Discriminating early- and late-stage cancers using multilayer perceptron

by

Marzieh Soleimanpoor

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Industrial Engineering

KOÇ ÜNİVERSİTESİ

April 17, 2020

Discriminating early- and late-stage cancers using multilayer perceptron

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Marzieh Soleimanpoor

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Mehmet Gönen

_____

Assist. Prof. Barış Yıldız

_____

Assist. Prof. Hande Küçükaydın

Date: _____

[To my beloved husband]

ABSTRACT

Discriminating early- and late-stage cancers using multilayer perceptron
Marzieh Soleimanpoor
Master of Science in Industrial Engineering
April 17, 2020

Prediction of cancer stages is essential to early diagnosis of the tumours and to develop appropriate therapeutic strategies for cancer patients. Determining the biomarker genes is extremely important for understanding the progression of cancer from early-stages to late-stages. Standard machine learning algorithms could be adopted to build classifiers based on genomic characterizations to discriminate between early- and late-stage cancers. Some of these algorithms have limited knowledge extraction capability due to the strongly correlated nature of high-dimensional genomic data even if they have satisfactory prediction accuracies. In this thesis, to identify the relevant molecular mechanisms of the biological processes which might be driving cancer progression, we implemented an algorithm that integrates pathways/gene sets into our neural network model. We developed a multilayer perceptron (MLP) method on multiple pathways/genes to separate early- and late-stage of cancers. Our main goal was to determine biomarkers that have an effect on the progression of cancer and to obtain comparable predictive performance. We compared the performance of our MLP model with three existing classification algorithms namely, random forests, support vector machines and multiple kernel learning on 20 cancer types for two experiment types. Our MLP algorithm obtained better or comparable predictive performance against other baseline algorithms on most of the TCGA datasets. The results showed that our algorithm can extract meaningful and disease-specific key mechanisms of cancer prognosis.

ÖZETÇE

Çok katmanlı perceptron kullanarak erken ve geç evre kanserleri ayırma
Marzieh Soleimanpoor
Endüstri Mühendisliği, Yüksek Lisans
Nisan 2020

Kanser hastaları için uygun tedavi edici stratejiler ve prognostik tahminler geliştirmek için kanser evrelerinin sınıflandırılması büyük önem taşır. Biyobelirteçlerin içerdiği genlerin belirlenmesi, kanserin erken aşamalardan geç aşamalara ilerlemesini izlemede son derece önemlidir. Erken ve geç evre kanserler arasında ayrım yapmak amacıla genomik karakterizasyonlara dayanan sınıflandırıcılar oluşturmak için çeşitli standart yapay öğrenme algoritmaları benimsenebilir. Bu algoritmalar sınırlı bilgi çıkarma kabiliyetine sahiptir, ancak yüksek boyutlu genomik verilerin güçlü bir şekilde ilişkili doğası nedeniyle kabul edilebilir tahmin doğruluklarına sahiptir. Bu nedenle, bu biyolojik süreçlerle ilgili moleküler mekanizmaları tanımlamak için, yolakları/gen kümelerini öğrenme modelimize entegre eden bir algoritma önerdik. Bu tezde, kanserlerin erken ve geç aşamalarını ayırt etmek için çok sayıda yolak /gen kümesi üzerinde bir "çok katmanlı perseptron" yaklaşımı geliştirdik. Bu çalışmanın temel amacı, kanserin oluşumu ve ilerlemesi üzerinde önemli bir etkiye sahip olan biyobelirteçleri belirlemek ve daha iyi veya kabul edilebilir bir tahmin performansı elde etmekti. Ayrıca, çok katmanlı perseptron modelimizin performansını, iki farklı deneyde ve 20 kanser tipinde mevcut üç farklı yapay öğrenme algoritması olan rassal orman, destek vektör makineleri ve çoklu çekirdek öğrenimi ile karşılaştırdık. Çok katmanlı perceptron metodumuz diğer temel algoritmalardan daha iyi performans gösterdi veya kıyaslanabilir sonuçlar elde etti. Sonuçlar, algoritmamızın kanser gelişiminin temel mekanizmalarına ilişkin anlamlı ve hastalığa özgü bilgileri çıkarabildiğini gösterdi.

ACKNOWLEDGMENTS

First, I would like to express my profound gratitude to my thesis advisor Assoc. Prof. Mehmet Gönen, for his supportive attitude, encouragement, enthusiasm, brilliant ideas and suggestions through this research project. He's the smartest people I know and I could not wish for a better advisor and mentor for my Master study. I also would like to sincerely thank the members of my advisory committee who generously gave their time to discuss and offer me precious comments toward improving my research Assist. Prof. Barış Yıldız and Assist. Prof. Hande Küçükaydın. I gratefully acknowledge the funding from the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant SBAG 216S461. I especially thank my family and friends who have always encourage me to pursue my dreams. Finally, I must thank my loving, wonderful and patient spouse as without his support I may never have completed this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANN:	Artificial Neural Network
AUROC:	Area Under the Receiver Operating Characteristic
LASSO:	Least Absolute Shrinkage and Selection Operator
MLP:	Multilayer Perceptron
MKL:	Multiple Kernel Learning
MSigDB:	Molecular Signatures Database
TCGA:	The Cancer Genome Atlas
RF:	Random Forest
ROC:	Receiver Operating Characteristic
RNA-seq:	RNA-sequencing
RLU:	Rectified Linear Unit
RMSE:	Root Mean Square Error
SVM:	Support Vector Machine
TNM:	Tumour, Nodes, and Metastases

Chapter 1

INTRODUCTION

Nowadays, cancer has become one of the most leading causes of death. Diagnosis of cancer in earlier stages can assist the cancer treatment procedure. In this context, many types of research conducted to study the cancer stage prediction problem. The genomic characterization that extracted from tumour biopsies of cancer patients is used to develop new approaches to diagnose, treat, and prevent cancer diseases. The aim of this study is classifying primary tumours into early-stage or late-stage. Machine learning techniques can make a significant contribution to the process of early diagnosis of cancer. A variety of machine learning methods including artificial neural networks (ANNs), support vector machines (SVMs) and multiple kernel learning (MKL) have been extensively used in cancer studies to develop predictive models. Applying these machine learning methods can facilitate the extraction of meaningful information of the biological mechanism from high-dimensional data. However, extraction of informative data is not easily achievable from high-dimensional and correlated datasets. Moreover, limited number of profiled tumours is another important restriction for machine learning methods to perform well on high-dimensional genomic characterizations. Treatment of cancer is quite easier in the early-stages rather than the late-stages. Hence, mortality rates increase from early-stages to late-stages of cancer. Therefore, understanding the cancer formation in the earlier stages of cancer is critical to identify and use optimal treatment.

In this thesis, we developed a classification algorithm that separates the early- and late-stage of cancers from each other by using gene expression profiles of cancer patients. We checked the predictive accuracy of the applied algorithm using the area under the receiver operating characteristic curve metric.

1.1 Previous studies

In this section, we reviewed the related literature to the problem of discriminating the stages of cancers and the recent studies that applied the multilayer perceptron (MLP) method for cancer classification problems.

1.1.1 Discriminating cancer stages

There are several studies that discussed the problem of separating cancer stages from each other using genomic measurements.

Broët et al. (2006) proposed a statistical score-based approach on clinical and microarray data and determined gene expression changes and crucial features that separate early- and late-stage from each other. This approach outperformed the classical Wald statistic taken from the Cox model in determining the prognosis effect of the disease. Jagga and Gupta (2014) presented a correlation-based algorithm to identify individual genes that discriminate between clinical tumour stages from each other just for a kidney renal clear cell carcinoma cohort. They employed various supervised machine learning algorithms such as random forests (RFs) and SVMs on the selected genes to develop classification models for classifying different tumours into early- and late-stage categories properly. Bhalla et al. (2017) developed threshold-based classification algorithms to categorize cancer patients into early-stage patients and late-stage patients for kidney renal clear cell carcinoma similar to the problem investigated by Jagga and Gupta (2014). They trained some supervised machine learning techniques such as RFs and SVMs on the expression-based genes to estimate the quality of identified gene expression features. Bhalla et al. (2018) described the key transcripts with an expression-based threshold and developed a multi-class machine learning model for separating early- and late-stage tumours of papillary thyroid cancer using their expression data. Additionally, they performed a correlation-based analysis to determine alterations in correlation among diverse transcripts in the cancer stages. Kaur et al. (2019) studied the problem of discriminating early- versus late-stage and cancer versus normal samples of liver hepatocellular carcinoma using machine learning techniques. They applied a threshold-based feature

selection method and developed models based on different algorithms namely, SVMs, Naive Bayes and RFs on genomics and epigenomics profiles of patients.

Although this sort of metrics in scoring-based and thresholding-based problems determine predictive gene expression signatures with a restricted number of features, actually they have two important limitations. First, difficult interpretation owing to high-dimensional and correlated input data space. Second, in high-dimensional space for a given prediction task, if machine learning classifiers use various subsets of the same patient cohort, they might choose diverse biomarkers as predictive (Ein-Dor et al., 2005, 2006).

Unlike the above studies, Rahimi and Gönen (2018, 2020) used pathways/gene sets to build a unified model on these selected gene sets. Therefore, they solved the problem of interpreting the results in high-dimensional and correlated input data. Their main contribution was to identify related biological processes and to train a classifier on the selected gene sets. They evaluated the predictive performance of their proposed method on various cohorts. To this aim, they applied three machine learning methods on gene expression data to build classifiers for separating early- and late-stage of 20 cancer types from each other. They compared the predictive performance of SVMs, RFs, MKL, and multitask MKL (MTMKL) for distinguishing the pathological stages of cancers using genomic information of profiled tumours and pathways/gene sets.

Singh et al. (2018) adopted a machine learning method to determine biomarkers on gene expression profiles for generating classifiers to segregate the stages of papillary renal cell carcinoma. They trained a machine learning workflow integrating various feature selection approaches and classification models to analyze the RNA sequencing data collection. Kumar et al. (2019) designed a classifier for differentiation of early- and late-stage of prostate cancer using RNA-seq based gene expression data. Various machine learning methods such as SVMs, Naive Bayes, RFs and MLP were applied with the maximum instructive subset of features taken from gene expression dataset.

1.1.2 Multilayer perceptron for classification

MLP has been extensively used in different cancer classification problems using gene expression dataset. Here, we briefly mention a number of studies, which used the MLP method for classification.

Wang et al. (2006) offered an initialization technique for MLP parameters and architecture to avoid the curse of dimensionality in the enormous genomic data collections for different diseases such as leukemia cancer, central nervous system cancer, and limb-girdle muscular dystrophy. They first applied the proposed technique to the multiclass classification problem, then evaluated the impact of the suggested method on prediction accuracy and training efficiency using different metrics such as the area under the receiver operating characteristic curve. Targonski et al. (2019) implemented a two-phase algorithm which is called Gene Oracle algorithm on curated gene sets taken from the Molecular Signatures Database (MSigDB). In the first phase, they applied the classification approach to find biomarker genes in all datasets using an MLP algorithm that contains an input layer, three hidden layers, and an output layer with softmax activation function. In the second phase, they used a combinatorial method to decompose different gene sets with high classification potential. They noted that classification accuracy improved by enhancing the number of genes that considered as input features.

Thakur et al. (2011) developed an MLP classification model that consists of a hidden layer with sigmoid activation function to distinguish ovarian cancer and presented a method based on feature selection that is called serum protein. They applied t-test and neural networks for differentiating the curve of patients and normal people. The accuracy of the MLP approach was evaluated using different metrics and then compared against the linear discriminant analysis classifier method. Nasser and Abu-Naser (2019) developed a prediction model based on an MLP algorithm using the primary tumour dataset. They first determined proper factors that have an impact on tumour classification and converted them into the specific format to build an MLP model for classifying the tumours of a given patient based on the patients' test. The implemented MLP method contains one input layer, one hidden

layer, and one output layer, and backpropagation is used to train the model and to find the optimal hyperparameters.

Sokouti et al. (2014) presented a Levenberg–Marquardt feedforward MLP to classify cervical cancer cells for unsegmented cell feature images that obtained from 100 patients into three categories which are healthy, low grade, and high grade intraepithelial squamous lesion. This diagnosis system includes two phases, which are preprocessing the images to reduce noises and then using feedforward MLP for classification. They correctly classified images with 100% classification rate. Dash and Dash (2014) proposed a symmetrical uncertainty and correlation-based feature selection technique and also MLP method to categorize high-dimensional microarray datasets. This approach identifies important genes in cancer classification problems using thresholds for relevancy and redundancy of features in large datasets. The results showed that implemented MLP algorithm obtained higher accuracy for small subsets of genes with multiple categories in comparison to the other feature selection algorithms.

Several MLP methods have been presented recently to address lung and breast cancers related classification problems, hence we reviewed the MLP models in these specific cancers problems.

Lung cancer prediction

The following studies applied MLP method for different lung cancer classification problems.

Chang et al. (2018) designed a unique lung cancer diagnosis tool as a sensor system, which contains seven metal oxide gas sensors, to analyze the volatile organic compounds to identify early-stage lung cancer. They investigated several well-known classification techniques namely, MLP, one-class SVM, and two-class SVM for pattern classification of multi-dimensional sensing data. Furthermore, the principal component analysis technique was performed for testing the distribution of sensing data that are in high-dimensional space. MLP model outperformed the SVM model

to specify lung cancer patients. Additionally, they used an MLP model to predict different categories, namely, lung cancer versus healthy control, stage, sex, tumour location.

Azzawi et al. (2016) developed a gene expression programming (GEP) model for lung cancer prediction using microarray data collection. They compared the prediction performance of the proposed method with three different machine learning algorithms, namely, SVMs, MLP consist of one hidden layer with logistic activation function, and radial basis function networks. It is observed that the GEP model outperformed the other methods using fewer features. Rafii et al. (2015) presented an approach based on MLP method with two hidden layers to classify lung cancer on the microarray dataset. The proposed approach used the principal component analysis technique to decrease the dimensionality of data and then performed classification using MLP network on the lung cancer gene expression data collection. The performance of the proposed MLP is measured using the accurate classification rate and then compared to the results of other methods on the same dataset.

Breast cancer prediction

We reviewed the following articles for breast cancer classification problems using MLP method.

Al-Shargabi et al. (2019) proposed a tuned MLP algorithm for classification of breast cancer tumours on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Tuning the hyper-parameters together with feature selection algorithms were utilized to improve the prediction accuracy of the MLP model. Their experimental results showed that the grid search approach is appropriate to find the best hyper-parameters to achieve a more accurate prediction compared to the basic MLP method. Ibrahim et al. (2015) introduced a multiobjective classification model using MLP for the diagnosis of the breast cancer. To solve the multiobjective optimization problems with tuning MLP parameters such as the number of hidden units, a differential evolution algorithm was applied that lead to a better accuracy. To evaluate the proposed classifier, ten-fold cross-validation method was performed.

1.2 Our contributions

In this thesis, we addressed the problem of extracting disease-related information from an extensive amount of redundant data and noise in high dimensional data space using pathways/gene sets and gene expression profiles in a unified model. Pathways/gene sets are lists of genes that define parts of genomic characterizations and map them to the interactions of genes and molecular mechanisms within cancer cells. Therefore, instead of learning a classifier on a list of gene expression features, we used pathways/gene sets to train a classifier on selected genes and determine relevant biological processes at the same time. For this purpose, we applied the MLP methodology (McCulloch and Pitts, 1943), which was used mainly for supervised machine learning problems. This approach is trained on a group of input-output pairs and learns to model the relationship between the inputs and outputs. The backpropagation algorithm is applied to train the network which finds the values of all weights using the chain rule to minimize the error of the trained network. We built kernel matrices on multiple views of each input pathways/gene sets that are extracted from gene expression profiles of primary tumours. These kernel matrices are the inputs of our MLP network. We summarized the important contributions of this research as follows: We implemented an MLP network on the Hallmark gene sets collection (Liberzon et al., 2015), which extracted from Molecular Signatures Database (MSigDB) to train a classifier and to determine the related biological processes that are important to differentiate the stages of cancers. We evaluated the prediction performance of the adopted MLP network on the problem of separating early- and late-stage primary tumours from each other using their gene expression data. Our results demonstrated the capacity of our algorithm to obtain biologically meaningful and disease-specific information for the prognosis of cancers.

