DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

BIOMARKER IDENTIFICATION FOR DISCRIMINATION OF CANCER TYPES

by

Cem Buğra ALKAN

October, 2019

İZMİR

BIOMARKER IDENTIFICATION FOR DISCRIMINATION OF CANCER TYPES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Computer Engineering

Cem Buğra ALKAN

October, 2019
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**BIOMARKER IDENTIFICATION FOR DISCRIMINATION OF CANCER TYPES**" completed by **CEM BUĞRA ALKAN** under supervision of **ASST. PROF. DR. ZERRİN IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

....

Asst. Prof. Dr. Zerrin IŞIK

Supervisor

....

Asst. Prof. Dr. Özlem AKTAŞ

Jury Member

....

Asst. Prof. Dr. Özlem ERDAŞ ÇİÇEK

Jury Member

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Assistant Professor Dr. Zerrin IŞIK for her enlightening wisdom and patience.

I would like to thank my friends Bengisu, Selin, Salih, and Emre for reminding me to keep going.

I would also like to thank my grandfather Cemal DEMİRBAŞ for his support and belief.

Of course, I would like to thank my father, Özkan ALKAN, who always gives me a different perspective on my ideas and projects.

Last but not least, I would like to thank my dear mother, Tülin ALKAN for being on my side in all the important moments of my life.

Cem Buğra ALKAN

BIOMARKER IDENTIFICATION FOR DISCRIMINATION OF CANCER TYPES

ABSTRACT

RNA-sequencing data provides measurements of mRNA (messenger RNA) levels of genes based on tissue or blood samples. The critical changes in transcriptome can be observed more accurately by using RNA-sequencing data that eventually helps to understand different behavior of the disease. In this study, different feature selection methods and machine learning algorithms were examined for accurate discrimination of cancer types by using RNA-sequencing data which was obtained from blood samples.

In the analysis, six cancer types were compared with each other and healthy samples. Correlation coefficient and information gain analyses are applied as main feature selection methods. The selected genes are provided as the input of Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF) machine learning algorithms, that were evaluated by applying 10-fold cross-validation.

In the experimental results, machine learning algorithms achieved higher than 0.85 accuracies in the discrimination of hepatobiliary, lung, and pancreatic cancer types. When machine learning models are evaluated in terms of accuracy, RF and SVM were more successful than NB for many cases. A literature-based validation revealed that some of the genes used in classifiers might be promising biomarkers for discrimination of hepatobiliary and pancreatic cancers.

Keywords: Cancer detection, RNA-sequencing data, support vector machine, naïve Bayes, random forest

KANSER TÜRLERİNİ AYIRT EDEBİLMEK İÇİN BİYOİŞARETÇİ TANIMLAMASI

ÖZ

RNA sıralama verileri, doku veya kan örneklerine dayanarak mRNA gen seviyelerinin ölçümlerini sağlar. Transkriptomdaki kritik değişiklikler, RNA dizileme verisi ile daha iyi incelenerek hastalığın davranışını daha doğru şekilde gözlemlemeye yardımcı olur. Bu çalışmada, kan örneklerinden elde edilen RNA dizileme verileri kullanılarak kanser türlerinin doğru şekilde ayırt edilebilmesi için farklı özellik seçim yöntemleri ve makine öğrenme algoritmaları incelenmiştir.

Analizde altı kanser türü birbiriyle ve sağlıklı örneklerle karşılaştırılmıştır. Özellik seçim yöntemleri olarak korelasyon katsayısı ve bilgi kazanımı analizleri uygulanmıştır. Seçilen genler, 10 katlı çapraz doğrulama uygulanarak değerlendirilen Destek Vektör Makinesi (SVM), Naif Bayes (NB) ve Rastgele Orman (RF) makine öğrenme algoritmalarına girdi olarak verilmiştir.

Deney sonuçlarında, makine öğrenme algoritmaları, hepatobiliyer, akciğer ve pankreas kanseri tiplerinin ayırt edilmesinde 0,85 doğruluk elde etmiştir. Makine öğrenim modelleri doğruluk açısından değerlendirildiğinde, RF ve SVM'nin birçok durumda NB'den daha başarılı olduğu görülmüştür. Literatüre dayalı bir doğrulama, sınıflandırıcılarda kullanılan bazı genlerin, hepatobiliyer ve pankreas kanserlerinin ayırt edilmesinde ümit verici biyobelirteçler olabileceğini ortaya koymuştur.

Anahtar Kelimeler: Kanser tespiti, RNA sekanslama verisi, destek vektör makinesi, naif Bayes, rastgele orman

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Motivation

Cancer research is widely valued around the world due to constantly increasing disease rates. The most important outcome of cancer research is proven to be an early diagnosis. Treatment results are expected to be increased upon early detection of cancerous cells. The common procedure for early detection is heavily based on medical imaging systems, biopsy, and physical symptoms. Although these diagnosis techniques are very reliable and proven-over-time methods, there can be certain downfalls such as unnecessary amounts of exposure to radiation, high costs of different medical imaging modalities and time-consuming for medical staff, invasiveness of biopsy. In that sense, less invasive and more cost-effective modalities are needed to further investigate the genetic or epigenetic alterations in malignant cells.

The search for less invasive methods leads to a liquid biopsy which relies on biomarkers. A liquid biopsy requires bodily fluids such as blood, CSF (cerebrospinal fluid), the lymphatic fluid that are accessed far less invasively (Perakis & Speicher, 2017). Biomarkers are limited in numbers and need to be perfected for more accurate outcomes. Researching and perfecting these biomarkers require the correct computational methods. Recent studies focus on different computational modalities and their interactions on biological data obtained from gene microarrays (Abdel Samee, Solouma, & Kadah, 2012).

RNA-sequencing is a relatively new experiment that can take the place of microarray technology in the future among many other gene expression technologies. There are many resources that can be used to produce gene samples for RNA-sequencing such as tissue and blood. RNA-sequencing can help to differentiate between gene expressions of normal and treated cells. The main principle of RNA-sequencing is high-throughput sequencing while microarrays use hybridization. RNA-sequencing has more technical advantages compared to microarrays resulting in a higher capacity for gene expressions, less background noise in the image, a need for less RNA (ribonucleic acid) sample and lower cost (“RNA-seq,” n.d.). Such experiments enable scientists to compare normal and disease genes based on transcriptome; mRNA, tRNA (transfer RNA), rRNA (ribosomal RNA). Understanding the changes in transcriptome provides information regarding the function of genes, therefore it helps to recognize different behavior of cells.

1.2 Problem Definition

The understanding of the cancer-causing genes is still a challenging problem. The discrimination of cancer types without applying biopsy is still not practical in clinic applications. If some marker proteins can be identified in blood samples instead of using tissue samples, the diagnostic time and cost would be decreased dramatically.

1.3 Contribution

This thesis aims to evaluate different feature selection and machine learning methods to discriminate different cancer types by using RNA-sequencing data obtained from blood samples of patients. The found genes, which can effectively discriminate two types of cancer, would be suggested as diagnostic biomarkers for further clinical studies. The original data set was taken from the study of Zhang et al. which applied an mRMR (minimum redundancy maximum relevance) for feature selection and SVM for modeling (Zhang et al., 2017). In this study two feature selection methods, which are less complex than mRMR, were used. After that, the genes selected as features are fed to three different machine learning algorithms and results were compared.

1.4 Organization of Thesis

This thesis consists of five chapters organized as follows:

In Chapter 2, I provide detailed background information and a literature review to describe some essential concepts such as biomarkers, RNA-sequencing data analysis.

In Chapter 3, I introduce our general road map in six main sections including the RNA-sequencing data, data pre-processing, normalization, feature selection methods, machine learning algorithms, and evaluation metrics used to extract information from this data.

In Chapter 4, I present the results and the biological interpretation of these results to compare different machine learning algorithms and different feature selection methods.

In Chapter 5, I conclude the study and offer future work.

CHAPTER TWO
LITERATURE REVIEW

2.1 Biomarkers

A biomarker is described as any biological molecule which is found in body fluids or tissues, can be used to distinguish a disease by giving a normal or abnormal sign according to National Cancer Institute (Henry & Hayes, 2012). When diagnosing cancer by pathological techniques, a sample has to be taken from the suspected tissue and has to be examined. However, when the case comes to the stage of sample-taking, most of the time cancer has already grown enough to cause the tissue to malfunction (Srinivas, Kramer, & Srivastava, 2001). Biomarkers can help to earlier diagnose the disease before it causes any defects (Srinivas et al., 2001). Proteins, protein-metabolite conjugates, small-molecule metabolites, nucleotides, and lipids can be examples for those molecules (Srivastava & Creek, 2019). In cancer researches, those biomarkers can be produced by the cancer cell or produced by the body against the cancer cells (Srivastava & Creek, 2019). In this study, the mRNA measurements are used as a biomarker.

2.2 RNA-Sequencing Data

The genetic code of an organism is collected in DNA (deoxyribonucleic acid) as a huge collection of genes and this coding data is transcribed into RNA to synthase proteins. RNA is a molecule that has a vital duty in diverse biological processes. The set of all RNA molecules in a cell is called the transcriptome. A deep view of the transcriptome can be obtained by RNA-sequencing (Byron, Van Keuren-Jensen, Engelthaler, Carpten, & Craig, 2016). Also, any next-generation sequencing technique which is used to study RNA technique is named as RNA-sequencing (Chu & Corey, 2012). Diverse areas related to human health include the application of RNA-based measurements which consist of diagnosis of diseases, prognosis and therapeutic selection (Byron et al., 2016). Tools to determine the presence and amount of RNA molecules in biological samples can be grouped under RNA-sequencing.

