IOANNIS TORAKIS

Machine Learning Methods for the Evaluation of Biomolecular Markers of Pancreatic Cancer and its Correlation with Embryogenesis

TECHNICAL UNIVERSITY OF CRETE
SCHOOL OF ELECTRICAL & COMPUTER ENGINEERING
DIGITAL SIGNAL & IMAGE PROCESSING LABORATORY

Machine Learning Methods for the Evaluation of Biomolecular Markers of Pancreatic Cancer and its Correlation with Embryogenesis

Diploma Thesis
by

Ioannis Torakis
Supervisor: Prof. Michael Zervakis

Thesis Committee
Professor Michael Zervakis
Associate Professor Michail G. Lagoudakis
Associate Professor Georgios Chalkiadakis

Chania, Greece
October 2019

ABSTRACT

Pancreatic cancer is a highly lethal disease, accounting for many deaths every year. It is considered as one of the most aggressive types of cancer, and one of the major problems is the lack of early detection. A patient is diagnosed with pancreatic cancer only in advanced stages, when the possibility of developing a metastases is high. There is no standard procedure to diagnose high risk patients, since they remain asymptomatic in the cancer's early stages. Surgical resection is regarded as the only potentially curative treatment, and adjuvant chemotherapy with gemcitabine or S-1, an oral fluoropyrimidine derivative, is given after surgery. Therefore, researchers focus on the procedure of its creation, at a molecular level. There are four major driver genes for pancreatic cancer: KRAS, CDKN2A, TP53, and SMAD4. KRAS mutation and alterations in CDKN2A are early events in pancreatic tumorigenesis.

Recent researches suggest that there is a correlation of some critical signaling pathways that are activated during pancreatic cancer tumorigenesis with the procedure of embryogenesis. Though, the lack of an analysis that will be able to extract these genes involved in the pathways suggested, both in pancreatic cancer patients and embryogenesis samples is crucial. The aim of this thesis is to apply machine learning methods to find the biomolecular markers that are deferentially expressed on pancreatic cancer patients and correlate them with markers from embryogenesis. Since these markers are extracted, we will use them as classifiers on different machine learning methods, to try and classify if they refer to patient or healthy subjects.

Our thesis contributes a " 25 gene signature" of biomolecular markers which are involved in signaling pathways found in both embryogenesis and pancreatic carcinogenesis, obtained via feature extraction and feature selection methods. These markers are used as classifiers for pancreatic cancer classification, and two machine learning classification models are proposed as well. The classification models achieved high accuracy levels, and we support the notion that our “25 gene signature” in its entirety can play a classification role in discriminating patients with pancreatic cancer from healthy controls.

It is the same God which worketh all in all.
—A Corinthians 12:6

ACKNOWLEDGEMENTS

First of all, I would like to thank my parents Dimitris and Chrysanthi for their continuous support, help and patience in every decision that I have made in all these years. I would also like to thank my brothers Sotiris and Stefanos for their patience and tolerance towards my strange character.

Special thanks to my supervisor, Prof. Michael Zervakis for his useful guidance and valuable help, due to which I turned my interest in Biomedical Engineering. I would also like to thank Associate Prof. Michail G. Lagoudakis and Associate Prof. Georgios Chalkiadakis for their participation in my thesis committee.

Furthermore, I would like to thank Dr. Stelios G. Sfakianakis for his helpful suggestions and great interest for my work.

Last, but not least, I would like to express my special graditude and appreciation to Dr. Katerina Bei for her useful insights, suggestions and continuous support and guidance throughout all these months.

CONTENTS

LIST OF FIGURES

Part I

INTRODUCTION AND BACKGROUND

1

INTRODUCTION

Pancreatic cancer continues to be a major unsolved health problem, despite all the efforts and technological advances in cancer treatment, which have little impact on disease course. Almost all patients dignosed with pancreatic cancer develop metastases and die. Pancreatic cancer arises when exocrine or endocrine cells in the pancreas, a glandular organ behind the stomach, begin to multiply out of control and form a mass. These cancerous cells have the ability to invade other parts of the body.There are a number of types of pancreatic cancer. The most common, pancreatic adenocarcinoma (PDAC), accounts for about 95% of cases, and the term "pancreatic cancer" is sometimes used to refer only to that type. It also lies among the most aggressive types of pancreatic cancer, giving survival rates under 10% in its final stages.

The main factors causing this type of cancer are smoking, age and some genetic disorders. Yet, little do we know about it's primary causes. Advances in molecular biology have, however, helped us understand deeper the pathogenesis of pancreatic cancer. Many patients have mutations of the K-ras oncogene, and various tumour-suppressor genes are also inactivated. Pancreatic ductal adenocarcinoma (PDAC), is characterized by near-universal mutations in K-ras and frequent deregulation of crucial embryonic signalling pathways, including the Hedgehog (Hh) and Wnt–β-catenin cascades. [3] [4]

The main concern with pancreatic cancer is that disease prognosis is extremely poor. Signs and symptoms of the most-common form of pancreatic cancer may include yellow skin, abdominal or back pain, unexplained weight loss, light-colored stools, dark urine, and loss of appetite. Yet, these symptoms do not seem to appear in the disease's early stages, and symptoms that are specific enough to suggest pancreatic cancer typically do not develop until the disease has reached an advanced stage. By the time of diagnosis, pancreatic cancer has often spread to other parts of the body, thus making it one of the most dangerous and aggressive types of cancer.

According to latest studies [5], it is suggested that tumors often display inappropriate activation of signaling pathways which are essential for embryonic development and tissue homeostasis. Cancer may arise because the developmental programs that create the dramatic alterations in form and structure in embryonic development are potentially corrupted. The cells in our bodies retain memories of these processes and cancer can occur on the following years, if imperfections occur in the fidelity of these pathways. [4] Embryonic development (or embryogenesis) stands for the procedure by which an embryo forms and develops. Embryonic development begins with the fertilization of the egg cell (ovum) by a sperm cell, (spermatozoon). Once fertilized, the ovum is referred to as a zygote, a single diploid cell. The zygote undergoes multiple mitotic divisions without any significant growth (a process known as cleavage) and cellular differentiation, leading to the development of a multicellular embryo.

It is suggested that the requirements of cellular proliferation and differentiation are considerably similar in the signaling pathways that govern embryogenesis and PDAC. These pathways, which will be analyzed on next chapters, are critical for both embryogenesis and PDAC, and thus have been targeted for cancer therapy.

It is a great challenge to find new ways to cure pancreatic cancer, since in only 20% of the cases it is resectable. Therefore, we focus on finding what causes it, in molecular level. In these types of studies,when studying cancer types in a molecular level, the proposed technique is analysis of high dimensional gene expression microarrays.

Gene expression microarrays are widely used for gene expression profiling and to study these profiles in human cancers. Microarray gene expression analysis is a promising method for studying, classifying and even proposing disease treatment, for tumor related genes, amongst various types of human cancers. [5]

Microarray analysis is strongly correlated with machine learning methods. Various gene extraction, gene selection and classification methods are proposed in the literature, which focus on reducing the high dimensional feature space to a lower one, by removing the irrelevant, redundant and noisy genes, in order to achieve accurate classification of cancer types [6]. The scientific fields of data mining, statistical analysis and machine learning, provide us with a variety of methods and tools for analyzing microarray datasets, which will be the main field of our study in this thesis.

1.1 THESIS CONTRIBUTION

As we conclude from above, it appears that there is a correlation of critical pathways in the process of embryogenesis and pancreatic cancer. The aim of this thesis is to extract and evaluate biomolecular markers from samples of patients with pancreatic cancer, as well as their correlation with embryogenesis. Implementing machine learning methods, this thesis contributes some models that have the discrimination ability to identify the genes involved in these pathways. Furthermore, after the statistical analysis and the extraction of these markers, some other models are proposed that suggest classifiers for the diagnosis of disease at molecular level.

In particular, our analysis implements some machine learning methods, in R language for statistical computing. Examined datasets of pancreatic cancer and embryogenesis are high-dimensional microarray gene expression datasets, and they are extracted from the platform of Gene Epression Omnibus (GEO) [7].

After applying some preprocessing on the examined datasets, we extract the significantly differentially expressed genes (DEG), in human embryos, PDAC tissue and PDAC peripheral blood datasets, by implementing feature extraction methods. To visualize these genes, heatmap and clustering is also used. We then evaluate these markers, examining if they are involved in the critical pathways suggested above.

Subsequently, some classification machine learning methods are proposed (as Support Vector Machines and k-Nearest Neighbours) on the extracted genes, in order to use them as predictors and classify samples as patients or healthy.

Finally, in order to further reduce the number of predictors and result to a lower dimensional feature space, a feature selection is implemented, where the used machine learning method is Support Vector Machines - Recursive Feature Elimination (SVM-RFE).

1.2 THESIS OVERVIEW

Chapter 2 describes the necessary background for this thesis. Datasets and machine learning methods used for this thesis are extensively described. In Chapter 3 the significance of extracting the correlated molecular markers is discussed, as well as their matching to the critical pathways proposed, while in Chapter 4 we briefly refer to the related work of others. In Chapter 5 we describe our work and we present our models implementation from a technical point of view. In Chapter 6 we present the results of our proposed models. Finally, in Chapter 7 we discuss the results of this thesis and in Chapter 8 we suggest some possible future research enhancements and directions.

2

THEORETICAL BACKGROUND

2.1 DESCRIPTION OF DATASETS

We decided to work with 12 high-dimensional microarray gene expression datasets, which are briefly mentioned below. All datasets were obtained from the Gene Expression Omnibus repository of the National Center for Biotechnology Information [7], and they are publicly available.