We organized the rest of this dissertation as follows: In Chapter 2, we provided a brief background to cancer biology and deep learning. Chapter 3 presented the methodology used for this study and the baseline methods. The computational experiments and the material of the proposed method were described in Chapter 4. In Chapter 5, we proposed the analysis of results for two groups of experiments.

The conclusions were summarized in Chapter 6.

Chapter 2

BACKGROUND

In this chapter, we presented the background of our research. First, we described cancer-related definitions and information. We then explained different cancer types, hallmarks of cancer, cancer risk factors, and cancer staging systems. Next, we discussed ANNs and the notion of perceptron. Finally, we described MLP methodologies, different regularization, and activation functions used in the MLP methods.

2.1 Cancer

Cancer is a devastating disease and has become one of the most important causes of death across the world. Cancer is a group of diseases distinguished by the excessive growth of the abnormal cells. The cells build a tumour, or travel inside the blood vessels or in the lymph nodes. Categorizing the cancers is based on the type of cells or organs from which they originated and first developed.

2.1.1 Histological kinds of cancer

There are more than 100 different types of cancers because malignant can grow in almost all parts of the body. There are six major categories based on the histological type which are explained as follows.

Carcinoma begins in the epithelial tissue of the skin or the tissue lining of organs, such as lungs, kidney, and liver. According to the National Cancer Institute (2019), most of the diagnosed cancer instances are carcinomas. Carcinomas are divided into some common subtypes, including adenocarcinoma, squamous cell carcinoma, basal cell carcinoma, ductal carcinoma in situ, and invasive ductal carcinoma. Adenocarcinoma (which evolves in an organ or gland) and squamous cell carcinoma (that starts in the squamous epithelium) are two main subtypes of carci-

noma.

Sarcoma develops in the connective or supportive tissues, namely, bone, cartilage, and soft tissues. Fat, muscle, and blood vessels are different examples of soft tissues. There are virtually 50 diverse types of soft tissue sarcomas. Although sarcomas are mostly originated in the arms or legs, but they may be found in the other parts of the body, e.g., the neck, head area, and internal organs.

Myeloma is arising from plasma cells of the bone marrow, which produces the body's blood cells. The main part of the body's immune system is plasma cell. Myeloma starts when healthy plasma cells change and grow out of control. Instead of creating helpful antibodies, the cancer cells generate abnormal proteins. Additionally, it is also known as *multiple myeloma* which is a blood cancer and affects diverse parts of the body, namely, the skull, spine, pelvis, and ribs. Multiple myelomas form many tumours called plasmacytomas by collecting myeloma cells in many bones.

Leukemia begins in the blood and bone marrow. The disease often causes the overproduction of abnormal and defective white blood cells, and it also has influence on red blood cells. Leukemia prevents the marrow from creating normal red and white blood cells and platelets. It is also called "liquid"cancer because it exists in the body fluids. The four main types of leukemia are acute lymphocytic leukemia, chronic lymphocytic leukemia, acute myeloid leukemia, and chronic myeloid leukemia.

Lymphoma is related to a type of cancer which develops in the nodes or glands of the lymphatic system, that has the role of generating white blood cells and purify bodily fluids. It is also called "solid cancer" because it appears in solid organs for instance breast or prostate. Lymphoma is subclassified into two group: Hodgkin's lymphoma and non-Hodgkin's lymphoma. In fact, lymphoma and myeloma are two different cancers of the immune system.

Brain and spinal cord cancers start in the cells of the brain or spinal cord and they are called as central nervous system cancers. The brain controls the body's functions by sending electrical signals through nerve cells, called neurons. Signals transfer information along nerve fibers to other neurons and control organs and

muscles. The spinal cord is composed of nerve fibers that transmit signals from the body to the brain and vice versa. It also contains a supportive cell called glial cells which is a connective tissue that surrounds neurons and supports them to conduct signals. Glioma is the most prevalent type of brain tumours which originates in the glial cells. Some of the brain and spinal cord tumors are benign tumours which are noncancerous and slow-growing, and others are malignant tumours that are cancerous and capable to spread into different parts of the body.

The cancer types information are collected and retrieved from National Cancer Institute (2019) which is publicly available at `https://www.cancer.gov/about-cancer/`.

2.1.2 Hallmarks of cancer

Cancer is a highly diverse and complicated disease, but a set of features are common between approximately all malignancies. Those features are called hallmarks of cancer, and they are an integrated set of abilities that are obtained during the long process of tumorigenesis and malignant progression (Lazebnik, 2010). Hanahan and Weinberg (2000) proposed the six major hallmarks of cancer which are (1) sustained proliferative signaling, (2) evading growth suppressors, (3) activating invasion and metastasis, (4) enabling replicative immortality, (5) including angiogenesis, and (6) resisting cell death. Additionally, Hanahan and Weinberg (2011) published an update in 2011 and proposed another two novel hallmarks (7) deregulating cellular energetics and metabolism, and (8) avoiding immune destruction. Ruddon (2007) described two enabling features which are genome instability with subsequent gene mutation and tumour-promoting inflammation.

2.1.3 Cancer risk factors

Since cancer is a genetic disease, it starts when one or more genes in a cell mutate. Genetic information determines the construction and function of the organisms passed from one generation to the offspring by genes. Genes are made from DNA and control the function of cells by producing proteins. Proteins play many critical roles in the body such as enzyme, messenger, transport/storage, structural

component, and antibody. A mutation is a random change that occurs in the DNA sequences, and it has different types and small-scale mutations can be categorized into base substitutions, deletions and insertions, inversion, and point mutation categories. Mutations often happened for two main reasons. First, DNA fails to copy accurately during the replication and transcription process. Second, mutations can also occur by external influences such as exposure to environmental factors, for instance, specific chemicals or radiation, tobacco smoking and UV radiation in sunlight (Lodish et al., 2000).

In general, anything that causes uncontrolled growth for normal cells and develops abnormal cells can cause cancer. According to the scientific researches, cancer was produced by the interaction of many factors with each other and there is no single cause for cancer. Common risk factors can be included in internal and external factors (Vineis and Wild, 2014). Some of the most important factors are as follows.

Internal factors are biological and intrinsic factors that are related to the system inside the body such as age, gender, certain hormones, inherited genetic defects, and skin type.

External factors are the factors that come from outside of the body and cause cancers. These external factors can be divided into environmental exposures, occupational exposures, and lifestyle.

The environmental exposure such as sunlight and other ionizing radiation comes from different sources that can increase the risk of cancer creation. UV radiation, sunlamps, fine particulate matter and ionizing radiation from radioactive fallout, radon gas, and x-rays are some examples of the environmental exposures.

Specific occupational exposures including carcinogens can increase cancer risk. They refer to the harmful effects of working in an unsafe place such as chemical industries, mining/metal industry, painting industries, leather, rubber, and woodworking industries. There are many occupational exposures which are certain agents of cancers for instance chemicals, radioactive materials, asbestos, cadmium, and nickel. Lifestyle-related factors and personal habits such as smoking tobacco, alcohol consumption, high-fat diet, obesity, and physical inactivity can also increase the risk of cancers.

2.2 Cancer staging system

Cancer staging is one of the basic activities in oncology that refers to the process of specifying the extent to which cancer is located. Diagnosis of cancer stages is also a critical factor in deciding proper treatments for the patients. Precise staging is required to examine the outcomes of treatments to facilitate the collection and exchange of information between cancer treatment centers to create a reference guide for cancer research (Gress et al., 2017). There are diverse cancer staging systems, but the tumour, node, and metastasis (TNM) system is the most commonly used and useful staging system for the most kinds of cancers. The American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control (UICC) developed the TNM staging system. The TNM system categorizes cancers via the size and extent of the main tumour (T), and the involvement number of nearby lymph nodes that have cancer (N). It determines the existence of distant metastases, which means that cancer has spread from the primary tumour to distant sites (M). Cancer staged at several time points and these time points are termed classification of the TNM staging system and different categories are assigned to a particular classification. There are five types of classification in the TNM system which are clinical, pathological, posttherapy, recurrence, and autopsy. Clinical and pathological classification are more predominant between the mentioned classification categories (Gress et al., 2017).

2.2.1 Clinical stage information

Clinical stage is an evaluation of the extent of cancer to determine the amount or spread of cancer in the body in regard to patient history, physical investigation and several imaging tests, which are done before starting the treatment. Indeed, it is a pretreatment stage and the main part of determining the most efficient treatment to be used based on the available information on biopsy.

2.2.2 Pathological stage information

The pathological or surgical stage relies on the examination of the resected specimens by removing tissue samples during surgery. Furthermore, this stage depends on the results of exams and tests from the clinical stage. Surgery is performed before initiation of systemic therapy and determines the difference between normal cells and the removed samples.

Clinical annotation for cancer patients that have recorded tumours is also available in the TCGA project. Since our goal was segregating the early- and late-stage of cancers from each other, we examined `pathologic_stage` annotation of every patient based on the latest freeze (i.e., `Data Release 22 - January 16, 2020`), and there were 6,958 patients with this information.

Cancer stage grouping

TNM system explains cancers with complete details and TNM scores determine the stages for most of the cancers. However, generally, all types of cancer have four stages that ranked by Roman numerals I, II, III, and IV and some of them also have stage 0 to characterize the progression of cancers (Compton et al., 2012).

Stage 0 : In this stage, abnormal cells are remained in the area they originated and have not spread to the nearby tissues. This stage describes cancer that is called carcinoma in situ which means all cancer cells are in their place.

Stage I : This stage usually includes small primary tumours that have not grown extremely into any lymph nodes or another areas of the body. It is frequently named early-stage cancer and the tumour can be removed by surgery.

Stages II and III : In these stages, primary tumours are more extensive, locally advanced and have spread into nearby tissues or lymph nodes. They cannot spread to the another area of the body. Stage III is more advanced spread than stage II. They can be cured by radiation, chemotherapy, and surgery.

Stage IV : Cancers have spread to the other organs and distant parts of the body and mortality rates are higher in this stage. This stage is also called metastatic cancer or advanced cancer.

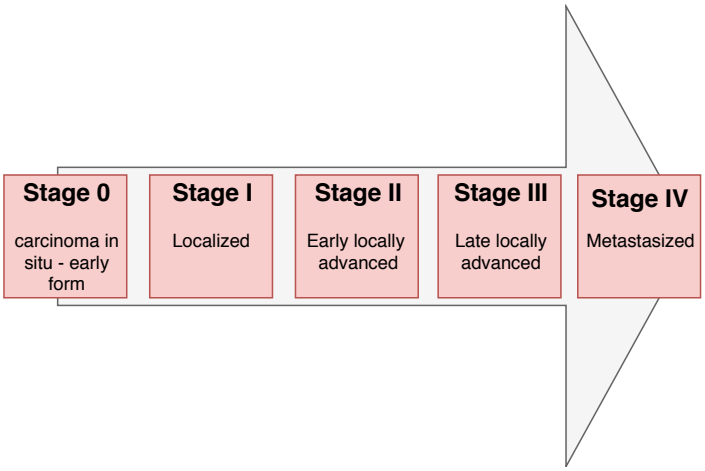

Figure 2.1: Stages of cancer

There are different staging systems for different types of cancer groups and they differ based on the purpose of staging. For instance, another staging system that has five major categories is usually used by cancer registries is as follows.

In situ cancer is also called stage 0. In this stage, cancer cells are in the external layer of the skin and have not spread to deeper areas of the body and nearby tissues.

Localized cancer is a kind of cancer that is located in the place where it started and has not spread to nearby lymph nodes or to another areas of the body.

Regional cancer is another category of cancer and it has grown further on the tumour to nearby lymph nodes, tissues, or organs.

Distant cancer is a kind of cancer which spreads to distant organs or distant lymph nodes of the body. This cancer stage is also known as distant metastasis.

2.3 Artificial neural network

ANNs also known as connectionist systems, are inspired by biological neural networks. These systems are designed based on the brain process information and biological systems that contain neurons (McCulloch and Pitts, 1943). An ANN is a network, which is consisted of units (nodes) and connections (edges). An ANN unit is called an *artificial neuron* and models neurons in a biological brain. A connec-

tion in an ANN can transmit a signal between the artificial neurons (similar to the synapses in a biological brain).

An ANN learns to implement tasks with considering instances, normally without being designed with task-specific rules. ANNs have diverse applications in machine learning research for example, image processing and character recognition, speech recognition, computer vision, and text processing.

2.3.1 Perceptron

ANN networks take their inspiration from the structure of brain. A perceptron is an ANN model of a biological neuron and a basic processing component (Rosenblatt, 1958). A perceptron can be viewed as a technique for supervised learning of binary classification models. Perceptrons are building block of ANNs and each input of a perceptron has its relevant connection weight and the output is a weighted function of the inputs. There are two different input types for a perceptron: (1) inputs which are directly from the environment and (2) inputs which are outputs of the other perceptrons.

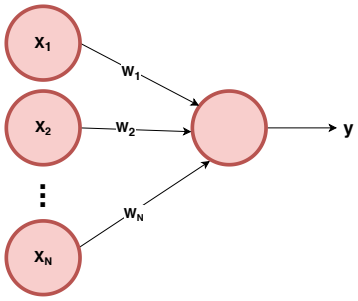

Figure 2.2: A perceptron model

As an algorithm, a perceptron receives inputs, moderates them with connection weight values, then uses an activation function to produce the output. Figure 2.2 represents a perceptron model, where $x_i \ \ \forall i \in i = 1, ..., N$ are the inputs, $w_i \in R$ are connection weights, and y is the output which is produced by weighted sum of the inputs. Perceptron has two types which are single-layer and multilayer perceptron.

2.3.2 Single-layer perceptron

A single-layer perceptron has just two layers, i.e. the input layer and the output layer (Figure 2.3). The neurons in the input layer are fully connected to a neuron or several neurons in the next layer. A perceptron with one layer of weights is just applicable for linear functions and it is not suitable for solving nonlinear problems. As a linear classification model, the single-layer perceptron is the most basic feedforward neural network. Note that a *feed-forward neural network* is the first and simplest kind of ANN designed wherein the links between the nodes do not form a cycle.

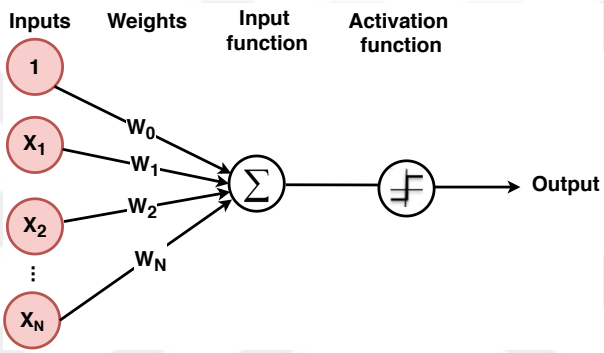

Figure 2.3: A single layer perceptron

2.3.3 Multilayer perceptron

An MLP is a class of feedforward ANN architecture and is a non-parametric estimator that can be applied for classification and regression (Alpaydın, 2009). As opposed to other statistical techniques, MLP does not consider any data distribution assumption. It can model considerably non-linear functions and can be learned to approximate virtually any smooth, measurable function. Therefore, these characteristics make this approach an appealing alternative for designing numerical models among statistical approaches. A sample MLP network is illustrated in Figure 2.4.

2.4 MLP architecture

The MLP network consists of the input layer, the output layer, and one or more hidden layers between them. Each hidden layer contains multiple perceptrons which

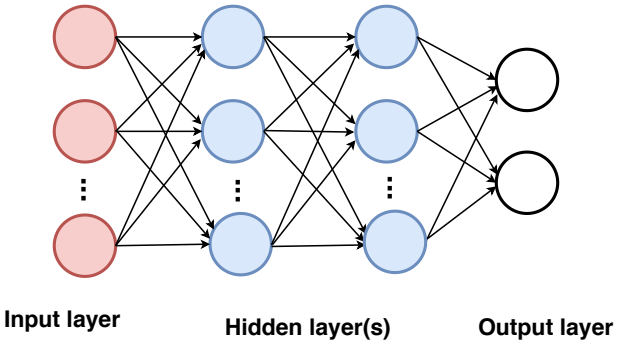

Figure 2.4: A multilayer perceptron

are known as hidden units. An MLP connects several layers in a directed graph, it shows that the signal path across the neurons only flows one way and these paths are organized in layers. There are directed connections among layers of a network from lower layers to upper layers, and there is not any interconnection between neurons in the same layer. Excluding the input neurons, each node in the network is a neuron that can use a linear or a nonlinear activation function. The number of measurements for the pattern problem is equal to the number of nodes in the input layer. In addition, the number of nodes in each layer and the number of layers are parameters of the problem. The objective of the MLP method is minimizing the error and optimizing the model with sufficient parameters and good generalization for classification or regression tasks.