Mutated cells act differently than normal cells. To understand the different mechanism of them the gene expression causing those differences have to be found and examined. To do that first the structure of the genes has to be understood. Each cell contains chromosomes, and all chromosomes are formed by genes in them. Some of those genes more active than others. Which genes are active and how much they are transcribed can be answered by the high throughput sequencing data. RNA-seq can be used to measure the gene activity of normal and mutated cells. Then those two can be compared to figure out what is the difference between them. To do that first, the sequencing library has to be prepared. Then, sequencing has to be done. And finally, data analysis will be made.

2.2.1 Preparing RNA-seq Library

In the first step, RNA is isolated from the cell. Then, since the reading capacity of the sequencing device is limited the RNA has to be broken down to small fragments. The RNA will be converted to double-stranded DNA since it is more stable than RNA and easily modified and amplified. After this, the sequencing adaptor will be added to the fragments. With the adaptors, the sequencing device recognizes the fragments. The fragments with the adapter will be PCR (Polymerase chain reaction) amplified. With the quality control step which includes library concentration and fragments lengths verification the library preparation finishes (“Whole Transcriptome and mRNA Sequencing Guide,” n.d.)

2.2.2 Sequencing The Library

The DNA fragments wanted to be sequenced are put on the chosen sequencer. The recently used one is Illumina which labels the nucleotide with fluorescent (Kukurba & Montgomery, 2015). The fluorescent probes of the device attached to each nucleotide then take a picture to map. This process continues until all bases are read. End of this process the raw data has been created. The data has to be filtered by removing the garbage reads and aligning the high-quality reads to the genome(Kumar et al., 2012, p.). The choice of reference genome affects the complexity of the alignment process (“RNA-seq,” n.d.) Genome is split into small fragments then the index and the location of each fragment are created. Also, the read is split into small fragments. Then, read fragments are matched to the genome fragments. The matched fragments determine the location on the genome. Even if the reference genome is not matching fully with the fragments, by breaking them into small pieces a partial match can be made. After the matches for genes will be counted and this gives a matrix with genes and number of matches for each sample cell.

2.2.3 Analyzing The Data

This part is the last step of RNA sequencing. The separation between mutated and normal cells can be done here through analyses of the data obtained. This profiling gives high-resolution of the entire transcription (Kukurba & Montgomery, 2015).

2.2.4 Applications of RNA-Seq

There are several studies focusing on cancer detection with the usage of RNA-sequencing. In one study, RNA-sequencing was used to identify biomarkers from different tissues for the cancer types which lead metastases commonly. In that study, CUP (colorectum, kidney, liver, lung, ovary, pancreas, prostate, and stomach) metastasis has been studied with 17471 transcripts from 3244 samples and 26 different tissue types taken from International Cancer Genome Consortium and The Cancer Genome Atlas. The researchers used 10-fold cross-validation on the log-transformed and quantile normalized data. The overall accuracy of the algorithm was 90.5% and generated signatures for the top eight cancer types causing CUP (Wei, Shi, Jiang, Kumar-Sinha, & Chinnaiyan, 2014).

Another study integrated RNA-sequencing, PPI (protein-protein interaction) data, and RPPA (reverse phase protein array) data to detect the survival times of cancer patients and obtain prognostic biomarkers. To identify the biomarkers random walk-based algorithm was used. After that, with selected biomarkers gene expression measurement a classifier was trained to predict the survival times of patients. On average the accuracy rate of this method was from 66% to 78% for three datasets (Isik & Ercan, 2017).

One study focuses on evaluating the performance of four clustering algorithms and twelve distance measures commonly used for gene expression analysis with 15 different RNA-sequencing datasets. The study results show clustering cancer samples on gene quantification can be useful. However, the usage of non-specific filtering causes superior results. Also, these researchers suggest using log-transformation on the data before clustering (Jaskowiak, Costa, & Campello, 2018).

In another research, the RNA-sequencing data, which was obtained from kidney biopsies, was used to understand kidney rejections caused by T-cells. The SVM and RF algorithms were trained with kidneys with stable function and T-cell-mediated rejection data (Liu, Tseng, Wang, Huang, & Randhawa, 2019).

Moreover, a classifier was developed with the help of RNA sequencing data to identify the UIP (usual interstitial pneumonia) pattern to predict idiopathic pulmonary fibrosis. The authors of this study mention even though the limited sample size, disease heterogeneity and technical batch effects they developed a model with 70% sensitivity and 88% specificity (Choi et al., 2018).

Also, a machine learning model was built with RNA-Seq to identify differentially expressed transcripts linked with prostate cancer. In this study, the authors mention that prostate cancer has a high number of unexplained variables and says it is one of the most common cancer types in the world. For that reason, finding biomarkers for this disease can be promising to improve the survival rates of the high-risk patient population. They used 106 prostate cancer samples with various states of disease. 44 transcripts related to the different stages of the disease were detected (Singireddy et al., 2015).

Furthermore, another study used RNA-sequencing data was used to predict the cancer types. Here, the RNA-Seq data obtained from TCGA (The Cancer Genome Atlas) used in the with five machine learning algorithms which are DT (decision tree), kNN (k nearest neighbor), linear SVM, poly SVM and ANN (artificial neural network). They are compared according to training time, precision, recall, F1-score. Among them, with 95.8% accuracy linear SVM was the best (Y.-H. Hsu & Si, 2018).

In addition, a study focused to distinguish cancer patients and healthy persons with the help of deep learning. They used deep learning by ensemble approach which is a method aggregates the results of different algorithms and decides by the collective result of them. First, five different classification algorithms were used and their outputs fed to a deep learning algorithm (Xiao, Wu, Lin, & Zhao, 2018).

Lastly, another study applied the minimum redundancy and maximum relevance feature selection method and SVM model to distinguish seven sample types from each other; they obtained the highest 75% accuracy for the discrimination of cancer types with different specificity and sensitivity scores (Zhang et al., 2017). This project took as an example of this thesis and higher accuracy, sensitivity, and specificity were aimed.

CHAPTER THREE
METHOD AND MATERIALS

In this study, we used two different feature selection methods with three different machine learning algorithms to discriminate seven groups which are six subtypes of cancer and a healthy group. The processed performed in this project are shown in Figure 3.1. After the cleaning phase of genes, the feature selection methods created specific subsets of genes based on the various thresholds. Then, the selected genes were given to the machine learning algorithms. By considering significant results the disease-causing genes were detected. The details of these steps are explained in this chapter.

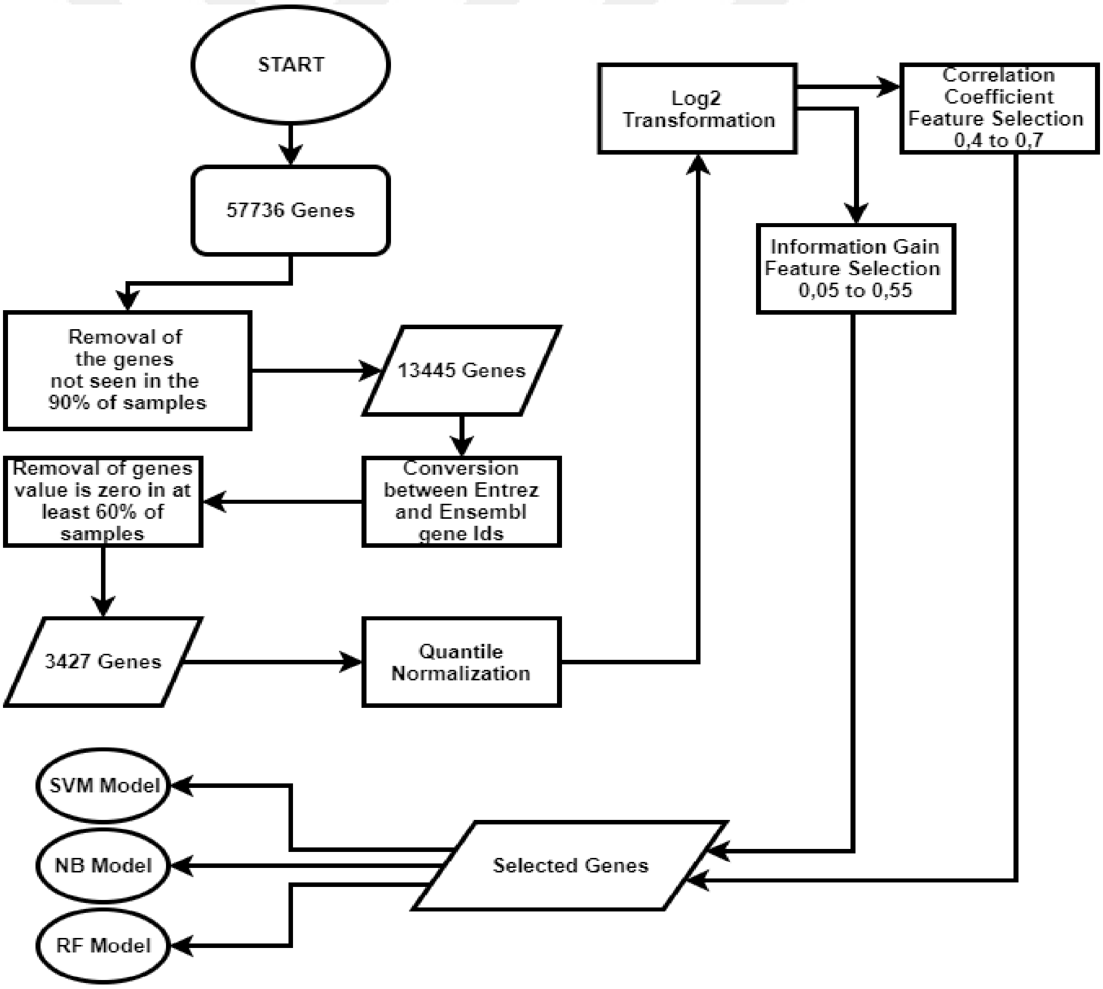
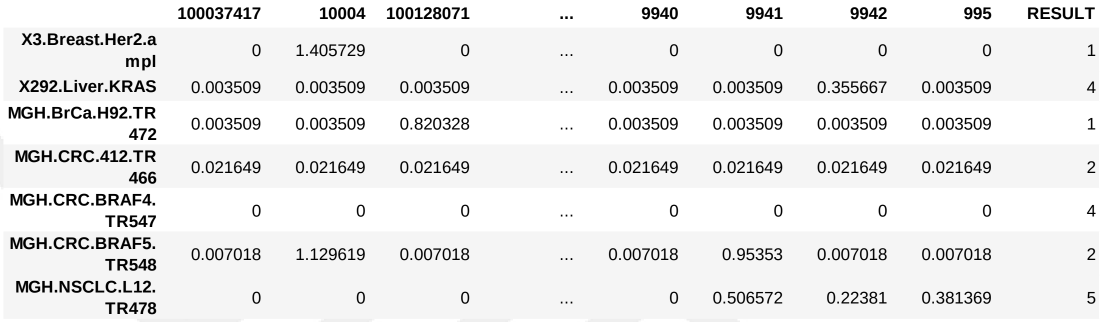