- Human Embryos. This dataset describes the development of human embryos from week 4 through 9.

    1. Gene Expression Atlas for Human Embryogenesis (GSE15744) [8]. It contains expression levels of 54,675 RNAs of 18 human embryos samples, 3 samples for each week.

- PDAC Human Tissue. These datasets contain samples of human pdac tissue cells vs. normal cells.

    1. Integrative Survival-Based Molecular Profiling of Human Pancreatic Cancer [mRNA] (GSE32676) [9]. It contains expression levels of 54,675 mRNAs of 25 human PDAC tumors and 7 non-malignant pancreas samples.
    2. The gene expression of normal pancreatic and PDAC tisssues (GSE71989) [10]. It contains expression levels of 54,675 mRNAs of 14 human PDAC tumors and 8 non-malignant pancreas samples.

3. S100P is a metastasis-associated gene that facilitates transendothelial migration of pancreatic cancer cells (GSE19281) [11]. It contains expression levels of (22,283 and 22,645) mRNAs of 4 human PDAC tumors and 3 non-malignant pancreas samples (run on two different platforms).

4. Whole-Tissue Gene Expression Study of Pancreatic Ductal Adenocarcinoma (GSE15471) [12]. It contains expression levels of 54,675 mRNAs of 36 human PDAC tumors and 36 non-malignant pancreas samples.

5. Expression data from Mayo Clinic Pancreatic Tumor and Normal samples (GSE16515) [13]. It contains expression levels of 54,675 mRNAs of 36 human PDAC tumors and 16 non-malignant pancreas samples.

6. Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma (GSE28735) [14]. It contains expression levels of 33,297 mRNAs of 45 human PDAC tumors and 45 non-malignant pancreas samples.

7. Microarray gene-expression profiles of 69 pancreatic tumors and 61 adjacent non-tumor tissue from patients with pancreatic ductal adenocarcinoma (GSE62452) [15]. It contains expression levels of 33,297 mRNAs of 69 human PDAC tumors and 61 non-malignant pancreas samples.

- PDAC Human Peripheral Blood. These datasets contain samples of human pdac peripheral blood cells vs. normal cells.

  1. Expression profiling of PBMC from patients with hepatocellular carcinoma (GSE49515) [16]. It contains expression levels of 54,675 mRNAs of 3 human peripheral blood mononuclear cell (PBMC) and 10 normal samples.

2. Gene expression data from CD14++ CD16- classical monocytes from healthy volunteers and patients with pancreatic ductal adenocarcinoma (GSE60601) [17]. It contains expression levels of 54,675 mRNAs of 9 human peripheral blood mononuclear cell (PBMC) and 3 normal samples.

3. Blood biomarkers of pancreatic cancer associated diabetes identified by peripheral blood-based gene expression profiles (GSE15932) [18]. It contains expression levels of 54,675 mRNAs of 8 human peripheral blood mononuclear cell (PBMC) and 8 normal samples.

4. Expression data from peripheral blood in pancreatic ductal adenocarcinoma (PDAC) patients (GSE49641) [19]. It contains expression levels of 33,297 mRNAs of 18 human peripheral blood mononuclear cell (PBMC) and 18 normal samples.

All the mentioned datasets are high-dimensional microarray gene expression datasets. High-density DNA/RNA microarrays, are able to project thousands of genes simultaneously, producing the gene expression profiles. [20]. Microarray technology is a hybridization technique that aims on gene expression profiling or assessing the genome content of closely related cells or organisms. It allows monitoring the quantity of mRNA present in a cell, by collecting it and attaching it to a solid surface. [21]

Over the years, a variety of microarrays and chips has been introduced, with the most important of them being cDNA microarrays and GeneChip arrays (oligo arrays), developed at Affymetrix. [22] These techniques are based on the differential-hybridization strategy, where the cDNA plaques are replaced with spotted cDNAs or oligos, and radioactive labels are replaced with fluorescent ones. The potential of these methods is their ability to simultaneously analyze the expression of mRNAs from thousands of genes in a single experiment, producing some raw data which will be further analyzed in a computer environment [1].

High-density gene expression microarrays use oligonucleotides containing 25 base pairs used to probe genes. Each gene is represented by 16-20 pairs of oligonucleotides, forming a probe set. These pairs contain the perfect match (PM) probes, which are paired with the mismatch (MM) probes. MM probes are created by changing the 13th (middle) base of the probe set, in order to measure non-specific binding. After the RNA samples are labeled and hybridized, images are produced and alalyzed, resulting to an intensity value for each probe. These intensities contain information about the amount of hybridization occurred for each oligonucleotide probe and they are the final gene expression values produced by the microarray. [23]. The process is described in figure 1.

There is a variety of different platforms that can be used with microarrays. Our datasets run on 4 different GeneChips, $[HG - U133plus_2]$ Affymetrix Human Genome U133 Plus 2.0 Array, $[HuGene10stv1_{H}s_{E}NSG]$ Affymetrix GeneChip Human Gene 1.0 ST Array, $[HG - U133A]$ Affymetrix Human Genome U133A Array and $[HG - U133B]$ Affymetrix Human Genome U133B Array. All 4 chips are constructed by Affymetrix, and they follow the in situ oligonucleotide technology type [22]. However, the gene expression results, and the number of gene expression levels can be inconsistent due to the different probes these platforms use. Data inconsistency can also be introduced due to tissue or sample heterogeneity amongs experiments, different data preprocessing methods or the different background each sample comes from [5].

2.2 PREPROCESSING

High dimensional microarray gene expression datasets, produce raw data which contain the measured intensities and locations of the hybridized array, the information relating probe pair sets to locations on the array, and the information relating the probe sequences to locations on the array. [24] These raw data have to be preprocessed in order to give us the final gene expression

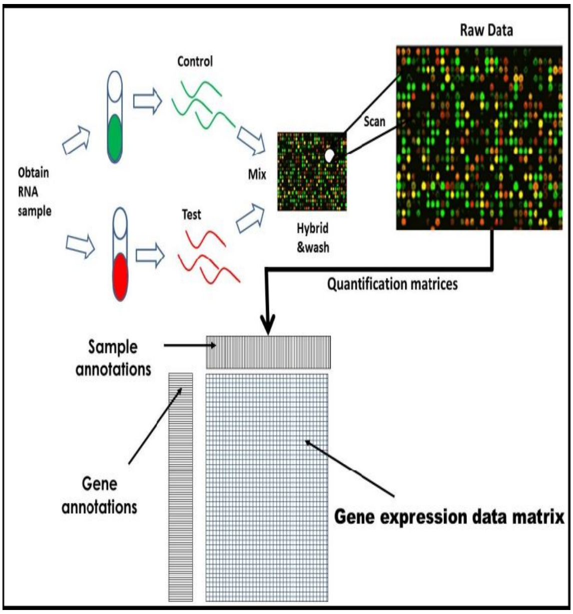


Figure 1: The mRNAs that are expressed in the compared cells, are copied into the complementary DNAs using a reverse transcriptase and they are labelled fluorescently. The produced complex cDNA probes are used to hybridize to the cDNA templates or gene-specific oligos, either spotted on a glass surface or directly synthesized, to yield the expression of thousands of genes simultaneously. Red and green dots represent the cDNAs only expressed in normal or tumours cells respectively, while the yellow indicates the cDNAs expressed in both samples[1] .

values that will be then used for classification, regression, feature extraction and feature selection.

Preprocessing is an essential process, since the expression levels may suffer from unwanted variation. This variation is often introduced during sample preparation, construction of the arrays, and arrays processing (labeling,hybridization and scanning). These sources insert the so called "obscuring variation" , which has to be removed during the preprocessing stage, since it has different effects on data and can lead the analysis to misleading results.

The method used in this analysis for the data preprocessing stage is the robust multi-array average (RMA), one of the most commonly and widely used normalization methods. RMA is a method that is divided in 3 steps: (i) background-correcting, based on a model using the transformation $B(\cdot)$, (ii) data normalization which normalizes the arrays using quantile normalization, and (iii) data summarization which fits a linear model to the background-corrected, normalized and $\log_2$ transformed probe intensities for each probe set. A robust procedure such as median polish is being used to estimate model parameters, in order to protect against outlier probes.

RMA has major advantages, compared to other methods, since is has the smallest standard deviation across replicates and has the least noise than other measures at lower concentrations. Its major advantage is noticeable in low expression values, where the standard deviation is up to 10 times smaller than the other measures. Overall, RMA achieves greater sensitivity and specificity in detection of differential expression, providing the researchers working with GeneChip technology with a powerful tool. [23]

We used two variations of RMA in our analysis, provided by the dedicated R packages, gcrma and oligo. GCRMA differs from RMA in the step of background correcting. GCRMA method adjusts for background intensities in gene expression data which include optical noise and non-specific binding. It uses probe sequence information to estimates the probe affinity to non-specific binding. It then continues with the steps of normalization and summarization as described by rma [24]. The rma analysis provided by the oligo package, is a similar process, with the main difference being that it provides support to more platforms that lack the Mismatch probes (MM probes), since it does not require them for the background-correction step.

