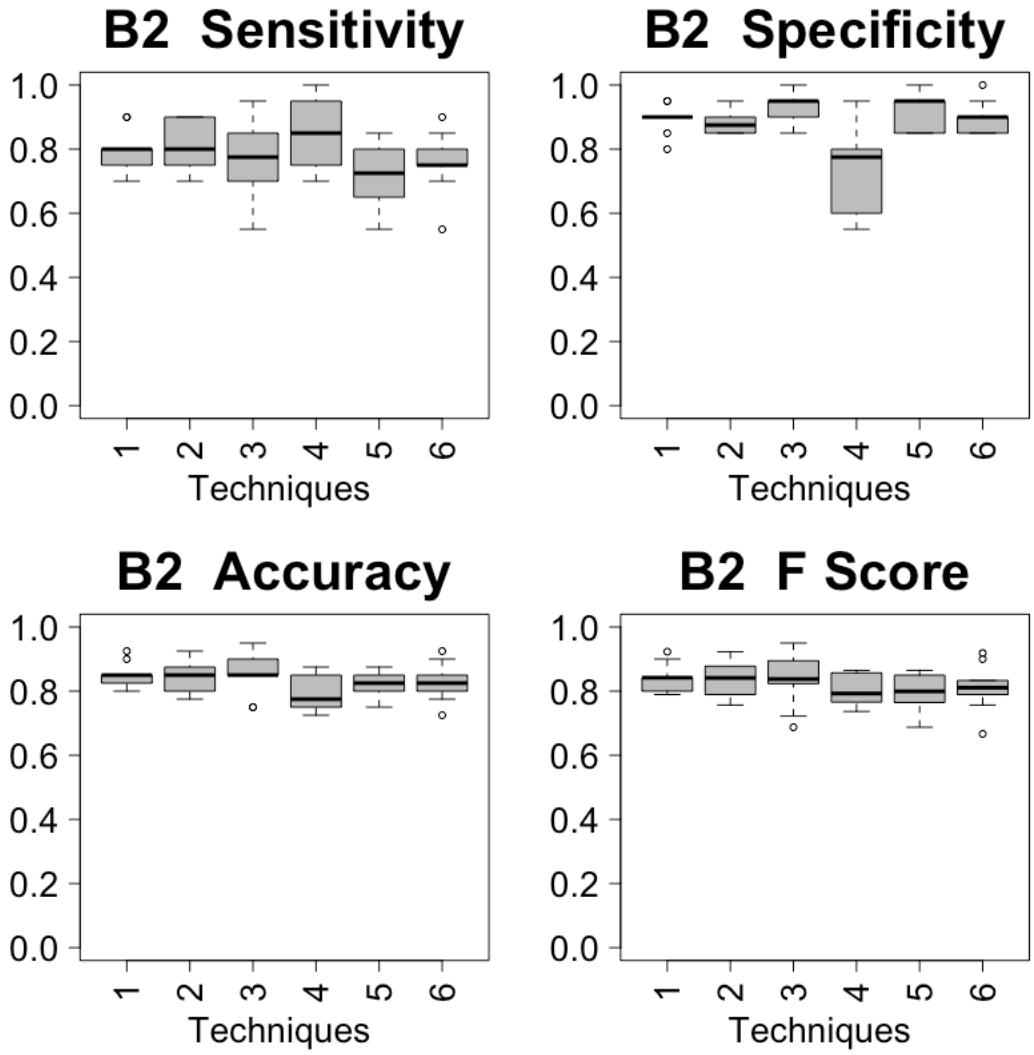

Figure 37: miRNA and mRNA Model Variances of Predictive Models of B2

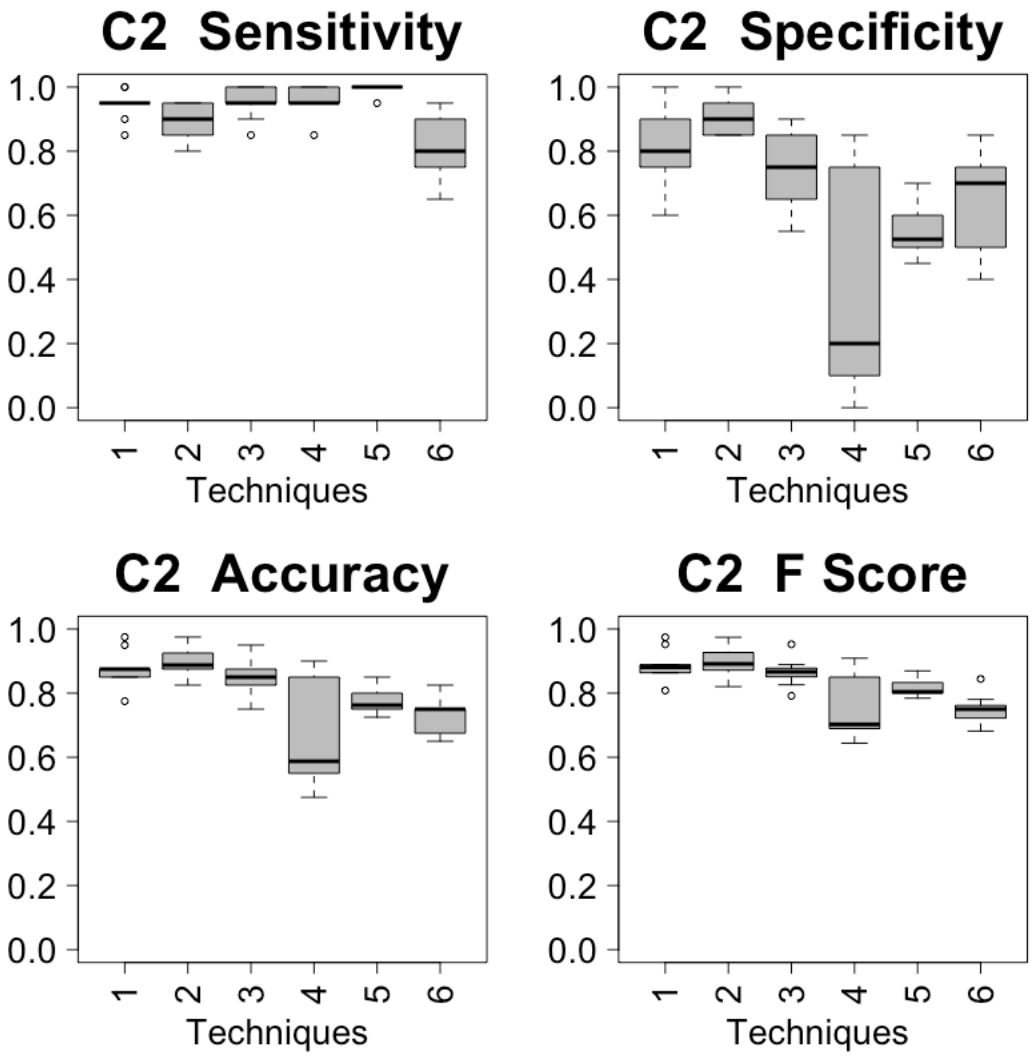

Figure 38: miRNA and mRNA Model Variances of Predictive Models of C2

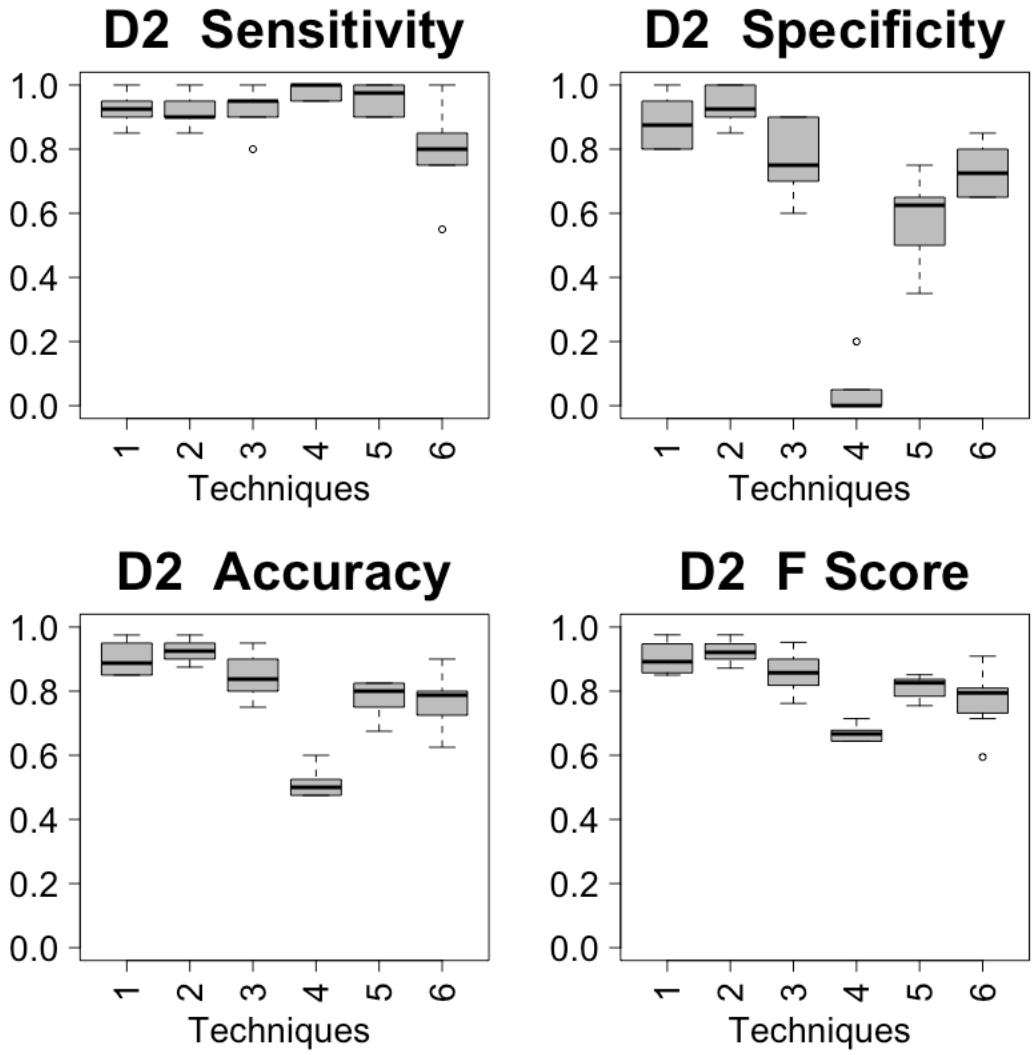

Figure 39: miRNA and mRNA Model Variances of Predictive Models of D2

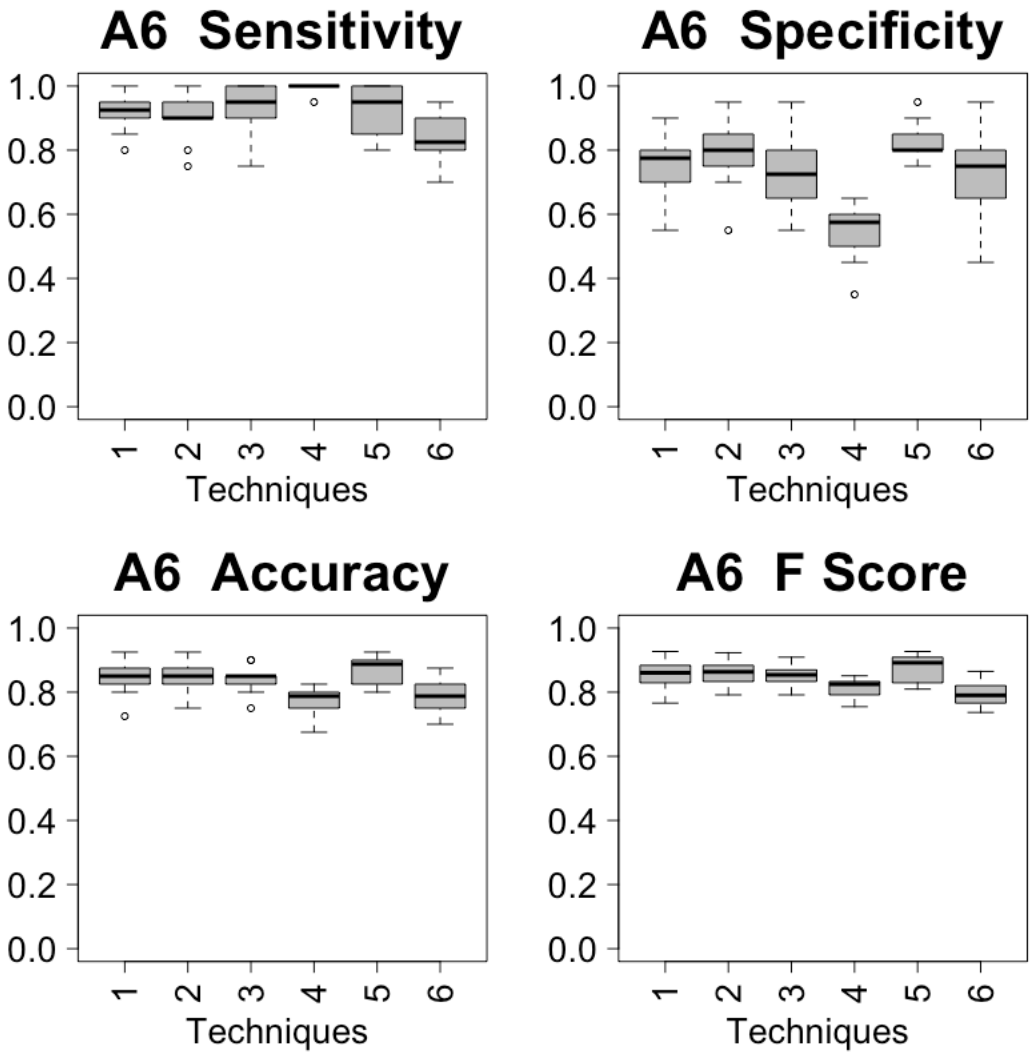

Figure 40: mRNA and Methylation Model Variances of Predictive Models of A6

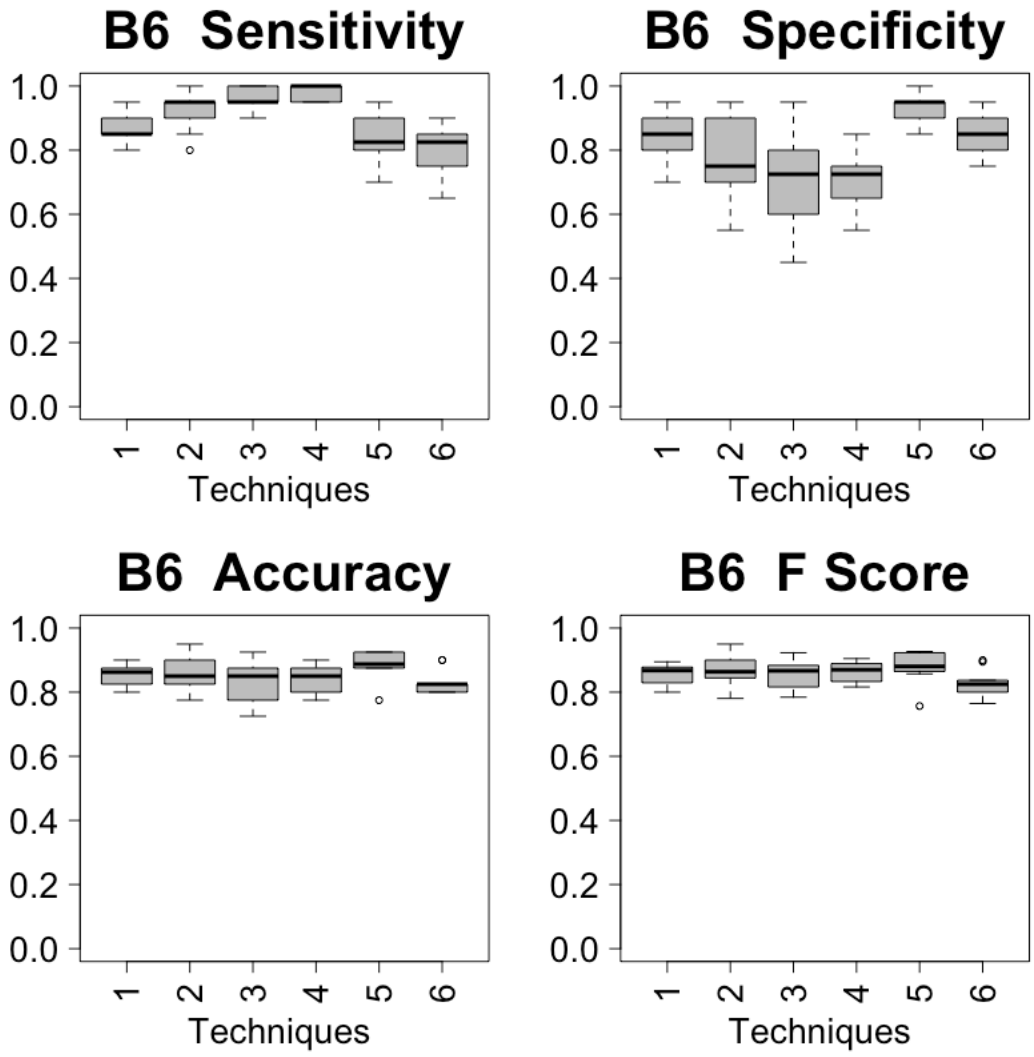

Figure 41: mRNA and Methylation Model Variances of Predictive Models of B6

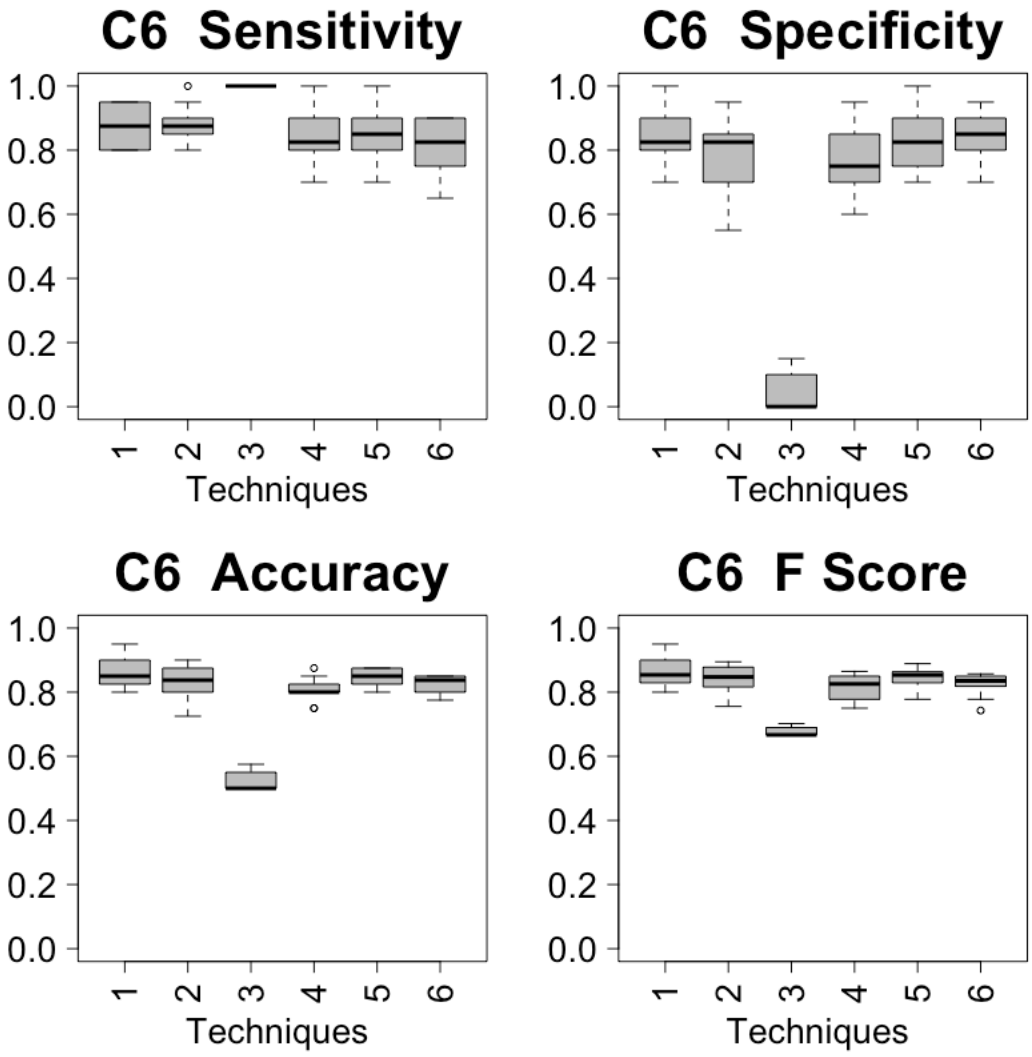