START
57736 Genes
Removal of the genes not seen in the 90% of samples
13445 Genes
Conversion between Entrez and Ensembl gene Ids
Removal of genes value is zero in at least 60% of samples
3427 Genes
Quantile Normalization
Log2 Transformation
Correlation Coefficient Feature Selection 0,4 to 0,7
Information Gain Feature Selection 0,05 to 0,55
Selected Genes
SVM Model
NB Model
RF Model

Figure 3.1 Flowchart of the process

3.1 Dataset and Preprocessing

The data was downloaded from GEO (Gene Expression Omnibus) with the access number of GSE68086 and including the gene expression of blood samples from 285 individuals. 39 samples belong to breast cancer, 42 samples belong to colorectal cancer, 40 samples belong to glioblastoma, 14 samples belong to hepatobiliary cancer, 60 samples belong to lung cancer, 35 samples belong to pancreatic cancer, and 55 samples belong to healthy controls (Zhang et al., 2017).



	100037417	10004	100128071	...	9940	9941	9942	995	RESULT
X3.Breast.Her2.ampl	0	1.405729	0	...	0	0	0	0	1
X292.Liver.KRAS	0.003509	0.003509	0.003509	...	0.003509	0.003509	0.355667	0.003509	4
MGH.BrCa.H92.TR472	0.003509	0.003509	0.820328	...	0.003509	0.003509	0.003509	0.003509	1
MGH.CRC.412.TR466	0.021649	0.021649	0.021649	...	0.021649	0.021649	0.021649	0.021649	2
MGH.CRC.BRAF4.TR547	0	0	0	...	0	0	0	0	4
MGH.CRC.BRAF5.TR548	0.007018	1.129619	0.007018	...	0.007018	0.95353	0.007018	0.007018	2
MGH.NSCLC.L12.TR478	0	0	0	...	0	0.506572	0.22381	0.381369	5

Figure 3.2 Example of initial data

In the preprocessing phase, some genes were eliminated which are the ones not available in 90% of the individuals. In the beginning, the total number of genes was 57736, after this elimination, the number is reduced to 13445. The Ensembl gene identifiers were translated to Entrez gene identifiers. Ensembl gene identification is an annotation system annotating different vertebrates in various genome projects (Aken et al., 2016). Also, the National Center for Biotechnology Information has another system that is called Entrez gene identifiers for a reliable annotation of gene names (Maglott, Ostell, Pruitt, & Tatusova, 2005). If the mRNA expression of a gene is measured as zero in the 60% of patient samples, this gene is also removed. At the end of those processes, 3427 genes remained for further analysis. In Figure 3.2, a small example of the cleaned data is shown. The column names show the gene identifiers, the result column represents the assigned cancer type of the sample, the index value in the first column is the given name for each patient sample. After gene cleaning, the size of the original data is 285 rows and 3428 columns.

3.2 Normalization

The large-scale experiments always come with a downfall of variations due to various reasons that ultimately affects the gene expression analysis. Minimizing many variations to obtain a more accurate comparison of different data samples is called normalization. Quantile normalization became standard for data analysis of high-throughput data to remove unwanted technical variations (Hicks & Irizarry, 2014). Although quantile normalization was developed for gene expression microarrays, currently it is used for RNA-sequencing and other data types (Hicks & Irizarry, 2014). Hence the quantile normalization is found to be appropriate for this study.

Quantile normalization aims for the statistical properties of two or more distributions to be exactly the same. To do that, each distribution is set to the mean value. This ensures that the new lowest value becomes the mean value of all the lowest values. In the same way, the highest and middle values are also set to their mean values. With this technique, the maximums align among themselves and minimums align among themselves. This method stretches all distributions together and the order of features in their own distributions never change but the distributions become in the same length. Therefore, quantile normalization ensures that gene expression levels for each sample are the same while gene orders are maintained. After normalization, a logarithm base 2 transformation was applied.

3.3 Feature Selection

I used the correlation coefficient and information-gain feature selection methods in the study. Even though the PCA (Principal Component Analysis) might give better results than the correlation coefficient or information gain, it was not used. Because PCA takes currently available features and creates more effective features combinations from them, however, this process cannot be reversible. The final features are linear weighted combinations of single features; hence the individual contribution of each gene cannot be obtained, eventually, the singleton biomarkers could not be driven. Due to all these reasons, the PCA was not used as the feature reduction method.

3.3.1 Correlation Coefficient

The correlation coefficient measures the relationship between the dependent variable (in this case cancer type) and the independent variable (each individual gene). If two variables are linearly dependent, their correlation is close -1 or 1 and they become strongly correlated (Hsu & Hsieh, 2010; Yu & Liu, n.d.). However, if the value is 0, then they are not related at all (H.-H. Hsu & Hsieh, 2010). Using the correlation coefficient as the feature selection method helps to remove non-related or uncorrelated features (Yu & Liu, n.d.). The correlation coefficient *r* is calculated in Equation 3.1.

$$r = \frac{n(\sum xy\ ) - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \tag{3.1}$$

where *n* is the number of variables, *x* is the independent variable, *y* is the dependent variable. If the small change on the independent variable causes a serious change in the dependent variable, there would be a strong correlation between those variables.

Here, I computed correlation by using two cancer types among the seven of them. The features (genes) having a correlation value higher than 0.4, 0.5, 0.6, or 0.7 were selected as significant ones. The selected features were given as the input of machine learning algorithms.

3.3.2 Information Gain

Although the correlation coefficient is a good way to choose features, in the real world there is no linear relation between variables all the time (Yu & Liu, n.d.). Hence, here information gain helps. Information gain reveals how much information a feature gives about the class variable. The important features supposed to have higher information gain value than the less important ones. Also, unrelated features should get zero value. This technique based on entropy, which splits the data into subsets that have representatives with similar values, so it measures the impurity of samples in a specific subset.

As similar to the correlation coefficient, the information gain was computed for two classes. A grid search was applied in which the features with an information gain score in the interval of [0.05, 1] were analyzed by applying a step size of 0.05. The selected features were given as the input of machine learning algorithms.

3.4 Machine Learning Algorithms

The machine learning algorithms applied in the study will be explained in this section. The data samples are labeled with one of six cancer types or healthy. Hence, totally there are seven types of labels and different supervised learning algorithms will classify them. The data were evaluated according to cancer type labels. When the multiclass classification setup is applied, the total number of samples for each cancer type should be higher, which is not the case in our dataset. So, a binary classification setup was applied to increase the success rate.

3.4.1 Support Vector Machine

Support Vector Machine aims to maximize the margin around the separation hyperplane and thereby creates the largest possible distance between different class instances. When the optimum separation hyperplane is found, the data points staying around the margin of that hyperplane are considered as support vectors of the classifier. For this reason, the complexity of an SVM model is not affected by the number of features, but the number of support vectors (i.e., samples). This makes the SVM a suitable candidate to be used in datasets with a large number of features and a low number of samples (Kotsiantis, Zaharakis, & Pintelas, 2006). The kernel functions provide an opportunity to solve non-linearly separable problems by using a linear classifier. For that purpose, kernel functions map the input vectors in a higher dimension in which original samples can be separated by a simple hyperplane.

The SVM library used in this project belongs to the scikit-learning library in Python. The C penalty parameter is set to 1; the Gaussian kernel was used. The shrinking optimization parameter is set to true. The probability scoring parameter is set to false. The stopping criteria are set to 0.001.

3.4.2 Naïve Bayes

The Bayesian theorem describes the conditional probabilities of events and the Naïve Bayes classifier is built upon this theory (VanderPlas, n.d.). The Naïve Bayes classifier assumes that all features are independent events. Even though this assumption is unrealistic for real-world problems, the resulting model is surprisingly successful when it is compared to alternative techniques (Rish, 2001). The probability of a Bayesian classifier is calculated by Equation 3.2.

$$P(\boldsymbol{X}|C) = \prod_{i=0}^{n} P(X_i|C) \tag{3.2}$$

where *C* is the class of a sample, *X* is the vector of features and the X_i is the element in the vector (Rish, 2001).

The Naïve Bayes library used in this project belongs to the scikit-learning library in Python. The prior probabilities of classes are set to the default value of none; the *var* smoothing is set to 10^{-9}.