2.3 FEATURE EXTRACTION METHODS

High throughput technologies generally produce large datasets of gene expression values. Typically, microarrays produce tens of thousands of gene features while some of the genes appear in more than one expressions. Furthermore, some datasets may contain gene expressions extracted from different platforms. A characteristic of gene expression microarray datasets is the small amount of patient samples which is significantly smaller than the number of genes, commonly known as "curse of dimensionality". However, among the large amount of genes, only a small fraction is related to specific diseases and can be used to extract information about them. The existence of a large number of irrelevant features inserts serious problems for machine learning methods, along with statistical and analytical challenges, since it strongly affects their computational time and seriously reduces their classification accuracy. [25]

Thus it is crucial to filter out this large number of irrelevant features that are not of interest, and work with smaller datasets containing genes relevant to our analysis. Furthermore, a common problem in all machine learning methods is the risk of overfitting. Data overfitting arises when the number of features (classifiers) is comparatively larger than the number of samples, which is exactly the case in high throughput gene expression microarray data, where we have thousands of genes but only less than 100 samples. When data overfitting happens, a decision function to seperate the training data can be found, but it will perform poorly when tested in an independent testing dataset[26]. These challenge can be alleviated by using two types of methods: Feature Extraction and Feature Selection. The aim of both methods is to extract a small subset of features with information useful to our analysis, in order to reduce processing time and ensure higher classification accuracy. Feature selection is described in its dedicated section [27].

Feature Extraction aims to transform a high-dimensional feature space into a low-dimensional space, in an appropriate way that the transformed variables contain information on the data relevant to our analysis, which is otherwise hidden in the large data set. Both feature extraction and feature selection are data mining methods. These methods include clustering, basic linear transforms of the input variables (Principal Component Analysis/Singular Value Decomposition, Linear Discriminant Analysis), spectral transforms, wavelet transforms or convolution of kernels. A large number of gene selection and extraction approaches exist, such as ttest, relief-F, information gain, and Principal Component Analysis (PCA), Linear Discriminant Analysis, independent component analysis (ICA). [20] Data mining can be performed with machine learning methods or with classical statistical approaches. In our analysis we are interested in measuring the expression change of each gene, in two class datasets (pdac-normal).

The most common difference statistical measure used to identify differentially expressed genes (DEGs) is the $\log_2$ fold-change. The $\log_2$ fold change, is a statistical measure describing how much an expression is changing between two distinct groups of samples.

$$\log_2 FC = \log_2 \frac{B - A}{A}$$

However, when analyzing high-dimensional datasets, with much larger number of features than samples, we also expect a high number of false positive test results. Therefore, another statistical measure has to be introduced to moderate the number of falsely called genes. In this analysis, the false discovery rate (FDR) is used, which was firstly introduced by Benjamini and Hochberg, as an expected proportion of false positive genes among all positive genes. The FDR is regulated by raw p-values, another statistical measure, which are user adjusted in order to control the FDR tolerance. The m raw p-values are first ordered by ascending order, then the adjusted p-values are given by

$$\bar{p} = \min_{k=j,\ldots,m} \{\min(\frac{m}{k} p_{r_k}, 1)\ \}$$

Differential expression analysis is completed by setting the FDR to a specific threshold and then calculating the fold change for each gene. The results of this statistical analysis are displayed in a matrix, where the significantly differential expressed genes are ordered by their $\log_2$ fold change values [28].

2.3.1 *Heatmaps and Clustering*

Heatmaps and clustering are widely used in gene expression analysis studies, for data visualization and quality control. More specifically, they are one of the most popular methods used in high throughput gene expression profiling, as they are produced by the technology of microarrays. A heatmap is a graphical representation of the input data in a matrix, where each value is described by a color.

Heatmaps are used in biomedical engineering to represent the levels of the gene expression data, across a number of comparable samples. In a typical gene expression heat map, the y-axis is assigned to the genes, while the x-axis is assigned to the samples. A gene expression heatmap, especially when combined with clustering, can provide the user with very useful insights about the quality, the distribution, the evolution and the features of the data, since it visualizes the data by pseudocoloring them from a predefined color spectrum.

Unsupervised Clustering is the process of grouping a set of genes or samples together, based on a similarity metric that is computed for features. Clustering methods aim on grouping the objects into a predetermined number of group, in way that a specific function is maximized. Cluster analysis will always produce the predetermined clusters. The quality of the clustering though, depends on the algorithm used to produce the current clustering. Examples of clustering algorithms are the k-mens algorithm, Farthest First Traversal Algorithm, Density-based clustering and Expectation Maximization (EM). [27]

Clustering is used to classify sample subtypes, or to identify outliers in the dataset. The majority of cancer gene expression datasets contains samples defined by a phenotype: disease and control groups. In the best case scenario, after the cluster is complete, the samples should be grouped into two subgroups, based on their phenotype. Though, many factors could affect the outcome of the clustering, leading to ambiguous results, and thus making the clustering a powerful tool to identify novel subtypes. [29]

Clustering can be applied to samples, genes, or both. We will be applying clustering only on genes, since we are interested in grouping the genes and identifying their outliers, and not in clustering the samples. The algorithm that we used to produce the clustering, is the k-means algorithm with pearson distances.

The k-means algorithm partitions an input dataset into k-clusters, a user predefined value. It starts with k random clusters, and then moves along samples in order to minimize the distance of each sample to their respective cluster centroid, and maximize distance between samples. Samples are moved to the cluster with the shortest relevant distance to the cluster centroid. The k-means algorithm is repeated a number of times, every time starting with a random set of initial clusters, until an optimal cluster solution is obtained. The distances are also recalculated on each repetition. There are several distance types used with clustering algorithms, with the most common of them being the euclidean, manhattan, pearson correlation, eisen cosine correlation, spearman correlation and kendall correlation distance. [27] We used the pearson correlation distance, as described in [30], which measures the strength of association between two variables X and Y. The Pearson coefficient is defined as the covariance of X and Y divided by the product of their respective standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Given vectors x and y, respectively sampling X and Y and each of length n, the sample Pearson coeffcient $r_{x,y}$ is obtained by estimating the population covariance and standard deviations from the samples:

$$r_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

After the clustering is produced, a dendrogram is also displayed, which contains information about the correlation of the involved genes. A sample heatmap with clustering follows. [29]

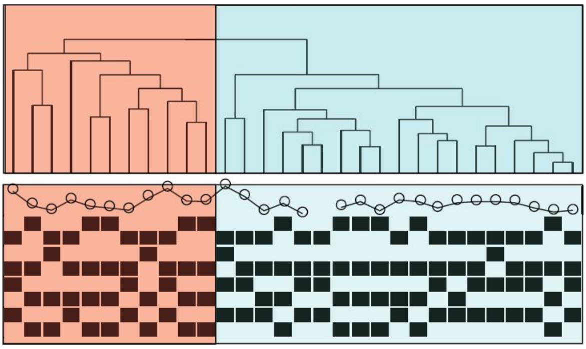


Figure 2: Sample heatmap with clustering. Dendrogram is displayed on top.

2.4 CLASSIFICATION METHODS

Supervised classification, also referred to as prediction, is defined as the process of developing algorithms in order to classify the input data to predefined categories. Algorithms have to undergo a training procedure, where they classify the features based on the samples of a training dataset, and then their accuracy is evaluated by testing the classifiers on a testing dataset. [27] Classification methods can be used to diagnose diseases or predict disease outcomes based on gene expression patters, extracted from microarray data. [6] Thus, developing reliable and accurate classifiers is essential for successful disease diagnosis and/or treatment. However, in the space of microarray data, where we have a multilevel feature space, the performance of most classification algorithms is poor, due to the excessive number of the classifiers. This problem is dealt with feature extraction and feature selection methods, which are described in their dedicated sections. Examples of classification methods are Support Vector Machines, k-Nearest Neighbours, Artificial Neural Networks (ANNs), Decision Trees, MLB Neural Networks, Bayesian, CART classification trees and Random Forrest. [27]

Artificial Neural Networks (ANNs) determine a network structure and learning parameters, by using various algorithms, which aim to produce sample weights, by minimizing an objective function. In each iteration, the estimation is compared to the real output, and then the local error is derived. This local error is used to adjust the input vector weights, according to a learning rule. The training stage is a time consuming process, since it iteratively trains and tests different networks on an independent test sample, eventually resulting to the network with the lowest error rate [31].

Decision Trees (or classification trees) on the other hand, are mainly used in data mining since they are able to discover hidden correlations among data. The aim of this method, is to create a binary tree by dividing the input vectors at each node, based on an evaluation function. One of the most popular decision trees method is the classification and regression trees (CART). CARTs begin by assigning all samples to root object and then they split each explanatory variable at all possible spliting points. Each sample is then split into two nodes, according to its corresponding splitting point. The explanatory variable and split point with the highest reduction of impurity are selected, and then are splitted according to the spliting point. The process is repeated until all nodes are set as parent nodes, and the tree reaches maximum size. The a tree-pruning is performed by using cross-validation, in order to result to the best-sized tree [31].

In order to evaluate the performance of the classification methods, some metrics are examined which give a thorough description about the classification ability of the examined method. The most commonly used metrics are accuracy, ROC curve, sensitivity (true positive rate), specificity (false positive rate) and mean square error (MSE).

- Accuracy, as the name suggests, is the ratio of the correct predictions to the total number of input samples, and is one of the most significant performance metrics.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ samples}$$

- ROC (Receiver Operating Characteristics) curve, is a curve that describes the model's ability to classify the input data correctly. It is a metric equivalent to accuracy, and it is sometimes used instead of the latter. Its values range from 0 to 1, with 1 being representing the maximum classification performance.

- Sensitivity or true positive rate, is defined as the ratio of positive samples that are correctly predicted as positive,with respect to all positive data samples.

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative}$$

- Specificity or false positive rate, is defined as the ratio of negative samples that are falsely predicted as positive, with respect to all negative data samples.

$$Specificity = \frac{False\ positive}{False\ positive + True\ negative}$$

- Mean squared error is defined as the average of the squared difference between the true values and the predicted values.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y}_j)^2$$

Amongst the various classification algorithms mentioned before, we decided to work with support vector machines (SVM) and k-nearest neighbours (KNN), which performed better in our analysis and are described below.