Figure 42 : mRNA and Methylation Model Variances of Predictive Models of C6

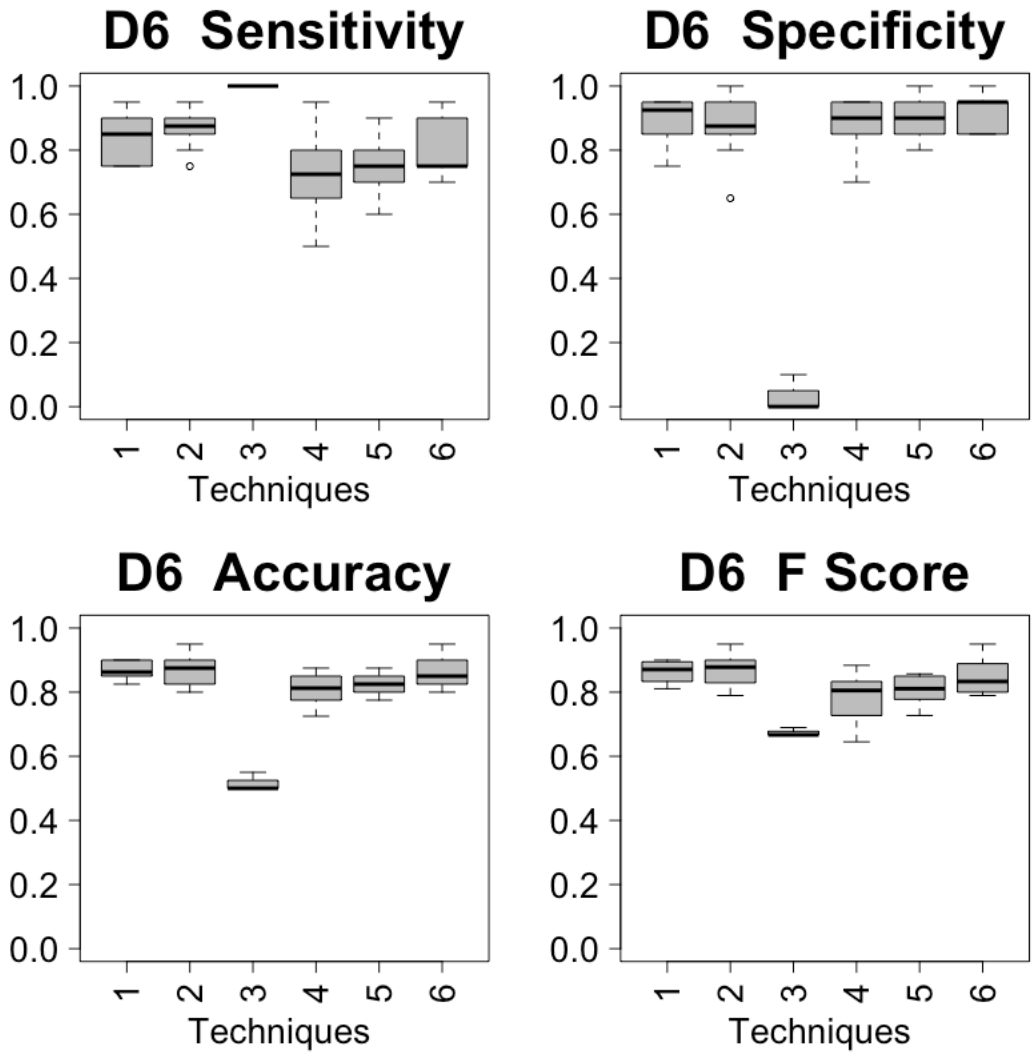

Figure 43: mRNA and Methylation Model Variances of Predictive Models of D6

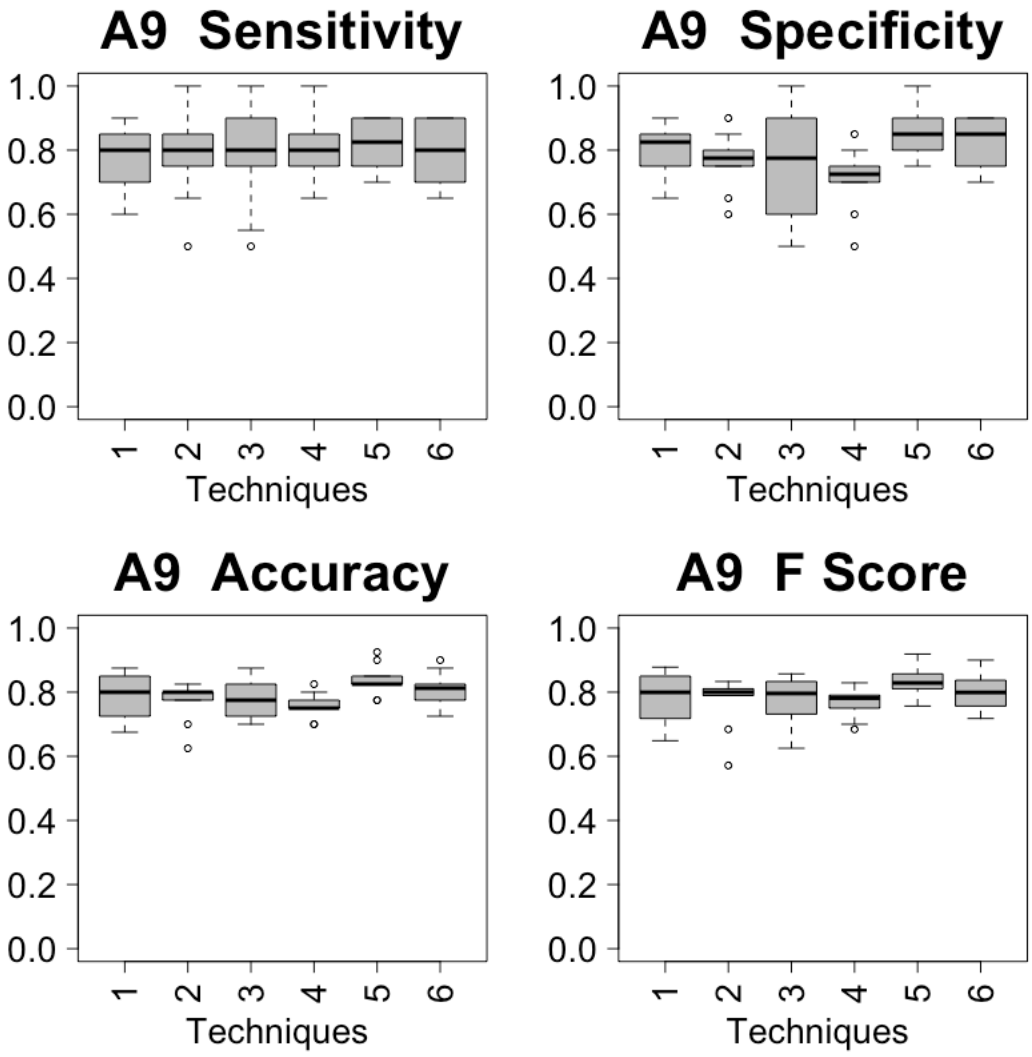

Figure 44: miRNA and Methylation Model Variances of Predictive Models of A9

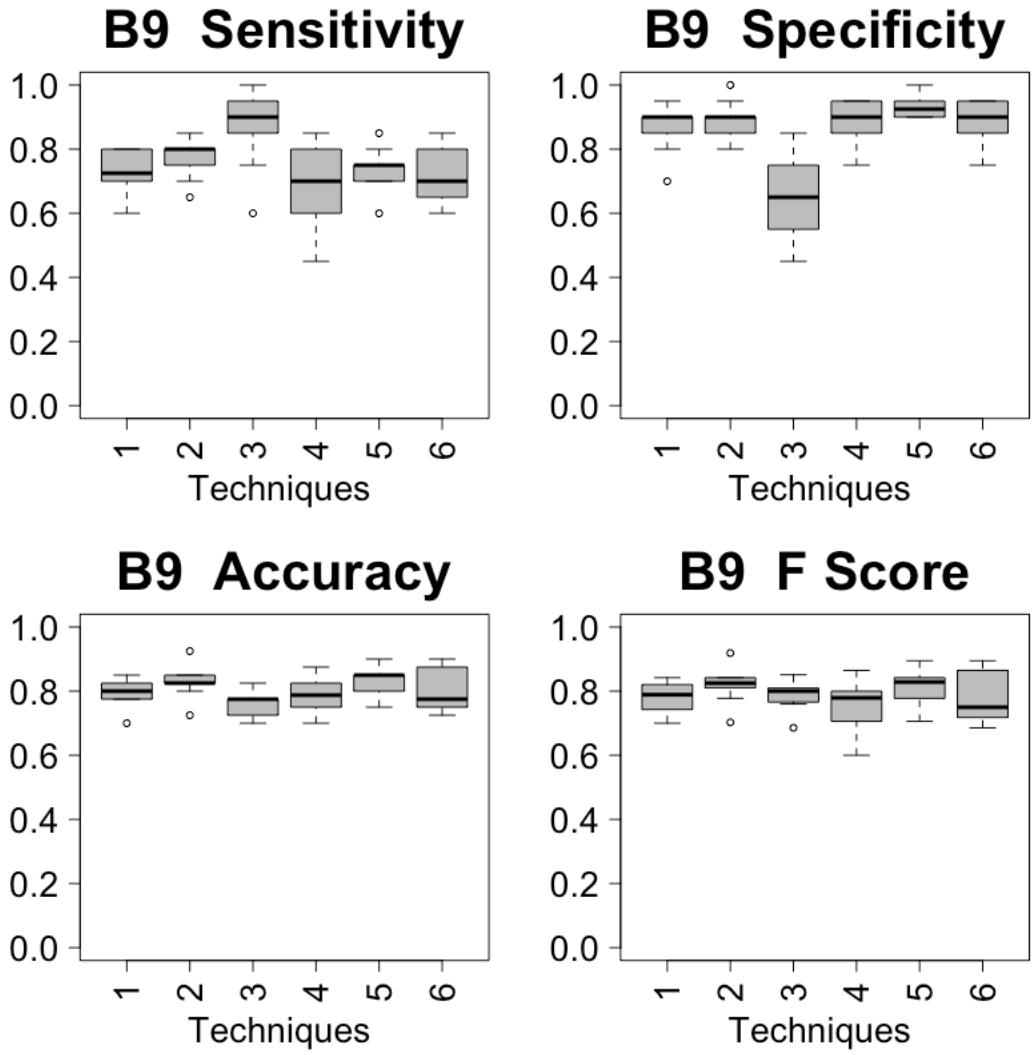

Figure 45: miRNA and Methylation Model Variances of Predictive Models of B9

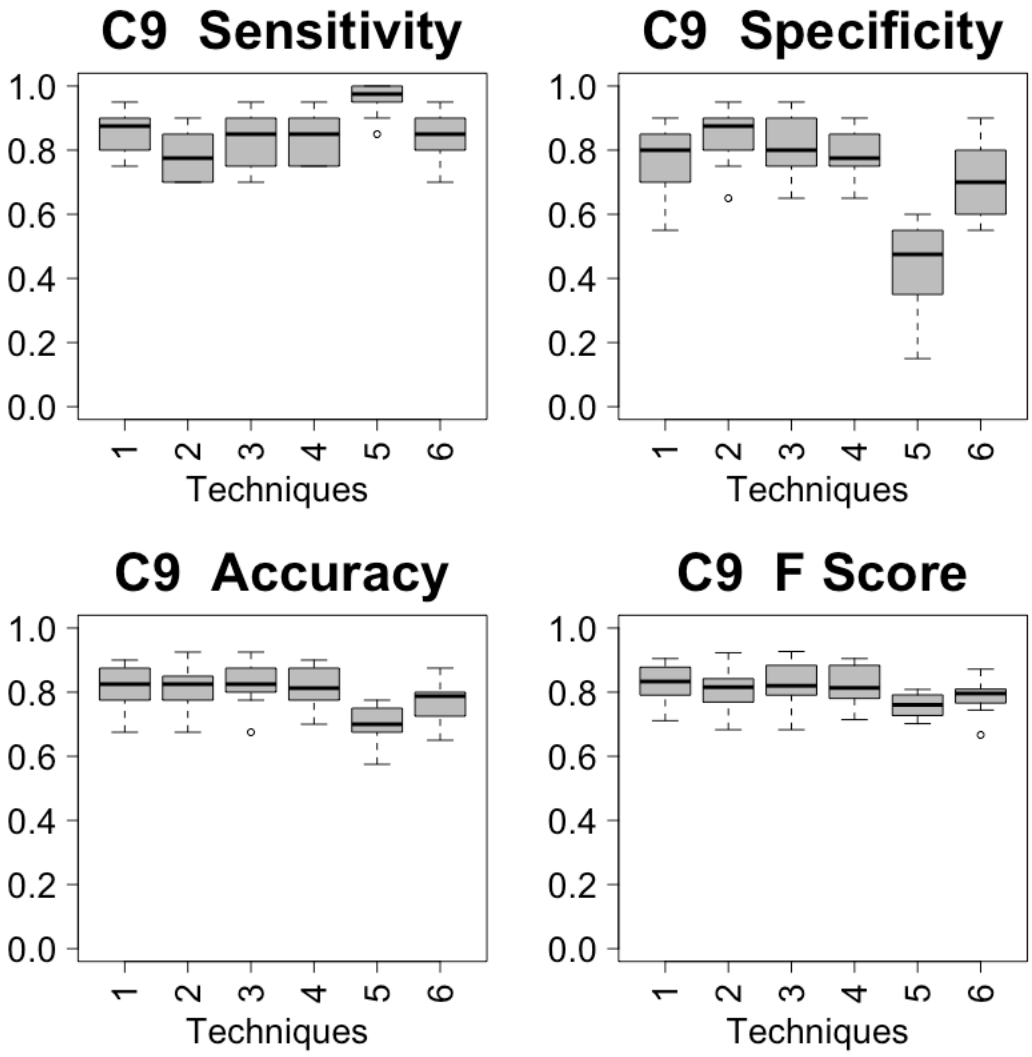

Figure 46: miRNA and Methylation Model Variances of Predictive Models of C9

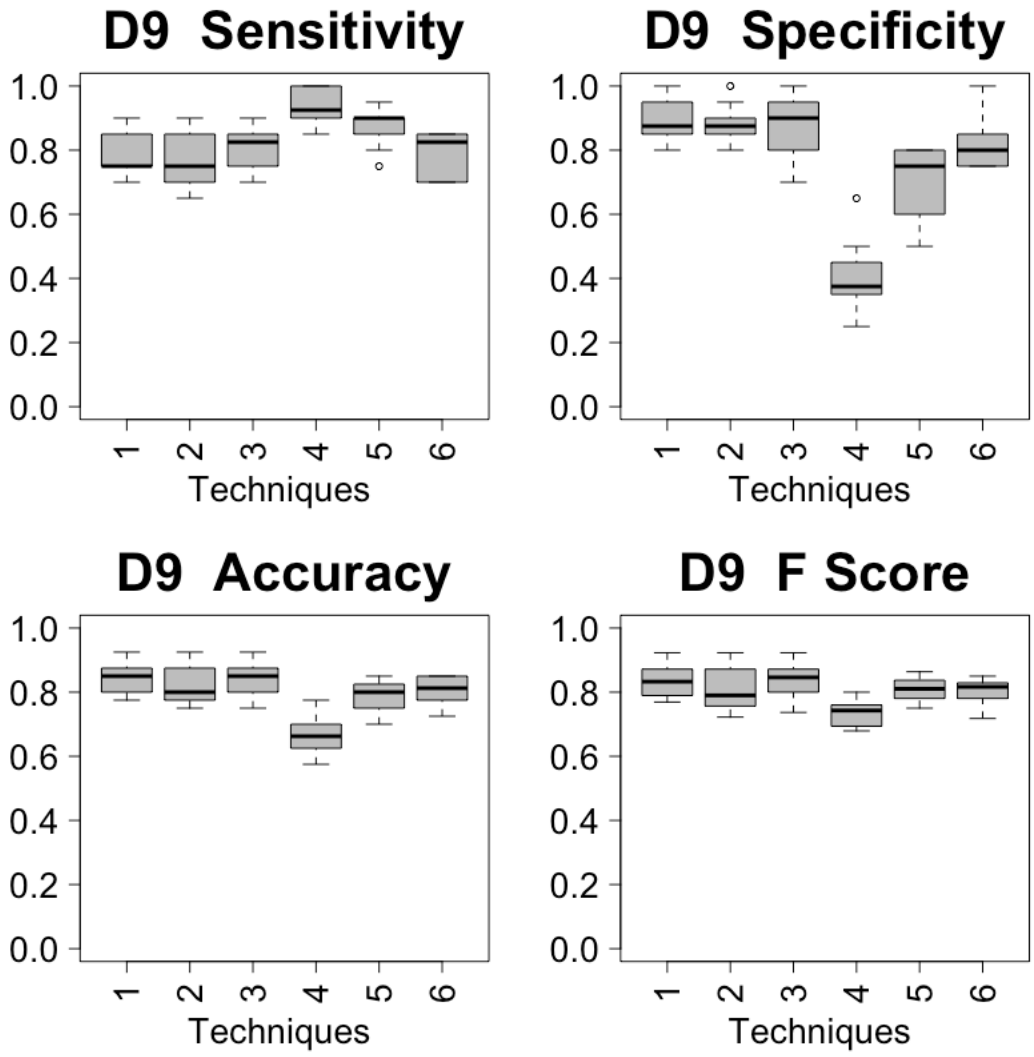

Figure 47: miRNA and Methylation Model Variances of Predictive Models of D9