3.4.3 Random Forest

Random forest is a machine learning algorithm that works with numerous decision trees and the statistical bagging method. A decision tree uses a threshold to decide if the input goes left or right side of the tree branch until reaching the end of the tree. In the deepest level of the tree, at leaves, the model gives an answer according to the flow in the entire path of the tree. Bagging is the process of creating new datasets from the original dataset by selecting elements randomly. It does random sampling while building trees and chooses random features to split nodes. Each decision tree in the random forest learns from random samples in the training set. By training the trees with different samples, trees might have a high variance for the samples they learn the whole forest will have lower variance. Each tree will have its own solution because of the dataset and features given to it in the training phase and this diverse forest will have the power to make more robust predictions (Breiman, 2001) ("Random Forest Regression model explained in depth," 2019). For the generalization phase, each tree votes for the classification label of a sample and the majority class label of these votes generates the final decision of the random forest algorithm.

The RF library used in this project belongs to the scikit-learning library in Python. The number of trees is set to 100 by default. The quality measurement of the split is the root mean square. The maximum depth of the tree is 2. The minimum number of samples for a split is 2. The maximum number of features for a split is a number of features and pruning is canceled.

3.5 Evaluation Metrics

The accuracy metric is sensitive when patient samples are imbalanced in different cancer types. Therefore, other evaluation metrics have to be used to guarantee the success of the project. The metrics should be insensitive to the sample numbers between cancer types. Sensitivity and specificity are the metrics belong to this type. The metrics are used to calculate those ratios are true positive (TP), false positive (FP), false negative (FN) and false positive (FP). TP means correctly prediction of the positive class, TN means correctly prediction of the negative class, FP means the positive class is predicted as false and FN means the negative class is predicted positive. The accuracy formula is given in Equation 3.3.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3.3)$$

Moreover, to be sure the accuracies of the models are not a coincidence, I applied the *k*-fold cross-validation method. This method divides the dataset to *k* subsets; *k-1* subsets are used to train the model and the remaining subset is used to measure the performance of the model. This operation is repeated *k* times and the mean accuracy of these *k*-folds gives the overall prediction performance of the model. In this study, 10-fold cross-validation was used to measure the accuracy of each model.

Sensitivity is the measure of how correctly measured the true classified results. It means that true-positive results divided by the total real positive results. Specificity is the measure of how correctly measured the false classified results. It means that the true negative results divided by total real negative results. Those metrics are independent of the number of class members. That is why they are a better way to use in cases where the classes are imbalanced (Tharwat, 2018).

After the obtain of sufficient sensitivity and specificity the genes provide those are researched in the literature. For this research, DAVID (the database for annotation, visualization, and integrated discovery)(Huang et al., 2007) and DisGeNet(Piñero et al., 2015) were used. The genes mentioned are given to those databases and the resulting pathways were examined to find relations between the genes and diseases.

CHAPTER FOUR
RESULTS AND DISCUSSION

This study applied machine learning algorithms, SVM, RF, and NB, to predict the cancer types by applying the given features that are selected with the correlation coefficient and information gain methods. A grid search was designed to find optimum values of thresholds. After setting different thresholds, these selection methods led the different number of features for optimum classification of cancer types. The evaluation of each model was performed by applying 10-fold cross-validation.

4.1 Performance of Correlation Coefficient Feature Selection

In the correlation coefficient analysis, I chose the features which have a low (0.4, 0.5) and mild correlation (0.6, 0.7) values. Since the higher correlation values did not leave any significant gene in the data set. Based on this analysis, the number of selected features varied from 4 to 111. The accuracy of each model is varying between 0.03 to 0.95 as given in Table 4.1.

Table 4.1 The performance of machine learning models for classifying different cancer types by using the correlation-based feature selection

Cancer Types	Accuracy	Method	Threshold	# of Features
Hepatobiliary vs Lung	0.78	SVM	0.40	33
Hepatobiliary vs Lung	0.86	SVM	0.50	7
Hepatobiliary vs Lung	0.77	NB	0.50	7
Hepatobiliary vs Lung	0.93	RF	0.40	33
Hepatobiliary vs Pancreatic	0.80	SVM	0.50	10
Hepatobiliary vs Pancreatic	0.72	NB	0.50	10
Hepatobiliary vs Pancreatic	0.85	RF	0.50	10
Hepatobiliary vs Pancreatic	0.95	RF	0.40	68
Breast vs Colorectal	0.66	SVM	0.40	4
Breast vs Colorectal	0.59	RF	0.40	4
Breast vs Lung	0.72	NB	0.40	3
Colorectal vs Healthy	0.71	NB	0.40	111

4.2 Performance of Information Gain Feature Selection

I performed feature selection with the information gain values between 0.05 and 1. However, there were no significant features for information gain value of higher than 0.55. Hence, only the features, which have information gain values between 0.05 and 0.55, were selected iteratively. After that analysis, the number of selected features varied from 2 to 875. The accuracy of each model is varying between 0.03 to 1.00, as similar to the correlation-based method (Table 4.2).

Table 4.2 The performance of machine learning models for classifying different cancer types by using the information gain-based feature selection

Cancer Types	Accuracy	Method	Threshold	# of Features
Hepatobiliary vs Lung	1.00	SVM	0.30	8
Hepatobiliary vs Lung	0.31	NB	0.30	8
Hepatobiliary vs Lung	0.87	RF	0.30	8
Hepatobiliary vs Pancreatic	0.71	SVM	0.25	8
Hepatobiliary vs Pancreatic	0.71	NB	0.25	8
Hepatobiliary vs Pancreatic	0.88	RF	0.25	8
Hepatobiliary vs Pancreatic	0.83	RF	0.20	21
Breast vs Colorectal	0.67	SVM	0.15	20
Breast vs Colorectal	0.75	RF	0.15	20
Breast vs Lung	0.78	NB	0.50	2
Colorectal vs Healthy	0.66	NB	0.05	875

4.3 Performance Comparison of Correlation Coefficient and Information Gain

When I evaluate all experimental results, I observed that the number of features does not usually have a strong effect on the classification success of machine learning models. One of the poorest results was obtained in the discrimination of breast and colorectal cancers. Neither information gain nor correlation coefficient cannot achieve higher than 0.66 accuracy. The NB had the lowest performance compared to RF and SVM. In terms of accuracy, RF and SVM were more successful than NB for many cases. The same models show that one case of information gain is better than the correlation coefficient as the feature selection method or vice versa. A previous study (Zhang et al., 2017) has applied the same RNA-sequencing data and achieved around 0.75 accuracy for classifying different cancer types. Our study provided better results with an average accuracy of 0.89, especially while differentiating hepatobiliary, lung, and pancreatic cancer types.

There are various evaluation techniques to measure the performance of a machine learning model. The most commonly used one is accuracy which measures the correctness of predictions. Figure 4.1 shows the accuracy level of hepatobiliary and lung cancers, Figure 4.2 shows the accuracy level of hepatobiliary and pancreatic cancers. The feature selection methods and machine learning algorithms are compared in these plots. Even though the threshold value for feature selection methods is between 0 and 1, the efficient results were standing in the interval of 0.05 to 0.5. Hence, the plots show only these efficient results.

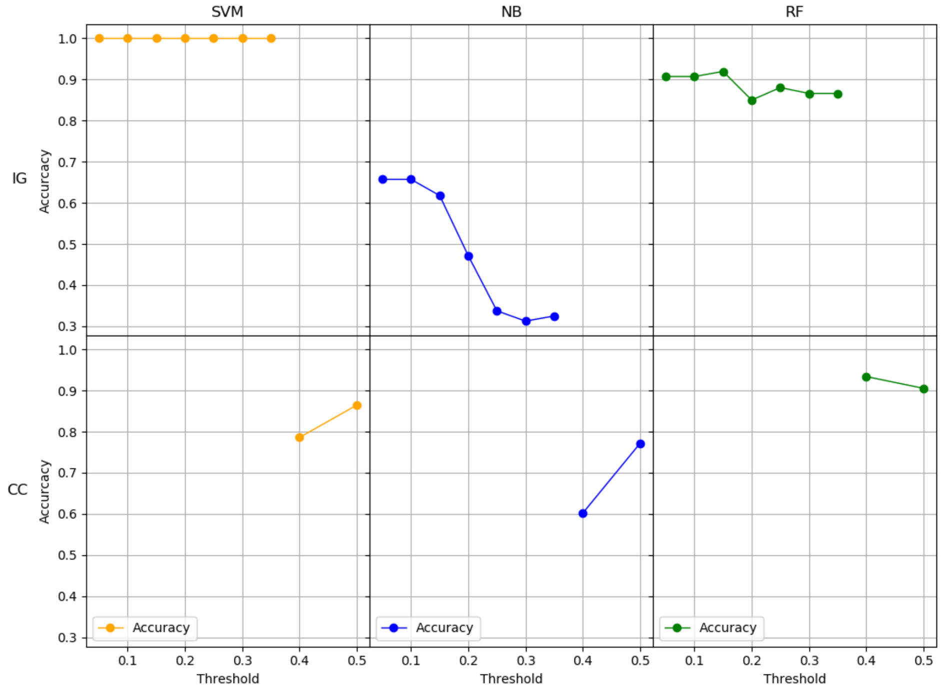


Figure 4.1 Accuracy level of hepatobiliary and lung cancers against feature selection methods with various threshold values