2.4.1 *Support Vector Machines*

Support vector machines (SVMs) is the first machine learning method that we will be using, for the classification process. SVMs belong to the group of supervised learning methods, and they can be used both for classification and regression.

Support vector machine is a powerful tool used for two-class classification and it targeted to be used as a non-linear mapping of the input vectors into a high-dimensional feature space. It relies on the idea of finding the maximum geometric margin between the two classes. One of the simplest types of support vector machines is linear classification, which attempts to set a straight line seperating data with two dimensions. A linear classifier is also reffered to as hyperplane. Various hyperplanes change achieve the same target, to seperate the two class data, but only one can achieve the maximum seperation. [27]

The basic principle of the learning procedure in SVM is to find a hyperplane which will seperate the data into two classes, and then try to maximize the margin between the two classes and the seperating hyperplane, whilst ensuring the accuracy of correct classification. The final binary classifier that is produced, is called optimal seperationg hyperplane. It does not suffer from local optima problem, i.e it works without a convex optimization problem. [20] [31]

SVM was initially designed for binary classifications problems, with many variations having been introduced for different purposes. [20] In case of linearly seperable data, the principle of SVM is described as follows. [20] The main goal of the training phase is to find the linear function :

$$f(\,x) \;= W^T X + b$$

which will be the line that will divide the data and the space to two different classes according to the condition:

$$W^T X + b > 0$$

$$W^T X + b < 0$$

These functions define the seperating plane, and the distance between the two parallel hyperplane equals to: $2/\|W\|^2$. This quantity is reffered to as the classification margin, as described in figure **??**. In order to maximize the classification margin, the algorithm is required to solved the following optimization problem:

- minimize $1/2\|W\|^2$
- subject to $Y_i(W^T X_i + b) \geq 1$

In case of non-linearly seperable data, SVM will have to work with more than two dimensions, and therefore will have to map the data from the input space into a high-dimensional feature space. The classes will then be seperated by an optimal hyperplane. [20] In order to perform this mapping, we will use a function called a kernel function. The four basic kernel functions are linear, polynomial, radial basic function (RBF) and sigmoid and they are described below:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T, x_j + r)^d, \gamma > 0$
- RBF: $K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2), \gamma > 0$
- Sigmoid: $K(x_i, x_j) = tanh(\gamma(x_i^T, x_j) + r)$

  where r, d and $\gamma = \frac{1}{2\sigma^2}$ are kernel parameters.

For non-linearly separable data, SVM requires the solution of the following optimization problem:

- minimize $1/2\|W^T\|^2 + C\sum_{i=1}^{n} \xi_i$

- subject to $Y_i(\ W^T X_i + b)\ \geq 1 - \xi_i$

- $\xi_i \geq 0$

In our approach, we used the RBF Kernel. A radial basis function is a real valued function, which only depends on the euclidean distance of a sample x_i and its center value x_j, and a tuning parameter σ. The centers are automatically computed by the SVM algorithm. The kernel's goal is to minimize the distance of each sample x_i from its center,which is achieved by calculating the value weights in each run. The successfulness of SVM strongly depends on the choice of the kernel function K, and of course the hyper parameters (σ is the case of RBF), therefore in order to adjust optimally these parameters we should perform a cross-validation procedure. [31]

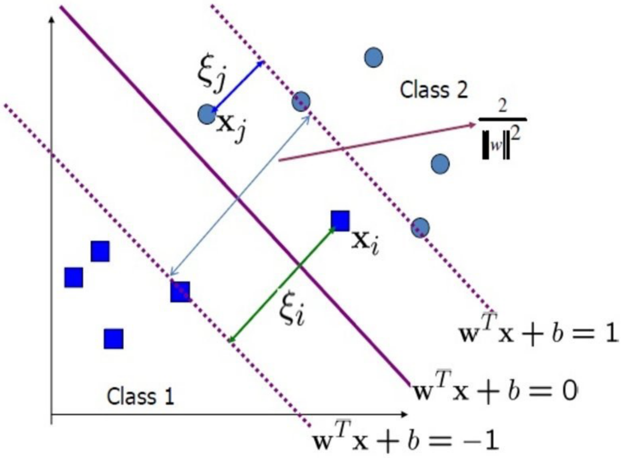

ξ_j
$\mathbf{x}_j$
Class 2
$\frac{2}{\|w\|^2}$
$\mathbf{x}_i$
ξ_i
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
Class 1

Figure 3: Maximum margin hyperplanes for SVM divides the plane into two classes

2.4.2 *K Nearest Neighbours*

K Nearest Neighbours (KNN) is the second machine learning method that we will be using, for the classification process. KNN belongs to the group of supervised learning methods, and it can be used for both classification and regression types of problems. KNN's target is to classify the outcome of a query point, by evaluating the values of a selected number of its nearest neighbours. The method estimates the outcome of a given query point, by finding k examples where their distance from the point is minimized (i.e. its neighbours). In case of classification problems, predictions are based on a majority of voting, while for regression problems, predections are determined by averaging the outcomes of the k nearest neighbours. While tuning the model, it is important to set the appropriate value of k (i.e. the neighbours taken into consideration), since it can strongly affect the accuracy of the predictions. For example, a small number of k will add large variance to predictions, while a large value of k may lead to a large model bias. The proper tuning of the k parameter is regulated by integrating a cross validation method, which will find the optimal k value.

The neighbours of a point are defined by a distance metric that we set beforehand. The most common is the Euclidean distance, while others possible metrics are Euclidean squared, City-block, and Chebychev distances. We will be using the Euclidian distance in this thesis, which is described by the following equation, according to Bishop, C. (1995) [32] :

$$D(x-p) = \sqrt{(x-p)^2}$$

where x is a query point and p is a case from the sample.

A widely used approach to improve the prediction accuracy is to introduce distance weights. This approach matches the closest to the point cases with large values of k. For every near neighbour, a new set of weights w is defined, where w refers to the relative closeness of each neighbour with respect to the query point. Weights are computed according to Bishop, C. (1995) [32] :

$$W(\,x, p_i) \;=\frac{exp(\,-D(\,x, p_i)\,)}{\sum_{i=1}^{k} exp(\,-D(\,x, p_i)\,)}$$

where $D(\,x, p_i)$ is the distance between the query point x and the ith instance of the sample p. All the weights sum to 1. In the classification problems, after the calculation of the weights, the case with the maximum weight w_{max} is set to the output value of the query point x. K nearest neighbours are able to achieve high classification accuracy and learn fast, even with a high-dimensional feature space, like the microarray data. [31]

2.4.3 *Cross Validation*

In order to validate our classification methods described earlier, it is necessary to have some test datasets, independent from the training datasets, that will be used to measure the classification error. However, since our datasets our significantly limited and hard to find, it is difficult to obtain independent datasets for testing, or weeken our training datasets by keeping out some samples for testing. A technique that will give a solution to this problem is V-fold cross validation. [27]

The concept behind cross-validation is the same as with a single holdout validation set, to estimate the model's predictive ability and performance on unseen data. It's basic principle is that it repeats the experiments multiple times by dividing the training dataset in "V" different parts every time, keeping one of them out for validation and using the others for learning. It does not require seperate test datasets, and it also does not reduce the training dataset. The training dataset is partitioned into "V" smaller datasets, called "folds". The default number of "V" is 10. In each repetition, 1 subset is kept out for testing and the remaining "V-1" are used for training. This procedure is repeated "V" times, resulting to a bigger test dataset and taking advantage of the full spectrum of the training dataset. It is worth mentioning that cross validation does not prevent overfitting in itself, but it may help in identifying a case of overfitting caused by the classification method. [27]

2.4.4 *Leave One Out Cross Validation*

As mentioned in previous sections, since we are working with microarray data, the number of samples are usually very small. Therefore, it is not the best practice to divide our subset our datasets to training and testing. Instead, the most commonly used technique to test our classification methods, is "V" fold cross validation, in it's most greedy case, called Leave One Out Cross Validation. [33] In Leave One Out Cross Validation (LOOCV) , the number of partitions equals to the number of dataset size (m). Each testing dataset consists of 1 sample, and each training dataset contains (m-1) samples, which are used to construct the classifier which is tested on the leftout sample. By repeating this process for (m) times, we use all samples as testing data samples, and we finally come up with m predictions. [20] The performance of the classifier is then evaluated by the average misclassification rate:

$$E_r = \frac{1}{m}\sum_{i=1}^{m} \delta(\ e_i, y_i)$$

where y_i is the true class label, for instance x_i , and

$$\delta(\ x_i, y_i) \ = \begin{cases} 0 & if x = y \\ 1 & if x \neq y \end{cases}$$

It is also worth mentioning that the variable selection needs to be performed on each repetition, with the remaining samples other than x_i, rather than pre-selecting the variables from the complete dataset and then validate on the test sample. This is important because the variable selection should rely only on the training and not the testing dataset. Performing the cross validation with preselected subset, could lead to misleading results. [33]

2.5 FEATURE SELECTION METHODS

Feature selection is the process of subseting a dataset with relevant and reduntant features, in order to improve the performance of the classification methods, regarding accuracy and time to construct the model. [6]It differs from the feature extraction process, as it selects a subset from already selected features, thus avoiding the drawback of the output interpretability. The feature selection methods are classified as filters, wrappers and embedded, depending on the methods used to evaluate the feature subsets. [20]

2.5.1 *Filter Methods*

Filter methods are widely used on gene ranking, as they have computational efficiency. They select the best subset by variable ordering, using variable ranking methods, implementing heuristic methods. They also use a ranking criterion of statistics, in order to score the variables and define a threshold value, discarding the variables under it. Their main drawback is that they are independent of the specific required prediction task. That means that they will select the features even if the latter don't fit in the classification model, thus making them unreliable. [6]