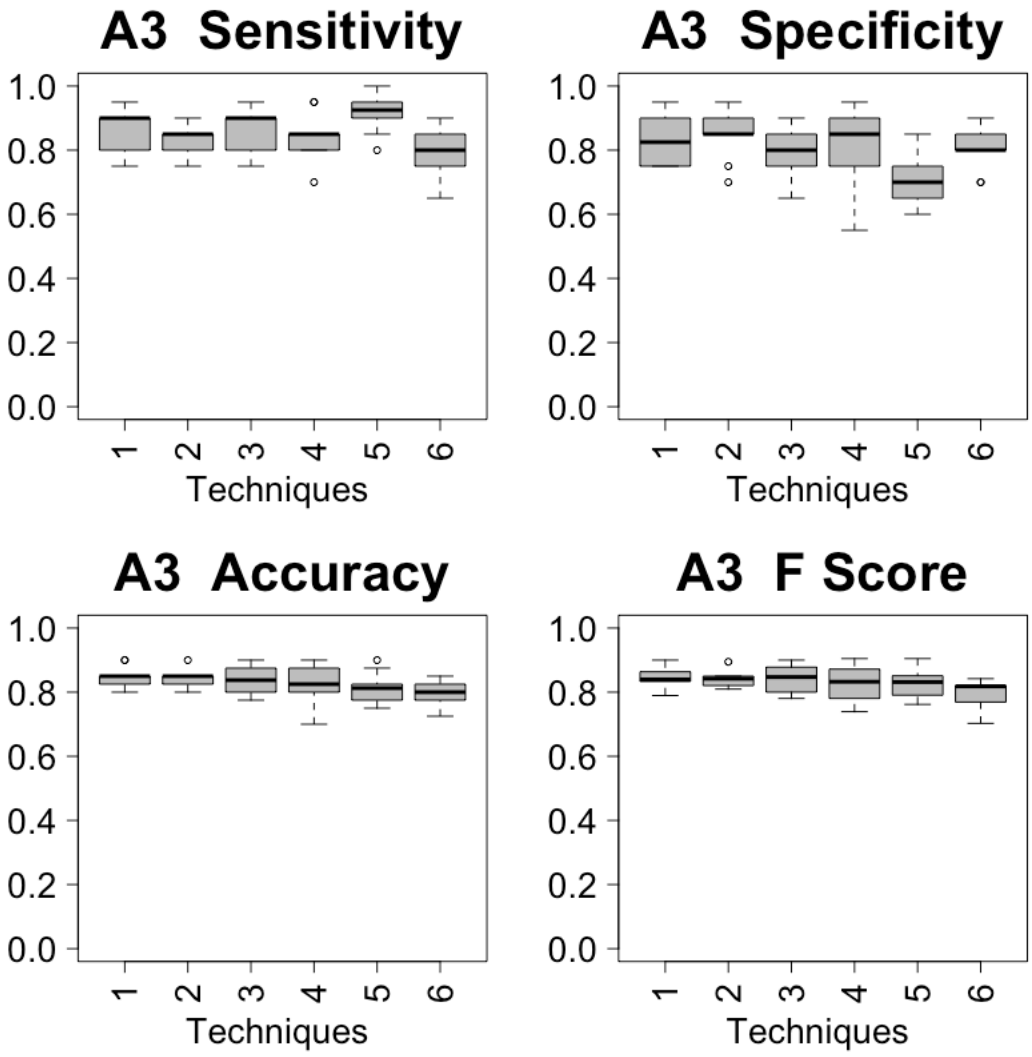

Figure 48: miRNA, mRNA and Methylation Model Variances of Predictive Models of A3

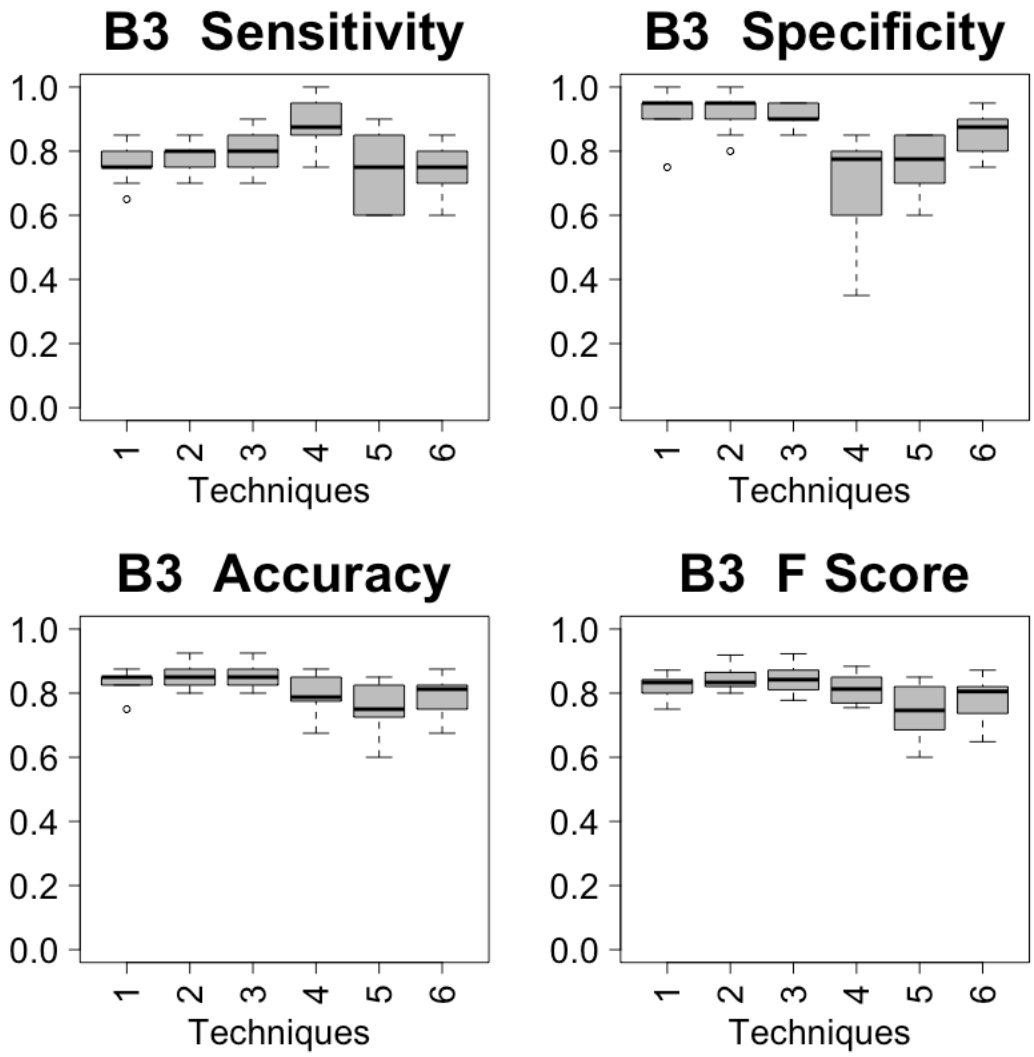

Figure 49: miRNA, mRNA and Methylation Model Variances of Predictive Models of A3, B3 , C3 And D3

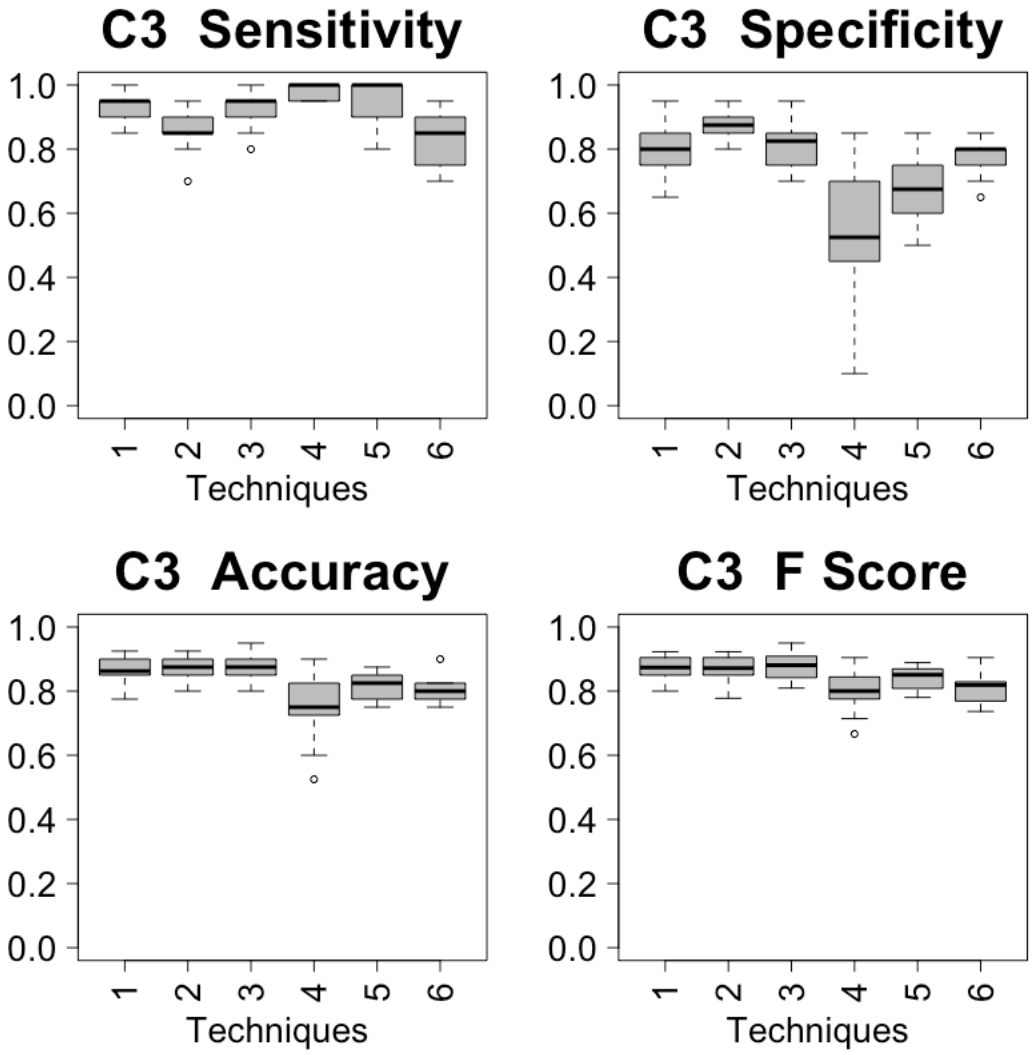

Figure 50: miRNA, mRNA and Methylation Model Variances of Predictive Models of C3

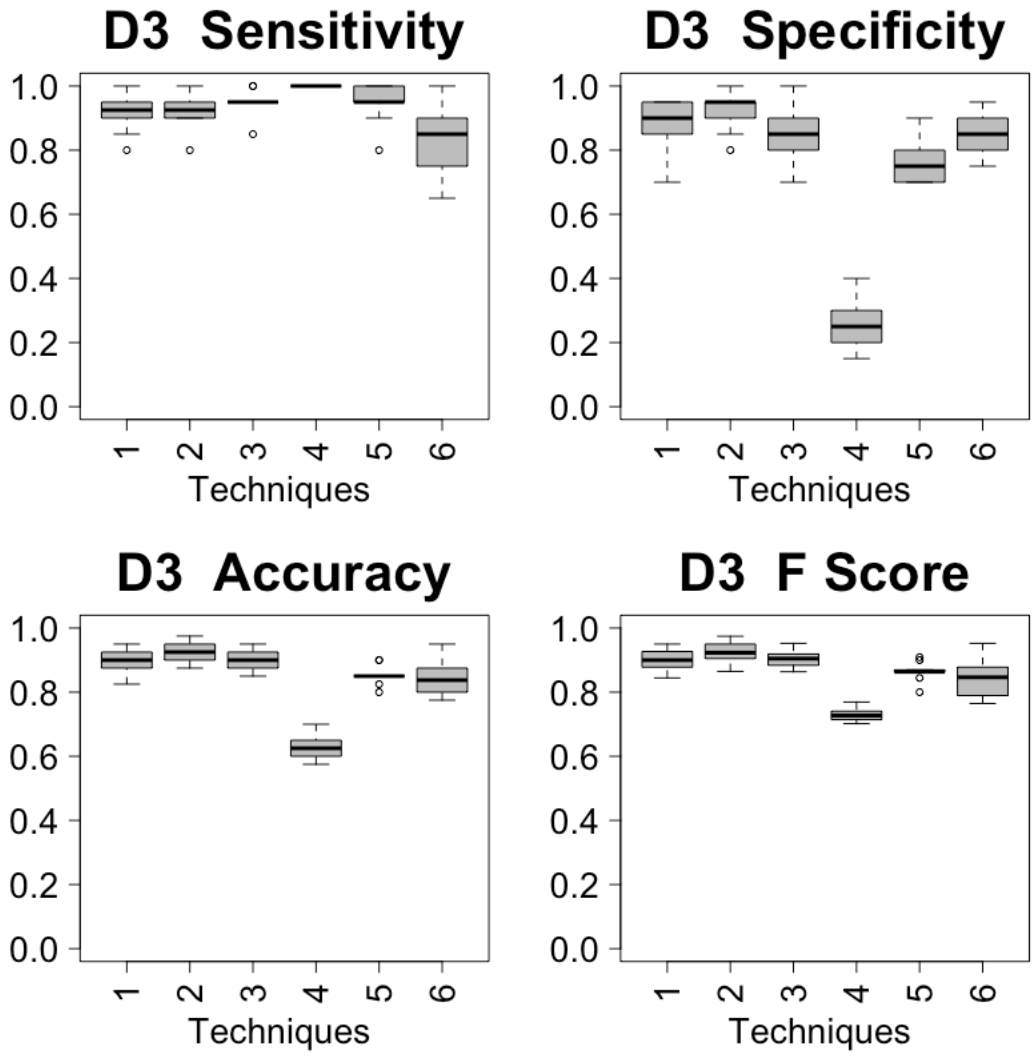

Figure 51: miRNA, mRNA and Methylation Model Variances of Predictive Models of D3

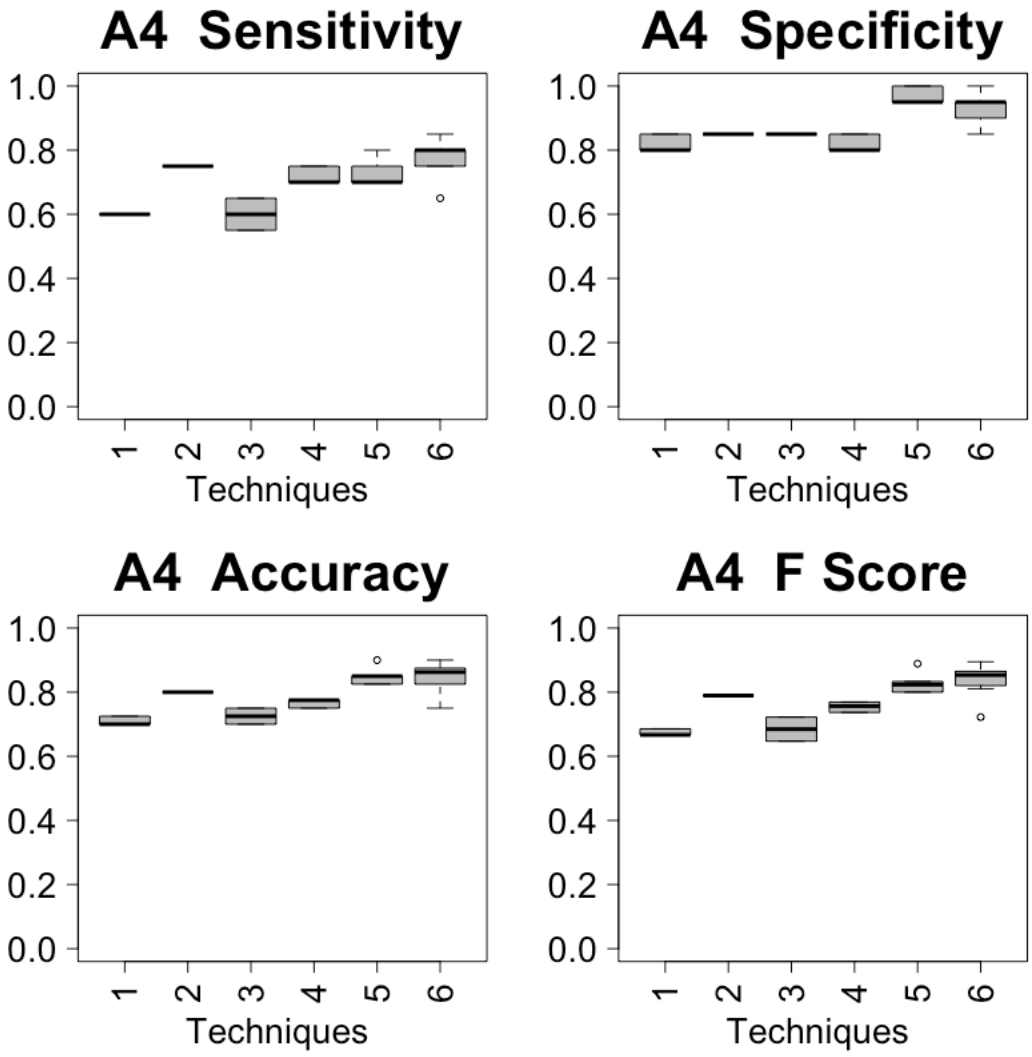

Figure 52: miRNA, mRNA and Hypo-Methylation Model Variances of Predictive Models of A4