MLP can be developed for binary and multiclass classifcation problems as well as nonlinear and multiple output regression. We will explain MLP method for classification in the following.

2.4.1 MLP for binary classification

In classification, MLP can implement nonlinear discriminants to calculate output units. In a two-class discrimination task (i.e. binary classification), *sigmoid* activation function is needed to applied (explained with more details in section 2.6.3). The general procedure is feeding the inputs to the input layer and the hidden units values are computed with respect to the bias. Each hidden node is a perceptron and it uses the nonlinear sigmoid activation function on its weighted sum. In MLP that

has been shown in Figure 2.5 the applied sigmoid activation function on the hidden layer can be formulated as

$$\hat{y}_i = sigmoid(v^T z_h) = \frac{1}{1 + e^{-v^T z_h}}, \qquad \forall i \in i = 1, ..., N \qquad (2.1)$$

where z_h represents the value of hidden unit h, v is the connection weight from hidden layer to output layer, and $\hat{y}_i$ is the prediction value of sample i which approximates the probability of class 1, $P(C_1|\boldsymbol{x}_i)$, and also the probability of class 2, $P(C_2|\boldsymbol{x}_i) = 1 - \hat{y}_i$, ($C_1$ and C_2 indicate class 1 and class 2 labels, respectively).

Likewise, the sigmoid activation function between hidden layer and input layer can be represented as

$$z_{ih} = sigmoid(w_h^T \boldsymbol{x}_i) = \frac{1}{1 + e^{-w_h^T \boldsymbol{x}_i}}, \qquad \forall i \in i = 1, ..., N \quad (2.2)$$

where z_{ih} is the value of h^{th} unit of hidden layer for sample i, w_h represents the connection weight between input layer and h^{th} unit of hidden layer, and $\boldsymbol{x}_i$ indicates the input value of sample i, where the error function in this case is

$$Error_i = -[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)], \qquad \forall i \in i = 1, ..., N \quad (2.3)$$

where y_i is the label of sample i.

Such a problem can be solved using gradient descent algorithm and the update equations can be represented as

$$\Delta v_h = \eta \sum_{i=1}^{N} (y_i - \hat{y}_i) z_{ih} \qquad (2.4)$$

$$\Delta w_{hd} = \eta \sum_{i=1}^{N} (y_i - \hat{y}_i) v_h z_{ih} (1 - z_{ih}) x_{id} \qquad (2.5)$$

where η is the learning factor that decreases gradually to convergence, v_h is the connection weight from h^{th} unit of hidden layer to output layer, w_{hd} indicates the connection weight between input unit d and hidden unit h, and x_{id} represents the d^{th} input value of sample i. Similar to the single perceptron, for regression and classification and the update equations are the same (this does not imply that their values are equal to each other). When there are only two classes, one output unit is sufficient as it is illustrated in Figure 2.5.

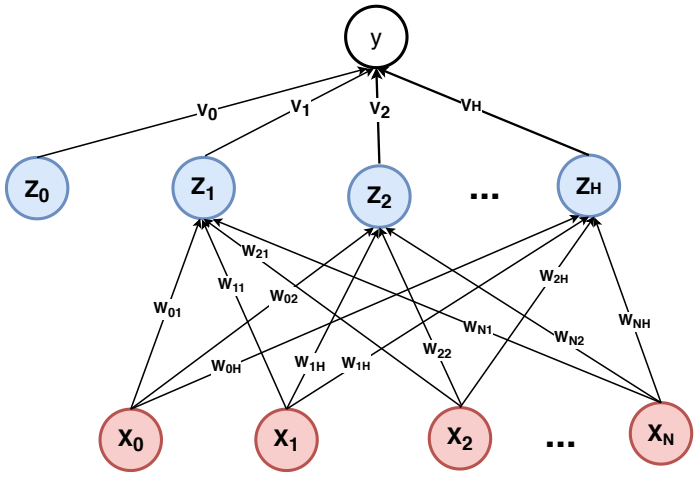

Figure 2.5: An MLP for binary classification

2.4.2 MLP for multiclass classification

In multiclass discrimination, every input data has to be assigned to one of the multiple output classes (Figure 2.6). The *softmax* activation function is used to indicate the dependency between classes (described in section 2.6.3 with more details).

Assuming that K denotes the number of output classes and the outputs are mutually exclusive, $\hat{y}_{ic}$ approximates the probability of class c for $c = 1, 2, ..., K$ of sample i based on the following formulation:

$$\hat{y}_{ic} = softmax(v_c^T z_i) = \frac{exp(v_c^T z_i)}{\sum_{o=1}^{K} exp(v_o^T z_i)}, \tag{2.6}$$

where z_i indicates the hidden unit value for sample i, and v_c is the connection weights between hidden unit and output unit for class c.

Activation function for hidden nodes can be linear or nonlinear such as sigmoid function. For having nonlinearity in the hidden nodes, sigmoid is one of the appropriate activation functions:

$$z_{ih} = sigmoid(w_h^T x_i), \tag{2.7}$$

where, z_{ih} is the value of hidden nodes of sample i. The error function for multiclass classification is:

$$Error_i = -\big[\sum_{c=1}^{K} y_{ic} log(\hat{y}_{ic})\big] \tag{2.8}$$

The update equations found using gradient descent procedure for multi-class classification are presented below.

$$\Delta v_{ch} = \eta \sum_{i=1}^{N} (y_{ic} - \hat{y}_{ic}) z_{ih}, \tag{2.9}$$

$$\Delta w_{hd} = \eta \sum_{i=1}^{N} (y_{ic} - \hat{y}_{ic}) v_{ch} z_{ih} (1 - z_{ih}) x_{id}. \tag{2.10}$$

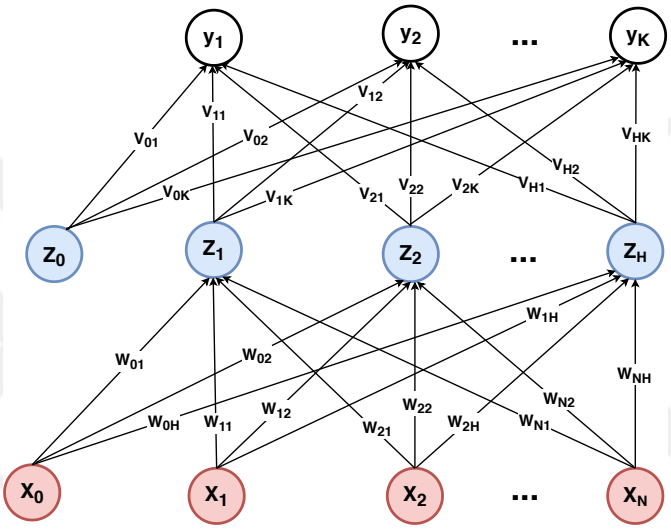

Figure 2.6: An MLP with multiple outputs

2.5 Backpropagation method

Backpropagation is an optimization algorithm in deep learning which is used to train a feedforward neural network for supervised learning. Backpropagation computes the gradient of the applied loss function with respect to the connection weights of the model and helps to tune these connection weights in the model. Therefore, it is important to apply gradient approaches for training multilayer networks and updating weights to reach the minimum value of loss. There are different ways to measure error, e.g., root mean squared error (RMSE). In the backpropagation algorithm, the gradient of the error function with respect to every weight is computed using the chain rule:

$$\frac{\partial Error}{\partial w_{hd}} = \frac{\partial Error}{\partial \hat{y}_c} \frac{\partial \hat{y}_c}{\partial z_h} \frac{\partial z_h}{\partial w_{hd}} \tag{2.11}$$

where w_{hd} is the first-layer weights, and z_h is the value of h^{th} unit of hidden layer. To prevent unnecessary calculations of middle numbers in the chain rule, the error propagates from the output y back to the inputs and this approach can iterate backwards by one layer at a time, starting from the last layer (Rumelhart et al., 1986).

A feedforward model starts to learn with random values of hyperparameters, and it makes a prediction using these learned values. Then, backpropagation is used in a backward pass, and differentiation gives a gradient, or a view of error. The parameters may be regulated as they move the MLP one step closer to the minimum error. It compares the predicted values with the real values and adjusts the initial hyperparameters for minimizing the error.

2.6 Activation functions

In ANNs, activation functions are used to learn non-linear mappings between inputs and outputs. Therefore, choosing a proper activation function plays a significant role in neural network performance improvement. Activation functions change an input signal of a neuron which is a nonlinear transformation to an output signal in a neural network. In particular, an ANN activation function $f(x)$ is applied on the sum of the products of inputs and biases to their corresponding weights to obtain the output of that layer and feed it as an input to the next layer. Essentially, activation functions determine whether a neuron should be activated or not. They activate a neuron when the receiving information is relevant, otherwise they ignore it. Since linear equations have limited capacity to solve a complex problem, it is necessary to have a nonlinear activation function in the neural network (Agostinelli et al., 2014).

Activation functions are divided into three categories: (1) threshold-based, (2) linear, and (3) nonlinear (Ramachandran et al., 2017).

2.6.1 Threshold-based activation functions

A threshold-based activation function is also known as a binary step function. It produces a binary output. The output produces 1 when the input sum is above

a threshold limit and it produces 0 when input sum is below the threshold limit (Plagianakos and Vrahatis, 2000).

2.6.2 Linear activation functions

The linear activation function is the unit function and the activation is the weighted sum from the neurons. Therefore, this activation is proportional to input and its output is not limited to any range. Using linear activation functions has one major drawback. An ANN has restricted power and capability to manage complexity of input dataset with linear activation functions.

2.6.3 Nonlinear activation functions

ANNs can model nonlinear processes to tackle many problems across different machine learning problems, namely, pattern recognition, classification, regression, cancer detection system and clustering. Recent ANNs algorithms utilize different nonlinear activation functions. They enable the network to build complicated mappings between the inputs and outputs of the network. Hence, nonlinear activation functions are extremely important for learning and modeling complex datasets that are nonlinearly seperable or has a nonlinear trend (Chung et al., 2016). There are diverse kinds of nonlinear activation functions, for instance, sigmoid or logistic, tanh or hyperbolic tangent, rectified linear unit (ReLU) and softmax activation function. In the current study, we applied sigmoid activation function for binary classification.

Sigmoid activation function

Sigmoid is an S-shaped and bounded function and it is known as squashing function. It requires an actual value as input and the value of its output is between 0 and 1. The main reason for using sigmoid is mapping the whole real axis into the finite interval. The sigmoid function has been utilized for binary classification problems in the logistic regression models. Some advantages of sigmoid activation function are as follows: it is nonlinear, continuously differentiable, monotonic, and has a determined output range. Some disadvantages of sigmoid activation function are as

follows: Its output is not zero centered and it increases the problem of vanishing gradients (Karlik and Olgac, 2011).

Softmax activation function

In ANNs, the softmax activation function computes the probability distribution of a class across "K" different classes and maps the output of a network to a probability distribution across expected output classes. This activation function is utilized to build a multiclass classifier. Softmax is used in terms of the output function in the last layer of models and assigns input data to one class whenever the number of possible classes is greater than two. Since the aim of the last layer is to change the score generated by the model into values that are interpretable by humans, therefore it is really important to determine the activation function of the last layer efficiently. The major benefit of using softmax is that the range of the outputs probabilities is between 0 and 1, and the summation of all the probabilities equals one. In multiclass classification, softmax function calculates the probabilities of each class and the true class is expected to have a higher probability (Alpaydın, 2009).

2.7 Regularization functions

Regularization is a standard technique that is used to avoid overfitting of the machine learning models by fitting a simpler function on training data. Once a classifier trains on detail and the noise in the training data, it negatively impacts its performance on new data, and overfitting occurs. Noisy data points do not consider the true characterization of a dataset. Learning these data points leads to training a more flexible classifier, and helps us prevent overfitting. Another way to control overfitting is increasing the size of the training dataset. Deep neural networks are highly complicated networks because they have many parameters and many non-linearities, so they are easy to overfit, whereas regularization is used to reduce the error by fitting a function appropriately on the given training set (Srivastava et al., 2014).

Regularization is applicable for tuning the function by allowing to add an addi-

tional penalty term in the error function. In deep learning, regularization actually applies penalties on layer parameters or layer activity during optimization and penalizes the weight matrices of the nodes. The network optimizes these penalties that are adopted in the loss function. The role of this additional term is controlling the excessively fluctuating function, so the coefficients cannot take extreme values. By optimizing the value of the regularization coefficient, a well-fitted model can be obtained. There are several regularization techniques that are useful in machine learning such as Lasso regularization, ridge regularization, early stopping, dropout, and elastic net. In this thesis, we used Lasso regularization technique which is described next.

2.7.1 Lasso regularization

Lasso (Least Absolute Shrinkage and Selection Operator) is also known as ℓ_1- norm, adds a penalty to the error function. In Lasso regulariztion the penalty of the error function is the summation of the absolute values of all weights. Lasso regularization minimizes the following objective function:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{P}|\beta_j| = RSS + \lambda \sum_{j=1}^{P}|\beta_j|, \tag{2.12}$$

which includes residual sum of square error (i.e, RSS) as the loss function and Lasso regularization function. The coefficients are chosen to minimize the function 2.12. β_0 is the intercept when all other predictors are equal to zero. This intercept measures mean value of the result. $\lambda \geq 0$ is a tuning or complexity parameter that examines the amount of shrinkage and decides how much we want to penalize the flexibility of the model. Coefficients should be learned based on the training dataset. The presence of noise in the training data causes to not estimate well-generalized coefficients for the future data. This is the reason for using regularization and shrink (Hastie et al., 2015).

2.8 Cross-validation

Cross-validation is a statistical method used to evaluate predictive performance of machine learning models. One way to avoid overfitting is utilizing cross-validation technique that assists in selecting proper parameters for model and evaluating the error through the test set. K-fold cross-validation is one type of cross-validation that we used in this thesis. It has one single parameter called K which is the number of partitions to split the selected dataset into.

Chapter 3

MULTILAYER PERCEPTRON WITH MULTIPLE KERNELS

Machine learning approaches are highly applicable for the prognosis of cancer. In this thesis, we attempted to implement a standard machine learning method for discriminating early- and late-stage of cancers. To predict the pathological stages of the samples, which are the primary tumours, we used their gene expression profiles. This problem can be addressed as a binary classification problem. There are several standard machine learning methods for solving classification problems such as SVMs (Cortes and Vapnik, 1995), neural networks (McCulloch and Pitts, 1943), and RFs (Breiman, 2001). High prediction accuracy of machine learning methods is not sufficient by itself. To this end, the knowledge extraction capability of the classifiers is also important for the problem of predicting the stages of cancer.

3.1 Problem formulation

We addressed the problem of separating the stages of primary tumours on the gene expression data formulated as a binary classification problem. Gene expression data is indicated as $\mathcal{X}$, and the related phenotype of these profiles tumours, which are early- and late-stage, is indicated as $\mathcal{Y}$. We arbitrarily named early-stage primary tumours as class 1 and late-stage primary tumours as class 0 in MLP algorithm. However, for the other methods, early-stage primary tumours are considered as positive class and late-stage primary tumours are considered as negative class. We represented the training dataset for a given problem as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where N indicates the number of patients with primary tumours, $\boldsymbol{x}_i$ indicates the gene expression profile of primary tumour for patient i, and $y_i \in \{0, 1\}$ in MLP algorithm. Note that in other competitive methods the class label of primary tumour of patient i is

represented by $y_i \in \{-1, +1\}$. This binary classification problem can be solved by learning a discriminant function to predict the phenotype from the gene expression profiles. This discriminant function can be shown as $f : \mathcal{X} \to \mathcal{Y}$. This function is learned over training data which is a subset of available data and is used to make predictions for out-of-sample primary tumours which are unseen during training the discriminant function.

3.2 Baseline algorithms

We compared the performance of our implemented MLP algorithm against three different well-known machine learning methods, namely, RF, SVM, and MKL on the Hallmark gene sets collection.