2.5.2 *Wrapper Methods*

Wrapper methods on the other hand, don't use feature relevant criteria like the filter methods. Instead, they depend on the performance of classifiers to obtain a feature subset. They use the predictive accuracy of a data mining method, to determine the fitness of a selected subset, by integrating the data mining method as a black box. The aim of this method is to find the subset with the maximum evaluation, by following a trial and error method. This approach forces the method to execute cross validation on small datasets in order to find the most accurate estimation, resulting in better overall performance. [6] On the downside, wrapper methods are very expensive regarding time and computations, when implemented on high dimensional feature space. [27]

2.5.3 *Embedded Methods*

The embedded methods were inspired as an attempt to combine the advantages of both filter and wrapper methods. Unlike the two previous ones, which seperate the feature selection and training process, the embedded methods integrate the feature selection methods into the costruction process of the classifier or regression model. [20] More specifically, embedded methods incorporate the feature selection as a part of the training process, while significantly reducing the computational time. The consider both relations between input and ouput features, and also search for features which allow better local discrimination. They use the independent statistical criteria used by filter methods, in order to obtain the optimal subsets of a known group of classifiers. After that, the classification method is used to select the optimal subset among the group of optimal subsets produced by the previous step. They can be categorized into three submethods, namely pruning method, built-in mechanism and regularization models. In the pruning method, all features are included in the training process initially, and then the ones with the smaller correlation coefficient values are recursively removed (pruned), using an SVM algorithm. In the built-in mechanism method, the features are selected by some supervised learning algorithms, in the training phase, while in the regularization method, the objective functions are used to minimize fitting errors and near zero regression coefficient features are eliminated.

Various feature selection techniques are suggested in the literature. LLDA based Recursive Feature Elimination (LLDA-RFE), kernel-penalized SVM (KP-SVM), discriminative least squares regression (LSR), Support Vector Data Description (SVDD) and Support Vector Machine - Recursive Feature Elimination (SVM-RFE) are some of the most significant ones. Feature selection methods are widely used in microarray data analysis due to their conceptual simplicity. However, as every algorithm, they come with some drawbacks. During the feature selection process, where most genes are eliminated, a large amount of information that is related to these genes is lost, while correlations between variables are not taken into consideration. These problems can be overcomed by selecting the optimal subsets according to a quality criterion instead of filtering out the redundant features. However, these methods will not perform as well on independent testing datasets, since they suffer from overfitting, and also implement some computational heavy algorithms ,which are difficult to integrate and interpret [6] .

2.5.4 *Support Vector Machine - Recursive Feature Elimination*

Support Vector Machine - Recursive Feature Elimination (SVM-RFE) is an embedded feature selection method. SVM-RFE is a widely used feature selection method specifically designed for microarray data analysis. The goal of this method is to determine a small subset of informative features that reduces processing time and provides higher classification accuracy. SVM-RFE uses the weight magnitude as ranking criterion. It works by repeadetly training an SVM classifier, with a subset of features, and in each iteration heuristically removing the features with the smaller feature weights. In each iteration, the parameters of the classification model (SVM) are reestimated, by implementing the method of cross validation. Also, a linear kernel is often used, with a proper parameter tuning, in order to achieve better classification accuracy. [25] An outline of the algorithm is presented below, in the case of a linear problem.

SVM-RFE Algorithm

1. Inputs:
   - Training examples: $X_0 = [x_1, x_2, ...x_k, x...x_l]^T$
   - Class labels: $y = [y_1, y_2, ..., y_k, y_l]^T$
2. Initialize:

- Subset of surviving features $s = [1, 2, ...n]$
- Feature ranked list $r = []$

3. Repeat until $s = []$
4. Restrict training examples to good feature indices:

$$X = X + 0(:, s)$$

5. Train the classifier: $\alpha = SVM - train(X, y)$
6. Compute the weight vector of dimension length(s):

$$w = \sum_k a_k y_k x_k$$

7. Compute the ranking criteria: $c_i = (w_i) ^2$, for all i
8. Find the feature with smallest ranking criterion: f = argmin(c)
9. Update feature ranked list: $r = [s(f) , r]$
10. Eliminate the feature with smallest ranking criterion:

$$s = s(1 : f - 1, f + 1 : length(s))$$

11. Output: Feature ranked list r.

More than one features can be eliminated in each step, in order to improve execution speed. [26]

Part II

PROBLEM AND APPROACH

3

PROBLEM STATEMENT

Problem statement follows. After the problem, the main points of our approach and implementation are exhibited as well.

3.1 PANCREATIC CANCER AND EMBRYOGENESIS

Pancreatic cancer begins when abnormal cells in the pancreas grow and divide out of control and form a tumor. The pancreas is a gland located deep in the abdomen, between the stomach and the spine. It makes enzymes that help digestion and hormones that control blood-sugar levels. Organs, like the pancreas, are made up of cells. Normally, cells divide to form new cells as the body needs them. When cells get old, they die, and new cells take their place. Sometimes this process breaks. New cells form when the body does not need them, or old cells do not die. The extra cells may form a mass of tissue called a tumor. A malignant tumor is called cancer. The cells grow out of control and can spread to other tissues and organs. A tumor is formed when DNA is subjected to changes. These changes happen according to some biological pathways. There are many types of biological pathways. Among the most well-known are pathways involved in metabolism, in the regulation of genes and in the transmission of signals.

Signal transduction pathways move a signal from a cell's exterior to its interior. Different cells are able to receive specific signals through structures on their surface called receptors. After interacting with these receptors, the signal travels into the cell, where its message is transmitted by specialized proteins that trigger a specific reaction in the cell. For example, a chemical signal from outside the cell might direct the cell to produce a particular protein inside the cell. In turn, that protein may be a signal that prompts the cell to move, or to replicate. Identifying what genes, proteins and other molecules are involved in a biological pathway can provide clues about what goes wrong when a disease strikes. Consequently, biological pathways are a significant field of study, to identify the cause of pdac and/or other types of cancer.

Embryogenesis, on the other hand, is a complex process that occurs during the first eight weeks after fertilization. The main idea is that a single cell is being transformed to an organism with a multi-level body plan. During these weeks, the embryo is undergoing some significant procedures, driven by some crucial signaling pathways. According to latest studies, there is a high relevance of the signaling pathways activated during embryogenesis, with those that cause pancreatic cancer, if the later get corrupted. That causes a major problem, since the cells retain memories of these processes, giving a high possibility of cancer to arise, if imperfections appear in these processes. [4]

3.2 GENE EXPRESSION ANALYSIS AND CANCER CLASSIFICATION

The lack of an analysis that will extract the genes involved in these processes and the correlated genes that are activated during embryogenesis and pancreatic cancer early stages, is a fact. It is of great interest to find some methods that will be able to do statistical analysis of the gene expression levels, and find the ones that are significantly differentiated. These process can be done by using mRNA gene expression microarray analysis. Analysis of mRNA gene expression is widely used to compare patterns of gene expression between cells or tissues of different kinds and under different conditions, for example between normal and cancer cells.

An important problem that arises is the huge number of genes included in the original datasets, as they are extracted from different platforms. This huge number of genes can affect the outcome of our work, as most of them are irrelevent to analysis, and it also inserts latency and worse accuracy to our prediction systems. Thus, it is crucial to filter out the genes that are not of interest, with data mining techniques, and work with smaller datasets that will contain genes relevant to our analysis. [27]

So the first goal of this thesis is to find the over/under expressed genes and consequently reduce the high-dimensional feature space, by applying feature extraction methods on various pdac and human embryos datasets. Subsequently, a cross-validation of the proposed pathways with the significant differentially expressed genes (DEG) will be performed.

The second goal of our analysis is to perform a classification of pancreatic cancer samples, based on the genes extracted from the feature extraction step. The proposed classification machine learning methods will be used in this step, with the extracted gene expression levels as classifiers. In order to impove the accuracy and the computational time of the classification algorithms, a feature selection will be performed as a last step.

No such study has been proposed yet, that will analyze the significant genes that participate in the signaling pathways of embryogenesis and pancreatic cancer, and this thesis comes to add to the literature an implementation of this process. Commonly used techniques in the fields of data mining, statistical analysis, machine learning on classification, feature extraction and feature selection will be implemented on gene expression microarray datasets, in an attempt to better identify and classify the incurable problem of pancreatic cancer in a molecular level.

4

RELATED WORK

The are a lot of independent, focused on different cancer types related studies that are worth mentioning. Yet, there are no similar studies that focus on the analysis of pancreatic cancer gene expression data and their correlation with the process of embryogenesis. Some of the most important studies regarding pancreatic cancer classification, feature extraction and feature selection techniques are mentioned in this section.

Wazir Muhammad et al [?], developed an artificial neural network (ANN) trained on various cancerous datasets, with pancreatic cancer patients among them. They focused on cancer prediction by incorporating features extracted from different cancer datasets to the neural network, and achieved high accuracy prediction.

Sarfaraz Hussein et al [34], proposed a deep learning approach, implementing both supervised and unsupervised learning methods, for lung and pancreatic cancer classification. They presented a framework for tumour determination with 3D screening based graph regularized sparse Multi-Task Learning (MTL), which can be used to obtain discriminative features for medical image analysis.

Eric Shadt et al [35], focused on developing some algorithms for performing feature extraction and normalization of high-density microarray gene expression datasets, in order to increase the sensitivity and specificity of detecting the presence of genes, and/or if they are marked as differentially expressed. They developed some feature extraction and normalization algorithms for the analysis of gene expression array data, and they achieved improved computation of gene intensities and expression ratios.