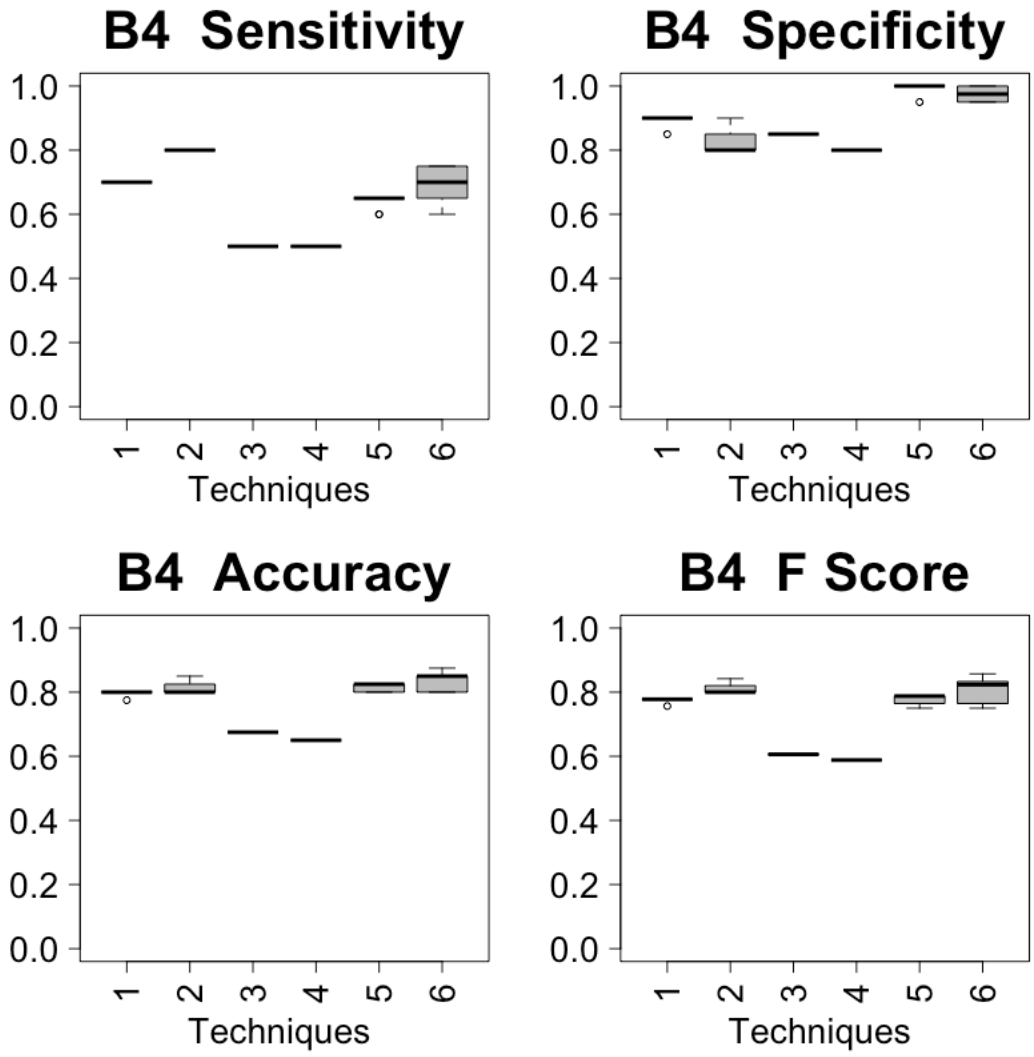

Figure 53: miRNA, mRNA and Hypo-Methylation Model Variances of Predictive Models of B4

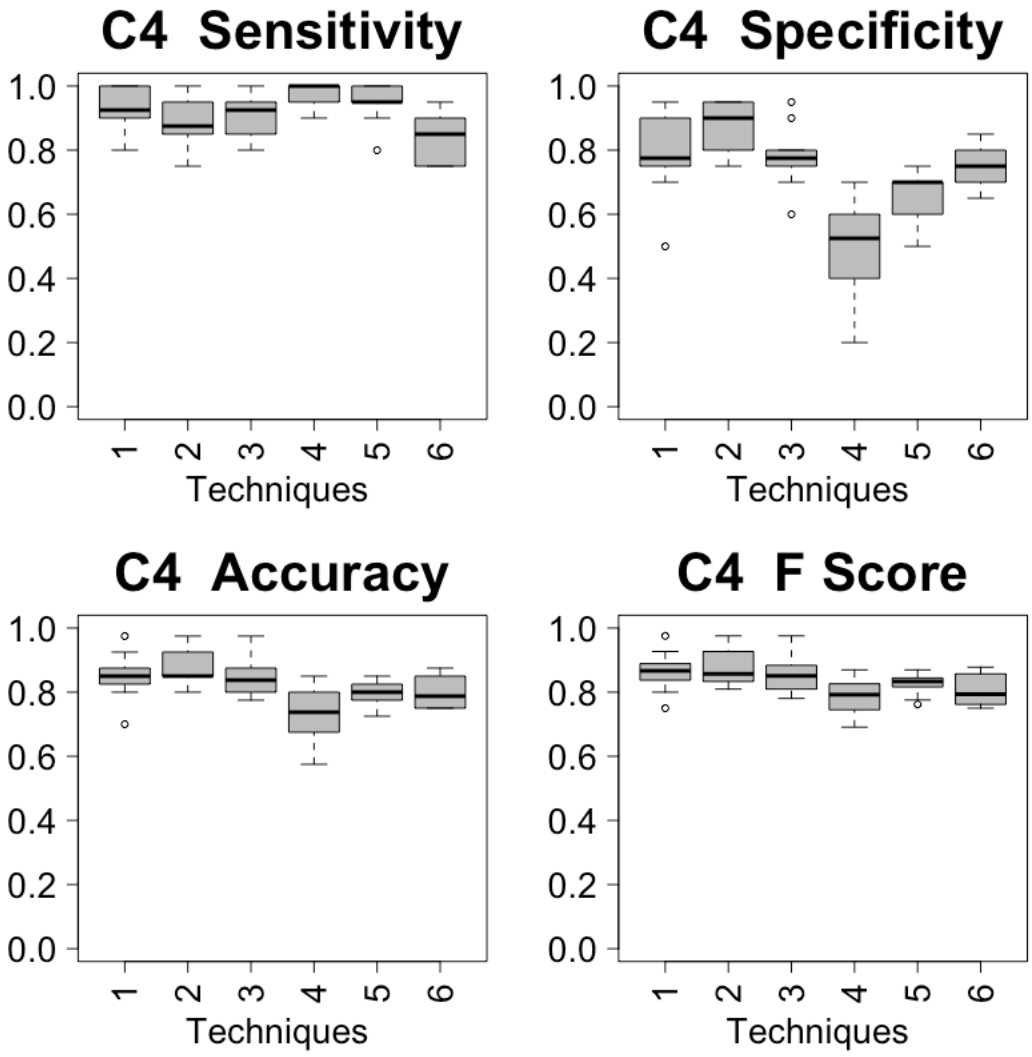

Figure 54: miRNA, mRNA and Hypo-Methylation Model Variances of Predictive Models of C4

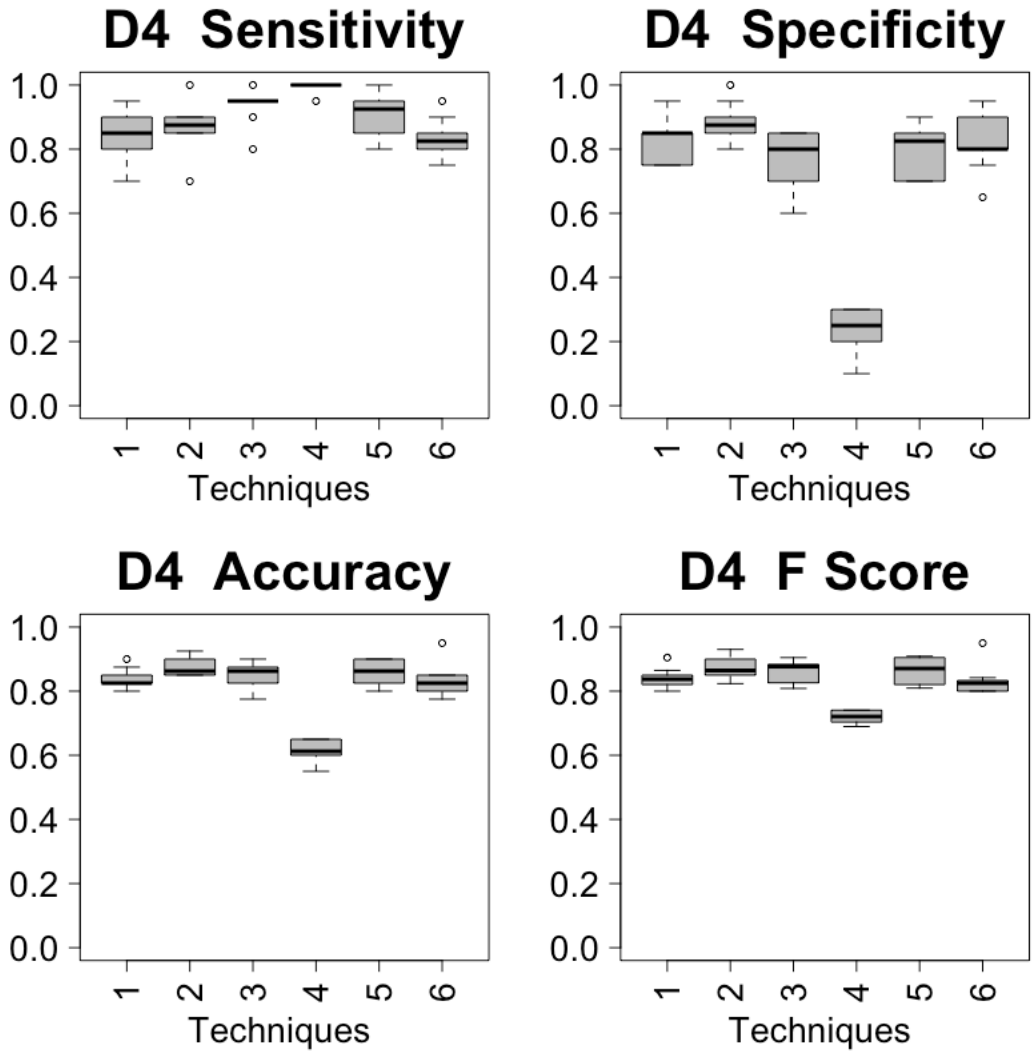

Figure 55: miRNA, mRNA and Hypo Methylation Model Variances of Predictive Models of D4

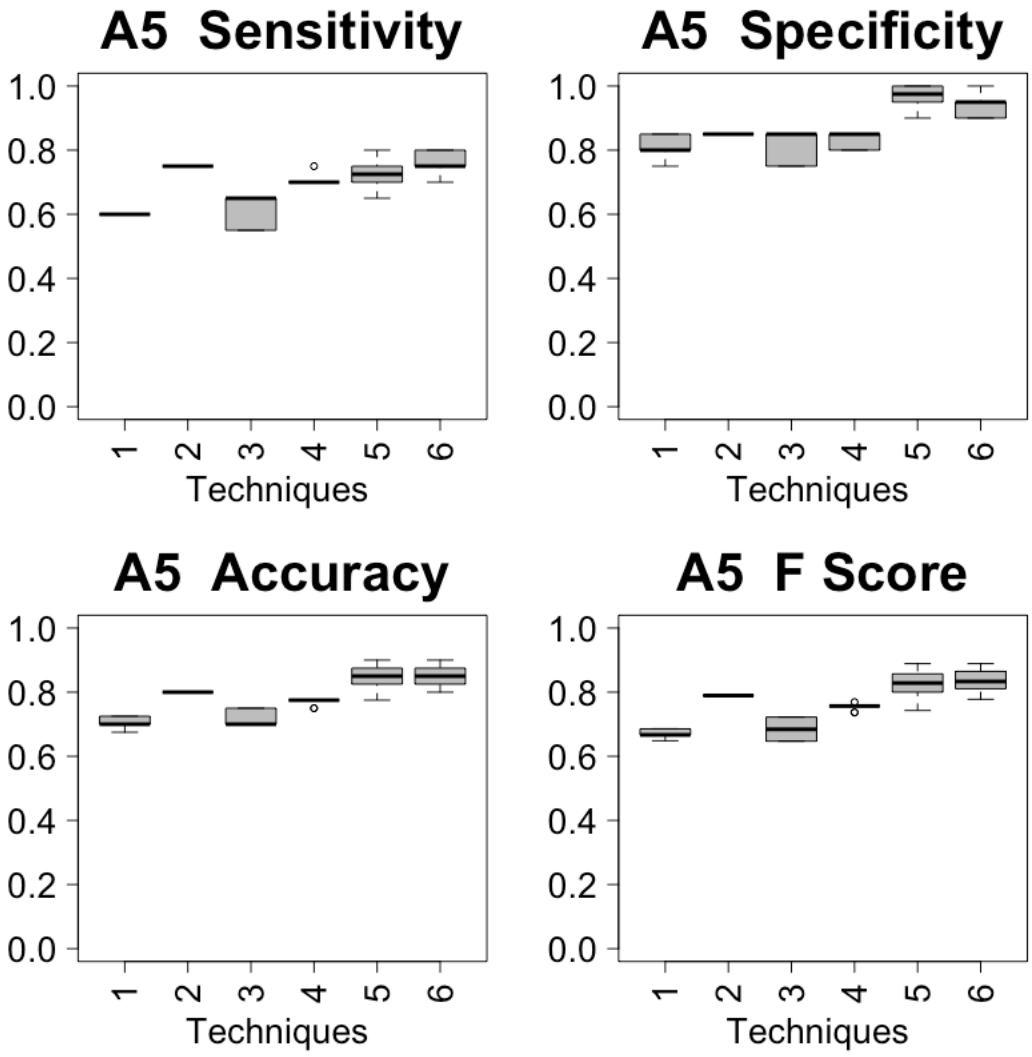

Figure 56: miRNA, mRNA and Hyper-Methylation Model Variances of Predictive Models of A5

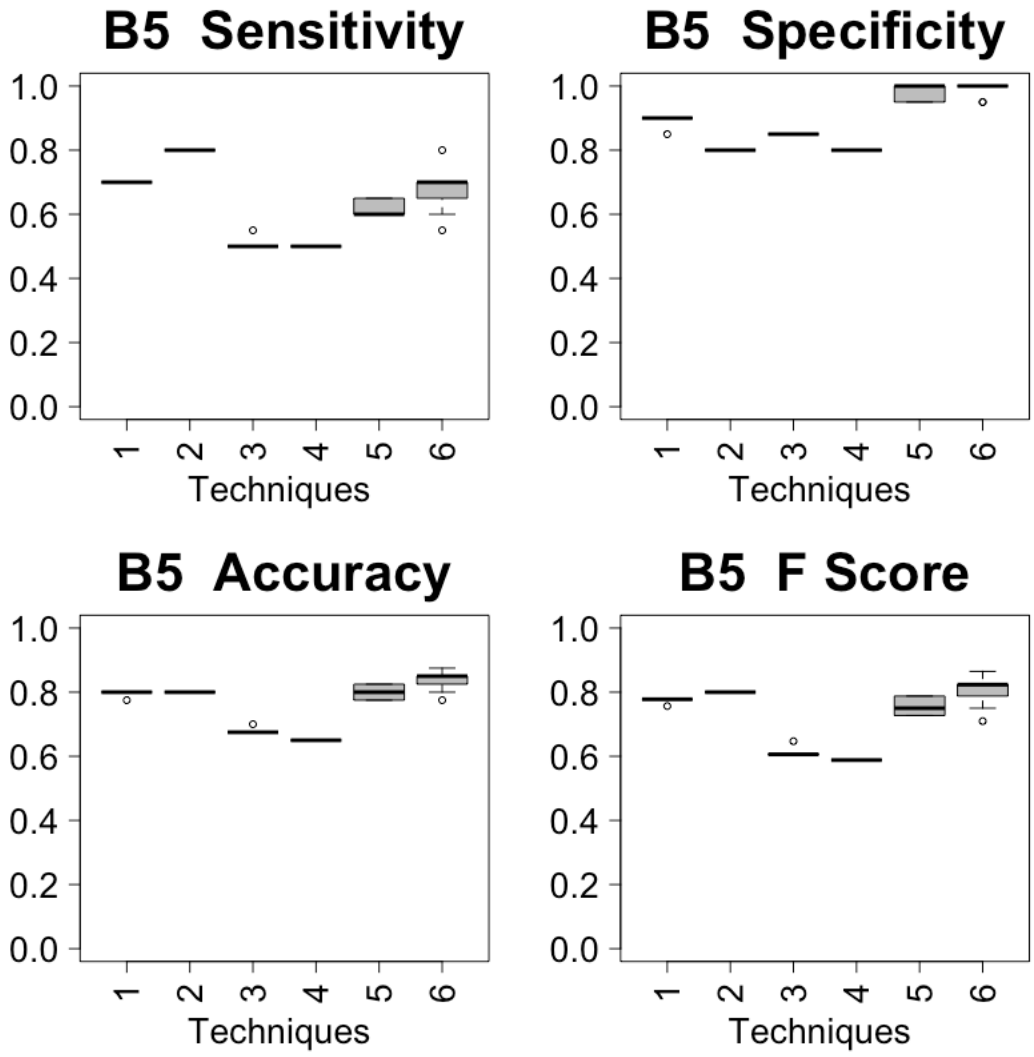

Figure 57: miRNA, mRNA and Hyper-Methylation Model Variances of Predictive Models of B5