3.2.1 Random forest

The RF is a classification technique that applies an ensemble strategy for combining several weak decision trees to obtain a more robust classifier compared to a single tree. These decision trees built on random input samples and split nodes on a random subset of features. Randomly selecting a subset of features increases the diversity. Each decision tree classifies samples by splitting the training set into subsets through a forking path of decision points. Each subset is a decision point and has a rule to decide which branch to take. We should continue and move down the tree and apply this rule at each decision point until the end of the branch (i.e., the leaf that has the class label) is reached. Therefore, every decision tree provides a classification by voting for a specific class and then the RF model aggregates the votes from all of the decision trees. Next, the RF model decides the final class of the test data to obtain a better prediction accuracy. Since RF algorithm has a proper ability for classification and generalization, it is widely used in many recent research projects and real-world applications in different areas. It is obvious that a set of multiple weak classifiers usually perform better than a single classifier given the same training dataset. Also, it is computationally faster than other ensemble methods. Misclassification/error rate of RF classifiers depends on the correlation

and dependency between trees, and the classification strength of each individual tree. Therefore, the RF algorithm is robust to noise and outliers. Also, it runs effectively on large datasets (Breiman, 2001). RFs commonly used for the problem of predicting stages of cancer which are the phenotype of the disease from the genomic measurements (Díaz-Uriarte and De Andres, 2006; Pang et al., 2006; Statnikov et al., 2008; Nam et al., 2009).

In this case, we selected RF as a baseline algorithm. Although the predictive performance of RFs are quite good, their capability to extract knowledge are very restricted. Since each decision tree in the RF algorithm built on a bootstrap sample of the training data, knowledge extraction of RFs depends on the bootstrapping step because these bootstrap samples are randomly selected subsets of variables.

3.2.2 Support vector machine

SVM is a type of supervised machine learning technique and it is applicable in classification or regression problems. SVM is a discriminant-based method used in both linear and nonlinear binary classification problems. The objective of the SVM method is to discover the hyperplane that maximizes the margin among different classes of the training data and distinctly classifies the data points. Misclassification in binary classification SVM happens when an instance lies on the incorrect side of the hyperplane. Kernel trick uses a specific transformation to convert input data into higher dimensional data. Kernel trick is used for linearly inseparable data, and it enables us to form nonlinear boundaries by mapping from the original input space into the feature space. SVM training problem can be commonly formulated as a linearly constrained convex quadratic optimization problem (Cortes and Vapnik, 1995).

The optimization model of learning binary classification SVMs can be presented

as

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{N}\xi_i \\ \text{with respect to} \quad & \boldsymbol{w} \in \mathbb{R}^D, \ \ \boldsymbol{\xi} \in \mathbb{R}^N, \ \ b \in \mathbb{R} \\ \text{subject to} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall i \\ & \xi_i \geq 0 \qquad \forall i, \end{aligned} \tag{3.1}$$

where $\boldsymbol{w}$ indicates the set of weights dedicated to features, C is a penalty factor as a positive regularization parameter, which identifies the trade-off between maximizing the margin and minimizing the slacks. When C is small, constraints can be easily ignored that leads to maximizing the margin and less importance is given to classification errors, whereas large C makes constraint hard to ignore, and it means that the focus is more on avoiding misclassification at the expense of keeping the margin small. $\boldsymbol{\xi}$ shows the set of slack variables that give the deviations from the margins. D indicates the number of input features, it illustrates the number of genes in gene expression profiles. Finally, b denotes the intercept parameter, which is the bias value of the separating hyperplane.

As an alternative, we can solve the corresponding Lagrangian dual optimization problem using the Lagrangian function. There are two significant advantages of solving the dual optimization problem in SVMs. First, the number of decision variables reduces. Second, solving the dual helps to integrate kernel functions into the classification problem to classify data that are not linearly separable in the initial feature space, which leads to nonlinear models. The Lagrangian function can be formulated as follows:

$$\mathcal{L} = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i\big) - \sum_{i=1}^{N}\beta_i\xi_i. \tag{3.2}$$

By taking the derivatives with respect to the decision variables of the primal problem which are the parameters and set them to 0, we get

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \beta_i \qquad \forall i.$$

Plugging these back into the Lagrangian function, the dual problem can be formulated as:

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{N} \alpha_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \\ \text{with respect to} \quad & \boldsymbol{\alpha} \in \mathbb{R}^N \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \qquad \forall i, \end{aligned} \tag{3.3}$$

where instead of $(D + N + 1)$ decision variables, there are N decision variables and to achieve nonlinear models $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ term can be substituted with a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

3.2.3 Multiple kernel learning

The MKL method has been extensively studied over the past decade. Instead of choosing a single kernel that might not be sufficient to capture the complexity of the problem, MKL applies a weighted (sum) combination of input kernels (Gönen and Alpaydın, 2011), where for each kernel the weight is optimized during the training procedure. Therefore, the MKL algorithm is used to decide the most effective combination of a set of kernels to form the best classifier. The aim of the underlying combination of kernels is to enhance the classification accuracy by adjusting the kernel to the features of heterogeneous data. Multiple linear and nonlinear kernel functions can be combined in the MKL algorithm and these kernels might be defined on the identical input representation or on diverse input data sources such as multi-view learning. There are many kernel-based machine learning techniques (e.g., SVMs) whose generalization performance extremely depend on the kernel function, which is used to measure the similarity between two instances. SVM classifiers' goal

is to achieve a trade-off between the algorithm complexity and predictive performance that is extensively depend on the penalty error C and kernel parameters. The MKL starts by extracting features from all available data sources and then calculate different kernel functions on the training dataset. Finally, it finds the optimal kernel combination that gives the highest predictive performance in the final kernel classifier. In computational biology applications, there are different types of data representations for the same problem. Therefore, bioinformatics is an appropriate area for using data integration techniques. There are three different ways to integrate different data sources into a unified model:

1. Early data integration concatenates all input representations into a single dataset before learning and then builds the classifier on this dataset.

2. Late data integration combines predictive models learned on each representation.

3. Intermediate data integration uses a joint model to combine the representations in the form of distance or kernel matrices.

Although early and late data integration have better predictive accuracy in some studies, intermediate data integration is often used due to its comparable predictive performance and additional interpretability. Therefore, instead of using early and late data integration for differentiation the stages of cancer, the intermediate integration is used. Kernel matrices are calculated and then the best combination of these kernels is found in the predictive algorithm during the training.
Rahimi and Gönen (2018) used an MKL algorithm with a group lasso regularizer to differentiate the stages of cancers from each other. Group Lasso MKL combines separate kernel matrices that constructed for each gene set. An ℓ_1-norm of the kernel weights is used to enhance the interpretability of MKL algorithm. This ℓ_1-norm results in building a sparse combination of the kernels, where most of the kernel weights are zero (Xu et al., 2010). Each kernel matrix built on a specific gene set, hence gene sets with nonzero kernel weights provide information for the discrimination problem between early- and late-stage cancers.

Group Lasso MKL is used to identify kernel weights together with other parameters of the problem. To this aim, the following optimization problem, which has the dual optimization problem for SVM within itself as an inner problem, needs to be solved.

$$\begin{aligned} \text{minimize} \quad & J(\boldsymbol{\eta}) \\ \text{with respect to} \quad & \boldsymbol{\eta} \in \mathbb{R}^P \\ \text{subject to} \quad & \sum_{m=1}^{P} \eta_m = 1 \\ & \eta_m \geq 0 \qquad \forall m, \end{aligned} \tag{3.4}$$

where $\boldsymbol{\eta}$ denotes the kernel weights, P represents the number of input kernels, and $J(\boldsymbol{\eta})$ represents the dual SVM optimization problem in (3.3). In equation (3.4) the objective function has changed by substituting $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j$ term with $\sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$. To achieve a sparse solution the unit simplex constraint which is the equality constraint in (3.4) is added to the optimization problem. Note that equality constraint in (3.4) is same as inducing ℓ_1-norm on the weights of kernels.

Rahimi and Gönen (2018) applied an optimization approach as an alternative which has been proposed by (Xu et al., 2010) to solve the optimization problem in (3.4) because it is not solvable with respect to both parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ together. The optimization problem in (3.4) is the outer minimization problem which is convex with respect to $\boldsymbol{\eta}$ and (3.3) is the inner minimization problem which is convex with respect to $\boldsymbol{\alpha}$. The alternative optimization approach is used in an iterative way to optimize the mentioned parameters. This approach is based on group Lasso MKL algorithm and developed for binary classification problems. This iterative algorithm starts by setting the kernel weights to uniform values, e.g., $\eta_m^{(1)} = 1/P$.

At every iteration t, to find the support vector coefficients $\boldsymbol{\alpha}^{(t)}$, the optimization problem in 3.3, which is the standard SVM model, is solved by utilizing existing kernel weights $\boldsymbol{\eta}^{(t)}$. To identify kernel weights of the next iteration $\boldsymbol{\eta}^{(t+1)}$ a closed-form update equation is applied as follows:

$$\eta_m^{(t+1)} = \frac{\eta_m^{(t)} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)}}{\sum_{o=1}^{P} \eta_o^{(t)} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i^{(t)} \alpha_j^{(t)} y_i y_j k_o(\boldsymbol{x}_i, \boldsymbol{x}_j)}} \qquad \forall m,$$

where the superscripts $(t+1)$ and (t) indicate the next and current iterations, respectively. This iterative methodology is used to solve binary classification problems (Xu et al., 2010). Important pathways/gene sets can be determined by checking their corresponding kernel weights in the model implemented by Rahimi and Gönen (2018).

3.3 Our proposed multilayer perceptron model

In this thesis, we implemented an MLP algorithm for binary classification problem to determine the biological procedures that differentiate early-stage cancers from late-stage cancers. To this aim, for each pathway/gene set, we used distinct kernel matrices which are calculated on the expression features of genes that exist in the specific pathway/gene set. These kernels capture the complexity and measure the similarity between patients. The procedure of building kernels on gene expression profiles is same as the procedure in the MKL algorithm. However, instead of combining diverse kernels with a weighted sum to obtain a more informative kernel using the MKL technique, we adopted an MLP algorithm for binary classification. MLP is the most popular class of feedforward neural networks, and it is suitable for both classification and regression problems. The MLP model contains at least three layers of nodes, which are an input layer, a hidden layer, and an output layer. These nodes in each layer are also known as *neurons*.

3.3.1 Structure of the proposed MLP

We designed a multi-input MLP algorithm for predicting the pathological stages of primary tumours (Figure 3.1). In this layout, subscripts i and m index the units in the input layers and hidden layer, respectively.

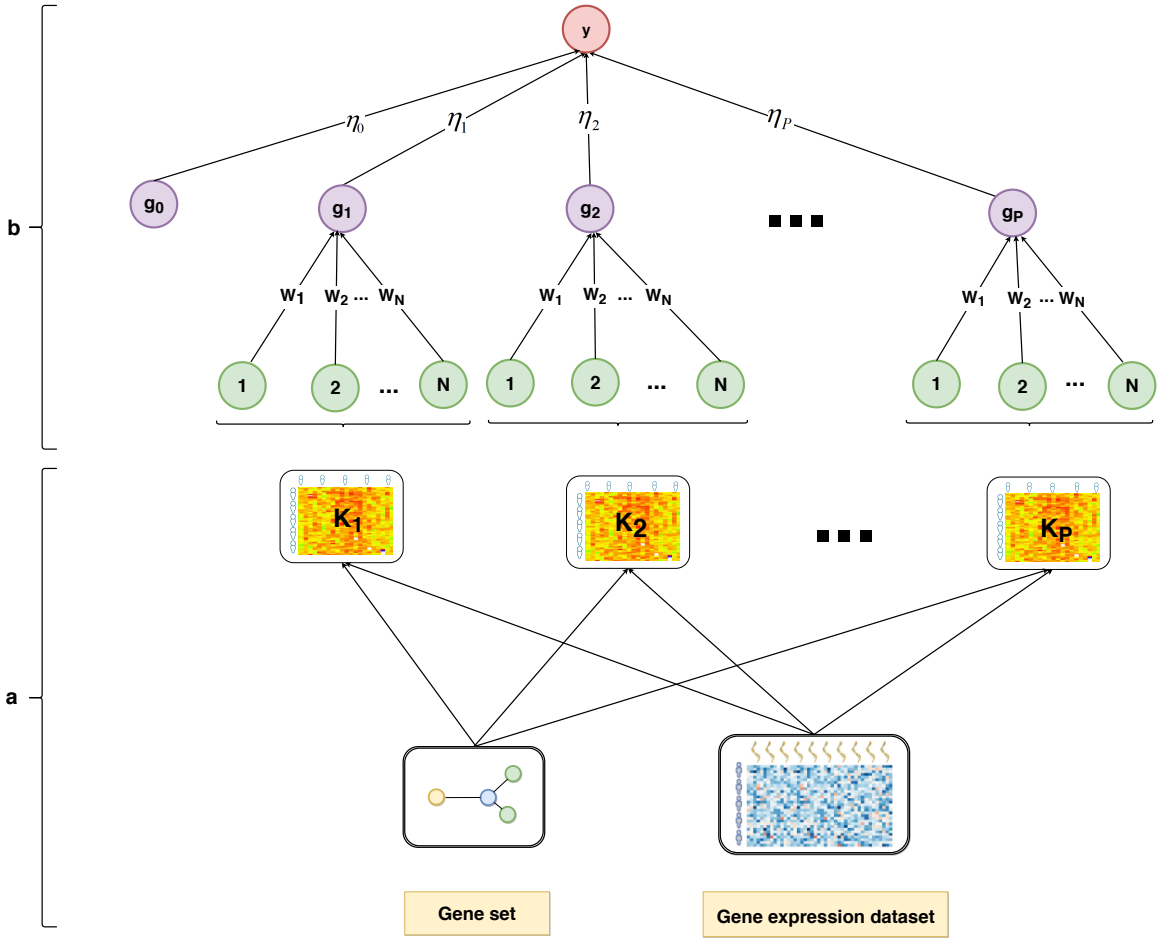

Figure 3.1: Overview of our MLP algorithm. (a) Calculating kernel matrices for integrating pathways/gene sets into the classifier. (b) The structure of our MLP algorithm.

We calculated the kernel matrices on gene expression profiles of patients for distinct pathways/gene sets. These kernels are the inputs to our proposed MLP network that can be represented as $\{k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{m=1}^{P}$, where P is the number of kernels and the dimension of each kernel matrix is $N \times N$. We integrated each kernel matrix into a separate input layer, thus, the constructed MLP has P different input layers for each input kernel (i.e., $K_1, ..., K_P$) and each input layer has N nodes. Every node in each input layer is connected to the corresponding kernel node in the hidden layer.

Since MLP is a feedforward neural network, the data in this network goes from the input layer to the last layer. Every input layer has its corresponding connection weights $w_1, ..., w_N$. These weights are shared because we want to use the same

weight value for different input kernels of each input layer to decrease the number of parameters. Therefore, applying weight sharing leads to decreased model complexity.

In the hidden layer, our implemented MLP algorithm has one hidden node for each input kernel $g_1, ..., g_p$ and one bias node g_0. The value of each hidden node is equal to the weighted sum of its inputs:

$$g_m = \sum_{i=1}^{N} w_i \boldsymbol{x}_{m,i} \qquad \forall m = 1, ..., P, \tag{3.5}$$

where N is the number of input units in each input layer, w_i is the connection weight among input layer i and hidden layer, and $\boldsymbol{x}_{m,i}$ is the value of the input entry which are the input kernels for $\boldsymbol{x}_i$ and kernel m. We defined the input entries for each input layer as:

$$\boldsymbol{x}_{m,i} = (k_m(\boldsymbol{x}_1, \boldsymbol{x}_i), ..., k_m(\boldsymbol{x}_N, \boldsymbol{x}_i)) \qquad \forall m = 1, ..., P \tag{3.6}$$

where $k_m(.,.)$ measures the similarity between samples for pathway m.

We computed the value of each hidden unit by propagating a linear activation function in the forward path from the corresponding input layer. Therefore, the values of hidden units become the weighted sum of the corresponding input kernel matrices:

$$g_m = \sum_{i=1}^{N} w_i k_m(\boldsymbol{x}_i, \boldsymbol{x}) \qquad \forall m = 1, ..., P. \tag{3.7}$$

Bias node is an extra parameter in the MLP algorithm, which is utilized to adapt the output of the network together with the weighted sum of the input values to the neurons of the network. Bias produces constant value 1 (or another constant value) and is not connected to the previous layer. Therefore, bias is a constant that assists the network to fit properly for the inputs data. Our implemented MLP network does not have a bias node in the input layer, however, in the hidden layer it has a node for the bias term.