Terrence s. Furey et al [36], developed a new way to analyze high-dimensional microarray data, using SVMs. Their analysis involved classification of the tissue samples, and an review of the data for mislabeled or ambiguous tissue results. After computational analysis, the mislabeled tissue samples were detected, and perfect classification of tissues was achieved (but without high confidence), upon correction and removal of the mistaken outliers.

Chris Ding et al, [37], proposed a minimum redundancy - maximum relevance feature selection implementation. The selected genes cover the feature space is a more balanced way, while capturing broader characteristics of phenotypes. Improvements were observed among 4 machine learning methods (Naive Bayes, Linear discriminant analysis, Logistic regression, and Support vector machines), that were used for cancer classification.

Isabelle Guyon et al, [26] also proposed a new feature selection method, based on support vector machines - recursive feature elimination. The managed to yield better classification performance and biological relevance to cancer type. They improved significantly the baseline method which makes implicit orthogonality assumptions, and managed to verify the biological relevance of the selected genes by SVMs with cancer diagnosis.

5

OUR APPROACH

5.1 WORKFLOW OVERVIEW

This thesis contributes an analysis of pancreatic cancer gene expression microarray data, along with their correlation with data from human embryos. Common machine learning methods will be used, in an attempt for data mining (feature extraction/feature selection) and cancer classification, on gene expression microarray datasets.

The main tool used in this study is the R project, a language and environment for statistical computing and graphics [38]. R provides the user with a command line (cli), without any graphical user interface (GUI). The user communicates with the software via R scripts, or simple commands. A workspace is also available, where all variables and data are stored, in an .Rdata file. R also provides a package manager, from where the user can install packages which include various implemented functions, in order to use different functionality. Packages are downloaded from repositories, and can be installed and used locally. In our analysis, we downloaded and used functions from the packages: gcrma, oligo, dplyr, limma, gplots and caret. Some sample scripts used in our analysis are described in Appendix A.4.

The steps of our analysis are mentioned below. Each step is described in detail, in their dedicated sections.

1. Data acquisition from GEO database.
2. Data preprocessing.

3. Feature extraction.
4. Data visualization with heat maps / clustering.
5. Cancer classification.
6. Feature selection.

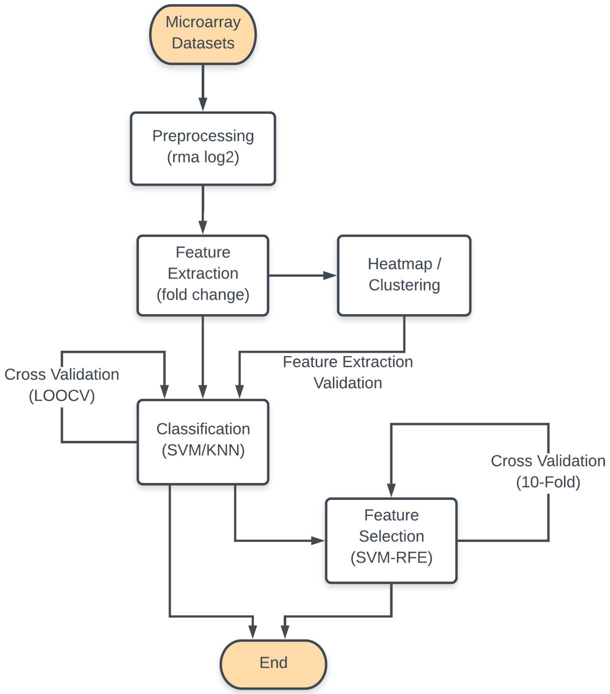
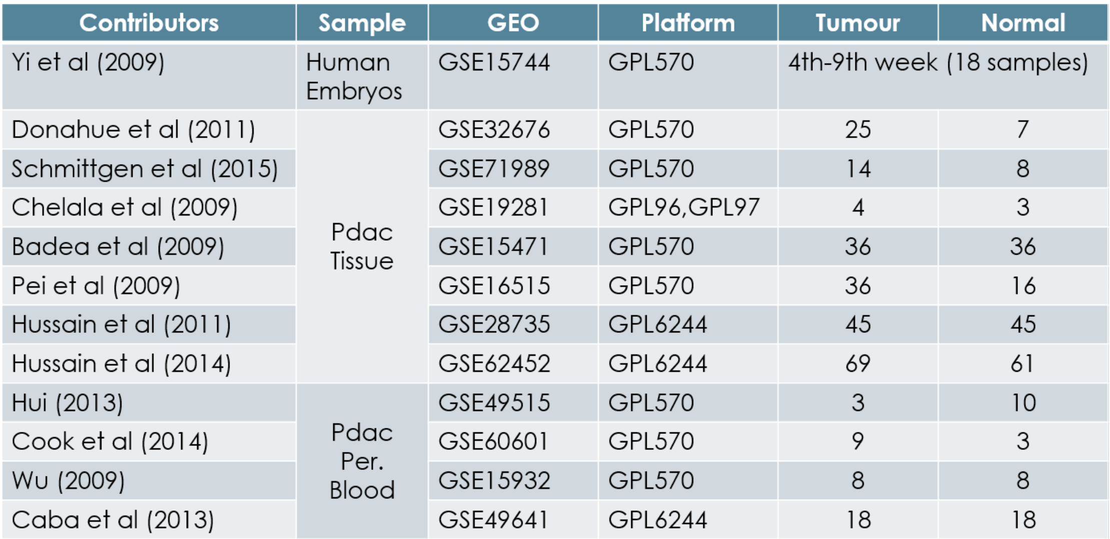

Microarray Datasets
Preprocessing (rma log2)
Feature Extraction (fold change)
Heatmap / Clustering
Cross Validation (LOOCV)
Feature Extraction Validation
Classification (SVM/KNN)
Cross Validation (10-Fold)
Feature Selection (SVM-RFE)
End

Figure 4: Workflow chart

5.2 ANALYSIS OF DATASETS

We will work with high dimensional gene expression microarray datasets in our study. All datasets are obtained from the GEO database [7]. 12 datasets will be used in total, with 11 of them containing gene expression levels from human tissue and peripheral blood of patients with pancreatic cancer, and 1 of them containing gene expression levels from the development of human embryos from week 4 through 9. They have run on 4 different Affymetrix

GeneChips® : GPL570, GPL96, GPL97 and GPL6244. [22]. Figure 5 gives an overview of the used datasets.



Contributors	Sample	GEO	Platform	Tumour	Normal
Yi et al (2009)	Human Embryos	GSE15744	GPL570	4th-9th week (18 samples)	
Donahue et al (2011)	Pdac Tissue	GSE32676	GPL570	25	7
Schmittgen et al (2015)		GSE71989	GPL570	14	8
Chelala et al (2009)		GSE19281	GPL96,GPL97	4	3
Badea et al (2009)		GSE15471	GPL570	36	36
Pei et al (2009)		GSE16515	GPL570	36	16
Hussain et al (2011)		GSE28735	GPL6244	45	45
Hussain et al (2014)		GSE62452	GPL6244	69	61
Hui (2013)	Pdac Per. Blood	GSE49515	GPL570	3	10
Cook et al (2014)		GSE60601	GPL570	9	3
Wu (2009)		GSE15932	GPL570	8	8
Caba et al (2013)		GSE49641	GPL6244	18	18

Figure 5: Overview of datasets

1. GSE15744 - Human embryos. This dataset contains gene expression levels of 3 embryos for each of the 4th, 5th, 6th, 7th, 8th, and 9th week of human embryonic development. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.
2. GSE32676 - Pdac tissue. This dataset contains samples from tumour tissue of 25 patients with pancreatic cancer (pdac) and 7 control (non-malignant) samples. The tissue samples come from early stage pdac patients. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.
3. GSE71989 - Pdac tissue. This dataset contains samples from tumour tissue of 14 patients with pancreatic cancer (pdac) and 8 control (non-malignant) samples. The tissue samples come from advanced stage pdac patients. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.

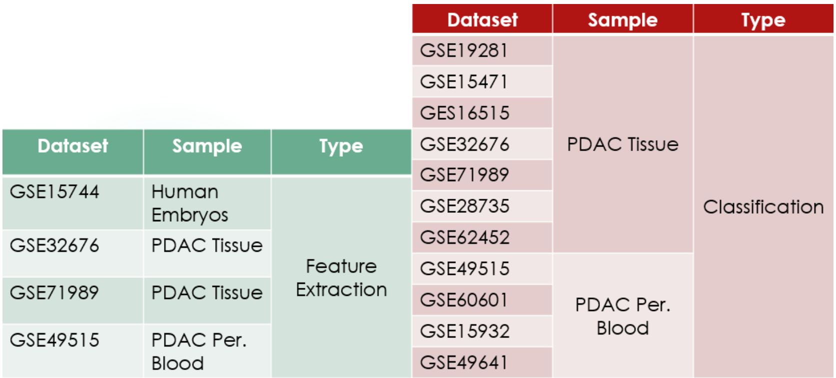
4. GSE19281 - Pdac tissue. This dataset contains various disease samples from tumour tissues. We have selected 4 pancreatic cancer (pdac) samples and 3 normal (non-malignant) pancreas samples. All samples run on two different Affymetrix GeneChips®, GPL96 and GPL97, which contain 22,283 and 22,645 gene expression levels respectively. The samples where combined, resulting to 44,928 gene expression levels.
5. GSE15471 - Pdac tissue. This dataset contains samples from tumour tissue of 36 patients with pancreatic cancer (pdac) and 36 control (non-malignant) samples. The samples were obtained at the time of surgery from resected pancreas patients. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.
6. GSE16515 - Pdac tissue. This dataset contains samples from tumour tissue of 36 patients with pancreatic cancer (pdac) and 16 control (non-malignant) samples. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.
7. GSE28735 - Pdac tissue. This dataset contains samples from tumour tissue of 45 patients with pancreatic cancer (pdac) and 45 control (non-malignant) samples. The experiment run on GPL6244 Affymetrix GeneChip® which contains 33,297 gene expression levels.
8. GSE62452 - Pdac tissue. This dataset contains samples from tumour tissue of 69 patients with pancreatic cancer (pdac) and 61 adjacent control (non-malignant) samples. The experiment run on GPL6244 Affymetrix GeneChip® which contains 33,297 gene expression levels.
9. GSE49515 - Pdac peripheral blood. This dataset contains samples from Peripheral blood mononuclear cell (PBMC) of 3 patients with pancreatic cancer (pdac) and 10 control (non-malignant) samples. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.