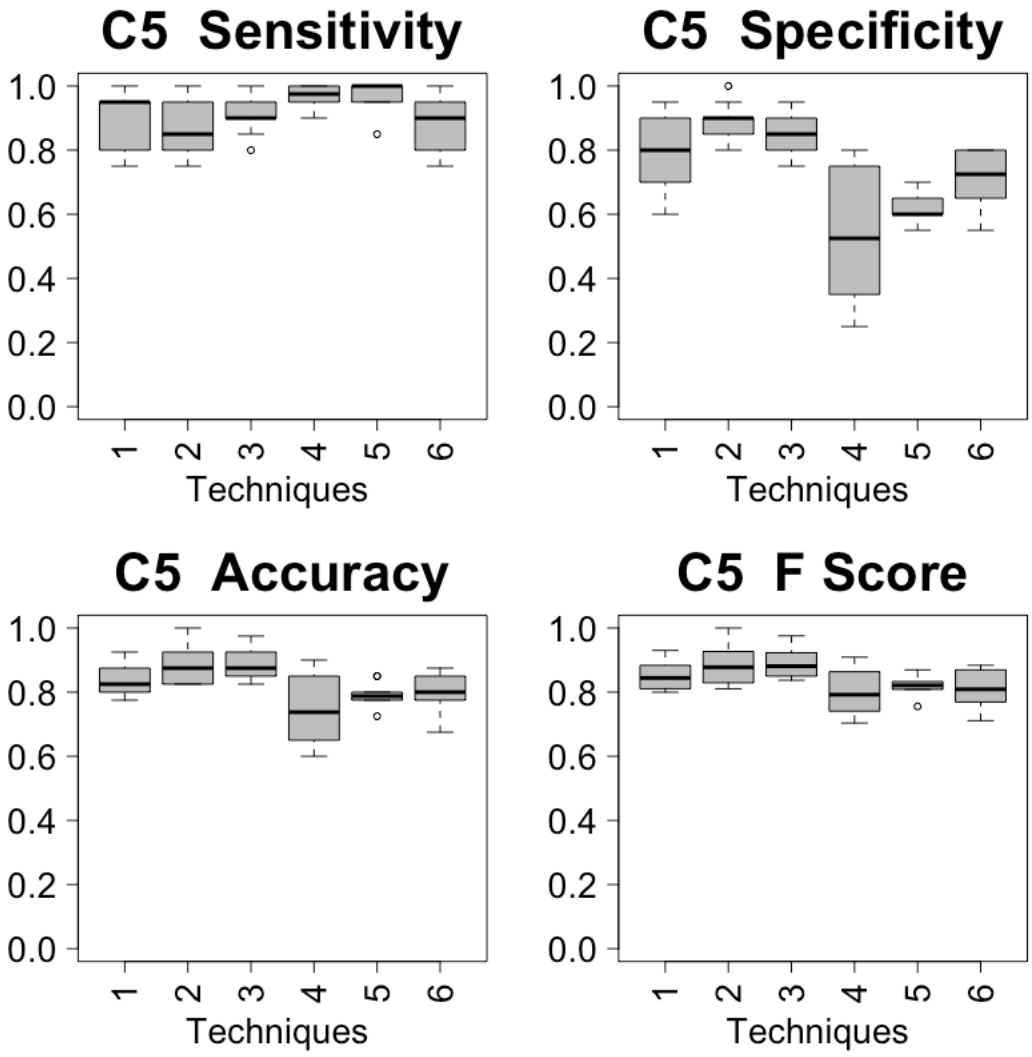

Figure 58: miRNA, mRNA and Hyper-Methylation Model Variances of Predictive Models of C5

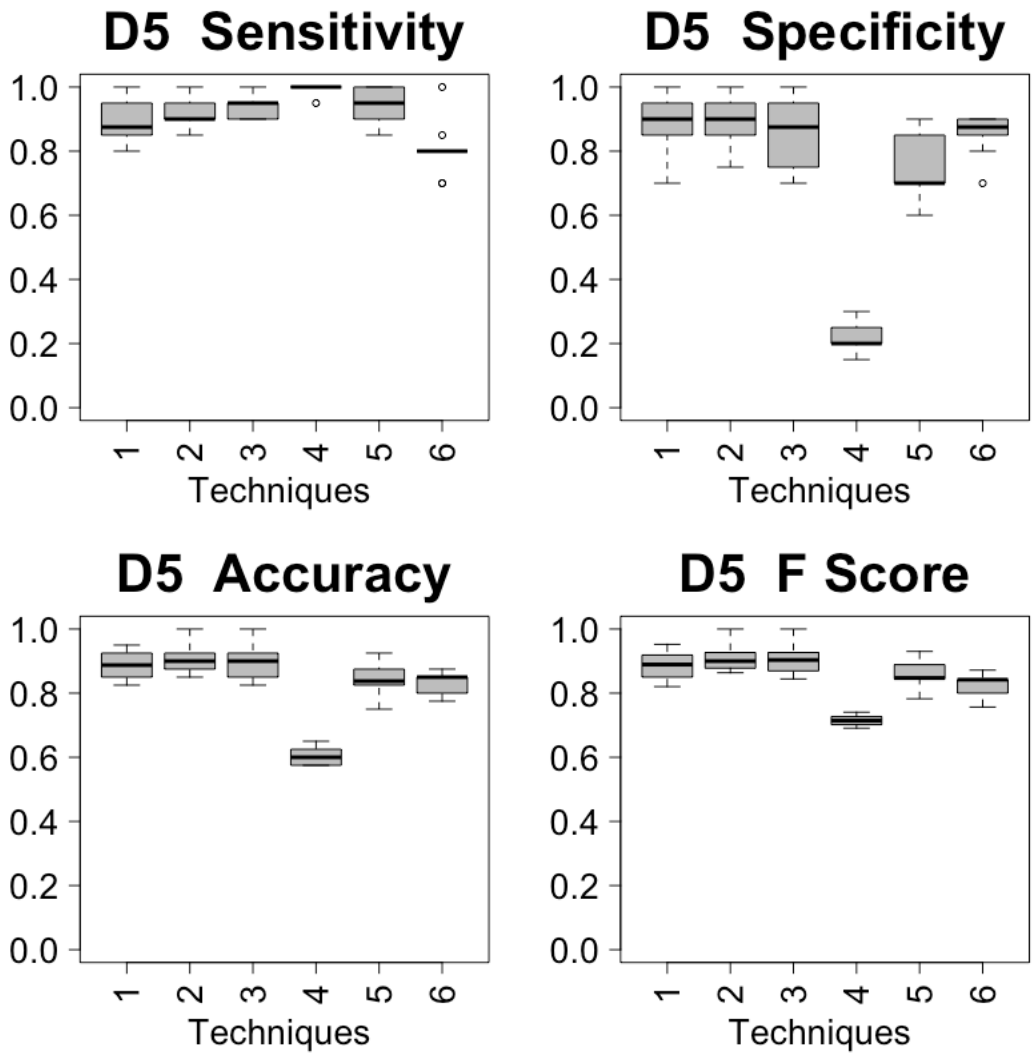

Figure 59: miRNA, mRNA and Hyper Methylation Model Variances of Predictive Models of D5

APPENDIX B

BIO-MARKERS USED IN MODELS

Table 5: miRNA and Methylated Gene relations presented Triple Model

In the final model , 7 miRNA biomarkers where presented with target differentially expressed methylated genes.

	miRNA-Methylation	*#*	*HYPER METHYLATED GENES*	*#*	*HYPO-METHYLATED GENES*
H	hsa-mir-142	3	SLC30A6, S100A11, CBD3		
	hsa-mir-3124	2	SENP1, CYP2U1		
	hsa-mir-326			1	C8orf17
	hsa-mir-331	1	IRF4		
	hsa-mir-4419b	2	TRUB2, SLC30A6		
	hsa-mir-766	5	ORC6, NDUFB5, GMPR, TMEM251, POLR1D		
L	hsa-mir-203a	1	SLC7A6OS		

Table 6: miRNA and mRNA Gene relations presented Triple Model

In the final model , miRNA biomarkers and their target mRNA where presented together.

	miRNA-mRNA	*#*	*upregulated mRNA*	*#*	*Down regulated mRNA*
H	hsa-mir-142	57	ENSG00000055208.16, ENSG00000065809.12, ENSG00000068366.18, ENSG00000069431.9, ENSG00000070961.13, ENSG00000084093.14, ENSG00000087460.22, ENSG00000088179.7, ENSG00000091009.7, ENSG00000096717.10, ENSG00000100934.13, ENSG00000101966.11, ENSG00000104689.8, ENSG00000105851.9, ENSG00000108064.9, ENSG00000112208.11, ENSG00000116062.13, ENSG00000116406.17, ENSG00000118058.19, ENSG00000122877.12, ENSG00000128585.16, ENSG00000128585.16, ENSG00000129450.7, ENSG00000134352.18, ENSG00000134644.14, ENSG00000134982.15, ENSG00000135913.9, ENSG00000136436.13, ENSG00000138398.14, ENSG00000140836.13, ENSG00000142512.13, ENSG00000143924.17, ENSG00000145365.10, ENSG00000147894.13, ENSG00000148516.20, ENSG00000152683.13, ENSG00000154217.13, ENSG00000158050.4, ENSG00000164164.14, ENSG00000164463.11, ENSG00000165209.17, ENSG00000165997.4, ENSG00000166211.7, ENSG00000166501.11, ENSG00000168944.14, ENSG00000171150.7, ENSG00000172493.19, ENSG00000175414.6, ENSG00000178695.5, ENSG00000179331.2, ENSG00000181450.16, ENSG00000181722.14, ENSG00000183918.13, ENSG00000185862.6, ENSG00000196233.10, ENSG00000197714.7, ENSG00000198162.11	0	-
	hsa-mir-29c	17	ENSG00000078304.18, NSG00000084093.14, ENSG00000086758.14, ENSG00000096717.10, ENSG00000117020.15, NSG00000135426.13, ENSG00000139218.16, ENSG00000143384.11, ENSG00000143970.15, ENSG00000144468.15, ENSG00000156675.14, ENSG00000158050.4, ENSG00000159873.8, ENSG00000162924.12, NSG00000164164.14, ENSG00000169756.15, ENSG00000179454.12	0	-
	hsa-mir-3124	23	ENSG00000001561.6, ENSG00000006459.9, ENSG00000028277.19, ENSG00000101347.8, ENSG00000101966.11, ENSG00000105866.12, ENSG00000111727.10 ENSG00000112486.13, ENSG00000113441.14, ENSG00000121067.16, ENSG00000134460.14, ENSG00000151789.8, ENSG00000152601.16, ENSG00000153201.14, ENSG00000154305.15, ENSG00000157827.18, ENSG00000160593.16, ENSG00000163635.16,	0	-

		ENSG00000164463.11, ENSG00000186265.8, ENSG00000196233.10, ENSG00000205542.9, ENSG00000273274.1		
hsa-mir-3130	12	ENSG00000089916.16, ENSG00000091009.7, ENSG00000115947.12, ENSG00000119820.9, ENSG00000130584.9, ENSG00000143751.9, ENSG00000162924.12, ENSG00000173473.9, ENSG00000174749.5, ENSG00000185155.10, ENSG00000185215.7, ENSG00000197714.7	0	-
hsa-mir-326	10	ENSG00000009307.14, ENSG00000114331.11, ENSG00000120885.18, ENSG00000129003.14, ENSG00000143970.15, ENSG00000165997.4, ENSG00000171051.7, ENSG00000171988.16, ENSG00000172943.17 ENSG00000179813.5	0	-
hsa-mir-331	24	ENSG00000065526.9, ENSG00000079819.15, ENSG00000085224.19, ENSG00000087460.22, ENSG00000096717.10, ENSG00000100055.19, ENSG00000103187.7, ENSG00000112658.7, ENSG00000115977.17, NSG00000116285.11, ENSG00000117500.11, ENSG00000117713.16, ENSG00000131931.7, ENSG00000136436.13, ENSG00000138443.14, ENSG00000144381.15, ENSG00000148606.11, NSG00000156976.13, NSG00000160352.14, ENSG00000162924.12, ENSG00000163960.10, ENSG00000165821.10, ENSG00000172943.17, ENSG00000241644.2	0	-
hsa-mir-4419b	39	ENSG00000021574.10, ENSG00000074660.14, ENSG00000074706.12, ENSG00000076108.10, ENSG00000080345.16, ENSG00000081320.9, ENSG00000090659.16, ENSG00000090924.13, ENSG00000101974.13, NSG00000106948.15, ENSG00000108061.10, ENSG00000112658.7, ENSG00000117593.9, ENSG00000119820.9, ENSG00000120693.12, ENSG00000120868.12, ENSG00000125637.14, ENSG00000125730.15, ENSG00000131931.7, ENSG00000134070.4, ENSG00000134371.9, ENSG00000136052.8, ENSG00000136731.11, ENSG00000139146.12, ENSG00000151320.9, ENSG00000152683.13, ENSG00000158473.6, ENSG00000160049.10, ENSG00000161405.15, ENSG00000162692.9, ENSG00000163792.6, ENSG00000166822.11, ENSG00000167984.15, ENSG00000174579.3, ENSG00000178607.14, ENSG00000181896.10, ENSG00000186265.8, ENSG00000269743.2, ENSG00000273274.1	0	-
hsa-mir-4444	1	ENSG00000143603.17	0	-
hsa-mir-4474	8	ENSG00000096717.10, ENSG00000116406.17, ENSG00000137478.13, ENSG00000159217.8, ENSG00000168214.19, ENSG00000168685.13, ENSG00000171150.7, ENSG00000181450.16	0	-
hsa-mir-4491	3	ENSG00000108061.10, ENSG00000174749.5, ENSG00000181896.10	0	-
hsa-mir-4523	1	ENSG00000198113.2	0	-
hsa-mir-625	13	ENSG00000076108.10, ENSG00000100647.7, ENSG00000101109.10, ENSG00000102245.6, ENSG00000127124.12, ENSG00000130584.9, ENSG00000159217.8, ENSG00000159873.8, ENSG00000162885.11 , ENSG00000179454.12, ENSG00000180370.9, ENSG00000187239.15, ENSG00000254087.6	0	-
hsa-mir-766	61	ENSG00000003436.13, ENSG00000051108.13, ENSG00000065923.8, ENSG00000074660.14, ENSG00000077044.8, ENSG00000087589.15, ENSG00000090382.5, ENSG00000090659.16, ENSG00000090924.13, ENSG00000100625.8, ENSG00000101109.10, ENSG00000101966.11, ENSG00000103522.14, ENSG00000105246.5, ENSG00000108061.10, ENSG00000108819.10, ENSG00000110395.4, ENSG00000114331.11, ENSG00000122188.11, ENSG00000124507.9, ENSG00000125107.15, ENSG00000125730.15, ENSG00000126003.6, ENSG00000126003.6, ENSG00000129450.7, ENSG00000130935.8, ENSG00000131471.5, ENSG00000132704.14, ENSG00000134070.4, ENSG00000134371.9, ENSG00000138134.10, ENSG00000138439.11, ENSG00000146112.10, ENSG00000146955.9, ENSG00000148730.6, ENSG00000150630.3, ENSG00000153147.5, ENSG00000161405.15, ENSG00000163349.20, ENSG00000164164.14, ENSG00000165821.10, ENSG00000166250.10, ENSG00000166822.11, ENSG00000167077.11, ENSG00000171408.12, ENSG00000175414.6, ENSG00000176390.11, NSG00000178385.12, ENSG00000178502.5, ENSG00000180616.7, ENSG00000181450.16,	0	-

			ENSG00000181896.10, ENSG00000183484.10, ENSG00000186265.8, ENSG00000188994.11, ENSG00000197057.7, ENSG00000197714.7, ENSG00000239264.7, ENSG00000243811.6, ENSG00000273274.1, ENSG00000278195.1		
L	hsa-mir-203a	40	ENSG00000004468.11, ENSG00000049249.7, ENSG00000068796.15, ENSG00000087460.22, ENSG00000100985.7, ENSG00000104432.11, ENSG00000106483.10, ENSG00000108854.14, ENSG00000116473.13, ENSG00000116473.13, ENSG00000117713.16, ENSG00000118260.13, ENSG00000119004.13, ENSG00000120693.12, ENSG00000121879.3, ENSG00000125630.14, ENSG00000126003.6, ENSG00000126778.8, ENSG00000128512.18, ENSG00000139146.12, ENSG00000142875.18, ENSG00000143190.20, ENSG00000148516.20, ENSG00000152601.16, ENSG00000153561.11, ENSG00000157827.18, ENSG00000158985.12, ENSG00000160791.13, ENSG00000163848.17, ENSG00000173166.16, ENSG00000175471.18, ENSG00000179454.12, ENSG00000185551.11, ENSG00000187800.12, ENSG00000188452.12, ENSG00000196233.10, ENSG00000196428.11, ENSG00000196428.11, ENSG00000205268.9 , ENSG00000266412.4	0	-

Table 7: Differentially Methylated Genes Selected in Triple Model
In the final model there was 70 Hypermethylated genes and 75 Hypomethylated genes that does not target any miRNA