Similar to each input layer, we applied the linear activation function to the weighted sum of the input values for each hidden unit. We converted this weighted

sum to the output by a linear function such that the output value consists of the transformed weighted sum from the previous layer.

The linear activation function applied to the weighted sum of the hidden layer produces the output of the hidden layer as follows:

$$\sum_{m=1}^{P} \eta_m g_m + \eta_0 g_0 \tag{3.8}$$

where $\eta_1, ..., \eta_P$ are the connection weights among the hidden layer and the output layer for every hidden unit.

There is also a bias node in the hidden layer which is indicated by g_0, where η_0 is the bias weight. Finally, in the output layer, the number of nodes is equal to the number of classes of the original dataset. Note that in our MLP algorithm, there was one perceptron as output that is sufficient for two classes. We applied a nonlinear sigmoid activation function to the input values of the output neuron. Since we solved a binary classification problem and sigmoid is an appropriate activation function for this kind of problems, we used sigmoid in the last layer.

The output layer is a fully connected layer which means that all nodes in the hidden layer linked to the output layer which has one node. We enforced Lasso regularization on top of all the connection weights between hidden layer nodes and output layer node. Using Lasso leads to sparsity in the kernel weight vectors. Therefore, we obtained mostly zero weights due to ℓ_1-norm and reduced overfitting risk by eliminating irrelevant kernels. Since negative kernel weights are not meaningful, we also imposed non-negativity constraints on the connection weights of the last layer to improve the interpretability of our MLP.

We obtained the output of the applied MLP model which is the prediction value for primary tumours by applying sigmoid activation function in the last layer:

$$\hat{y} = sigmoid(\sum_{m=1}^{P} \eta_m g_m + \eta_0 g_0), \tag{3.9}$$

where $\hat{y}$ can be written as follows:

$$\hat{y} = \frac{1}{1 + e^{-(\sum_{m=1}^{P} \eta_m g_m + \eta_0 g_0)}} \tag{3.10}$$

The objective function of the MLP algorithm for the binary classification problem is to minimize the error (loss), which is the difference among the desired output and the predicted output. We used binary cross-entropy loss function to find optimum parameters (weights) values as follows:

$$f(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - (y_i)) \log(1 - (\hat{y_i}))], \tag{3.11}$$

where the y is the actual value of the output and $\hat{y}$ is the predicted value of the output. The process of training an MLP is to determine a set of parameters that reduce the error (among the outputs of the MLP algorithm and the target value) to minimum error value using gradient descent that also known as backpropagation. Updating the weight values to decrease the amount of the loss function is the purpose of training MLP by backpropagation algorithm. Updating parameters includes two steps: (1) computing the gradient of the applied loss function with respect to the parameters and then (2) updating the parameters using the gradients. This procedure is repeated until the minimum value of the loss function is obtained.

We modified the weights using an optimization function that calculates the gradient of the loss function. Backpropagation algorithm uses the chain rule to calculate the gradient of connection weights from hidden to output layer. The update equation for η_m can be written as follows:

$$\Delta\eta_m = \mu \sum_{i=1}^{N}(y_i - \hat{y}_i)g_{m,i} \qquad \forall m = 1, ..., P, \tag{3.12}$$

where μ is the learning rate which represents the step size to update the parameters. The subscript i is the indicator for i^{th} sample. The updated equations for the connection weights between input layer and hidden layer after implementing the gradient descent procedure can be formulated as follows:

$$\Delta w_i = \mu \sum_{i=1}^{N}\sum_{m=1}^{P}(y_i - \hat{y}_i)g_{m,i}\boldsymbol{x}_{m,i} \qquad \forall i = 1, ..., N. \tag{3.13}$$

Since we implemented our MLP network in Tensorflow (Abadi et al., 2015), the entire backpropagation technique was performed automatically using a specific optimizer and the loss function for calculating the output of the MLP model.

We illustrated the overview of our MLP algorithm for predicting pathological stage of tumours based on their available gene expression profiles in Figure 3.1. In part (a) distinct kernel matrices were calculated on the gene expression profiles of primary tumours for different pathways/gene sets that considered as inputs. Kernel matrices denoted as $\mathbf{K}_m \ \forall m = 1, ..., P$, integrated into the constructed MLP model for integrating kernel information into the classification algorithm. Part (b) demonstrates the structure of our conducted MLP network, which contains P input layers with N number of nodes, one hidden layer with $P + 1$ nodes, and one output layer with one node.

Chapter 4

COMPUTATIONAL EXPERIMENTS

In this chapter, we discussed our datasets and presented the results of our computational experiments.

4.1 Datasets

Staging determines the extent of cancer in the body and accurate staging is helpful to predict progression of cancer and begins the effective medical treatment of patients. Obviously the cancer prognosis in the early-stage is an essential element in the efficient treatment of the disease. There are different techniques for the staging of cancer including imaging procedure and TNM staging system. However, these techniques have limitations and since the precise prediction of the disease outcome is really important; thus it is required to establish alternative methods for classification. We developed a classification algorithm for predicting the stages of cancer using genomic data and RNA sequencing-based gene expression data prepared from The Cancer Genome Atlas (TCGA) via adopting state-of-art supervised machine learning methods. The importance of classifying stages of cancer has lead many scientists, from the biomedical and the bioinformatics field, to study the application of machine learning algorithms. These techniques are able to determine and detect patterns and relationships between them from huge and complex datasets. Therefore, machine learning techniques are useful to predict the stages of cancer. We separated early-stage from late-stage cancers using genomic data with the MLP method.

This study was performed using clinical information and genomic characterizations of preprocessed gene expression datasets by RNA-seq and pathological stage of primary tumours from publicly available TCGA project.

4.1.1 TCGA datasets

TCGA is a comprehensive consortium that offers remarkable genomic characterizations along with clinical and histopathological information of more than 10,000 cancer patients across 33 different cancer types. It assists researchers to determine the main cancer-causing genomic alterations. In this dissertation, to understand the differentiation between stages of cancers, we used different cancer cohorts from the TCGA dataset.

We also utilized preprocessed gene expression profiles of tumours, which were generated by the unified RNA-Seq pipeline of the TCGA project. For each cancer type, for all primary tumours, we downloaded HTSeq-FPKM files from the latest data freeze (i.e., `Data Release 22 - January 16, 2020`), results in 9,911 files in overall. Metastatic tumours are not included considering that their fundamental biology would be completely dissimilar to primary tumours. We represented cancer types with at least 20 samples in each stage of the cancer, and that is why 20 cancer types were included in our experiments. Detailed information for two groups of the experiments can be found in Table 4.1.

All of the cohorts in our experiments are publicly accessible in the GDC data portal at `https://portal.gdc.cancer.gov`.

4.1.2 Pathway/gene set databases

There exist different approaches for identifying gene-level molecular signatures such as gene set analysis and predictive modeling methods. Due to strongly correlated nature of gene expression datasets, gene molecular signatures derived from gene expression datasets are not robust when the training dataset size is small (Ein-Dor et al., 2005, 2006). These dependencies and correlation between gene expression data would lead to obtain diverse molecular signatures from various subsets of the identical training dataset. To identify the specific molecular signatures, we integrated the prior information about genes into a conjoint modeling approach in the form of pathways/gene sets.

To predict the stage of a tumour, understanding the biological mechanisms that

separate the stages of cancers from each other is important. To achieve this aim, we utilized pathways and/or gene sets. These collections of data present detail information about sets of genes which are similar or dependent in terms of their functions. The MSigDB is used to extract the Hallmark gene sets which are publicly available at `http://software.broadinstitute.org/gsea/msigdb`. Hallmark gene sets represent precise biological states and convey a specific process. Additionally, every gene set displays coherent expression in different cancers. These gene sets were created through a computational procedure based on determining overlaps among the gene sets that exist in another MSigDB collections and the keeping genes that present dependent expression (Liberzon et al., 2015). The Hallmark database consists of 50 gene sets, and the sizes of these gene sets alter between 32 and 200. We chose this collection since it was proposed especially for cancer research.

All of the competitive algorithms employed gene expression profiles of tumours and Hallmark pathway/gene set database to predict pathological stages of primary tumours. However, RFs and SVM did not take multiple kernels as inputs, therefore, it is not possible to integrate pathway or gene set information in their original formulation. RFs and SVMs require just one single matrix that is gene expression profiles of primary tumours in their original formulation. Using the Hallmark gene set leads to more robust and reliable computational results. MKL and MLP algorithms integrate pathways/gene set in their inputs. Furthermore, MKL and our MLP methods extract extra information about the differentiation among stages of cancers.

4.1.3 List of datasets

Our experiments can be divided into two categories with 15 and 18 TCGA cohorts, respectively. These two groups of experiments are constructed out of 20 cohorts. In Table 4.1, we reported TCGA cohorts abbreviation, name of the disease and the number of primary tumour samples in each pathological stage of cancer for each cohort. We also reported the number of sample primary tumours in early-stage and late-stage cancers, and the total number of samples in two groups of the experiments,

namely, E1 and E2. The total number of patients that used for both groups of the experiments, E1 and E2, were 5,547 and 6,038, respectively.

Table 4.1: Summary of 20 TCGA used cancer types in our two groups of the experiments

Cohort	Disease name	Stage I	Stage II	Stage III	Stage IV	Early (E1)	Late (E1)	Total (E1)	Early (E2)	Late (E2)	Total (E2)
ACC	Adrenocortical carcinoma	9	37	16	15	—	—	—	46	31	77
BLCA	Bladder urothelial carcinoma	2	130	140	134	—	—	—	132	274	406
BRCA	Breast invasive carcinoma	181	619	247	20	181	886	1,067	800	267	1,067
COAD	Colon adenocarcinoma	75	176	128	64	75	368	443	251	192	443
ESCA	Esophageal carcinoma	16	69	49	8	16	126	142	85	57	142
HNSC	Head and neck squamous cell carcinoma	25	70	78	259	25	407	429	95	337	429
KICH	Kidney chromophobe	20	25	14	6	20	45	65	45	20	65
KIRC	Kidney renal clear cell carcinoma	265	57	123	82	265	262	527	322	205	527
KIRP	Kidney renal papillary cell carcinoma	172	21	51	15	172	87	259	193	66	259
LIHC	Liver hepatocellular carcinoma	171	86	85	5	171	176	347	257	90	347
LUAD	Lung adenocarcinoma	274	121	84	26	274	231	505	395	110	505
LUSC	Lung squamous cell carcinoma	244	162	84	7	244	253	497	406	91	497
MESO	Mesothelioma	10	16	44	16	—	—	—	26	60	86
PAAD	Pancreatic adenocarcinoma	21	146	3	4	21	153	174	—	—	—
READ	Rectum adenocarcinoma	30	51	51	24	30	126	156	81	75	156
SKCM	Skin cutaneous melanoma	2	66	27	3	—	—	—	68	30	98
STAD	Stomach adenocarcinoma	53	111	150	38	53	299	352	164	188	352
TGCT	Testicular germ cell tumours	55	12	14	0	55	26	81	—	—	—
THCA	Thyroid carcinoma	281	52	112	55	281	219	500	333	167	500
UVM	Uveal melanoma	0	39	36	4	—	—	—	39	40	79
					Total	1,883	3,664	5,547	3,738	2,300	6,038

4.1.4 Constructing datasets

In this part, we determined the datasets for two groups of the experiments. RNA-seq and gene set data sources were required for training a classifier to determine the pathological stage of tumours using their gene expression profiles. Therefore, we extracted primary tumours with accessible `HTSeq-FPKM` file and `pathologic_stage` annotation for every cancer type. Localized cancers are considered early-stage because they are only found in the tissue where cancer originated before it spreads to the other areas of the body. Regional and distant metastasis cancers are often considered late-stage due to the fact that cancer has spread to nearby lymph nodes and/or to a different part of the body respectively.

At first, we identified primary tumours that have `Stage I` annotation to be early-stage which is also known as localized cancers and the other tumours which have `Stage II`, `III`, or `IV` annotations to be late-stage which are regional or distant

spreads cancer. There are some primary tumours with `Stage X` and `IS` annotations, although their gene expression profiles are available they were not included in our analyses, since `Stage X` is the terminal stage of cancer and `IS` means in situ and contains samples that did not turn into tumour yet. Moreover, these two stages are quite different from the other stages of tumours in terms of biology and they have a very limited number of samples for each cohort. For instance, there are 12 primary tumours in cohort `BRCA` with `Stage X` annotation and 46 primary tumours with `IS` annotation that we did not consider in our experiments.

In the next step, we incorporated TCGA cohorts that have at least 20 tumours from both early-stage and late-stage classifications in our last dataset list. Our first and second groups of the experiments specified as E1 and E2 contained 15 and 18 cohorts, respectively. The list of 15 cohorts that were used in our first group of experiments along with details of the number of samples in early-stage and late-stage cancers were given in Table 4.1.

In total, we used 5,547, primary tumours in E1. The cohort `KICH` is the smallest dataset with 65 primary tumours and the cohort `BRCA` is the largest dataset with 1,067 primary tumours. The size of other cohorts alter between 65 and 1,067. The lowest percentage of primary tumours in early-stage is 5.79%, which corresponds to `HNSC` cohort and the highest percentage of primary tumours in early-stage is 67.90%, which corresponds to `TGCT` cohort. The percentage of early-stage tumours in `HNSC` (25/432) indicates that among 432 samples in `HNSC` cohort just 25 of them are in early-stage and also the percentage of early-stage tumours in `TGCT` (55/81) indicates that among 55 samples in `TGCT` cohort 55 number of them are in early-stage (early-stage in the first group of the experiment corresponds to `Stage I` annotation).

In addition to E1, we tested another categorization of early-stage and late-stage cancers by assigning primary tumours annotated with `Stage I` or `II` to early-stage and primary tumours annotated with `Stage III` or `IV` to late-stage. In this case, localized and regional spreads cancers refer to early-stage and distant spreads cancers indicate late-stage. In this group of the experiments, we included 18 datasets with at least 20 tumours from both categories. The list of 18 datasets that were utilized in our second group of experiments along with details about the number of samples

in early-stage and late-stage cancers are also shown in Table 4.1.

In total, we used 6,038, primary tumours in E2. The lowest percentage of primary tumours in early-stage is 21.99%, which corresponds to `HNSC` cohort and the highest percentage of primary tumours in early-stage is 81.69%, which corresponds to `LUSC` cohort. The percentage of early-stage tumours in `HNSC` (95/432) indicates that among 432 samples in `HNSC` cohort just 95 of them are in early-stage and also the percentage of early-stage tumours in `LUSC` (406/497) indicates that among 497 samples in `LUSC` cohort 406 of them are in early-stage (early-stage in the second group of the experiments corresponds to `Stage I` and `Stage II` annotation). To evaluate the predictive performance of our MLP on gene sets, we conducted two groups of experiments (E1 and E2) on 15 and 18 datasets, respectively. The detailed information of these datasets are represented in Table 4.1. These datasets are taken from TCGA cohorts for performing the computational experiment on two different labeling procedures which are: `Stage I` versus `Stage II` or `Stage III` or `IV` and the other labeling is `Stage I` or `Stage II` versus `Stage III` or `IV`. We compared the performance of MLP method against three baseline methods, namely, RFs, SVMs, and MKL on gene sets.

4.2 Experimental setting

In this study, the four-fold inner cross-validation method is used to pick the hyper-parameters of RFs, SVMs, MKL and MLP models. In this technique, we divided the data points which are the profiled primary tumours of cancer patients into two sets by randomly selecting 80% as the training dataset and the remaining 20% as the test dataset. We kept the same proportion of the early-stage and late-stage classes in training and test datasets during the validation. This approach is called stratified cross-validation. To obtain Gaussian distribution among gene expression profiles of primary tumours, and since the input data is count data and can take only non-negative values, we applied $\log_2$-transformation on them before building classifiers. Each feature in the training dataset was standardized to have a mean of zero and a standard deviation of one, then each feature in the test dataset was

normalized by mean and the standard deviation over the training set.

We repeated the whole procedure of cross-validation with stratification and normalization 100 times to reliably measure the performance of machine learning models.