10. GSE60601 - Pdac peripheral blood. This dataset contains samples from Peripheral blood mononuclear cell (PBMC) of 9 patients with pancreatic cancer (pdac) and 3 control (non-malignant) samples. Classical CD14++ CD16- monocytes were isolated from the peripheral blood of healthy volunteers and patients with pancreatic ductal adenocarcinoma. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.

11. GSE15932 - Pdac peripheral blood. This dataset contains samples pancreatic cancer-associated diabetes mellitus. We only used the pancreatic cancer-without diabetes mellitus samples, from Peripheral blood mononuclear cell (PBMC) of 8 patients with pancreatic cancer (pdac) and 8 control (non-malignant) samples. The experiment run on GPL570 Affymetrix GeneChip® which contains 54,675 gene expression levels.

12. GSE49641 - Pdac peripheral blood. This dataset contains samples from Peripheral blood mononuclear cell (PBMC) of 18 patients with pancreatic cancer (pdac) and 18 control (non-malignant) samples. 18 patients with unresectable PDAC were recruited. Instead of extracting tumour tissue, (PBMC) was obtained for study purposes. The experiment run on GPL6244 Affymetrix GeneChip® which contains 33,297 gene expression levels.

Four of these datasets are used for the process of feature extraction, while all of the pancreatic cancer datasets are used for the classfication process, as shown in figure 6.



Dataset	Sample	Type
GSE15744	Human Embryos	Feature Extraction
GSE32676	PDAC Tissue	
GSE71989	PDAC Tissue	
GSE49515	PDAC Per. Blood	

Dataset	Sample	Type
GSE19281	PDAC Tissue	Classification
GSE15471		
GES16515		
GSE32676		
GSE71989		
GSE28735		
GSE62452		
GSE49515	PDAC Per. Blood	
GSE60601		
GSE15932		
GSE49641		

Figure 6: Feature extraction/Classification datasets

We obtain the raw data for each dataset. The raw data come in a .zip format, which contains the CEL files regarding each sample. A CEL file is a data file created by Affymetrix DNA microarray image analysis software. It contains the data extracted from the probes on an Affymetrix GeneChip®. However, the gene expression levels are raw values, as extracted by the each platform, which need to be preprocessed in order to be further analyzed. The preprocessing step follows.

5.3 PREPROCESSING

Raw gene expression microarray data, contain information about the measured intensities and probe locations on the microarray. However, in order to obtain the gene expression values that will be used later in our analysis, a preprocessing stage is required. Each obtained dataset is also available with preprocessed values, where different methods or R packages have been used i.e quantile normalization, global scaling etc. Yet, we cannot use these normalized expression levels, since the have not been normalized with the same methods, thus they cannot be compared. Instead, we apply robust multi-array average (RMA) on all datasets, a widely used normalization method on microarray gene expression data. RMA works in three steps: it background corrects, in normalizes using quantile normalization and it log_2 transforms.

We preprocess our data using the package gcrma (R/Bioconductor) [39]. However, GCRMA does not support GPL6244 since the mismatch probes are missing on this platform. Package oligo (R/Bioconductor) [40] comes of use in this case, since it does not require the presence of mismatch probes, and is able to background correct the gene expression levels without them. Therefore, we used both gcrma and oligo packages, repeating the preprocessing stage for datasets that ran on GPL570 and GPL96,GPL97 Affymetrix GeneChips®.

RMA starts with background correction of the gene expression values. Instead of storing proble intensities, the probe affinities are computed and stored. Each probe affinity is computed by obtaining the base-position profiles from nonspecific binding data, where the base-position profiles refer to the contribution of each base type at each position along the probe. Subsequently a quantile normalization of the data is taking place, followed by a log_2 transformation. Gene expression data are log_2 transformed, in order to model proportional chances rather than additive changes, which is typically more biologically relevant [39]. Log transformation has also the advantage of producing a continuous spectrum of values.

After preprocessing of the data is complete, we proceed with feature extraction techniques.

5.4 FEATURE EXTRACTION

Feature extraction is a necessary step in microarray gene expression dataset analysis, and the first of our two goals for this thesis. It aims on reducing the high-dimensional feature space to a lower one, by filtering out all the irrelevant to the analysis features. Statistical methods are widely used on feature extraction in microarray datasets, and especially the identification of differentially expressed genes (DEGs). We will perform a DEGs analysis on our datasets, and then confirm the results by visualizing the extracted genes with heatmaps and clustering.

5.4.1 *Differentially Expressed Genes Analysis*

Identification of differentially expressed genes is commonly performed with the statistical measure log_2 fold change. It measures how much an expression is changing between two distinct groups, pdac and normal in our case. Fold change is examined along with the false discovery rate (FDR), which is tuned by the p-value.

The datasets that will be used in the feature extraction process are two pdac tissue datasets (GSE32676 and GSE71989), one peripheral blood dataset (GSE49515) and the human embryos dataset (GSE15744). All the used functions are provided by the limma (R/Bioconductor) package[41].

We start off by loading the normalized data, for each of the three pdac datasets. We then create two factor levels (pdac and normal) for our model. A linear model is then fit on each gene given a series of arrays, using the function *lmFit*. Contrasts are created for all the levels, which express the difference of the two factors (pdac-normal), by using the function *makeContrasts*. We then compute moderated t-statistics, moderated F-statistic and log_2 fold change of the differential expressed values, by the empirical Bayes moderation of the standard errors towards a common value, using the function *eBayes*.

After we have calculated log_2 fold change for all gene expression levels, we sort the genes by their log_2 fold change (lfc) values in an descending order. We also take into consideration the false discovery rate (FDR), which is an important statistical measure that will filter out the falsely DEG called genes. We adjust the FDR by the Benjamini and Hochberg method (BH). FDR is regulated by the p-value, which we set to 0.01, or 1% FDR tolerance. That means that genes with FDR > 1% , which are probably considered as DEG, are filtered out making the feature extraction process more accurate. We examine the genes with lfc > 2 ($2log_2$ fold change). P-value and lfc attributes are passed to the *topTable* function, which returns the $2log_2$ fold change genes for each of the three pdac datasets.

Thing differ for the human embryos dataset (GSE15744), where we have more than two factor levels. This particular dataset contains 3 human embryos samples for each of the weeks 4 through 9 of embryonic development. Thus, we examine the progress of the gene expression levels, and we need a factor level for each week compared to its next i.e week 4-week 5, week 5-week 6 etc. Subsequently, we follow the same steps as in the pdac datasets.

5.4.2 *Heatmaps/Clustering*

In order to confirm the validity of our DEGs, we proceed with data visualization. Heatmaps and clustering are a widely used method for gene expression data visualization, since they point out the expression level differences and combine the genes in clusters, giving the user the ability to extract useful information about the quality and the characteristics of the input data.

We use the functions *hclust* and *heatmap.2*, provided by the R package gplots[42]. We start by calculating the clusters of the input data, where the pearson distance is used to calculate the distances between genes. We then set the number of clusters that we want our data divided in, by cutting the clustering tree to a specific height. Pseudocoloring follows, where all gene expressions are matched in a red-green spectrum, with red referring to the overexpressed values and green to the underexpressed ones. The heatmap is finally drawn by using the *heatmap.2* function. A dendrogram is also displayed, showing the correlations between genes that led to the current clustering. We create the heatmaps of the 4 datasets used for the feature extraction, from which we assess the produced DEGs. The mentioned heatmaps are under section 6.1.

5.4.3 *Extracted Genes*

The last step in the feature extraction process is to combine the DEGs extracted from all 4 examined datasets. We conclude to the final list of DEGs for the gcrma analysis, by examining the intersection of the datasets, as described in the VENN diagram below.