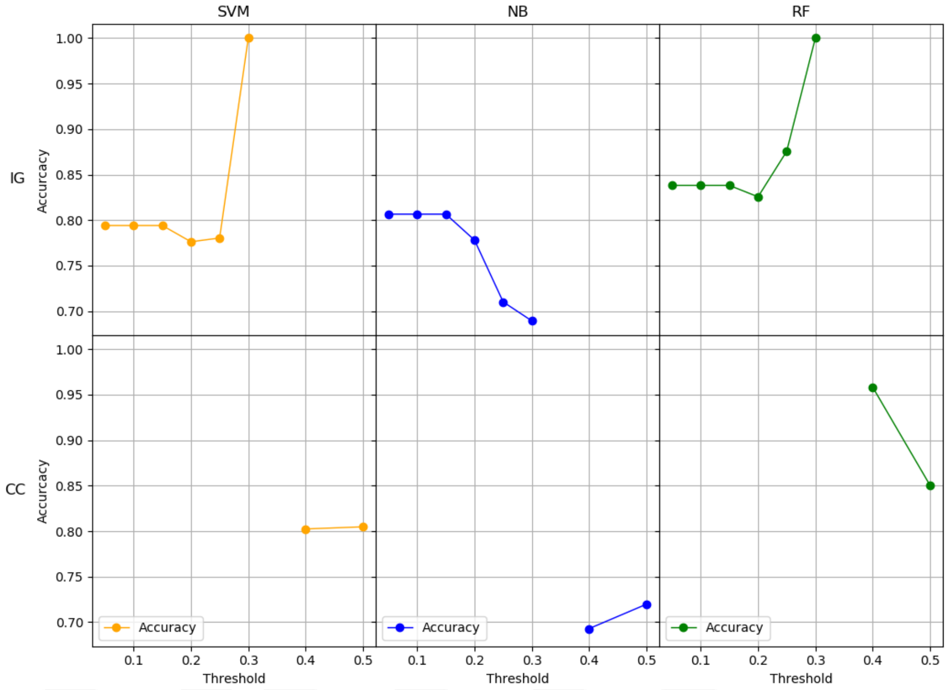


Figure 4.2 Accuracy level of hepatobiliary and pancreatic cancers against feature selection methods with various threshold values

After the application of feature selection methods and the machine learning algorithms, 85 of various binary comparisons (one cancer vs. another) led an accuracy value of higher than 75%. Since the sensitivity and specificity are among the evaluation metrics, better filtering has to be done. When this filtering was applied, the best-performing ones remained and those are the ones separating hepatobiliary and lung and, hepatobiliary and pancreatic cancers.

Figure 4.3 shows hepatobiliary and lung cancers sensitivity and specificity curves; Figure 4.4 shows hepatobiliary and pancreatic cancers sensitivity and specificity curves. The successful ones were selected by taking high sensitivity and specificity values after the filtration of accuracy rates higher than 75%. After this filtration process, the feasible ones were examined. The best-performing ones are the one which separates hepatobiliary and lung cancer with 0.5 correlation coefficient threshold by using an NB; the one separates hepatobiliary and lung cancer with 0.4 correlation coefficient threshold by using an SVM; the one separates hepatobiliary and pancreatic cancer with 0.4 correlation coefficient threshold by using an SVM; the one separates hepatobiliary and pancreatic cancer with 0.25 information gain threshold by using an SVM. The genes involving in those models were further analyzed in the following sections.

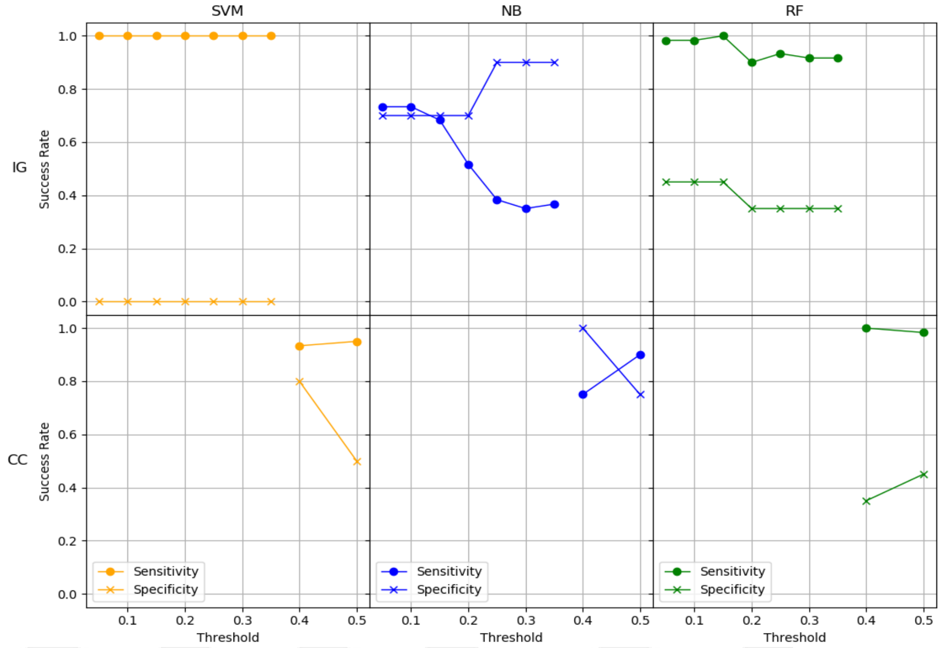


Figure 4.3 Hepatobiliary vs lung cancers sensitivity and specificity against feature selection methods with various threshold values

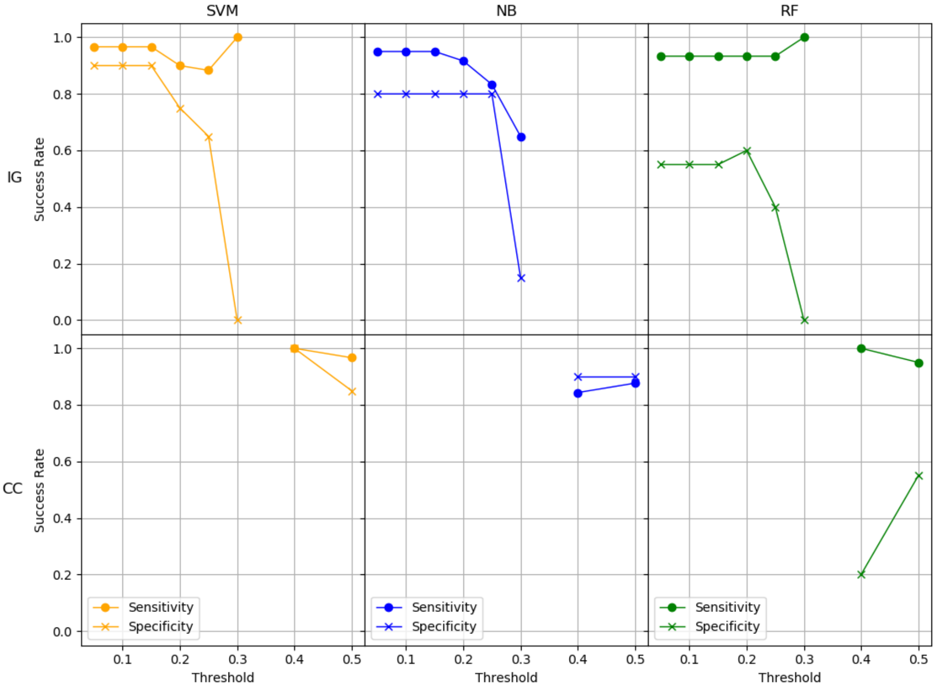


Figure 4.4 Hepatobiliary vs pancreatic cancers sensitivity and specificity against feature selection methods with various threshold values

4.4 Biological Evaluation

I will explain the biological evaluation of significant cancer discriminations (i.e., hepatobiliary vs. lung, hepatobiliary vs. pancreatic cancers) in this section.

4.4.1 Differentiation of Hepatobiliary and Lung

4.4.1.1 Correlation Coefficient with 0.5 Threshold Value

I observed that the models can discriminate hepatobiliary and lung cancers also with high sensitivity. In the case of correlation coefficient with threshold 0.5, the model created with seven genes ACPP (activatable cell-penetrating peptides), BMX (non-receptor tyrosine kinase), EGR1 (Early growth response factor 1), PLD1 (Phospholipase D1), MGAM (Maltase-glucoamylase), SEC31B (SEC31 homolog B), ARAP3 (ArfGAP with RhoGAP domain). These genes provided 0.77 accuracy, 0.9 sensitivity, and 0.75 specificity with the Naïve Bayes model.

Table 4.3 The GO terms related with hepatobiliary and lung with 0.5 threshold valued correlation coefficient

Category	Term	P.value	Genes
GOTERM_BP_ALL	GO:0030217~T cell differentiation	0.0707	BMX, EGR1
GOTERM_BP_DIRECT	GO:0016192~vesicle-mediated transport	0.0531	ARAP3, SEC31B
GOTERM_MF_DIRECT	GO:0003824~catalytic activity	0.0545	MGAM, EPLD1

Table 4.3 shows the genome ontologies (GO) annotations of these seven genes. Among those GO terms, GO:0030217~T cell differentiation was considered as important. A previous study explains that lung tumor growth was associated with activation of impaired T cells (Heim et al., 2018). However, this study did not give a solid relationship between the pathway and the disease relations.

Even though the pathway and disease relation did not give a good result, ACPP gene was related with malignant neoplasm of prostate, EGR1 was related with malignant neoplasm of prostate, malignant neoplasm of lung, lung neoplasms, and ARAP3 was related with malignant neoplasm of breast, colorectal cancer in DisGeNet (Piñero et al., 2015) which is a public discovery platform to research about human diseases.

In another study, effects of ACPP on human intrahepatic bile duct epithelial cell was examined and as a result, it has been seen the intensity of the intracellular signal is increasing with the ACPP incubation time in a certain range (Tu et al., 2016).

EGR1 was associated with HCC (hepatocellular carcinoma) and it was observed the level of EGR1 is significantly increased in HCC tissue (Bi et al., 2019). Another study shows that the level of EGR1 is significantly important for the survival of the NSCLC (non-small-cell lung cancer) patients (Zhu, Webster, Flower, & Woll, 2004).