Hyperparameters of RFs consist of the number of decision trees to grow and the number of considered features for splitting each node in each decision tree. The RF algorithm is implemented in the `randomForestSRC` R package version 2.5.1 (Ishwaran and Kogalur, 2017). We selected the number of `ntree` from the set $\{500, 1000, \ldots, 2500\}$ by applying the four-fold inner cross-validation technique that we explained.

In SVM, the hyperparameters are the kernel function and subsequently the regularization parameter C. For SVM and MKL on gene sets algorithms, we utilize the implementation proposed by Rahimi and Gönen (2018) in R that utilizes `MOSEK` version 8.1.0.34 to solve optimization problems (MOSEK ApS, 2019). Kernels show the similarity measure between two patients. Similarity measure among gene expression profiles of primary tumours was calculated using the Gaussian kernel defined as

$$k_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\boldsymbol{x}_i - \boldsymbol{x}_j)}{2\sigma^2}\right),$$

where the kernel width parameter was selected as the mean of pairwise Euclidean distances among training samples. C is the regularization parameter that selected from the set $\{10^{-4}, 10^{-3}, \ldots, 10^{+5}\}$ using our four-fold inner cross-validation approach.

Rahimi and Gönen (2018) executed 200 iterations for MKL on gene sets method to guarantee the convergence since the convergence usually happens in tens of iterations. The Gaussian kernel matrices were computed on subsets of genes in gene expression profiles, which were included in the related pathway/gene sets, and also kernel width parameters were chosen subsequently.

We constructed our multi-input MLP network in R using Tensorflow framework (Abadi et al., 2015), which is an open-source software for deep learning. We used Keras high-level library that can be used on top of Tensorflow, and it is appropriate for conducting complex networks such as multi-input models and networks with a

shared layer (Chollet et al., 2015). Our implemented MLP network includes multi-input layers and one hidden layer that is fully connected to the output node. The inputs of MLP algorithm are Gaussian kernels, which are fed into multiple input layers. Backpropagation is used automatically in Keras for training MLP networks. It calculates the output error through a *forwardpass* and then it computes the error gradients and updates the weights of the network through a *backwardpass*. These two passes are performed on input data of one mini-batch at a time and after updating weights the next batch of training dataset fed into the neural network algorithm. This procedure is repeated until all of the data points are used for training the classifier.

In our proposed MLP method, some of the hyperparameters including the number of hidden layers, the number of hidden units in the hidden layer, and activation function are predefined before training, and we did not change them for choosing the best parameters of MLP algorithm. Some of the hyperparameters such as the number of nodes in the output layer and activation function in output layer are fixed since we are solving a binary classification problem. The number of hidden nodes in the hidden layer is fixed because we have to determine the weights of input kernel for extracting specific cancer information. Our MLP approach consists of P input layers which are defined for different Gaussian kernels that are integrated as inputs and each of these input layer has N input neurons. There are also, P hidden neurons that are equal to the number of input kernels and an additional neuron for the bias term in the hidden layer part. Lasso regularization and non-negativity constraint are used on the connection weights of the hidden layer to train the MLP classifier. In the hidden layer, linear activation function has been used. Sigmoid activation function has been used for binary classification and applied on the output layer that has one neuron as output unit.

We used random initialization to break symmetry and provide better accuracy. Therefore, we randomly initialized the network and trained the MLP model employing Adam optimizer with a learning rate of 0.001. In this thesis, we used binary cross-entropy loss function to optimize weight parameters and the area under the receiver operating characteristic curve (AUROC) metric for classification to evalu-

ate the predictive performance of the trained model. We used ℓ_1-norm regularizer on kernel weights to reduce the complexity of the network to avoid overfitting. For our MLP, to guarantee the convergence of the model, we used 10,000 epochs, which corresponds to the number of passes over the training dataset.

The optimal batch size, which refers to the number of training examples in one batch, is selected from the set $\{25, 50, 75, 100\}$ for each replication using our four-fold inner cross-validation.

Table 4.2 list the parameters of our MLP network. Some of the hyperparameters, are predefined, and we did not change them during the training procedure.

Table 4.2: Parameters of our MLP network

Parameters	**Value**	**Status**
number of input layers	P	fixed
number of input neurons in each input layer	N	fixed
number of hidden layers	1	fixed
number of hidden neurons	P	fixed
number of output neurons	1	fixed
activation function (hidden layer)	linear	fixed
activation function (output layer)	sigmoid	fixed
loss function	binary cross-entropy	free
optimizer	Adam	free
ℓ_1 regularization parameter	0.001	free
number of epochs	10,000	free
batch size	(25, 50, 75, 100)	free

4.2.1 Performance metric

We used the standard area under the receiver operating characteristic curve (AUROC) to measure the predictive performance of four machine learning algorithms. ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR)

for several candidate threshold values varying between 0.0 and 1.0 to predict the class label.

TPR and FPR, also known as sensitivity and specificity, are calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}},$$

where TP, FN, TN, and FP are the abbreviations of the number of samples correctly classified as positive, samples incorrectly classified as negative, samples correctly classified as negative, and samples incorrectly classified as positive, respectively.

4.3 Predictive performance comparison of classifiers on TCGA datasets

We compared four machine learning methods in our computational experiments in terms of their predictive performance:(i) random forest (RF), (ii) support vector machine (SVM) , (iii) multiple kernel learning (MKL), and (iv) multilayer perceptron (MLP)

We considered RFs as a competitor method because they have been commonly used in phenotype classification problems in the bioinformatics area. Furthermore, SVMs are used as a competitor method since they can capture more complex relationship between datasets. Rahimi and Gönen (2018) compared the impact of integrating pathways/gene sets information on the prediction performance of MKL classifier. Note that our MLP algorithm uses kernels obtained by integrating pathways/gene sets as inputs of the neural network. MKL algorithm also computes kernels by integerating pathways/gene sets. That is why we used MKL algorithm as a competitor method against MLP algorithm. We also evaluated the information extraction capability of the algorithms on gene expression profiles of primary tumours.

We used Hallmark gene set collection for MKL and MLP algorithms that represent specific distinct biological processes. The box-and-whisker plot is used for each cohort to compare the AUROC values of the competitive methods across 100 replications. A two-tailed paired t-test is used to compare SVM, MKL, and MLP

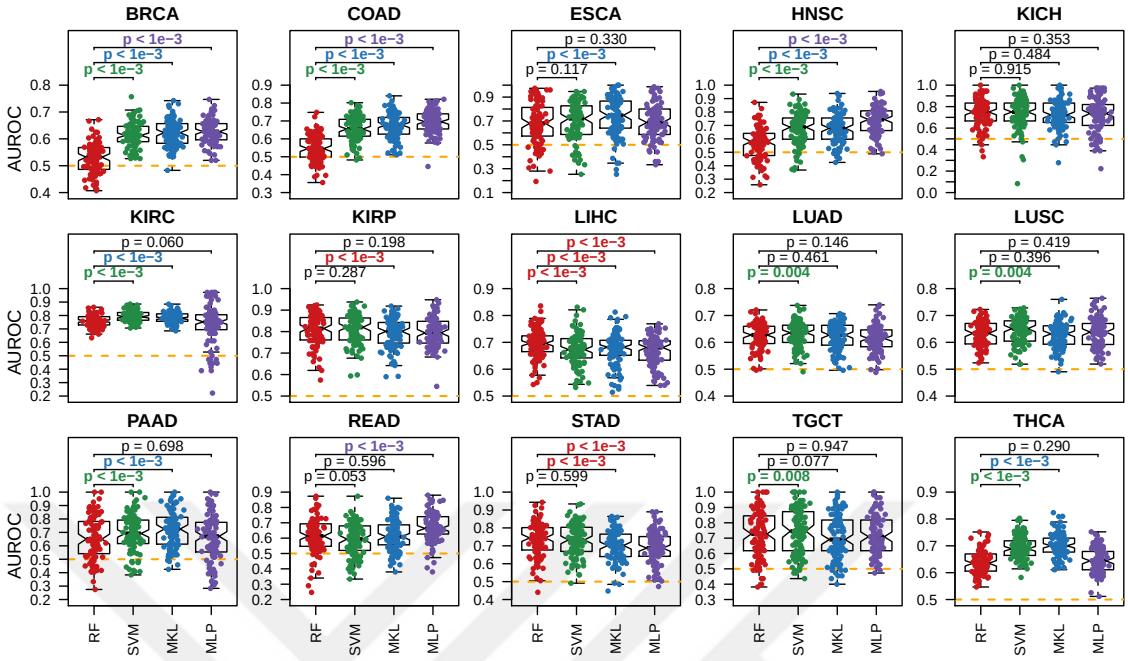

Figure 4.1: Predictive performances of RF, SVM, MKL, and MLP on 15 datasets constructed from TCGA cohorts

against RF to evaluate if the difference in the mean predictive performance between them is statistically significant. *P*-value is used to quantify the amount of difference among RF and each of the other three methods separately.

4.3.1 First group of the experiments

On 15 datasets created from TCGA cancer types for the first group of the experiments, E1, to predict the stages of cancers, we measured and compared the predictive performance of RF, SVM, MKL, and MLP on the Hallmark gene sets.

The predictive performance comparison of four algorithms for the task of separating the stages of cancer using the corresponding gene expression profiles for each cohort individually have been shown in Figure 4.1. The baseline performance is shown as a dotted line on 0.5 AUROC value, and we observed that median performance of all four classifiers are superior to the baseline performance. Therefore, this is a significant reason to conclude that meaningful and valuable information can be extracted from gene expression profiles of primary tumours.

P-values determine whether there is a significant difference between the results

of these three algorithms SVM, MKL, and MLP with RF algorithm separately. In Figure 4.1, *P*-value results are shown in different colors and each of these **red**, **green**, **blue**, and **purple** colors indicate that RF, SVM, MKL, and MLP have significantly better performance, respectively. Also, black color means that there is not a significant difference among the two algorithms.

According to Figure 4.1, we observed that MLP method significantly outperformed the RF method on 4 out of 15 datasets (i.e, `BRCA`, `COAD`, `HNSC`, and `READ`), whereas RF method outperformed the MLP method on 2 cohorts, (i.e, `LIHC` and `STAD`). In 9 out of 15 cohorts there is not significant difference between MLP and RF methods. MLP method significantly improved the value of predictive performance by 10.05% on `BRCA`, 14.93% on `COAD`, 16.21% on `HNSC`, 5.78% on `READ`, whereas the greatest drop in performance was 4.77% on `STAD`.

Table 4.3 indicates the average AUROC values obtained over 100 replications on all datasets for the first group of the experiments. Table 4.3 shows that MLP method has better average AUROC values over 100 replications, compared to RF method on 9 out of 15 cohorts which are `BRCA`, `COAD`, `ESCA`, `HNSC`, `KIRC`, `LUSC`, `READ`, `THCA`, and `PAAD`.

Table 4.3 indicates that MLP method obtained greater average AUROC value over 100 replications than SVM method on 5 out of 15 cohorts. These 5 cohorts are `BRCA`, `COAD`, `ESCA`, `HNSC`, and `READ`. We also observed that MLP method has better average AUROC value over 100 replications on 7 out of 15 cohorts in comparison with the MKL method. These 7 cohorts are `BRCA`, `COAD`, `HNSC`, `KIRP`, `LUSC`, `READ`, and `TGCT`.

4.3.2 Second group of the experiments

In the second group of the experiments, E2, we thoroughly compared the predictive performance of RF, SVM, MKL, and MLP on 18 different datasets from TCGA project using Hallmark gene sets collection.

Figure 4.2 shows the comparison of the predictive performance between RF, SVM, MKL, and MLP algorithms on 18 cohorts for cancer stages prediction task on

Table 4.3: Average AUROC of all algorithms over 100 replications on different cohorts in E1

Cohort	RF	SVM	MKL	MLP
BRCA	0.5275	0.61646	0.6217	**0.62790**
COAD	0.5473	0.6572	0.6716	**0.6966**
ESCA	0.6747	0.6933	**0.7349**	0.6961
HNSC	0.5696	0.6733	0.6788	**0.73173**
KICH	**0.7378**	0.7364	0.7280	0.7180
KIRC	0.7580	**0.7941**	0.7814	0.7621
KIRP	0.8079	**0.8118**	0.7933	0.7955
LIHC	**0.6949**	0.6729	0.6746	0.6673
LUAD	0.6234	**0.6294**	0.6212	0.6141
LUSC	0.6305	**0.6430**	0.6271	0.6361
PAAD	0.6569	0.6914	**0.7235**	0.6664
READ	0.6133	0.62	**0.6724**	0.6711
STAD	**0.7364**	0.7333	0.7021	0.6886
THCA	0.6397	0.6920	**0.7022**	0.6464
TGCT	0.7219	**0.7425**	0.7067	0.7206
Average	0.6668	0.6920	0.6925	0.6900

gene expression profiles of primary tumours. In the second group of experiments, we also observed that the median performance of all classification models are greater than 0.5 AUROC value, which is the baseline performance value and defined as dashed line.

Similar to the first group of the experiments, in the second group of the experiments, P-value is also used to identify the statistically significant differences between the performance of SVM, MKL, and MLP algorithms and the performance of the RF algorithm. Figure 4.2 shows that MLP method has significantly better performance than RF method on 6 out of 18 cohorts whose P-values are shown in **purple**.

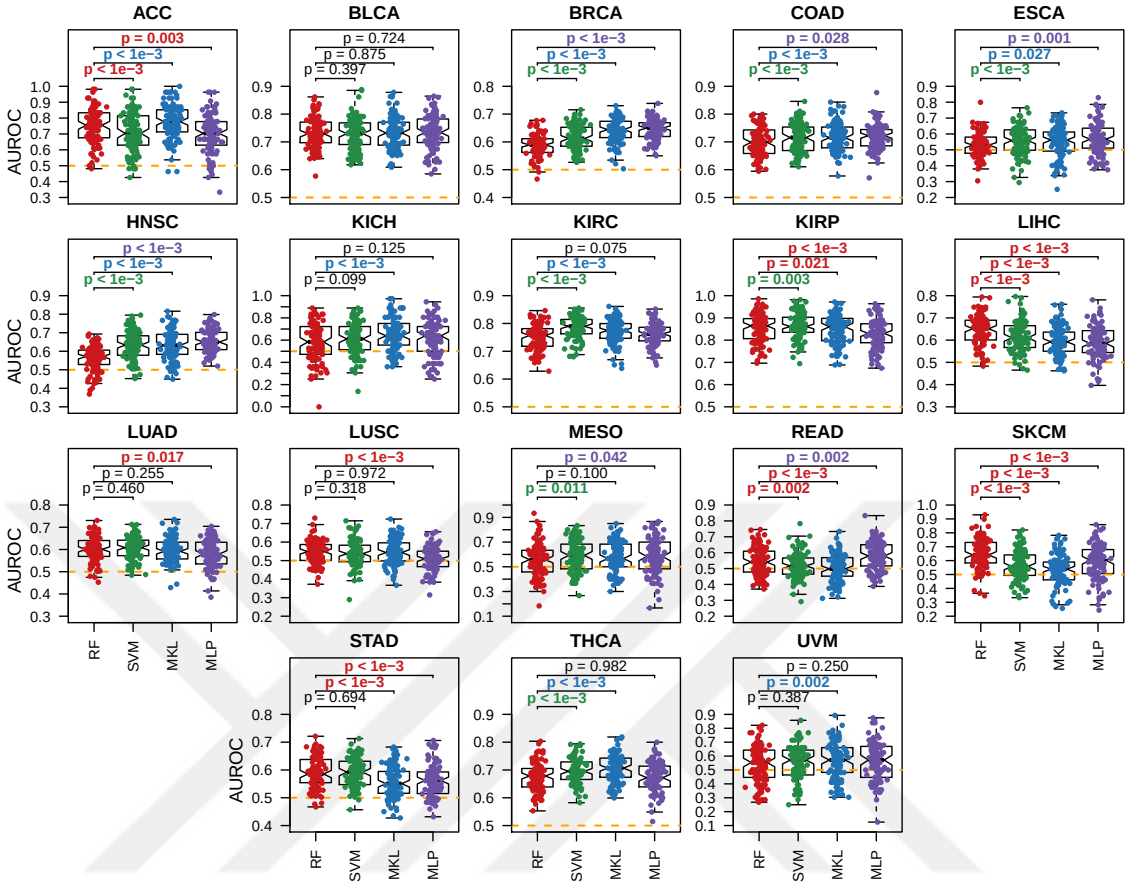

Figure 4.2: Predictive performances of RF, SVM, MKL, and MLP on 18 datasets constructed from TCGA cohorts

These six cohorts include BLCA, BRCA, COAD, ESCA, HNSC, and MESO. MLP improved the value of predictive performance by 5.65% on BRCA, 3.94% on READ, 4.23% on ESCA, 9.15% on HNSC, 3.95% on MESO, whereas the greatest drop in performance was 5.9% on SKCM.