	#	
Hyper Methylated Genes	70	AC004381.6, FAM214B, HMG20B, RABL2B, P2RX7, MCAT, RP1-191J18.66, UBXN8, VPS4B, CAAP1, ZNFX1-AS1_2, USP8, AP1AR, TPRKB, SLC30A5, ZC3H12C, Z98744.1, UEVLD, AAED1, BAP1, UBAP1, FAM161A, ETFDH, GOLGB1, CTD-2026K11.4, ZNF223, PTTG1IP, RPL23AP82, PIGP, AMZ2, FAM179B, SNORD1B, RNU6-560P, SNORD14B, HIST2H2BC, MIR636, MIR124-3, MIR3074, SNORD12C, ZNF845, AC079305.10, AC105247.1, NSUN5P1, Z93930.1, PET100, CTD-2192J16.11, CTB-187M2.1, AC013264.2, NDUFS5P2, RPL29P30, ARHGEF38, SMARCA5-AS1, CRNDE, ZNF564, AC004066.3, AC083843.2, AF186192.1, AF186192.5, CYP4A44P, LINC01003, SNORD3B-2, MIR3655, MIR4795, AC005625.1, CTB-129P6.11, CTD-2371O3.3, AC003956.1, CTC-339O9.2, CTD-2562G15.3, MIR6892
Hypo Methylated Genes	75	CTD-2555A7.3, SPO11, ZP2, SLAMF1, TTF1, HRH4, FBXO10, MS4A1, CCL15-CCL14, CRYGC, MS4A5, CCL23, C7orf33, OR5A2, ANKS4B, RPL24P4, KRTAP6-1, PRG2, KRTAP22-1, KRTAP6-2, KRTAP19-4, KRTAP15-1, KRTAP21-1, MPEG1, DCLRE1A, RNU6-940P, RN7SKP249, GCNT6, KRTAP5-6, KRTAP20-3, KRTAP20-4, KRTAP271, TRAJ43, TRAJ42, TRAJ29, TRAJ24, AC005229.7, NUTF2P5, AP000357.4, ARL5AP4, SUCLA2P1, AC004899.3, IGBP1P1, THRAP3P1, AC013436.6, LINC00692, RP11-101C11.1, AC113610.1, SPA17P1, KLF7-IT1, RN7SL337P, ERVFRD-1, KRTAP20-1, TRAJ60, BRD9P2, CTD-2083E4.5, RTEL1P1, RNU6ATAC32P, RNU6-327P, SERPINE3, KB-1107E3.1, KB-1615E4.2, CTC-535M15.2, LINC00520, CTD-2033D15.1, HLA-P, AF001550.7, MIR4288, RN7SL612P, CCL15, GS1-21A4.2, MIR7848, CTD-2373N4.3, FLJ42393, LINC01584

Table 8: Differentially Expressed miRNA Selected in Triple Model
In the final model there was 1529 upregulated genes and 9 downregulated mRNA which is not target of any miRNA

	#	
Down regulated mRNA	9	ENSG00000137675.4, ENSG00000164687.9, ENSG00000168703.5, ENSG00000178184.14, ENSG00000180921.6, ENSG00000198695.2, ENSG00000226145.6, ENSG00000262133.1, ENSG00000276023.3,
Up regulated RNA	1529	ENSG00000000938.11, ENSG00000002933.6, ENSG00000003096.12, ENSG00000003400.13, ENSG00000003402.18, ENSG00000005020.11, ENSG00000005700.13, ENSG00000005844.16, ENSG00000007129.16, ENSG00000007312.11, ENSG00000007314.10, ENSG00000008277.13, ENSG00000008952.15, ENSG00000009790.13, ENSG00000010610.8, ENSG00000010671.14, ENSG00000010818.7, ENSG00000011590.12, ENSG00000011600.10, ENSG00000012779.9, ENSG00000013725.13, ENSG00000015133.17, ENSG00000015285.9, ENSG00000019169.10, ENSG00000022267.15, ENSG00000023902.12, ENSG00000024048.9, ENSG00000025434.17,

ENSG00000026103.18, ENSG00000026297.14, ENSG00000026751.15, ENSG00000027075.12, ENSG00000028137.15, ENSG00000030066.12, ENSG00000030419.15, ENSG00000032219.17, ENSG00000033170.15, ENSG00000033178.11, ENSG00000033800.12, ENSG00000035720.6, ENSG00000037749.10, ENSG00000038427.14, ENSG00000038945.13, ENSG00000039123.14, ENSG00000039537.12, ENSG00000040199.17, ENSG00000040933.14, ENSG00000042980.11, ENSG00000043462.10, ENSG00000047410.12,
ENSG00000047457.12, ENSG00000048462.9, ENSG00000048991.15, ENSG00000049768.13, ENSG00000054219.10, ENSG00000054267.19, ENSG00000054282.14, ENSG00000055163.17, ENSG00000055917.14, ENSG00000056097.14, ENSG00000056277.14, ENSG00000057657.13, ENSG00000058063.14, ENSG00000058091.15, ENSG00000058272.14, ENSG00000060982.13, ENSG00000061918.11, ENSG00000062650.16, ENSG00000064218.4, ENSG00000064989.11, ENSG00000065328.15, ENSG00000065534.17, ENSG00000065613.12, ENSG00000065675.13, ENSG00000065717.13, ENSG00000065882.14, ENSG00000066056.12, ENSG00000066294.13, ENSG00000066336.10, ENSG00000066422.4, ENSG00000066583.10, ENSG00000067369.12, ENSG00000068784.11, ENSG00000068831.17, ENSG00000068976.12, ENSG00000069122.17, ENSG00000069275.12, ENSG00000070190.11, ENSG00000070915.8, ENSG00000071054.14, ENSG00000071073.11, ENSG00000071246.9, ENSG00000072401.13, ENSG00000072736.17, ENSG00000072818.10, ENSG00000072858.9, ENSG00000073614.10, ENSG00000073754.5, ENSG00000073849.13, ENSG00000073861.2, ENSG00000074370.16, ENSG00000074966.9, ENSG00000075151.18, ENSG00000075213.9, ENSG00000075420.11, ENSG00000076662.8, ENSG00000076770.13, ENSG00000077097.12, ENSG00000077420.14, ENSG00000078177.12, ENSG00000078269.12, ENSG00000078589.11, ENSG00000078674.16, ENSG00000079263.17, ENSG00000079335.16, ENSG00000080200.8, ENSG00000080298.14, ENSG00000081019.12, ENSG00000081189.12, ENSG00000081237.17, ENSG00000082074.14, ENSG00000083168.8, ENSG00000083454.20, ENSG00000083799.16, ENSG00000083828.14, ENSG00000084070.10, ENSG00000084676.14, ENSG00000085265.9, ENSG00000085276.16, ENSG00000085514.14, ENSG00000086200.15, ENSG00000086730.15, ENSG00000087303.15, ENSG00000088205.11, ENSG00000088340.14, ENSG00000088827.11, ENSG00000089012.13, ENSG00000089505.16, ENSG00000089639.9, ENSG00000089820.14, ENSG00000090060.16, ENSG00000091106.17, ENSG00000091317.7, ENSG00000091490.9, ENSG00000092051.15, ENSG00000092871.15, ENSG00000093072.14, ENSG00000093217.8, ENSG00000095002.11, ENSG00000095370.18, ENSG00000095574.10, ENSG00000095951.15, ENSG00000096654.14, ENSG00000099250.16, ENSG00000099308.9, ENSG00000099715.13, ENSG00000100060.16, ENSG00000100079.6, ENSG00000100122.5, ENSG00000100234.11, ENSG00000100281.12, ENSG00000100336.16, ENSG00000100346.16, ENSG00000100365.13, ENSG00000100368.12, ENSG00000100385.12, ENSG00000100578.13, ENSG00000100580.7, ENSG00000100600.13, ENSG00000100628.10, ENSG00000100629.15, ENSG00000100731.14, ENSG00000100815.11, ENSG00000101017.12, ENSG00000101082.12, ENSG00000101307.14, ENSG00000101310.13, ENSG00000101336.11, ENSG00000101916.11, ENSG00000101972.17, ENSG00000102043.14, ENSG00000102096.9, ENSG00000102401.18, ENSG00000102524.10, ENSG00000102755.9,
ENSG00000102879.14, ENSG00000102893.14, ENSG00000103365.14, ENSG00000103479.13, ENSG00000103540.15, ENSG00000103657.12, ENSG00000104043.13, ENSG00000104133.13, ENSG00000104213.11, ENSG00000104814.11, ENSG00000104894.10, ENSG00000104903.4, ENSG00000104972.13, ENSG00000104974.9, ENSG00000105122.11, ENSG00000105329.8, ENSG00000105339.9, ENSG00000105366.14, ENSG00000105369.8, ENSG00000105501.10, ENSG00000105639.17, ENSG00000105738.9, ENSG00000105967.14, ENSG00000106415.11, ENSG00000106511.5, ENSG00000106537.7, ENSG00000106560.9, ENSG00000106565.16, ENSG00000106952.6, ENSG00000107099.14, ENSG00000107290.12, ENSG00000107562.15, ENSG00000107581.11, ENSG00000107625.11, ENSG00000107669.16, ENSG00000107736.18, ENSG00000107742.11, ENSG00000107798.16, ENSG00000107864.13, ENSG00000107890.15, ENSG00000108055.9, ENSG00000108370.14, ENSG00000108405.3, ENSG00000108506.10, ENSG00000108510.8, ENSG00000108622.9, ENSG00000108798.7, ENSG00000108946.13, ENSG00000109320.10, ENSG00000109436.7, ENSG00000109684.13, ENSG00000109943.7, ENSG00000110077.13, ENSG00000110079.15, ENSG00000110324.8, ENSG00000110422.10, ENSG00000110448.9, ENSG00000110777.10, ENSG00000110799.12, ENSG00000110848.7, ENSG00000110876.9, ENSG00000110934.9, ENSG00000111144.8, ENSG00000111262.4, ENSG00000111348.7, ENSG00000111644.6, ENSG00000111647.11, ENSG00000111679.15, ENSG00000111729.11, ENSG00000111796.3, ENSG00000111817.15, ENSG00000111879.17, ENSG00000111885.6, ENSG00000111912.17, ENSG00000112195.8, ENSG00000112214.9, ENSG00000112303.12, ENSG00000112406.4, ENSG00000112624.11, ENSG00000112782.14, ENSG00000112799.7, ENSG00000112936.17, ENSG00000112964.12, ENSG00000113088.5, ENSG00000113263.11, ENSG00000113269.12, ENSG00000113368.10, ENSG00000113532.11,

ENSG00000113555.5, ENSG00000113595.13, ENSG00000113810.14, ENSG00000114013.14, ENSG00000114127.9, ENSG00000114439.17, ENSG00000114850.5, ENSG00000114978.16, ENSG00000115020.15, ENSG00000115085.12, ENSG00000115159.14, ENSG00000115165.8, ENSG00000115232.12, ENSG00000115271.9, ENSG00000115355.14, ENSG00000115421.11, ENSG00000115464.13, ENSG00000115604.9, ENSG00000115607.8, ENSG00000115760.12, ENSG00000115816.12, ENSG00000115935.15, ENSG00000115956.9, ENSG00000115966.15, ENSG00000115970.17, ENSG00000116017.9, ENSG00000116127.16, ENSG00000116584.16, ENSG00000116678.17, ENSG00000116701.13, ENSG00000116747.11, ENSG00000116748.18, ENSG00000116824.4, ENSG00000116852.13, ENSG00000116984.11, ENSG00000117000.8, ENSG00000117091.8, ENSG00000117114.18, ENSG00000117115.11, ENSG00000117215.13, ENSG00000117335.17, ENSG00000117523.14, ENSG00000117594.8, ENSG00000117697.13, ENSG00000118007.11, ENSG00000118292.7, ENSG00000118308.13, ENSG00000118407.13, ENSG00000118412.11, ENSG00000118495.17, ENSG00000118816.8, ENSG00000118849.8, ENSG00000118873.14, ENSG00000118922.15, ENSG00000119285.9, ENSG00000119397.15, ENSG00000119535.16, ENSG00000119699.6, ENSG00000119778.13, ENSG00000119844.13, ENSG00000119927.12, ENSG00000120063.8, ENSG00000120071.11, ENSG00000120262.9, ENSG00000120279.6, ENSG00000120280.5, ENSG00000120436.3, ENSG00000120519.13, ENSG00000120659.13, ENSG00000120802.12, ENSG00000120899.16, ENSG00000120907.16, ENSG00000121104.6, ENSG00000121210.14, ENSG00000121281.11, ENSG00000121361.3, ENSG00000121380.11, ENSG00000121481.9, ENSG00000121486.10, ENSG00000121577.12, ENSG00000121594.10, ENSG00000121797.9, ENSG00000121807.5, ENSG00000121895.7, ENSG00000121966.6, ENSG00000121988.16, ENSG00000122008.14, ENSG00000122025.13, ENSG00000122122.9, ENSG00000122223.11, ENSG00000122224.16, ENSG00000122482.19, ENSG00000122862.4, ENSG00000122986.12, ENSG00000123066.6, ENSG00000123329.16, ENSG00000123338.11, ENSG00000123411.13, ENSG00000123607.13, ENSG00000123636.16, ENSG00000124019.9, ENSG00000124191.16, ENSG00000124196.5, ENSG00000124203.5, ENSG00000124256.13, ENSG00000124496.11, ENSG00000124789.10, ENSG00000125245.11, ENSG00000125354.21, ENSG00000125384.6, ENSG00000125686.10, ENSG00000125735.9, ENSG00000125810.9, ENSG00000125900.11, ENSG00000125910.5, ENSG00000126264.8, ENSG00000126353.3, ENSG00000126759.11, ENSG00000126777.16, ENSG00000126860.10, ENSG00000126970.14, ENSG00000127084.16, ENSG00000127311.8, ENSG00000128218.7, ENSG00000128262.7, ENSG00000128271.18, ENSG00000128313.2, ENSG00000128340.13, ENSG00000128604.17, ENSG00000128815.16, ENSG00000128917.6, ENSG00000128923.9, ENSG00000129173.11, ENSG00000129515.17, ENSG00000129534.12, ENSG00000129675.14, ENSG00000130024.13, ENSG00000130038.8, ENSG00000130224.13, ENSG00000130592.12, ENSG00000130755.11, ENSG00000130830.13, ENSG00000131042.12, ENSG00000131378.12, ENSG00000131401.10, ENSG00000131725.12, ENSG00000131979.17, ENSG00000132334.15, ENSG00000132465.9, ENSG00000132514.12, ENSG00000132965.8, ENSG00000133116.7, ENSG00000133216.15, ENSG00000133246.10, ENSG00000133302.11, ENSG00000133422.11, ENSG00000133561.14, ENSG00000133574.8, ENSG00000133878.7, ENSG00000134061.5, ENSG00000134072.9, ENSG00000134109.9, ENSG00000134242.14, ENSG00000134470.18, ENSG00000134516.14, ENSG00000134539.15, ENSG00000134602.14, ENSG00000134744.12, ENSG00000134987.10, ENSG00000135077.7, ENSG00000135297.14, ENSG00000135338.12, ENSG00000135362.12, ENSG00000135439.10, ENSG00000135636.12, ENSG00000135829.15, ENSG00000135837.14, ENSG00000135932.9, ENSG00000135968.18, ENSG00000135999.10, ENSG00000136167.12, ENSG00000136237.17, ENSG00000136250.10, ENSG00000136286.13, ENSG00000136367.13, ENSG00000136404.14, ENSG00000136560.12, ENSG00000136603.12, ENSG00000136628.16, ENSG00000136634.5, ENSG00000136709.10, ENSG00000136840.17, ENSG00000136869.13, ENSG00000137078.7, ENSG00000137101.11, ENSG00000137414.5, ENSG00000137462.6, ENSG00000137491.13, ENSG00000137757.9, ENSG00000137841.10, ENSG00000137868.17, ENSG00000137962.11, ENSG00000138061.10, ENSG00000138071.12, ENSG00000138078.14, ENSG00000138160.5, ENSG00000138182.13, ENSG00000138190.15, ENSG00000138336.8, ENSG00000138378.16, ENSG00000138411.9, ENSG00000138449.9, ENSG00000138615.5, ENSG00000138684.6, ENSG00000138688.14, ENSG00000138735.14, ENSG00000138755.5, ENSG00000138767.11, ENSG00000138778.10, ENSG00000138792.8, ENSG00000138814.15, ENSG00000138964.15, ENSG00000139182.12, ENSG00000139187.8, ENSG00000139193.3, ENSG00000139194.6, ENSG00000139278.8, ENSG00000139436.19, ENSG00000139567.11, ENSG00000139610.1, ENSG00000139626.14, ENSG00000139725.6, ENSG00000139910.18, ENSG00000139985.6, ENSG00000140030.5, ENSG00000140199.10, ENSG00000140285.8, ENSG00000140368.11, ENSG00000140396.11, ENSG00000140548.8, ENSG00000140678.15, ENSG00000140749.8, ENSG00000140835.9, ENSG00000140968.9, ENSG00000141068.12, ENSG00000141161.10, ENSG00000141293.14, ENSG00000141480.16,