4.4.1.2 Correlation Coefficient with 0.4 Threshold Value

Although the results with those seven features are better than the reference study (Zhang et al., 2017), I reduced the correlation coefficient threshold to 0.4 and change algorithm to SVM. Then I have thirty-one genes which are ACPP, BMX, CCR1 (C-C motif chemokine receptor 1), EGR1, CXCR1 (C-X-C motif chemokine receptor 1), PLAGL1 (PLAG1 like zinc finger 1), PLD1, PTGS2 (Prostaglandin-endoperoxide synthase 2), NRP1 (Neuropilin-1), MGAM, RRP9 (Ribosomal RNA processing 9), ZNF235 (Zinc finger protein 235), DUSP10 (Dual specificity phosphatase 10), ATG2A (Autophagy related 2A), SEPT8 (Septin 8), SEC31B, FAM198B (Family with sequence similarity 198 member B), ARAP3, SLC37A3 (Solute carrier family 37 member 3), MSANTD4 (Myb/SANT DNA binding domain containing 4 with coiled-coils), AGPAT9 (Glycerol-3-phosphate acyltransferase 3), REL (REL proto-oncogene), CSNK1A1L (Casein kinase 1 alpha 1 like), AFMID (Arylformamidase), CLEC4C (C-type lectin domain family 4 member C), SLC39A11 (Solute carrier family 39 member 11), CLEC4G (C-type lectin domain family 4 member G), LIPN (Lipase family member N), METTL12 (Citrate synthase lysine methyltransferase), C3ORF62 (Chromosome 3 open reading frame 62), LILRB3 (Leukocyte immunoglobulin like receptor B3). These genes provided 0.78 accuracy, 0.93 sensitivity, and 0.8 specificity with the Support Vector Machine model.

Table 4.4 The GO terms related with hepatobiliary and lung with 0.4 threshold valued correlation coefficient

Category	Term	PValue	Genes
GOTERM_BP_ALL	GO:0006952~defense response	0.0016	CCR1, BMX, DUSP10, PTGS2, CLEC4C, RELT, CXCR1, LILRB3, EGR1
GOTERM_BP_ALL	GO:0098759~cellular response to interleukin-8	0.0046	CXCR1, EGR1

Table 4.4 continues

GOTERM_BP_ALL	GO:0098758~response to interleukin-8	0.0046	CXCR1, EGR1
GOTERM_BP_ALL	GO:0006955~immune response	0.0077	CCR1, BMX, DUSP10, CLEC4C, CLEC4G, RELT, LILRB3, EGR1
GOTERM_BP_ALL	GO:0044710~single-organism metabolic process	0.0118	CCR1, DUSP10, AFMID, NRP1, MGAM, PLD1, ACPP, PTGS2, LIPN, RELT, CXCR1, AGPAT9, EGR1

Table 4.4 continues

GOTERM_BP_ALL	GO:0007165~signal transduction	0.0125	ARAP3, BMX, CCR1, DUSP10, NRP1, PLD1, ACPP, PTGS2, CLEC4C, RELT, PLAGL1, CXCR1, CSNK1A1L, LILRB3, AGPAT9, EGR1
GOTERM_BP_ALL	GO:0035556~intracellular signal transduction	0.0148	CCR1, ARAP3, BMX, DUSP10, NRP1, PLD1, PTGS2, RELT, PLAGL1, AGPAT9
GOTERM_BP_ALL	GO:0007166~cell surface receptor signaling pathway	0.0152	CCR1, BMX, NRP1, ACPP, CLEC4C, RELT, CXCR1, CSNK1A1L, LILRB3, EGR1
GOTERM_BP_ALL	GO:0006954~inflammatory response	0.0158	CCR1, DUSP10, PTGS2, RELT, CXCR1,
GOTERM_BP_ALL	GO:0090335~regulation of brown fat cell differentiation	0.0184	DUSP10, PTGS2,

Table 4.4 continues

GOTERM_BP_ALL	GO:0007154~cell communication	0.0279	CCR1, ARAP3, BMX, DUSP10, NRP1, PLD1, ACPP, PTGS2, CLEC4C, RELT, PLAGL1, CXCR1, CSNK1A1L, LILRB3, AGPAT9, EGR1
GOTERM_BP_ALL	GO:0002376~immune system process	0.0284	CCR1, BMX, DUSP10, CLEC4C, RELT, CXCR1, CLEC4G, LILRB3, EGR1
GOTERM_BP_ALL	GO:0034097~response to cytokine	0.0334	CCR1, PTGS2, RELT, CXCR1, EGR1
GOTERM_BP_ALL	GO:0002521~leukocyte differentiation	0.0359	CCR1, BMX, LILRB3, EGR1
GOTERM_BP_ALL	GO:0061437~renal system vasculature development	0.0380	NRP1, EGR1
GOTERM_BP_ALL	GO:0061440~kidney vasculature development	0.0380	NRP1, EGR1
GOTERM_BP_ALL	GO:0050793~regulation of developmental process	0.0388	CCR1, ARAP3, DUSP10, NRP1, PTGS2, CSNK1A1L, LILRB3, EGR1

Table 4.4 continues

GOTERM_BP_ALL	GO:0006950~response to stress	0.0414	CCR1, BMX, DUSP10, NRP1, PTGS2, CLEC4C, RELT, PLAGL1, CXCR1, LILRB3, EGR1
GOTERM_BP_ALL	GO:1902531~regulation of intracellular signal transduction	0.0433	CCR1, ARAP3, DUSP10, NRP1, PTGS2, RELT, AGPAT9
GOTERM_BP_ALL	GO:0002042~cell migration involved in sprouting angiogenesis	0.0484	NRP1, PTGS2
GOTERM_BP_ALL	GO:0042180~cellular ketone metabolic process	0.0486	AFMID, PTGS2, EGR1
GOTERM_BP_ALL	GO:0009966~regulation of signal transduction	0.0486	CCR1, ARAP3, DUSP10, NRP1, PTGS2, ACPP, RELT, AGPAT9, EGR1
GOTERM_MF_ALL	GO:0019955~cytokine binding	0.0108	CCR1, NRP1, CXCR1
GOTERM_MF_ALL	GO:0016298~lipase activity	0.0190	CCR1, PLD1, LIPN
GOTERM_MF_ALL	GO:0042578~phosphoric ester hydrolase activity	0.0238	CCR1, DUSP10, PLD1, ACPP

Table 4.4 continues

GOTERM_MF_ALL	GO:0019956~chemokine binding	0.0343	CCR1, CXCR1
GOTERM_MF_ALL	GO:0016788~hydrolase activity, acting on ester bonds	0.0382	CCR1, DUSP10, PLD1, ACPP, LIPN
GOTERM_MF_ALL	GO:0001637~G-protein coupled chemoattractant receptor activity	0.0423	CCR1, CXCR1
GOTERM_MF_ALL	GO:0030246~carbohydrate binding	0.0706	257335 (MGAM, 198178 (CLEC4C, 182566 CLEC4G
REACTOME_PATHWAY	R-HSA-1483166:R-HSA-1483166	0.0483	PLD1, AGPAT9

Table 4.4 lists GO-term annotations for the thirty-one genes. Among those GO terms, when the GO:0098759~cellular response to interleukin-8 is searched and according to studies it was observed that non-small cell lung cancer was causing the production of IL-8 (Interleukin-8) with middle or high levels (Zhu et al., 2004) (Wang et al., 1996). Hence this GO term has a relationship with lung cancer. Moreover, in another study, the HCC cells were found the main producer of the IL-8 expression (Akiba, Yano, Ogasawara, Higaki, & Kojiro, 2001).

Another GO term is GO:0006955~immune response. The immune system has a vital role in the integrity and the maintenance of the organism. While it keeps protecting the organism against pathogens, it also has a role in cancer prevention. Generally, the abnormal proteins known as tumor antigens are the result of damaged DNA in cancer cells. Those tumor antigens make the cell different from others. On a daily bases, the immune system destroys cancer cells. The existence of contrivances that allow cancer cells to escape from immune responses preventing the development of malignant tumors is obvious and it (Australia, 2014). Cancer is induced by genetic and epigenetic changes (Welsh, 2013, p. 4). Many of these changes control signaling pathways that control cell death, cell division, cell growth, cell fate, and cell mobility, and may allow for the establishment of wider signal networks that promote cancer progression (Sever & Brugge, 2015).

With the correlation coefficient threshold 0.4 twenty-four more genes came out in addition to the ones found in the threshold 0.5. These genes are BMX, which is related to large cell carcinoma of lung, CCR1 which is related to liver carcinoma, PLAGL1 which is related to malignant neoplasm of stomach, NRP1 which is related to malignant neoplasm of prostate and pancreas, RRP9 which is related to malignant neoplasm of breast and stomach, CSNK1A1L which is related to colorectal cancer and finally PTGS2 which is related to many diseases according to research in DAVID (Huang et al., 2007).

I could not find any previous study about the relation between PLAGL1, RRP9, CSNK1A1L and hepatobiliary or lung cancer.

Although there is no study shows the relation between BTX and hepatobiliary, some studies show the relation between BTX and lung cancer. BMX is playing a crucial role in tumorigenesis and cancer progression within the PI3K/BMX/STAT3 signaling pathway (Peng et al., 2016).

Overexpression of NRP1 can be seen in many cancers including pancreatic and lung. However, depending on the cancer type, the inhibition of NRP1 expression has different effects (Vivekanandhan et al., 2017).

PTGS2 is an enzyme induced by proinflammatory stimuli. It is also known as COX2. It is often overexpressed in malignant tissues. In many malignancies including lung, its overexpression has been observed (Khorshidi et al., 2014).