The average AUROC values acquired through 100 replications on 18 datasets for the second group of the experiments can be found in Table 4.4.

Table 4.4 shows that MLP approach has better average AUROC values over 100 replications, compared to RF on 11 out of 18 cohorts, which are BLCA, BRCA, COAD, ESCA, HNSC, MESO, KICH, KIRC, READ, THCA, and UVM.

In Table 4.4, we see that our MLP algorithm obtained better average AUROC over 100 replications compared to SVM algorithm on 10 out of 18 datasets, which are BLCA, BRCA, COAD, ESCA, HNSC, MESO, READ, SKCM, UVM, and KICH. We also observed

Table 4.4: Average AUROC of all algorithms over 100 replications for different cohorts in E2

Cohort	**RF**	**SVM**	**MKL**	**MLP**
ACC	0.7492	0.7146	**0.7748**	0.6967
BLCA	0.7320	0.7289	0.7315	**0.7346**
BRCA	0.5866	0.6159	0.6411	**0.6430**
COAD	0.7005	0.7111	**0.7191**	0.7171
ESCA	0.5312	0.5577	0.5485	**0.5735**
HNSC	0.5626	0.6307	0.6335	**0.6541**
KICH	0.5844	0.6082	**0.6497**	0.6189
KIRC	0.75094	**0.7853**	0.7707	0.7619
KIRP	0.8552	**0.8651**	0.8473	0.8285
LIHC	**0.6454**	0.6170	0.5963	0.5903
LUAD	0.6024	**0.6063**	0.5957	0.5804
LUSC	0.5459	0.5396	**0.5462**	0.5089
MESO	0.5512	0.5860	0.5763	**0.5907**
READ	0.5400	0.5173	0.5031	**0.5794**
SKCM	**0.6491**	0.5668	0.5324	0.5902
STAD	**0.5904**	0.5886	0.5562	0.5602
THCA	0.6749	0.6977	**0.7081**	0.6751
UVM	0.5383	0.5468	**0.5718**	0.5605
Average	0.6328	0.6381	0.6390	0.6369

that our MLP algorithm has improved the value of average AUROC through 100 replications compare to MKL algorithm in 8 cohorts out of 18, which are `BLCA`, `BRCA`, `ESCA`, `HNSC`, `MESO`, `READ`, `SKCM`, and `STAD`.

In the second group of the experiments, based on Table 4.4, the average AUROC of MLP algorithm outperformed the other algorithms on 6 out of 18 cohorts. MKL, SVM, and RF algorithms outperformed the other algorithms separately on 6, 3, and

3 out of 18 cohorts, respectively.

Total number of gene expression features that were used by RF, SVM, MKL, and MLP algorithms is 4,357 for the genes that exist in the Hallmark gene set collection.

Figure 4.1 shows that in the first group of experiments SVM, MKL, and MLP algorithms have better performance against RF algorithm. However, integrating kernels in different forms by MKL and MLP algorithms lead to use a few numbers of gene expression features even less than 4,357 input features. Therefore, MKL and MLP algorithms trained the classifiers by discarding the ineffective and redundant gene sets. Knowledge extraction capability and interpretability of all algorithms are not same, because RF and SVM algorithms did not use kernels as inputs to train final classification model. RF and SVM did not integrate pathways/gene sets information in their original formulation. However, integrating kernels in different forms by MKL and MLP methods makes them more informative to specify the related and effective gene sets for the cancer stage prediction task.

4.3.3 Biological mechanisms determined by MLP method

To show the biological relevance of our MLP algorithm, we checked the pathways/gene sets chosen on all datasets. This process helped us to analyze the capability of our MLP algorithm to determine related gene sets based on weights of input kernels calculated during training. We counted the number of replications for which the kernel weight for each cohort and gene set pair is nonzero. For every gene set, we decided to include it in the final classifier if the corresponding kernel weight is nonzero.

Biological mechanisms determined by our MLP in the first group of experiments

We repeated the training procedure 100 times. In each cohort, some gene sets have been selected by our MLP algorithm with higher frequencies. The selection frequencies across 100 replication for the first group of experiments, E1, have been shown in Figure 4.3. The rows and columns of this selection frequencies matrix are the list of gene sets in the Hallmark collection and the list of cohorts in, E1,

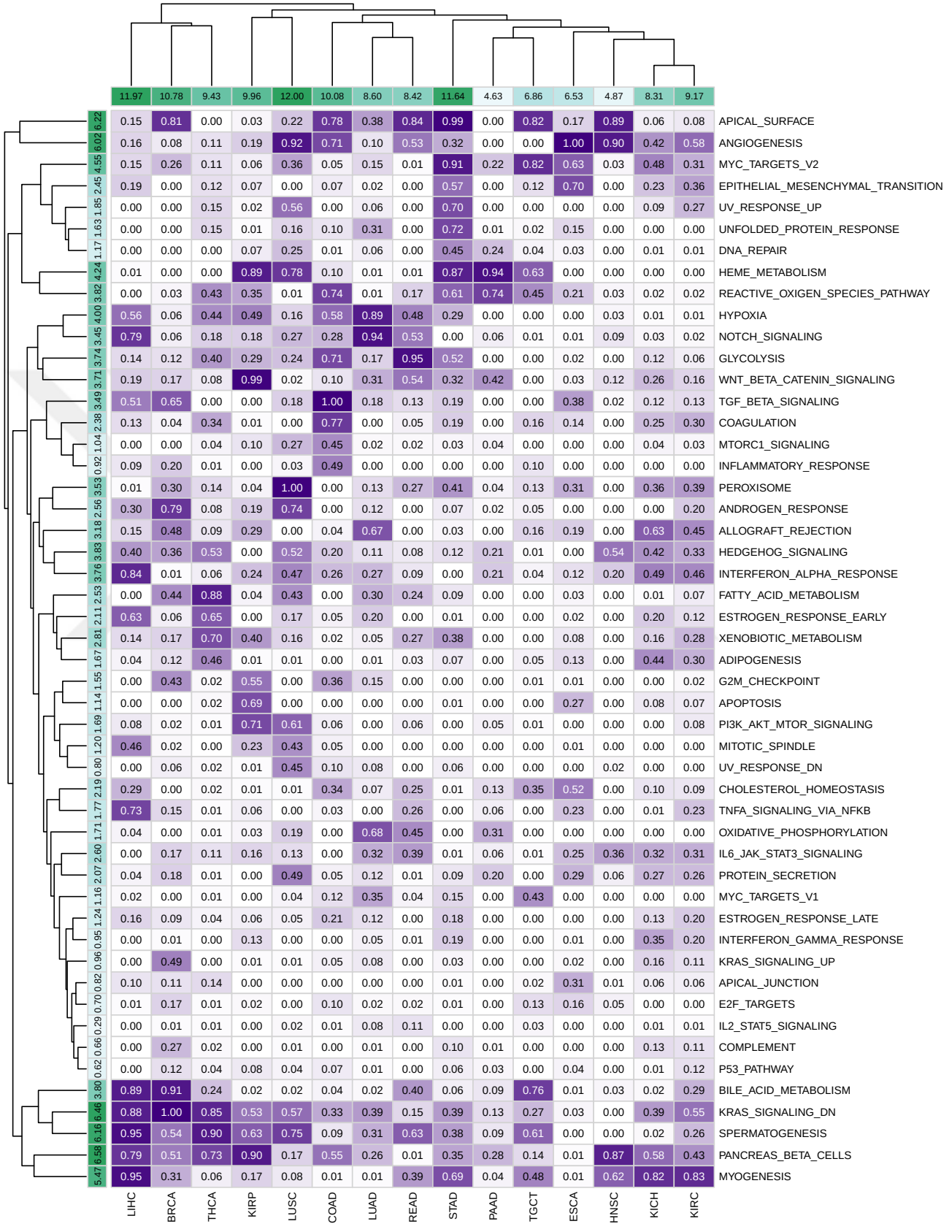

Figure 4.3: The selection frequencies of 50 gene sets in the Hallmark collection for 15 datasets in the first group of experiments

respectively. We used hierarchical clustering algorithm with Euclidean distance and full linkage functions to cluster the rows and columns of the selection frequencies matrix. Summation of each row indicates gene sets that are mostly chosen over all of the cohorts. Summation of each column indicates cohorts that used larger number of gene sets on the average.

There are 50 number of gene sets in the Hallmark collection. The selection frequencies of these gene sets across 100 replications for all cohorts in the first group of experiments have been shown in Figure 4.3. In `BRCA`, `COAD`, `LIHC`, `LUSC`, and `STAD` cohorts our MLP used more than 10 gene sets on the average to separate the stages of cancers from each other, whereas, in `ESCA`, `TGCT`, `HNSC`, and `PAAD` cohorts, MLP used less than 7 gene sets on the average to separate the cancer stages from each other. However, among the cohorts that MLP used less number of gene sets for classification, `HNSC` cohort has significantly improved prediction performance in comparison with RF, SVM, and MKL methods. Therefore, we note that separating the stages of cancers is more difficult in some cohorts.

To determine uninformative and informative gene sets for predicting the stages of cancer, we looked at the row sum of the selected frequencies. For instance, some of the gene sets which correspond to the formation of cancer in the early-stage were employed in more than 5 out of 15 cohorts and these gene sets are `ANGIOGENESIS`, `KRAS_SIGNALING_DN`, `MYOGENESIS`, and `SPERMATOGENESIS` (Bergers and Benjamin, 2003).

Each gene set carries information about a particular biological process. In the Hallmark collection, there are some gene sets that are related to metabolism such as `BILE_ACID_METABOLISM` (Chiang, 2013), `HEME_METABOLISM` (Liberzon et al., 2015), and `PANCREAS_BETA_CELLS` (Rashidi et al., 2009), and, in the final classifier, our MLP has selected these gene sets in more than 3 out of 15 cohorts on the average. `LIHC`, `PAAD`, and `THCA` diseases correspond to liver, pancreas, and thyroid gland organs. The most important role of these organs in the body is keeping the metabolism of body under control. Note that the tissues of `LIHC`, `PAAD`, and `THCA` cohorts carry out metabolic related functions. We noticed that, in most of the replications, our MLP selected the metabolism related gene sets for `LIHC`, `PAAD`, and `THCA` cohorts

with higher frequencies.

Epithelial cells are widespread in many organs such as breast, colon, lung, and stomach, and their corresponding disease abbreviations are `BRCA`, `COAD`, `LUAD`, and `STAD`, respectively. Figure 4.3 shows that, for these cohorts, our MLP selected the `APICAL_SURFACE` and `HYPOXIA` gene sets with higher frequencies.

In Hallmark collection, there are diverse gene sets that have important effect on epithelial cells, functions such as `APICAL_SURFACE` (Kotha et al., 2015) and `HYPOXIA` (Zeitouni et al., 2016) gene sets. Our MLP selected these gene sets for more than 4 out of 15 cohorts.

Cell cycle-related gene sets control growth and division of the cell and they are not related to discriminating the stages of cancers from each other. We observed that in different cohorts MLP picked cell cycle gene sets, namely, `DNA_REPAIR`, `G2M_CHECKPOINT`, `E2F_TARGETS`, `MYC_TARGETS_V1`, and `MTORC1_SIGNALING` with lower frequencies. This is another meaningful achievement of our MLP.

Biological mechanisms determined by our MLP in the second group of experiments

Similarly, in the second group of experiments, E2, we repeated the training procedure 100 times, and the gene sets selection frequencies by our MLP for 18 cancer types have been shown in Figure 4.4.

Summation of columns in Figure 4.4 shows us, in some cohorts, our MLP used more gene sets to classify the primary tumours, e.g., `KIRC`, `READ`, and `STAD`. For these cohorts, in the final trained model, our MLP employed more than 11 out of 50 gene sets on the average, whereas there are some other cohorts that used fewer than 8 out of 50 gene sets on the average, e.g., `UVM` and `MESO`. Although the selection frequency of `MESO` cohort is low, it has significantly better prediction performance in comparison with RF, SVM, and MKL methods. Therefore, summation of columns show us, separating early-stage and late-stage cancers from each other is more difficult in some cohorts for our MLP algorithm.

Row sum in Figure 4.4 enables us to understand which gene sets are informative for separating the stages of cancer in the second group of experiments. Summation of

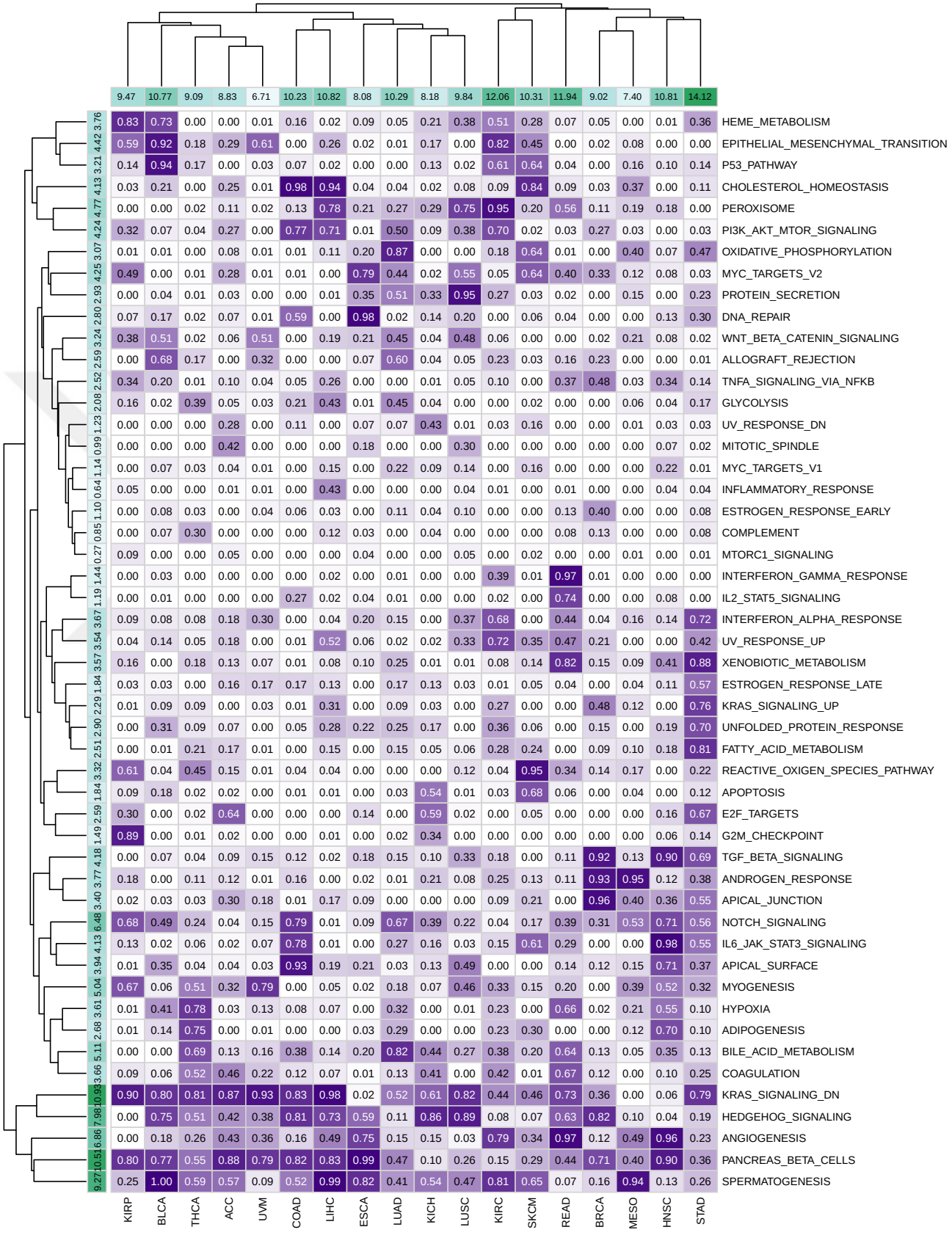

Figure 4.4: The selection frequencies of 50 gene sets in the Hallmark collection for 18 datasets in the second group of experiments

rows in Figure 4.4 shows us, our MLP selected some gene sets with higher frequencies, namely, KRAS_SIGNALING_DN and SPERMATOGENESIS that were used in the final model for more than 9 out of 18 datasets on the average, and also ANGIOGENESIS and MYOGENESIS were used in the final classifier for more than 5 out of 18 datasets on the average. These gene sets are correspond to formation of cancer in the early-stage (Bergers and Benjamin, 2003).