ENSG00000141506.12, ENSG00000141968.6, ENSG00000142347.15, ENSG00000143110.10, ENSG00000143119.11, ENSG00000143156.12, ENSG00000143157.10, ENSG00000143162.7, ENSG00000143167.10, ENSG00000143185.3, ENSG00000143195.11, ENSG00000143226.12, ENSG00000143297.17, ENSG00000143344.14, ENSG00000143498.16, ENSG00000143815.13, ENSG00000143851.14, ENSG00000144028.13, ENSG00000144130.10, ENSG00000144228.7, ENSG00000144426.17, ENSG00000144476.5, ENSG00000144711.12, ENSG00000144909.7, ENSG00000145041.14, ENSG00000145088.7, ENSG00000145246.12, ENSG00000145287.9, ENSG00000145332.12, ENSG00000145416.12, ENSG00000145649.7, ENSG00000145703.14, ENSG00000145715.13, ENSG00000145730.19, ENSG00000145734.17, ENSG00000145779.7, ENSG00000145850.7, ENSG00000146070.15, ENSG00000146094.12, ENSG00000146192.13, ENSG00000146247.13, ENSG00000146476.9, ENSG00000146966.11, ENSG00000147010.16, ENSG00000147124.11, ENSG00000147138.1, ENSG00000147168.11, ENSG00000147251.14, ENSG00000147443.11, ENSG00000147570.8, ENSG00000148700.12, ENSG00000148835.10, ENSG00000148948.6, ENSG00000149781.11, ENSG00000150337.12, ENSG00000150636.14, ENSG00000150637.7, ENSG00000150681.8, ENSG00000150961.13, ENSG00000150977.10, ENSG00000150995.16, ENSG00000151461.18, ENSG00000151490.12, ENSG00000151612.14, ENSG00000151702.15, ENSG00000151779.11, ENSG00000151835.12, ENSG00000152061.20, ENSG00000152213.3, ENSG00000152315.4, ENSG00000152404.14, ENSG00000152495.9, ENSG00000152804.9, ENSG00000152969.15, ENSG00000153012.10, ENSG00000153015.14, ENSG00000153107.10, ENSG00000153214.8, ENSG00000153234.12, ENSG00000153283.11, ENSG00000153551.12, ENSG00000153563.14, ENSG00000153574.8, ENSG00000153827.12, ENSG00000154001.12, ENSG00000154016.12, ENSG00000154065.15, ENSG00000154122.11, ENSG00000154451.13, ENSG00000154736.5, ENSG00000154822.14, ENSG00000154978.11, ENSG00000155304.5, ENSG00000155307.16, ENSG00000155330.8, ENSG00000155465.17, ENSG00000155629.13, ENSG00000155640.6, ENSG00000155659.13, ENSG00000155849.14, ENSG00000155926.12, ENSG00000155962.11, ENSG00000156136.8, ENSG00000156218.11, ENSG00000156531.15, ENSG00000156650.11, ENSG00000156869.11, ENSG00000156876.9, ENSG00000157107.12, ENSG00000157303.9, ENSG00000157450.14, ENSG00000157554.17, ENSG00000158270.11, ENSG00000158517.12, ENSG00000158714.9, ENSG00000158717.9, ENSG00000158850.13, ENSG00000159189.10, ENSG00000159314.10, ENSG00000159459.10, ENSG00000159618.14, ENSG00000159640.13, ENSG00000159753.12, ENSG00000159904.10, ENSG00000160185.12, ENSG00000160219.10, ENSG00000160224.15, ENSG00000160255.15, ENSG00000160654.8, ENSG00000160856.19, ENSG00000160883.9, ENSG00000161640.14, ENSG00000161929.13, ENSG00000161940.9, ENSG00000162434.10, ENSG00000162511.7, ENSG00000162614.17, ENSG00000162618.11, ENSG00000162654.8, ENSG00000162687.15, ENSG00000162711.15, ENSG00000162739.12, ENSG00000162775.13, ENSG00000162999.11, ENSG00000163029.14, ENSG00000163110.13, ENSG00000163125.14, ENSG00000163131.9, ENSG00000163154.5, ENSG00000163214.19, ENSG00000163219.10, ENSG00000163249.8, ENSG00000163297.15, ENSG00000163322.12, ENSG00000163376.10, ENSG00000163510.12, ENSG00000163518.9, ENSG00000163519.12, ENSG00000163563.7, ENSG00000163564.13, ENSG00000163600.11, ENSG00000163606.9, ENSG00000163611.10, ENSG00000163625.14, ENSG00000163638.12, ENSG00000163823.3, ENSG00000163947.10, ENSG00000164056.9, ENSG00000164088.16, ENSG00000164116.15, ENSG00000164134.11, ENSG00000164167.8, ENSG00000164330.15, ENSG00000164430.14, ENSG00000164483.15, ENSG00000164691.15, ENSG00000164749.10, ENSG00000165140.8, ENSG00000165168.7, ENSG00000165178.9, ENSG00000165288.10, ENSG00000165406.14, ENSG00000165457.12, ENSG00000165632.7, ENSG00000165685.7, ENSG00000165694.8, ENSG00000165813.15, ENSG00000166002.5, ENSG00000166086.11, ENSG00000166128.11, ENSG00000166263.12, ENSG00000166341.7, ENSG00000166439.5, ENSG00000166448.13, ENSG00000166478.8, ENSG00000166734.17, ENSG00000166927.11, ENSG00000166963.11, ENSG00000167083.5, ENSG00000167208.13, ENSG00000167261.12, ENSG00000167286.8, ENSG00000167613.14, ENSG00000167618.8, ENSG00000167635.10, ENSG00000167664.7, ENSG00000167850.3, ENSG00000167851.12, ENSG00000167861.14, ENSG00000167874.6, ENSG00000167895.13, ENSG00000168016.12, ENSG00000168071.20, ENSG00000168081.7, ENSG00000168229.3, ENSG00000168310.9, ENSG00000168386.17, ENSG00000168404.11, ENSG00000168405.13, ENSG00000168421.11, ENSG00000168438.13, ENSG00000168497.4, ENSG00000168813.15, ENSG00000168918.12, ENSG00000168995.12, ENSG00000169403.10, ENSG00000169413.2, ENSG00000169442.7, ENSG00000169508.6, ENSG00000169896.15, ENSG00000170458.12, ENSG00000170476.14, ENSG00000170485.15, ENSG00000170525.17, ENSG00000170571.10, ENSG00000170776.18, ENSG00000170909.12, ENSG00000170989.8, ENSG00000171049.8, ENSG00000171105.12, ENSG00000171115.3, ENSG00000171227.6, ENSG00000171522.5, ENSG00000171631.13, ENSG00000171643.12, ENSG00000171657.5, ENSG00000171659.12, ENSG00000171777.14,

ENSG00000171860.4, ENSG00000172007.5, ENSG00000172071.10, ENSG00000172116.20, ENSG00000172197.10, ENSG00000172215.5, ENSG00000172243.16, ENSG00000172292.13, ENSG00000172322.12, ENSG00000172403.9, ENSG00000172469.13, ENSG00000172575.10, ENSG00000172578.10, ENSG00000172673.9, ENSG00000172724.10, ENSG00000172794.18, ENSG00000172795.14, ENSG00000172845.12, ENSG00000173145.10, ENSG00000173198.5, ENSG00000173200.11, ENSG00000173208.3, ENSG00000173221.12, ENSG00000173281.4, ENSG00000173369.14, ENSG00000173372.15, ENSG00000173391.7, ENSG00000173451.5, ENSG00000173578.7, ENSG00000173585.14, ENSG00000173626.8, ENSG00000173706.11, ENSG00000173757.8, ENSG00000173762.6, ENSG00000173889.14, ENSG00000174004.5, ENSG00000174123.9, ENSG00000174125.6, ENSG00000174197.15, ENSG00000174255.6, ENSG00000174485.13, ENSG00000174500.11, ENSG00000174600.12, ENSG00000174718.10, ENSG00000174799.9, ENSG00000174837.13, ENSG00000174885.11, ENSG00000174944.7, ENSG00000174946.6, ENSG00000175463.10, ENSG00000175489.9, ENSG00000175538.9, ENSG00000175841.8, ENSG00000175857.7, ENSG00000176160.8, ENSG00000176371.12, ENSG00000176907.4, ENSG00000176986.13, ENSG00000177076.5, ENSG00000177272.8, ENSG00000177311.9, ENSG00000177374.11, ENSG00000177455.10, ENSG00000177575.11, ENSG00000177590.7, ENSG00000177688.6, ENSG00000177721.4, ENSG00000177885.12, ENSG00000178175.10, ENSG00000178199.12, ENSG00000178343.4, ENSG00000178562.16, ENSG00000178789.7, ENSG00000179144.4, ENSG00000179163.11, ENSG00000179456.10, ENSG00000179715.11, ENSG00000179833.4, ENSG00000179840.5, ENSG00000179841.8, ENSG00000179921.13, ENSG00000179934.6, ENSG00000180061.8, ENSG00000180096.10, ENSG00000180139.11, ENSG00000180353.9, ENSG00000180448.9, ENSG00000180644.6, ENSG00000180884.9, ENSG00000181036.12, ENSG00000181631.6, ENSG00000181744.7, ENSG00000181804.13, ENSG00000181847.10, ENSG00000182022.16, ENSG00000182134.14, ENSG00000182162.8, ENSG00000182183.13, ENSG00000182463.14, ENSG00000182487.11, ENSG00000182568.15, ENSG00000182578.12, ENSG00000182866.15, ENSG00000182919.13, ENSG00000183023.17, ENSG00000183508.4, ENSG00000183542.5, ENSG00000183765.19, ENSG00000183801.6, ENSG00000183807.7, ENSG00000183813.6, ENSG00000184117.10, ENSG00000184156.14, ENSG00000184293.6, ENSG00000184619.3, ENSG00000184682.5, ENSG00000184922.12, ENSG00000185070.9, ENSG00000185245.7, ENSG00000185261.12, ENSG00000185271.6, ENSG00000185482.6, ENSG00000185650.9, ENSG00000185669.5, ENSG00000185739.12, ENSG00000185811.15, ENSG00000185900.8, ENSG00000185905.3, ENSG00000185947.13, ENSG00000186063.11, ENSG00000186074.17, ENSG00000186152.6, ENSG00000186198.3, ENSG00000186479.4, ENSG00000186517.12, ENSG00000186583.10, ENSG00000186635.13, ENSG00000186766.7, ENSG00000186818.11, ENSG00000186827.9, ENSG00000187037.7, ENSG00000187116.12, ENSG00000187210.11, ENSG00000187474.4, ENSG00000187513.8, ENSG00000187554.10, ENSG00000187653.11, ENSG00000187764.10, ENSG00000187796.12, ENSG00000187808.4, ENSG00000187862.10, ENSG00000188033.8, ENSG00000188107.12, ENSG00000188263.9, ENSG00000188389.9, ENSG00000188404.7, ENSG00000188559.12, ENSG00000188641.11, ENSG00000188820.11, ENSG00000188848.14, ENSG00000189144.12, ENSG00000189233.10, ENSG00000189350.11, ENSG00000196159.10, ENSG00000196209.11, ENSG00000196329.9, ENSG00000196371.3, ENSG00000196405.11, ENSG00000196468.7, ENSG00000196498.12, ENSG00000196504.14, ENSG00000196505.9, ENSG00000196664.4, ENSG00000196684.11, ENSG00000196839.11, ENSG00000196865.4, ENSG00000196911.8, ENSG00000196950.12, ENSG00000197142.9, ENSG00000197258.5, ENSG00000197272.2, ENSG00000197323.9, ENSG00000197405.6, ENSG00000197471.10, ENSG00000197548.11, ENSG00000197629.5, ENSG00000197872.10, ENSG00000197880.7, ENSG00000197943.8, ENSG00000197992.5, ENSG00000198075.8, ENSG00000198178.9, ENSG00000198246.7, ENSG00000198286.8, ENSG00000198369.8, ENSG00000198399.13, ENSG00000198420.8, ENSG00000198586.12, ENSG00000198589.9, ENSG00000198624.11, ENSG00000198707.13, ENSG00000198771.9, ENSG00000198785.4, ENSG00000198821.9, ENSG00000198833.6, ENSG00000198851.8, ENSG00000198873.11, ENSG00000198879.10, ENSG00000198890.7, ENSG00000198900.5, ENSG00000198919.11, ENSG00000198924.6, ENSG00000198945.6, ENSG00000198959.10, ENSG00000203497.2, ENSG00000203710.9, ENSG00000203747.8, ENSG00000204131.7, ENSG00000204136.9, ENSG00000204161.12, ENSG00000204406.10, ENSG00000204472.11, ENSG00000204475.8, ENSG00000204482.9, ENSG00000204577.10, ENSG00000204745.3, ENSG00000204872.3, ENSG00000205089.6, ENSG00000205302.5, ENSG00000205436.6, ENSG00000205537.2, ENSG00000205683.10, ENSG00000205744.8, ENSG00000205784.2, ENSG00000205809.8, ENSG00000205810.7, ENSG00000205885.6, ENSG00000205930.7, ENSG00000211592.5, ENSG00000211593.2, ENSG00000211632.3, ENSG00000211669.2, ENSG00000211677.2, ENSG00000211688.1, ENSG00000211689.5, ENSG00000211694.2, ENSG00000211695.2, ENSG00000211698.2,

ENSG00000211710.3, ENSG00000211713.3, ENSG00000211714.3, ENSG00000211716.2,
ENSG00000211720.3, ENSG00000211724.3, ENSG00000211734.3, ENSG00000211746.3,
ENSG00000211751.6, ENSG00000211752.3, ENSG00000211753.3, ENSG00000211764.1,
ENSG00000211765.1, ENSG00000211766.1, ENSG00000211767.1, ENSG00000211768.1,
ENSG00000211771.1, ENSG00000211772.7, ENSG00000211776.2, ENSG00000211778.2,
ENSG00000211779.3, ENSG00000211780.3, ENSG00000211785.1, ENSG00000211786.3,
ENSG00000211787.1, ENSG00000211788.2, ENSG00000211789.2, ENSG00000211790.2,
ENSG00000211792.2, ENSG00000211793.2, ENSG00000211794.3, ENSG00000211795.3,
ENSG00000211796.1, ENSG00000211797.2, ENSG00000211799.3, ENSG00000211801.3,
ENSG00000211803.2, ENSG00000211804.3, ENSG00000211805.1, ENSG00000211806.2,
ENSG00000211807.3, ENSG00000211809.2, ENSG00000211810.3, ENSG00000211814.1,
ENSG00000211816.2, ENSG00000211818.1, ENSG00000211820.1, ENSG00000211821.2,
ENSG00000211850.1, ENSG00000211873.1, ENSG00000211875.1, ENSG00000211876.1,
ENSG00000211878.1, ENSG00000211879.1, ENSG00000211882.1, ENSG00000211885.1,
ENSG00000211886.1, ENSG00000211896.5, ENSG00000211911.1, ENSG00000211949.3,
ENSG00000211955.2, ENSG00000211959.2, ENSG00000211962.2, ENSG00000211972.2,
ENSG00000213062.4, ENSG00000213203.2, ENSG00000213262.3, ENSG00000213809.7,
ENSG00000213876.4, ENSG00000214077.4, ENSG00000214212.7, ENSG00000214269.3,
ENSG00000215571.5, ENSG00000215910.6, ENSG00000216490.3, ENSG00000217258.2,
ENSG00000217527.1, ENSG00000217643.1, ENSG00000220008.3, ENSG00000221043.2,
ENSG00000221535.1, ENSG00000223459.5, ENSG00000223466.1, ENSG00000223552.1,
ENSG00000223612.3, ENSG00000223750.1, ENSG00000223946.1, ENSG00000223969.4,
ENSG00000224137.1, ENSG00000224220.1, ENSG00000224383.6, ENSG00000224460.1,
ENSG00000224675.1, ENSG00000224875.2, ENSG00000225079.2, ENSG00000225205.4,
ENSG00000225234.1, ENSG00000225325.1, ENSG00000225422.4, ENSG00000225460.1,
ENSG00000225490.1, ENSG00000225731.1, ENSG00000225825.1, ENSG00000225885.5,
ENSG00000225938.1, ENSG00000225974.1, ENSG00000226004.1, ENSG00000226423.1,
ENSG00000226539.1, ENSG00000226660.2, ENSG00000226751.2, ENSG00000226777.6,
ENSG00000226979.7, ENSG00000227007.1, ENSG00000227032.1, ENSG00000227145.1,
ENSG00000227155.6, ENSG00000227191.5, ENSG00000227217.1, ENSG00000227295.2,
ENSG00000227345.7, ENSG00000227449.7, ENSG00000227470.1, ENSG00000227507.2,
ENSG00000227508.5, ENSG00000227531.1, ENSG00000227550.2, ENSG00000227678.6,
ENSG00000227776.1, ENSG00000228005.1, ENSG00000228427.1, ENSG00000228763.1,
ENSG00000228800.1, ENSG00000228804.4, ENSG00000228839.4, ENSG00000228986.1,
ENSG00000229092.2, ENSG00000229153.4, ENSG00000229191.1, ENSG00000229228.1,
ENSG00000229425.1, ENSG00000229473.2, ENSG00000229590.3, ENSG00000229613.1,
ENSG00000229754.1, ENSG00000229816.1, ENSG00000229939.1, ENSG00000230006.6,
ENSG00000230099.2, ENSG00000230107.1, ENSG00000230138.1, ENSG00000230155.5,
ENSG00000230390.1, ENSG00000230499.1, ENSG00000230530.1, ENSG00000230709.1,
ENSG00000230838.1, ENSG00000231105.1, ENSG00000231123.1, ENSG00000231128.4,
ENSG00000231231.4, ENSG00000231241.1, ENSG00000231265.1, ENSG00000231435.1,
ENSG00000231621.1, ENSG00000231690.2, ENSG00000231964.1, ENSG00000232208.2,
ENSG00000232334.1, ENSG00000232613.5, ENSG00000232628.4, ENSG00000232698.1,
ENSG00000232869.2, ENSG00000232884.6, ENSG00000233038.4, ENSG00000233093.4,
ENSG00000233306.2, ENSG00000233387.1, ENSG00000233521.4, ENSG00000233673.5,
ENSG00000233922.2, ENSG00000234142.1, ENSG00000234174.1, ENSG00000234184.4,
ENSG00000234332.1, ENSG00000234389.1, ENSG00000234515.1, ENSG00000234568.3,
ENSG00000234663.4, ENSG00000234816.2, ENSG00000235052.1, ENSG00000235151.1,
ENSG00000235304.1, ENSG00000235419.4, ENSG00000235499.1, ENSG00000235532.1,
ENSG00000235568.5, ENSG00000235586.1, ENSG00000235636.1, ENSG00000235659.1,
ENSG00000235802.1, ENSG00000235831.5, ENSG00000236213.1, ENSG00000236278.2,
ENSG00000236456.1, ENSG00000236469.1, ENSG00000236525.1, ENSG00000236846.1,
ENSG00000236876.3, ENSG00000236911.5, ENSG00000236935.1, ENSG00000237254.2,
ENSG00000237286.1, ENSG00000237470.3, ENSG00000237484.5, ENSG00000237513.1,
ENSG00000237522.1, ENSG00000237604.1, ENSG00000237638.1, ENSG00000237702.2,
ENSG00000237807.3, ENSG00000237914.4, ENSG00000237943.5, ENSG00000237955.1,
ENSG00000237980.1, ENSG00000238171.1, ENSG00000238241.1, ENSG00000238290.1,
ENSG00000239213.4, ENSG00000239281.2, ENSG00000239636.1, ENSG00000239941.1,
ENSG00000239961.2, ENSG00000239964.3, ENSG00000239998.4, ENSG00000240143.1,
ENSG00000240219.1, ENSG00000240292.1, ENSG00000240487.1, ENSG00000240505.7,
ENSG00000240535.7, ENSG00000240654.5, ENSG00000240787.1, ENSG00000240891.5,
ENSG00000240954.1, ENSG00000241106.5, ENSG00000241134.3, ENSG00000241158.4,