CCR1 is a member of the seven-transmembrane G-protein-coupled receptor family. It is involved in the activation and trafficking of immune cells and it is extensively expressed in many cell types (Shin et al., 2017).

4.4.2 Differentiation of Hepatobiliary and Pancreatic

4.4.2.1 Correlation Coefficient with 0.5 Threshold Value

I observed that the discrimination of hepatobiliary and pancreatic cancers is quite successful with ten features coming from the data with the correlation coefficient threshold 0.5. These genes are ALPL (Alkaline phosphatase), ERG (ETS transcription factor), MMP8 (Matrix metallopeptidase 8), DGAT2 (8 Diacylglycerol O-acyltransferase 2), SLC26A8 (Solute carrier family 26 member 8), TRABD2A (TraB domain containing 2A), VSIG10L (V-set and immunoglobulin domain containing 10 like), CCDC141 (Coiled-coil domain containing 141), RN7SL2 (cytoplasmic 2), TMEM233 (Transmembrane protein 233). They led 0.80 accuracy, 0.96 sensitivity, and 0.85 specificity by using an SVM model.

Table 4.5 The GO terms related with hepatobiliary and pancreatic with 0.5 threshold valued correlation coefficient

Category	Term	PValue	Genes
GOTERM_BP_ALL	GO:0007275~multicellular organism development	0.0094	DGAT2, ALPL, CCDC14, ERG, MMP8, SLC26A8
GOTERM_BP_ALL	GO:0048856~anatomical structure development	0.0161	DGAT2, ALPL, CCDC14, ERG, MMP8, SLC26A8
GOTERM_BP_ALL	GO:0044767~single-organism developmental process	0.0161	DGAT2, ALPL, CCDC14, ERG, MMP8, SLC26A8

Table 4.5 continues

GOTERM_BP_ALL	GO:0009888~tissue development	0.0179	DGAT2, ALPL, CCDC141, ERG, MMP8,
GOTERM_BP_ALL	GO:0044707~single-multicellular organism process	0.0235	DGAT2, ALPL, CCDC14, ERG, MMP8, SLC26A8

Table 4.5 shows GO-term annotations for ten genes. Among those GO terms, none of them is found to be significantly important. Hence, I focused on the genes. Some of the genes are ALPL, ERG, MMP8, DGAT2. ALPL is related to liver diseases, liver dysfunction, and tumoral calcinosis. ERG is related to leukemia, myelocytic, acute, malignant neoplasm of prostate, and Ewings sarcoma. MMP8 is related to melanoma, liver cirrhosis, lung diseases. DGAT2 is related to cholestasis, hepatitis, toxic, drug-induced liver disease, drug-induced acute liver injury.

The relationship between those genes and lung or hepatobiliary cancer was also analyzed. A study shows that ZEB2 represses transcription of a group of genes including ALPR. And ALPR is expressed in various types of tumors including pancreatic cancer (Katoh & Katoh, 2009).

Even though no important result was found about the MMP8, the results about MMP (matrix metalloproteinases) show its critical role in biliary cell migration (Terada, Okada, & Nakanuma, 1995).

4.4.2.2 Information Gain with 0.25 Threshold Value

I had successful results with the seven features coming from information gain with threshold 0.25 and SVM algorithm. The success rates were 0.78 accuracy, 0.88 sensitivity, and 0.65 specificity. The important thing here is the genes are TGFBR3 (Transforming growth factor beta receptor 3), TNR (Tenascin R), LIN28A (Lin-28 Homolog A), TRABD2A, FAM117B (Family with sequence similarity 117 member B), GAREML (GRB2 associated regulator of MAPK1 subtype 2), CCDC141, and only two of them are common with the ones in the correlation coefficient threshold 0.5 and SVM used case.

Table 4.6 The GO terms related with hepatobiliary and pancreatic with 0.25 threshold valued information gain

Category	Term	PValue	Genes
GOTERM_BP_ALL	GO:0048513~animal organ development	0.0065	TGFBR3, TNR, CCDC141, LIN28A
GOTERM_BP_ALL	GO:0022029~telencephalon cell migration	0.0105	TNR, CCDC141
GOTERM_BP_ALL	GO:0021885~forebrain cell migration	0.0110	TNR, CCDC141
GOTERM_BP_ALL	GO:0016477~cell migration	0.0142	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0022029~telencephalon cell migration	0.0105	TNR, CCDC141
GOTERM_BP_ALL	GO:0021885~forebrain cell migration	0.0110	TNR, CCDC141

Table 4.6 continues

GOTERM_BP_ALL	GO:0016477~cell migration	0.0142	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0022029~telencephalon cell migration	0.0105	TNR, CCDC141
GOTERM_BP_ALL	GO:0021885~forebrain cell migration	0.0110	TNR, CCDC141
GOTERM_BP_ALL	GO:0016477~cell migration	0.0142	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0048731~system development	0.0167	TGFBR3, TNR, CCDC141, LIN28A
GOTERM_BP_ALL	GO:0048870~cell motility	0.0178	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0051674~localization of cell	0.0178	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0040011~locomotion	0.0234	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0007275~multicellular organism development	0.0245	TGFBR3, TNR, CCDC141, LIN28A
GOTERM_BP_ALL	GO:0051240~positive regulation of multicellular organismal process	0.0205	TGFBR3, TNR, LIN28A

Table 4.6 continues

GOTERM_BP_ALL	GO:2000026~regulation of multicellular organismal development	0.0293	TGFBR3, TNR, LIN28A
GOTERM_BP_ALL	GO:0031099~regeneration	0.0297	TGFBR3, TNR
GOTERM_BP_ALL	GO:0006928~movement of cell or subcellular component	0.0318	TGFBR3, TNR, CCDC141
GOTERM_BP_ALL	GO:0044767~single-organism developmental process	0.0346	TGFBR3, TNR, CCDC141, LIN28A
GOTERM_BP_ALL	GO:0032502~developmental process	0.0377	TGFBR3, TNR, CCDC141, LIN28A
GOTERM_BP_ALL	GO:0021537~telencephalon development	0.0407	TNR, CCDC141
GOTERM_BP_ALL	GO:0050768~negative regulation of neurogenesis	0.0431	TNR, LIN28A
GOTERM_BP_ALL	GO:0050793~regulation of developmental process	0.0441	TGFBR3, TNR, LIN28A
GOTERM_BP_ALL	GO:0048468~cell development	0.0380	TGFBR3, TNR, LIN28A
GOTERM_BP_ALL	GO:0044707~single-multicellular organism process	0.0443	TGFBR3, TNR, CCDC141, LIN28A

Table 4.6 continues

GOTERM_BP_ALL	GO:0007399~nervous system development	0.0458	TNR, CCDC141, LIN28A

Table 4.6 lists the GO-terms related to given seven genes. Among those GO terms, GO:0016477~cell migration and GO:0048468~cell development was considered important since they have serious functions in tumor development and metastasis. The first step of tumor metastasis is a transgression of cancer cells into the surrounding tissue. To spread other organs in the body, cancer cells need blood vessel walls (Razidlo et al., 2015).

Two of the most important post-transcriptional regulatory proteins are RBPs (RNA binding proteins) and miRNAs (microRNAs) that effects gene expression. The abnormal expression of them causes the growth of human malignancies. LIN28A was found to be related to malignant neoplasms, neoplasm invasiveness in DisGeNet (Piñero et al., 2015). When it is investigated, the level of LIN28A and MSI2 have a positive correlation with HCC. Those findings show that LIN28A might have the potential to be used as a therapeutic target for CSCs (liver cancer stem cells) (Fang et al., 2017). In another study, a direct association between LIN28A and pancreatic cancer was detected. Also, the LIN28A decrease causes malignant behaviors in PANC1 cells. Hence, LIN28A might have a critical aspect of pancreatic cancer progression (Xu et al., 2016, p. 2).

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

According to WHO (World Health Organization), cancer is the most lethal disease in the world. Hence it has been highly studied in the last decades. To find a better cure for it, the fundamental causes of the disease have to be researched in more detail. The fundamental causes can be understood by examining diseased tissues in the cellular or on a genetic basis. There are many technologies to do that researches to discover better biomarkers to recognize cancer earlier.

In this study, the RNA sequencing technique was used to discover biomarker genes for discrimination of different cancer types. RNA sequencing technology is one of the popular high throughput sequencing methods. The data set had seven sample groups; one of them is healthy and other groups have different cancer types. To find the disease-causing genes the input samples of patients were cleaned and normalized. After that preprocessing, the total number of genes was 3427, it is 6% of the initial mRNA reads. However, when multi and binary classifications were run with this number of genes, the initial results were not satisfying. Since the number of genes was 3427, however, the number of patient samples was 285. Hence, to improve results, the number of genes has to be reduced more. To apply this, genes were going through feature selection methods according to how good they can separate the two different cancer types. The study that has the same patient samples was applying mRMR feature selection, we experimented with other methods. Information gain and correlation coefficient feature selection methods were used in our experiments. For both feature selection methods, the threshold values were iteratively increased, and a grid search technique was applied. For the correlation coefficient, efficient classification results were found in 0.4 and 0.5 threshold values; for information gain, it is ranging from 0.05 to 0.55. The composed cancer groups were given to the SVM, NB, RF machine learning algorithms. The results, which are not overfitting and have satisfying sensitivity and specificity values, are the ones separating hepatobiliary-lung and hepatobiliary-pancreatic cancers.

When we analyzed the results of the current study, they are more successful than the study of the initial patient samples in terms of accuracy, sensitivity, and specificity. The successful results traced back to identify the biomarker genes, their biological functions and related diseases were searched in biological databases. The success of models was proved in biological manners. Since the hepatobiliary and pancreatic cancers are located close in the body, so successful discrimination between them can be considered vital in the diagnostic phase without getting surgery. The results of this study might help to diagnose a new person who has pancreatic cancer or hepatobiliary cancer.