In the Hallmark collection, some gene sets carries information about metabolism related process. For instance, PANCREAS_BETA_CELLS (Rashidi et al., 2009) was used in more than 10 out of 18 datasets on the average and BILE_ACID_METABOLISM (Chiang, 2013) was used in more than 5 out of 18 datasets on the average. Our MLP selected these gene sets with higher frequencies. Moreover, some cohorts are related to organs that can control the metabolism, for instance, kidney, thyroid gland, and liver whose corresponding diseases are KIRC, THCA, and LIHC, respectively. We noticed, in these cohorts, our MLP selected the metabolism related gene sets with higher frequencies.

NOTCH_SIGNALING pathway is essential for cell proliferation, development, and differentiation of diverse organs such as head and neck, colon, lung, and kidney. The corresponding diseases for these organs are HNSC (Sun et al., 2014), COAD (Sikandar et al., 2010), LUAD (Galluzzo and Bocchetta, 2011), and KIRP (Xiao et al., 2017), respectively. Our MLP selected NOTCH_SIGNALING gene set with higher frequencies in more than 6 out of 18 cohorts, namely, HNSC, COAD, LUAD, and KIRP.

Similarly, HEDGEHOG_SIGNALING is an important signaling pathway and it is critical for cell differentiation in some organs, namely, breast, lung, and colon respectively and their corresponding diseases are BRCA (Souzaki et al., 2011), LUSC (Justilien et al., 2014), and COAD (Xu et al., 2012), respectively. These cohorts have chosen the HEDGEHOG_SIGNALING gene set with higher frequencies. Moreover, HEDGEHOG_SIGNALING gene set is selected with high frequencies in more than 7 out of 18 cohorts such as BRCA, LUSC, and COAD.

Epithelial cells separate internal and external parts of body and they exist in different parts of body such as colon, lung, and stomach whose corresponding disease abbreviation are COAD, LUAD, and STAD. By looking at Figure 4.4, we noticed

that for COAD, LUAD, and STAD cohorts our MLP selected the APICAL_SURFACE and HYPOXIA gene sets with high level of frequencies. APICAL_SURFACE (Kotha et al., 2015) and HYPOXIA (Zeitouni et al., 2016) gene sets have major impact on epithelial cells function and they were chosen for more than 3 out of 18 cohorts.

Among all genes that exist in Hallmark collection, our MLP did not choose gene sets that are unrelated to discriminating the stages of cancers from each other. Cell cycle-related gene sets' function is control growth and division of the cell and are not related to seperating cancer stages, namely, DNA_REPAIR, G2M_CHECKPOINT, E2F_TARGETS, MYC_TARGETS_V1, MYC_TARGETS_V2, and MTORC1_SIGNALING. Our MLP picked these gene sets with lower frequencies in 18 cohorts.

Chapter 5

CONCLUSION

Profiling tumour biopsies of cancer patients is essential to identify molecular characterization and biological processes that drive a tumour. Furthermore, profiles of tumours enable us to acquire information about diagnosis, formation, and development of cancer.

That is why, profiling tumour biopsies and determining active biological processes in a tumour has become an important problem. By understanding specific molecular mechanisms that cause cancer progression from early- to late-stage, we can employ efficient and appropriate therapeutic strategies to prevent this progression. In this thesis, we aimed to discriminate early- and late-stage of cancers from each other using gene expression profiles.

Machine learning algorithms have been applied in diverse areas such as cancer biology, retail, finance, manufacturing, medical diagnosis, telecommunications, speech recognition, robotics, healthcare, etc. We used supervised machine learning techniques for developing a binary classification model to classify cancer tumor stages based on their gene expression profiles. We applied an MLP algorithm to discriminate the stages of cancers from each other.

Our proposed MLP network (Figure 3.1) was tested on 20 different cancer types from TCGA with Hallmark gene set collection which is represented in Table 4.1 for two groups of the experiment. We compared the empirical results of MLP network on gene sets, against three widely used machine learning models, namely, RF, SVM, MKL. Like MKL algorithm, our MLP has the capability to integrate previous knowledge about pathways/gene sets into the model. Our algorithm is also able to derive cancer-specific importances of the input gene sets beside training a classifier for the classification problem.

From the result represented in Tables 4.3 and 4.4, it is observed that MLP model

was able to obtain better or comparable predictive performance values in terms of AUROC on different datasets compared to the other three standard machine learning algorithms. The major contribution of our empirical approach is implementing cancer stage classification and gene set selection conjointly by training ℓ_1-regularized MLP. Hence, we improved the performance of classifiers by removing noisy and irrelevant gene sets during training the classification model. Therefore, the measure of the robustness and the accuracy of the classifiers increased by eliminating redundant gene sets.

We used AUROC to measure the predictive power of our proposed MLP network in cancer stage prediction using gene expression profiles. We conducted two groups of experiments on 15 and 18 different TCGA cohorts. We noted that MLP on the Hallmark gene set was able to gain better or comparable predictive performance on the average against the competitive methods using fewer gene expression features in some of the datasets. We also found literature validation for the frequently selected gene sets by our MLP model in some of the datasets. Additionally, we observed that our MLP algorithm selected gene sets that are irrelevant for separating the stages of cancers with quite low frequencies.

There are various directions that we can consider for future research to enhance the predictive performance of the MLP algorithm on the cancer stage classification problem.

A possible future direction that would increase the accuracy of the classifier is model different types of cancer that have similar mechanisms together, which would increase the sample sizes for training, leading to more robust classifiers (i.e., multitask learning). As another future direction, we can address the cancer staging problem as a multiclass classification using MLP algorithm to classify the TNM stages of diverse cancers.

Although we cannot modify some of the hyperparameters in our MLP model such as number of output unit, sigmoid activation function in the last layer, and the number of hidden units before the output layer (P), but there exist some hyperparameters that can be modified to increase the accuracy of the proposed MLP model. To this aim, we can construct networks with more complex structures that

have more layers instead of one hidden layer. We can also solve our MLP with non-sharing weights in different input layers, which would increase the flexibility of our model with the cost of increasing the number of parameters, leading to slow convergence.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Agostinelli, F., Hoffman, M., Sadowski, P., and Baldi, P. (2014). Learning activation functions to improve deep neural networks. *arXiv*.

Al-Shargabi, B., Al-Shami, F., and Alkhawaldeh, R. (2019). Enhancing multi-layer perceptron for breast cancer prediction. *IJST*, 130:11–20.

Alpaydın, E. (2009). *Introduction to machine learning*. MIT press.

Azzawi, H., Hou, J., Xiang, Y., and Alanni, R. (2016). Lung cancer prediction from microarray data by gene expression programming. *IET*, 10(5):168–178.

Bergers, G. and Benjamin, L. E. (2003). Angiogenesis: tumorigenesis and the angiogenic switch. *Nat*, 3(6):401.

Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., and Raghava, G. P. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *SciRep*, 7:44997.

Bhalla, S., Kaur, H., Kaur, R., Sharma, S., and Raghava, G. (2018). Expression

based biomarkers and models to classify early and late stage samples of papillary thyroid carcinoma. *bioRxiv*, page 393975.

Breiman, L. (2001). Random forests. *Mach. Learn*, 45(1):5–32.

Broët, P., Kuznetsov, V. A., Bergh, J., Liu, E. T., and Miller, L. D. (2006). Identifying gene expression changes in breast cancer that distinguish early and late relapse among uncured patients. *Bioinformatics.*, 22(12):1477–1485.

Chang, J.-E., Lee, D.-S., Ban, S.-W., Oh, J., Jung, M. Y., Kim, S.-H., Park, S., Persaud, K., and Jheon, S. (2018). Analysis of volatile organic compounds in exhaled breath for lung cancer diagnosis using a sensor system. *SENSOR ACTUAT B-CHEM*, 255:800–807.

Chiang, J. Y. (2013). Bile acid metabolism and signaling. *Comprehensive Physiology*, 3(3):1191–1212.

Chollet, F. et al. (2015). *Keras*.

Chung, H., Lee, S. J., and Park, J. G. (2016). Deep neural network using trainable activation functions. In *IJCNN*, pages 348–352. IEEE.

Compton, C. C., Byrd, D. R., Garcia-Aguilar, J., Kurtzman, S. H., Olawaiye, A., and Washington, M. K. (2012). *AJCC cancer staging atlas: a companion to the seventh editions of the AJCC cancer staging manual and handbook*. SSBM.

Cortes, C. and Vapnik, V. (1995). Support-vector networks machine learning (pp. 237–297), vol. 20.

Dash, S. and Dash, A. (2014). A correlation based multilayer perceptron algorithm for cancer classification with gene-expression dataset. In *2014 14th International Conference on Hybrid Intelligent Systems*, pages 158–163. IEEE.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMCBioinformatics*, 7(1):3.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *ProcNatlAcad-SciUSA*, 103(15):5923–5928.

Galluzzo, P. and Bocchetta, M. (2011). Notch signaling in lung cancer. *Expert Rev. Anticancer Ther.*, 11(4):533–540.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Jmlr*, 12(Jul):2211–2268.

Gress, D. M., Edge, S. B., Greene, F. L., Washington, M. K., Asare, E. A., Brierley, J. D., Byrd, D. R., Compton, C. C., Jessup, J. M., Winchester, D. P., et al. (2017). Principles of cancer staging. *Springer*.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. ChHall/CRC.

Ibrahim, A. O., Shamsuddin, S. M., yahya Saleh, A., Abdelmaboud, A., and Ali, A. (2015). Intelligent multi-objective classifier for breast cancer diagnosis based on multilayer perceptron neural network and differential evolution. In *2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE)*, pages 422–427. ICCNEEE.

Jagga, Z. and Gupta, D. (2014). Classification models for clear cell renal carcinoma stage progression, based on tumor rnaseq expression trained supervised machine learning algorithms. In *BMC Proc*, volume 8, page S2. BioMed Central.

Justilien, V., Walsh, M. P., Ali, S. A., Thompson, E. A., Murray, N. R., and Fields, A. P. (2014). The prkci and sox2 oncogenes are coamplified and cooperate to activate hedgehog signaling in lung squamous cell carcinoma. *Cancer cell*, 25(2):139–151.

Karlik, B. and Olgac, A. V. (2011). Performance analysis of various activation functions in generalized mlp architectures of neural networks. *IJAIES*, 1(4):111–122.

Kaur, H., Bhalla, S., and Raghava, G. P. (2019). Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PloSone*, 14(9):e0221476.

Kotha, P. L., Sharma, P., Kolawole, A. O., Yan, R., Alghamri, M. S., Brockman, T. L., Gomez-Cambronero, J., and Excoffon, K. J. (2015). Adenovirus entry from the apical surface of polarized epithelia is facilitated by the host innate immune response. *PLoS Pathog.*, 11(3).

Kumar, R., Bhanti, P., Marwal, A., and Gaur, R. (2019). Gene expression-based supervised classification models for discriminating early-and late-stage prostate cancer. *Proc Natl Acad Sci India Sect B Biol Sci*, pages 1–25.

Lazebnik, Y. (2010). What are the hallmarks of cancer? *Nat Rev Cancer*, 10(4):232.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *CELL SYST*, 1(6):417–425.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Molecular cell biology 4th edition. *NCBI*.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull.Math.*, 5(4):115–133.

MOSEK ApS (2019). *MOSEK Optimization Suite Release Version 9.1.5.*

Nam, H., Chung, B. C., Kim, Y., Lee, K., and Lee, D. (2009). Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. *Bioinformatics*, 25(23):3151–3157.

Nasser, I. M. and Abu-Naser, S. S. (2019). Predicting tumor category using artificial neural networks.

National Cancer Institute (2005 (accessed october 30, 2019)). The cancer genome atlas. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer#types`.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E., and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, 22(16):2028–2036.

Plagianakos, V. P. and Vrahatis, M. N. (2000). Training neural networks with threshold activation functions and constrained integer weights. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5, pages 161–166. IEEE.

Rafii, F., Kbir, M. A., and Hassani, B. D. R. (2015). Mlp network for lung cancer presence prediction based on microarray data. In *2015 Third World Conference on Complex Systems (WCCS)*, pages 1–6. IEEE.

Rahimi, A. and Gönen, M. (2018). Discriminating early-and late-stage cancers using multiple kernel learning on gene sets. *Bioinformatics*, 34(13):i412–i421.

Rahimi, A. and Gönen, M. (2020). A multitask multiple kernel learning formulation for discriminating early-and late-stage cancers. *Bioinformatics*.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *arXiv*.

Rashidi, A., Kirkwood, T. B., and Shanley, D. P. (2009). Metabolic evolution suggests an explanation for the weakness of antioxidant defences in beta-cells. *Mech Ageing Dev*, 130(4):216–221.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol.Rev.*, 65(6):386.

Ruddon, R. W. (2007). *Cancer biology*. OUP.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Sikandar, S. S., Pate, K. T., Anderson, S., Dizon, D., Edwards, R. A., Waterman, M. L., and Lipkin, S. M. (2010). Notch signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. *Cancer Res.*, 70(4):1469–1478.

Singh, N. P., Bapi, R. S., and Vinod, P. (2018). Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. *COMPUT BIOL MED*, 100:92–99.

Sokouti, B., Haghipour, S., and Tabrizi, A. D. (2014). A framework for diagnosing cervical cancer disease based on feedforward mlp neural network and thinprep histopathological cell image features. *NEURAL COMPUT APPL*, 24(1):221–232.

Souzaki, M., Kubo, M., Kai, M., Kameda, C., Tanaka, H., Taguchi, T., Tanaka, M., Onishi, H., and Katano, M. (2011). Hedgehog signaling pathway mediates the progression of non-invasive breast cancer to invasive breast cancer. *CANCER SCI*, 102(2):373–381.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J MACH LEARN RES*, 15(1):1929–1958.

Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMCbioinformatics*, 9(1):319.

Sun, W., Gaykalova, D. A., Ochs, M. F., Mambo, E., Arnaoutakis, D., Liu, Y., Loyo, M., Agrawal, N., Howard, J., Li, R., et al. (2014). Activation of the notch pathway in head and neck cancer. *Cancer Res.*, 74(4):1091–1104.

Targonski, C. A., Shearer, C. A., Shealy, B. T., Smith, M. C., and Feltus, F. A. (2019). Uncovering biomarker genes with enriched classification potential from hallmark gene sets. *Sci Rep*, 9(1):1–10.

Thakur, A., Mishra, V., Jain, S. K., et al. (2011). Feed forward artificial neural network: tool for early detection of ovarian cancer. *SciPharm*, 79(3):493–506.

Vineis, P. and Wild, C. P. (2014). Global cancer patterns: causes and prevention. *Lancet*, 383(9916):549–557.

Wang, Z., Wang, Y., Xuan, J., Dong, Y., Bakay, M., Feng, Y., Clarke, R., and Hoffman, E. P. (2006). Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data. *Bioinformatics*, 22(6):755–761.

Xiao, W., Gao, Z., Duan, Y., Yuan, W., and Ke, Y. (2017). Notch signaling plays a crucial role in cancer stem-like cells maintaining stemness and mediating chemotaxis in renal cell carcinoma. *J Exp Clin Cancer Res*, 36(1):41.

Xu, M., Li, X., Liu, T., Leng, A., and Zhang, G. (2012). Prognostic value of hedgehog signaling pathway in patients with colon cancer. *Med. Oncol.*, 29(2):1010–1016.

Xu, Z., Jin, R., Yang, H., King, I., and Lyu, M. R. (2010). Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1175–1182. Citeseer.

Zeitouni, N. E., Chotikatum, S., von Köckritz-Blickwede, M., and Naim, H. Y. (2016). The impact of hypoxia on intestinal epithelial cell functions: consequences for invasion by bacterial pathogens. *Mol. Cell. Pediatr.*, 3(1):14.