ENSG00000241163.6, ENSG00000241351.2, ENSG00000241399.5, ENSG00000241490.1, ENSG00000241560.4, ENSG00000241717.1, ENSG00000241738.1, ENSG00000241962.8, ENSG00000242048.3, ENSG00000242258.1, ENSG00000242324.1, ENSG00000242574.7, ENSG00000242598.1, ENSG00000242736.1, ENSG00000243156.6, ENSG00000243232.4, ENSG00000243238.1, ENSG00000243544.3, ENSG00000243836.4, ENSG00000244227.4, ENSG00000244255.4, ENSG00000244273.1, ENSG00000244482.8, ENSG00000244661.1, ENSG00000244720.1, ENSG00000244968.5, ENSG00000245164.5, ENSG00000245648.1, ENSG00000245954.5, ENSG00000246084.2, ENSG00000246582.2, ENSG00000247199.3, ENSG00000247774.5, ENSG00000248383.4, ENSG00000248441.5, ENSG00000248571.1, ENSG00000248971.2, ENSG00000248996.1, ENSG00000249096.5, ENSG00000249129.1, ENSG00000249141.1, ENSG00000249334.1, ENSG00000249388.1, ENSG00000249437.6, ENSG00000249454.1, ENSG00000249667.1, ENSG00000249669.6, ENSG00000249835.2, ENSG00000249978.1, ENSG00000250050.1, ENSG00000250541.1, ENSG00000250629.1, ENSG00000250654.6, ENSG00000250687.5, ENSG00000250722.4, ENSG00000251009.2, ENSG00000251131.1, ENSG00000251301.5, ENSG00000251332.1, ENSG00000251363.2, ENSG00000251664.3, ENSG00000251922.1, ENSG00000251962.1, ENSG00000252503.1, ENSG00000253361.1, ENSG00000253409.1, ENSG00000253535.4, ENSG00000253647.1, ENSG00000253686.1, ENSG00000253690.1, ENSG00000253701.2, ENSG00000253755.1, ENSG00000253822.1, ENSG00000253837.1, ENSG00000253930.1, ENSG00000253957.1, ENSG00000254041.1, ENSG00000254100.1, ENSG00000254167.1, ENSG00000254198.1, ENSG00000254340.1, ENSG00000254415.3, ENSG00000254760.1, ENSG00000254802.1, ENSG00000254838.5, ENSG00000254887.1, ENSG00000254911.3, ENSG00000254959.5, ENSG00000255090.4, ENSG00000255163.1, ENSG00000255197.4, ENSG00000255340.1, ENSG00000255422.1, ENSG00000255441.1, ENSG00000255569.1, ENSG00000255733.4, ENSG00000255819.5, ENSG00000255833.1, ENSG00000255882.1, ENSG00000256262.1, ENSG00000256540.1, ENSG00000257093.5, ENSG00000257221.1, ENSG00000257315.1, ENSG00000257594.3, ENSG00000257894.2, ENSG00000257906.1, ENSG00000257924.1, ENSG00000258086.1, ENSG00000258181.1, ENSG00000258511.1, ENSG00000258546.1, ENSG00000258810.1, ENSG00000258867.4, ENSG00000258878.1, ENSG00000258926.1, ENSG00000259004.1, ENSG00000259005.1, ENSG00000259124.1, ENSG00000259436.1, ENSG00000259628.1, ENSG00000259772.5, ENSG00000259834.1, ENSG00000259847.1, ENSG00000260228.4, ENSG00000260244.1, ENSG00000260314.2, ENSG00000260496.3, ENSG00000260517.2, ENSG00000260719.1, ENSG00000260828.1, ENSG00000260861.5, ENSG00000261208.1, ENSG00000261218.4, ENSG00000261269.1, ENSG00000261371.4, ENSG00000261416.1, ENSG00000261471.1, ENSG00000261644.1, ENSG00000261757.1, ENSG00000262039.1, ENSG00000262097.1, ENSG00000262151.1, ENSG00000262823.1, ENSG00000263264.1, ENSG00000263413.2, ENSG00000263809.1, ENSG00000264188.1, ENSG00000264219.1, ENSG00000264773.1, ENSG00000264781.1, ENSG00000264869.1, ENSG00000265148.4, ENSG00000265517.1, ENSG00000265612.1, ENSG00000265714.1, ENSG00000265719.1, ENSG00000265975.1, ENSG00000266094.5, ENSG00000266283.1, ENSG00000266389.1, ENSG00000266750.1, ENSG00000266804.1, ENSG00000267045.1, ENSG00000267364.1, ENSG00000267653.1, ENSG00000267654.1, ENSG00000267764.1, ENSG00000268027.4, ENSG00000268041.1, ENSG00000268201.1, ENSG00000268510.1, ENSG00000268861.4, ENSG00000269220.1, ENSG00000269404.5, ENSG00000269800.1, ENSG00000269904.2, ENSG00000269919.1, ENSG00000269937.1, ENSG00000269967.1, ENSG00000270547.4, ENSG00000270550.1, ENSG00000270661.1, ENSG00000271680.1, ENSG00000271779.1, ENSG00000271820.1, ENSG00000272053.1, ENSG00000272211.1, ENSG00000272256.1, ENSG00000272382.1, ENSG00000272477.1, ENSG00000272498.1, ENSG00000272563.1, ENSG00000272567.1, ENSG00000272763.1, ENSG00000272886.4, ENSG00000272908.1, ENSG00000272917.1, ENSG00000273107.1, ENSG00000273123.1, ENSG00000273172.1, ENSG00000273341.1, ENSG00000273348.1, ENSG00000273433.1, ENSG00000273445.1, ENSG00000273669.1, ENSG00000273837.1, ENSG00000273855.1, ENSG00000273923.1, ENSG00000274008.1, ENSG00000274128.1, ENSG00000274134.1, ENSG00000274172.1, ENSG00000274752.1, ENSG00000274961.1, ENSG00000275052.3, ENSG00000275158.1, ENSG00000275302.1, ENSG00000275743.1, ENSG00000275772.1, ENSG00000276231.3, ENSG00000276317.1, ENSG00000276334.1, ENSG00000276405.1, ENSG00000276454.1, ENSG00000276557.1, ENSG00000276819.1, ENSG00000276842.1, ENSG00000276961.1, ENSG00000276980.1, ENSG00000277030.1, ENSG00000277117.3, ENSG00000277734.3, ENSG00000277855.1, ENSG00000277882.1, ENSG00000278030.1, ENSG00000279078.1, ENSG00000279082.2, ENSG00000279192.1, ENSG00000279311.1, ENSG00000279380.1, ENSG00000279406.1, ENSG00000279481.1, ENSG00000279541.1, ENSG00000279631.1, ENSG00000280008.1, ENSG00000280014.1, ENSG00000280143.1,

ENSG00000280194.1, ENSG00000280202.1, ENSG00000280304.1, ENSG00000280551.1, ENSG00000280734.1, ENSG00000281103.1, ENSG00000281741.1

APPENDIX C

PATHWAYS

Table 9: Top 15 Pathway of Genes for selected miRNA, mRNA and Methylation Markers

Category	Term	Count	LT	PH	PT	%	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	Chemokine signaling pathway (*)	84	1630	186	6879	1.9	1.60E-10	1.9	4.80E-08	1.60E-08	2.20E-07
KEGG_PATHWAY	Cytokine-cytokine receptor interaction	83	1630	243	6879	1.9	1.90E-04	1.4	5.50E-02	4.70E-03	2.50E-01
KEGG_PATHWAY	Endocytosis	83	1630	241	6879	1.9	1.40E-04	1.5	4.00E-02	4.10E-03	1.90E-01
KEGG_PATHWAY	Focal adhesion	82	1630	206	6879	1.9	2.80E-07	1.7	8.30E-05	1.00E-05	3.70E-04
KEGG_PATHWAY	HTLV-I infection	103	1630	254	6879	2.4	2.00E-09	1.7	6.00E-07	1.20E-07	2.70E-06
KEGG_PATHWAY	MAPK signaling pathway	84	1630	253	6879	1.9	4.90E-04	1.4	1.30E-01	1.00E-02	6.50E-01
KEGG_PATHWAY	Osteoclast differentiation (*)	72	1630	131	6879	1.7	2.90E-14	2.3	8.60E-12	8.60E-12	3.90E-11
KEGG_PATHWAY	Pathways in cancer	155	1630	393	6879	3.6	1.20E-12	1.7	3.60E-10	1.80E-10	1.60E-09
KEGG_PATHWAY	PI3K-Akt signaling pathway	116	1630	345	6879	2.7	1.90E-05	1.4	5.60E-03	6.30E-04	2.60E-02
KEGG_PATHWAY	Proteoglycans in cancer	83	1630	200	6879	1.9	2.70E-08	1.8	7.90E-06	1.30E-06	3.60E-05
REACTOME_PATHWAY	R-HSA-983168	93	2158	308	9075	2.2	7.30E-03	1.3	1.00E+00	9.90E-01	1.10E+01
KEGG_PATHWAY	Rap1 signaling pathway (*)	91	1630	210	6879	2.1	3.70E-10	1.8	1.10E-07	2.70E-08	5.00E-07
KEGG_PATHWAY	Ras signaling pathway	91	1630	226	6879	2.1	3.00E-08	1.7	8.80E-06	1.30E-06	4.00E-05
KEGG_PATHWAY	Regulation of actin cytoskeleton	74	1630	210	6879	1.7	1.50E-04	1.5	4.40E-02	4.10E-03	2.00E-01
KEGG_PATHWAY	Viral carcinogenesis	71	1630	205	6879	1.6	3.80E-04	1.5	1.00E-01	8.50E-03	5.00E-01

Table 10: Top 15 Pathway of Genes for selected miRNA, mRNA Markers

Category	Term	Count	LT	PH	PT	%	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR
EC_NUMBER	6.3.2.-	137	2506	205	4250	1.2	1.60E-02	1.1	1.00E+00	1.00E+00	2.20E+01
KEGG_PATHWAY	Chemokine signaling pathway	141	3958	186	6879	1.3	2.40E-07	1.3	7.30E-05	1.00E-05	3.30E-04
KEGG_PATHWAY	Endocytosis	187	3958	241	6879	1.7	6.30E-11	1.3	1.90E-08	6.20E-09	8.40E-08

KEGG_PATHWAY	Focal adhesion	152	3958	206	6879	1.4	1.40E-06	1.3	4.10E-04	5.10E-05	1.80E-03
KEGG_PATHWAY	HTLV-I infection	201	3958	254	6879	1.8	3.60E-13	1.4	1.10E-10	5.30E-11	4.70E-10
KEGG_PATHWAY	MAPK signaling pathway	195	3958	253	6879	1.7	6.70E-11	1.3	2.00E-08	5.00E-09	9.00E-08
KEGG_PATHWAY	Pathways in cancer	302	3958	393	6879	2.7	3.10E-16	1.3	1.00E-13	1.00E-13	4.40E-13
KEGG_PATHWAY	PI3K-Akt signaling pathway	238	3958	345	6879	2.1	8.00E-06	1.2	2.40E-03	2.40E-04	1.10E-02
KEGG_PATHWAY	Proteoglycans in cancer	157	3958	200	6879	1.4	5.70E-10	1.4	1.70E-07	3.40E-08	7.50E-07
REACTOME_PATHWAY	R-HSA-212436	234	5317	358	9075	2.1	6.00E-03	1.1	1.00E+00	9.90E-01	9.40E+00
REACTOME_PATHWAY	R-HSA-983168	215	5317	308	9075	1.9	3.80E-05	1.2	5.30E-02	5.30E-02	6.30E-02
KEGG_PATHWAY	Rap1 signaling pathway	153	3958	210	6879	1.4	4.30E-06	1.3	1.30E-03	1.40E-04	5.80E-03
KEGG_PATHWAY	Ras signaling pathway	168	3958	226	6879	1.5	1.50E-07	1.3	4.50E-05	7.50E-06	2.00E-04
KEGG_PATHWAY	Regulation of actin cytoskeleton	151	3958	210	6879	1.4	1.70E-05	1.2	5.00E-03	4.50E-04	2.20E-02
KEGG_PATHWAY	Viral carcinogenesis	142	3958	205	6879	1.3	5.10E-04	1.2	1.40E-01	1.30E-02	6.80E-01

Table 11: Top 15 Pathway of Genes for selected miRNA Markers

Category	Term	Count	LT	PH	PT	%	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR
KEGG_PATHWAY	Pathways in cancer	125	1183	393	6879	4, 0	4, 6E-13	1, 8	1, 3E-10	1, 3E-10	9, 6E-11
KEGG_PATHWAY	Proteoglycans in cancer	72	1183	200	6879	2, 3	2, 0E-10	2, 1	5, 7E-8	1, 9E-8	1, 4E-8
KEGG_PATHWAY	Hepatitis B	55	1183	145	6879	1, 7	4, 4E-9	2, 2	1, 3E-6	2, 1E-7	1, 5E-7
KEGG_PATHWAY	Focal adhesion	70	1183	206	6879	2, 2	6, 1E-9	2, 0	1, 8E-6	2, 5E-7	1, 8E-7
KEGG_PATHWAY	Rap1 signaling pathway	66	1183	210	6879	2, 1	4, 8E-7	1, 8	1, 4E-4	1, 1E-5	8, 2E-6
KEGG_PATHWAY	HTLV-I infection	76	1183	254	6879	2, 4	5, 1E-7	1, 7	1, 5E-4	1, 1E-5	8, 2E-6
KEGG_PATHWAY	Ras signaling pathway	67	1183	226	6879	2, 1	3, 8E-6	1, 7	1, 1E-3	5, 0E-5	3, 6E-5
KEGG_PATHWAY	Viral carcinogenesis	61	1183	205	6879	1, 9	9, 7E-6	1, 7	2, 8E-3	1, 1E-4	7, 8E-5
REACTOME_PATHWAY	R-HSA-212436	94	1599	358	9075	3, 0	3, 4E-5	1, 5	3, 8E-2	1, 7E-2	1, 7E-2
KEGG_PATHWAY	PI3K-Akt signaling pathway	88	1183	345	6879	2, 8	7, 0E-5	1, 5	2, 0E-2	6, 0E-4	4, 3E-4
KEGG_PATHWAY	Endocytosis	63	1183	241	6879	2, 0	4, 3E-4	1, 5	1, 2E-1	2, 7E-3	1, 9E-3
KEGG_PATHWAY	cAMP signaling pathway	53	1183	198	6879	1, 7	7, 4E-4	1, 6	1, 9E-1	4, 3E-3	3, 1E-3
KEGG_PATHWAY	MAPK signaling pathway	63	1183	253	6879	2, 0	1, 6E-3	1, 4	3, 8E-1	7, 8E-3	5, 6E-3

KEGG_PATHWAY	Regulation of actin cytoskeleton	54	1183	210	6879	1, 7	1, 8E-3	1, 5	4, 0E-1	8, 1E-3	5, 8E-3
REACTOME_PATHWAY	R-HSA-983168	74	1599	308	9075	2, 4	3, 6E-3	1, 4	9, 8E-1	1, 6E-1	1, 5E-1

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: KUTLAY, AYŞEGUL

Nationality: TURKISH

Date and Place of Birth: MUĞLA

Marital Status: SINGLE

Phone: 0 555 225 64 27

email: aysegul.kutlay@gmail.com, e133477@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	Medical Informatics, Informatic Institute, METU	2012
BS	Computer Education and Instructional Technologies, METU	2007
High School	Turgut Reis High School, Super Lise, MUĞLA	2001

WORK EXPERIENCE

Year	Enrollment
2020-Present	**Director of Software Development**
Frumatic Director of Software Development	Orchestrate a cross-functional, international, on-site and remote teams and managers through the execution and development of web and mobile applications as well as Computer Vision and Predictive AI solutions. Collaborate with key stakeholders such as customers, product managers, marketing and sales

teams. Provide detailed technical guidance while tracking progress and ensuring that deliverables that conform to the project schedule are delivered promptly.

Key Achievements:

- Managed cross regional international teams across multiple projects
- Lead and manage internal and external Software Development teams to deliver software, and architecture,

Defined software development process and Implement mechanisms to monitor, manage and provide progress on all software development activities,

2017-2020 **Software Development Group Manager**

Comodo Inc., Ankara, Turkey

Orchestrate a cross-functional team of more than 40 software engineers and managers through the execution and development of a large scale, customer-focused SaaS product with more than 80 components, NUCAL. Supply extensive technical direction while monitoring progress and ensuring timely completion of deliverables that adhere to the project plan. Collaborate with key stakeholders such as product and program managers and architects from requirements gathering and design through integration.

Key Achievements:

- Facilitated in the design of the technical architecture based on cloud computing.
- Coordinated several teams of engineers through all facets of the Agile development life cycle for ITARIAN , NUSAL and NuNOC projects.
- Masterminded whole product delivery cycles for all stages of a robotic process automation solution from product research and requirement analysis through software development and delivery.

Functioned as career manager for development team, conducting performance reviews and delivering progress reports to upper management.

20011-2017 **Product Research & Development Manager**

Akgün Software Co. Administrated multifaceted teams of professionals within the product research and development department,

executing complex projects and maintaining budgets while ensuring alliance with allotted time frames. Generated innovative project ideas and performed feasibility studies to gather and report results. Performed in-depth analysis of all project results to provide data-driven recommendations and actionable insights. Constructed the design and technical solution description of various research and innovation initiatives.

Key Achievements:

- Acquired a funding opportunity for two extensive projects from applications to research and innovation programs, accumulating over half of the required budget.
- Directed team through the development of several projects including SMS and Homecare modules, a Hospital Information System, and an MHRS HIS Integration.
- Created a new project proposal for both national and international funding programs.
- Delivered 13 projects to market including innovative solutions like the (laboratory results based) machine learning application , enterprise mobile hospital information system and the intensive care unit system.

2007-2011 **Software Engineer**

Havelsan Ehsim , Ankara, Turkey

Collaborated with a multi-disciplinary team consisting of system, mechanical, electric, and software engineers. Designed and developed a mobile control unit and specialized analytic software for I-Level and O-Level maintenance of a large-scale distributed system, CMDS.

2006-2007 **Student Assistant**

METU, Ankara, Turkey

Supporting Dean and Vise Dean for Implementation of management activities and management report,

- Performing academic literature search,
- Preparing brochures and posters of the faculty and the workshops conducted.

FOREIGN LANGUAGES

Native Turkish, Advanced English

PUBLICATIONS

- Integrative Predictive Modeling Of Metastasis In Melanoma Cancer Based On MicroRNA, mRNA And DNA Methylation Data, 23.09.2021, *Frontiers in Molecular Biosciences Biological Modeling and Simulation*
- Acceptance of Mobile Homecare Technologies: An Empirical Investigation on Patients with Chronic Diseases. In Current and Emerging mHealth Technologies (pp. 201–222). 2018, Springer International Publishing.
- Enterprise Mobile Health Information System: Approaches and Experiences, 2016, CEUR Workshop *Proceedings* (Vol. 1721).
- Analyses Of Factors Affecting Acceptance of Homecare Technologies by Patients With Chronic Diseases. 2012, Middle East Technical University MS Thesis Study. Middle East Technical University,