REFERENCES

Abdel Samee, N. M., Solouma, N. H., & Kadah, Y. M. (2012). Detection of biomarkers for Hepatocellular Carcinoma using a hybrid univariate gene selection methods. *Theoretical Biology and Medical Modelling*, *9*(1), 34.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., … Searle, S. M. J. (2016). The Ensembl gene annotation system. *Database: The Journal of Biological Databases and Curation*, *2016*, 1-19

Akiba, J., Yano, H., Ogasawara, S., Higaki, K., & Kojiro, M. (2001). Expression and function of interleukin-8 in human hepatocellular carcinoma. *International Journal of Oncology*, *18*(2), 257–264.

Australia, C. (2014). *The immune system and cancer [Text]*. Retrieved September 29, 2019, from http://edcan.org.au/edcan-learning-resources/supporting-resources/biology-of-cancer/defining-cancer/immune-system

Bi, J.-G., Zheng, J.-F., Li, Q., Bao, S.-Y., Yu, X.-F., Xu, P., & Liao, C.-X. (2019). MicroRNA-181a-5p suppresses cell proliferation by targeting Egr1 and inhibiting Egr1/TGF-β/Smad pathway in hepatocellular carcinoma. *The International Journal of Biochemistry & Cell Biology*, *106*, 107–116.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics*, *17*(5), 257–271.

Choi, Y., Liu, T. T., Pankratz, D. G., Colby, T. V., Barth, N. M., Lynch, D. A., … Huang, J. (2018). Identification of usual interstitial pneumonia pattern using RNA-Seq and machine learning: Challenges and solutions. *BMC Genomics*, *19*(2), 101.

Chu, Y., & Corey, D. R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Therapeutics*, *22*(4), 271–274.

Fang, T., Lv, H., Wu, F., Wang, C., Li, T., Lv, G., … Wang, H. (2017). Musashi 2 contributes to the stemness and chemoresistance of liver cancer stem cells via LIN28A activation. *Cancer Letters*, *384*, 50–59.

Heim, L., Friedrich, J., Engelhardt, M., Trufa, D. I., Geppert, C. I., Rieker, R. J., … Finotto, S. (2018). NFATc1 Promotes Antitumoral Effector Functions and Memory CD8+ T-cell Differentiation during Non-Small Cell Lung Cancer Development. *Cancer Research*, *78*(13), 3619–3633.

Henry, N. L., & Hayes, D. F. (2012). Cancer biomarkers. *Molecular Oncology*, *6*(2), 140–146.

Hicks, S. C., & Irizarry, R. A. (2014). When to use Quantile Normalization? *BioRxiv*, 012203.

Hsu, H.-H., & Hsieh, C.-W. (2010). Feature Selection via Correlation Coefficient Clustering. *Journal of Software*, *5*(12), 1371–1377.

Hsu, Y.-H., & Si, D. (2018). Cancer Type Prediction and Classification Based on RNA-sequencing Data. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2018*, 5374–5377.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., … Lempicki, R. A. (2007). DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, *35*(Web Server issue), W169–W175.

Isik, Z., & Ercan, M. E. (2017). Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients. *Computers in Biology and Medicine*, *89*, 397–404.

Jaskowiak, P. A., Costa, I. G., & Campello, R. J. G. B. (2018). Clustering of RNA-Seq samples: Comparison study on cancer data. *Methods*, *132*, 42–49.

Katoh, M., & Katoh, M. (2009). Integrative genomic analyses of ZEB2: Transcriptional regulation of ZEB2 based on SMADs, ETS1, HIF1alpha, POU/OCT, and NF-kappaB. *International Journal of Oncology*, *34*(6), 1737–1742.

Khorshidi, F., Haghighi, M. M., Nazemalhosseini Mojarad, E., Azimzadeh, P., Damavand, B., Vahedi, M., … Zali, M. R. (2014). The prostaglandin synthase 2/cyclooxygenase 2 (PTGS2/ COX2) rs5277 polymorphism does not influence risk of colorectal cancer in an Iranian population. *Asian Pacific Journal of Cancer Prevention: APJCP*, *15*(8), 3507–3511.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969.

Kumar, R., Ichihashi, Y., Kimura, S., Chitwood, D. H., Headland, L. R., Peng, J., … Sinha, N. R. (2012). A high-throughput method for illumina RNA-Seq library preparation. *Frontiers in Plant Science*, *3*, 202-214.

Liu, P., Tseng, G., Wang, Z., Huang, Y., & Randhawa, P. (2019). Diagnosis of T-cell-mediated kidney rejection in formalin-fixed, paraffin-embedded tissues using RNA-Seq-based machine learning algorithms. *Human Pathology*, *84*, 283–290.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *33*(Database Issue), D54–D58.

Peng, J., Wang, Q., Liu, H., Ye, M., Wu, X., & Guo, L. (2016). EPHA3 regulates the multidrug resistance of small cell lung cancer via the PI3K/BMX/STAT3 signaling pathway. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine*, *37*(9), 11959–11971.

Perakis, S., & Speicher, M. R. (2017). Emerging concepts in liquid biopsies. *BMC Medicine*, *15*(1), 75.

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., … Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database: The Journal of Biological Databases and Curation*, *2015*, 1-17

Random Forest Regression model explained in depth. (2019). Retrieved September 29, 2019, from GDCoder website: https://gdcoder.com/random-forest-regressor-explained-in-depth/

Razidlo, G. L., Magnine, C., Sletten, A. C., Hurley, R. M., Almada, L. L., Fernandez-Zapico, M. E., … McNiven, M. A. (2015). Targeting pancreatic cancer metastasis by inhibition of Vav1, a driver of tumor cell invasion. *Cancer Research*, *75*(14), 2907–2915.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*, 41–46.

RNA-seq: Basics, Applications and Protocol. (n.d.). Retrieved July 3, 2019, from Genomics Research from Technology Networks website: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

Sever, R., & Brugge, J. S. (2015). Signal Transduction in Cancer. *Cold Spring Harbor Perspectives in Medicine*, *5*(4).

Shin, S. Y., Lee, D. H., Lee, J., Choi, C., Kim, J.-Y., Nam, J.-S., … Lee, Y. H. (2017). C-C motif chemokine receptor 1 (CCR1) is a target of the EGF-AKT-mTOR-STAT3 signaling axis in breast cancer cells. *Oncotarget*, *8*(55), 94591-94605

Singireddy, S., Alkhateeb, A., Rezaeian, I., Rueda, L., Cavallo-Medved, D., & Porter, L. (2015). Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-Seq and machine learning techniques.

2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 1–5.

Srinivas, P. R., Kramer, B. S., & Srivastava, S. (2001). Trends in biomarker research for cancer detection. *The Lancet Oncology*, *2*(11), 698–704.

Srivastava, A., & Creek, D. J. (2019). Discovery and Validation of Clinical Biomarkers of Cancer: A Review Combining Metabolomics and Proteomics. *PROTEOMICS*, *19*(10), 1-22.

Terada, T., Okada, Y., & Nakanuma, Y. (1995). Expression of matrix proteinases during human intrahepatic bile duct development. A possible role in biliary cell migration. *The American Journal of Pathology*, *147*(5), 1207–1213.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. Retrieved November 16, 2019, from http://www.sciencedirect.com/science/article/pii/S2210832718301546

Tu, K., Zhao, L., Gu, J., Yan, P., Wang, F., & Cao, Y. (2016). The function of activatable cell-penetrating peptides in human intrahepatic bile duct epithelial cells. *Journal of Bioenergetics and Biomembranes*, *48*(6), 599–606.

VanderPlas, J. (n.d.). *Python Data Science Handbook*. 382–389.

Vivekanandhan, S., Yang, L., Cao, Y., Wang, E., Dutta, S. K., Sharma, A. K., & Mukhopadhyay, D. (2017). Genetic status of KRAS modulates the role of Neuropilin-1 in tumorigenesis. *Scientific Reports*, *7*(1), 1–11.

Wang, J., Huang, M., Lee, P., Komanduri, K., Sharma, S., Chen, G., & Dubinett, S. M. (1996). Interleukin-8 Inhibits Non-Small Cell Lung Cancer Proliferation: A Possible Role for Regulation of Tumor Growth by Autocrine and Paracrine Pathways. *Journal of Interferon & Cytokine Research*, *16*(1), 53–60.

Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., & Chinnaiyan, A. M. (2014). RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. *Neoplasia*, *16*(11), 918–927.

Welsh, J. (2013). Chapter 40—Animal Models for Studying Prevention and Treatment of Breast Cancer. In P. M. Conn (Ed.), *Animal Models for the Study of Human Disease* (997–1018). Boston: Academic Press.

Whole Transcriptome and mRNA Sequencing Guide. (n.d.). Retrieved November 16, 2019, from https://genohub.com/rna-seq-library-preparation/#workflow

Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine, 153*, 1–9.

Xu, M., Bian, S., Li, J., He, J., Chen, H., Ge, L., … Gong, A. (2016). MeCP2 suppresses LIN28A expression via binding to its methylated-CpG islands in pancreatic cancer cells. *Oncotarget*, *7*(12), 14476–14485.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03),* 856-863.

Zhang, Y.-H., Huang, T., Chen, L., Xu, Y., Hu, Y., Hu, L.-D., … Kong, X. (2017). Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. *Oncotarget*, *8*(50), 87494–87511.

Zhu, Y. M., Webster, S. J., Flower, D., & Woll, P. J. (2004). Interleukin-8/CXCL8 is a growth factor for human lung cancer cells. *British Journal of Cancer*, *91*(11), 1970–1976.