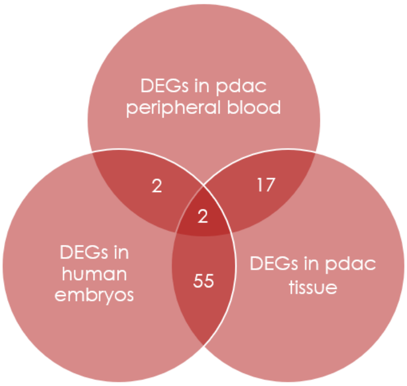
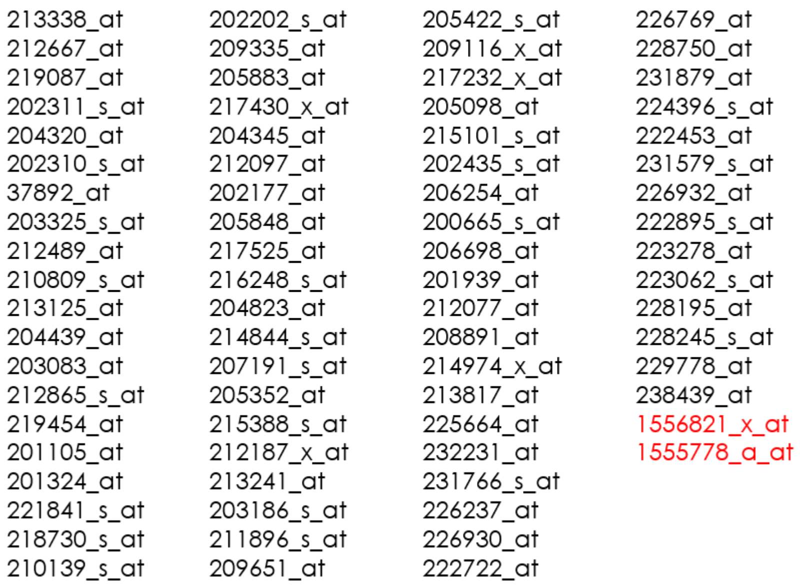


Figure 7: VENN diagram for extracted DEGs in gcrma analysis

- human embryos ∩ pdac tissue ∩ pdac per. blood : 2 DEGs
- human embryos ∩ pdac tissue : 55 DEGs
- human embryos ∩ pdac per. blood : 2 DEGs
- pdac tissue ∩ pdac per. blood : 17 DEGs

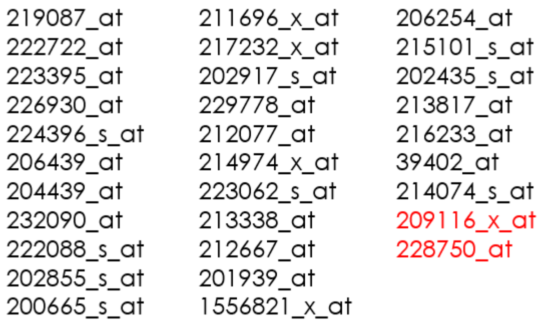
The final list of the extracted DEGs that we will use for the classification and feature selection methods, emerges from the combination of the 4 gene sets described above, which gives us a total of 76 DEGs. The same process is followed for the oligo analysis, which gives us a list of 31 DEGs.

With the process of feature extraction we have managed to reduce the high-dimensional feature space (of 50,000) genes to a lower one (of 100 genes), by filtering out the irrelevant to our analysis genes. The next step is to filter our datasets with the extracted lists of DEGs and keep the genes that will be used as predictors for the classification process. Since the experiments that produced our datasets have run on 4 different platforms, each one with a different number of genes, a problem that arises is that some genes are not present in all 4 platforms. Two DEGs included in the list of 76 genes (for the gcrma analysis),were not present on both GPL570, GPL96 and GPL97. Also two DEGs included in the list of 31 genes (for the oligo analysis) were not present on both GPL570, GPL96, GPL97 and GPL6244. Therefore, we exclude these 4 genes from our lists, and we proceed in the cancer classification methods with 74 and 29 DEGs respectively. These excluded genes that are not present in all platforms, are marked with red color in the following tables.



213338_at
212667_at
219087_at
202311_s_at
204320_at
202310_s_at
37892_at
203325_s_at
212489_at
210809_s_at
213125_at
204439_at
203083_at
212865_s_at
219454_at
201105_at
201324_at
221841_s_at
218730_s_at
210139_s_at
202202_s_at
209335_at
205883_at
217430_x_at
204345_at
212097_at
202177_at
205848_at
217525_at
216248_s_at
204823_at
214844_s_at
207191_s_at
205352_at
215388_s_at
212187_x_at
213241_at
203186_s_at
211896_s_at
209651_at
205422_s_at
209116_x_at
217232_x_at
205098_at
215101_s_at
202435_s_at
206254_at
200665_s_at
206698_at
201939_at
212077_at
208891_at
214974_x_at
213817_at
225664_at
232231_at
231766_s_at
226237_at
226930_at
222722_at
226769_at
228750_at
231879_at
224396_s_at
222453_at
231579_s_at
226932_at
222895_s_at
223278_at
223062_s_at
228195_at
228245_s_at
229778_at
238439_at
1556821_x_at
1555778_a_at

Figure 8: 76 DEGs as extracted from the gcrma analysis. The genes are described by their probe set identifiers on gpl570 platform (see also figure 55, 61, 62 of Appendix A.3).



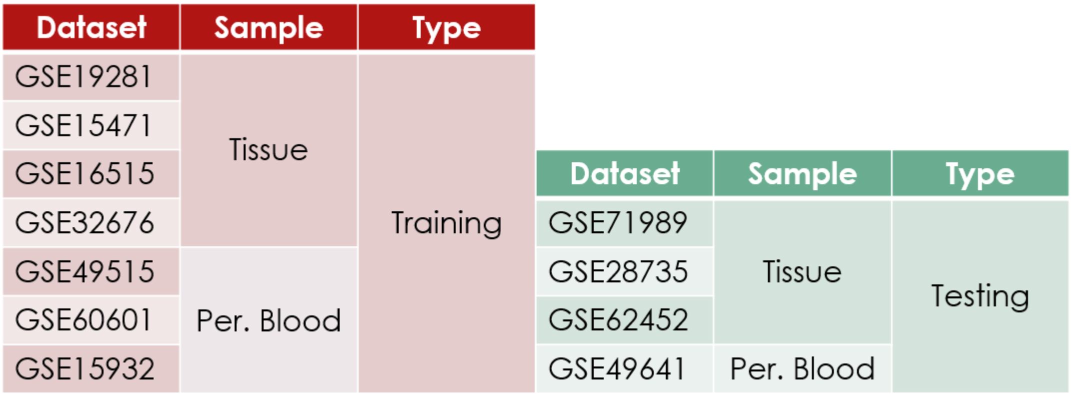
219087_at
222722_at
223395_at
226930_at
224396_s_at
206439_at
204439_at
232090_at
222088_s_at
202855_s_at
200665_s_at
211696_x_at
217232_x_at
202917_s_at
229778_at
212077_at
214974_x_at
223062_s_at
213338_at
212667_at
201939_at
1556821_x_at
206254_at
215101_s_at
202435_s_at
213817_at
216233_at
39402_at
214074_s_at
209116_x_at
228750_at

Figure 9: 31 DEGs as extracted from the oligo analysis. The genes are described by their probe set identifiers on gpl570 platform (see also figure 56, 63, 64 of Appendix A.3).

These extracted differentially expressed genes, are genes that are activated during the process of embryogenesis, and are also significantly transformed in pancreatic cancer patients. This analysis of examining the correlation of DEGs between pdac patients and human embryos, can lead us to conclusions about the genes that are involved in the common signaling pathways of embryogenesis and pancreatic cancer.

5.5 CLASSIFICATION

Pancreatic cancer classification is the second of our two goals for this thesis. In this step, we will try to classify the subjects into two classes, patient and healthy, based on the DEGs which will serve as classifiers. Various classification methods for microarray gene expression data are proposed in the literature, with some of them achieving high accuracy levels. After experimenting with different algorithms, like artificial neural networks, decision trees, deep learning and random forrest, we decided to work with support vector machines (SVM) and k-nearest neighbours (KNN), since they performed better on our data. We will also use leave one out cross validation (LOOCV) to calculate the accuracy of our classifiers. The LOOCV procedure involves keeping a sample out of the training dataset, building the decision function



Dataset	Sample	Type
GSE19281	Tissue	Training
GSE15471		
GSE16515		
GSE32676		
GSE49515	Per. Blood	
GSE60601		
GSE15932		

Dataset	Sample	Type
GSE71989	Tissue	Testing
GSE28735		
GSE62452		
GSE49641	Per. Blood	

Figure 10: Datasets used for training and testing

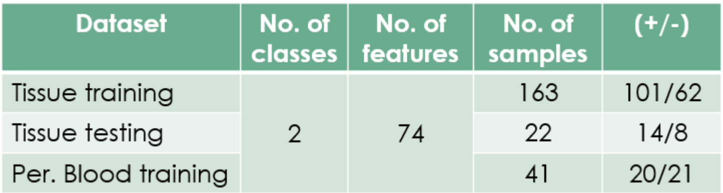
according to the remaining samples, and then testing on the removed sample. The process is repeated until all samples are kept out for testing, and the classification accuracy of the model is derived by averaging the classification rate of all repetitions.

These classification methods will be applied to 11 different two class (binary) microarray gene expression datasets. We divide the datasets to training and testing ones, according to figure 10. The features that will be used as classifiers for our models, are the DEGs genes extracted from the feature extraction step. We will perform a two class classification, where the two classes are *patient* and *healthy*.

We begin by creating the training datasets for the gcrma analysis. *GSE19281*, *GSE32676*, *GSE15471* and *GSE16515* are combined to create the tissue training dataset. *GSE49515*, *GSE60601* and *GSE15932* on the other hand, are composing the peripheral blood training dataset. The training and testing datasets for the gcrma analysis are described in figure 11. A binary variable *Patient* is also added to both training and testing datasets, which describes the quality of the samples.

$$Patient = \begin{cases} 0 & , Healthy \\ 1 & , Patient \end{cases}$$

This variable divides the data into two classes, based on which the classification will be performed.



Dataset	No. of classes	No. of features	No. of samples	(+/-)
Tissue training	2	74	163	101/62
Tissue testing			22	14/8
Per. Blood training			41	20/21

Figure 11: Training and testing datasets for the gcrma analysis. The positive class refers to patient subjects

We use the package *caret* (classification and regression training) [43], which provides us with the models and the functions for our classification methods. After the datasets are prepared for the classification process, we set the resampling process using the function *trainControl*. We use the leave one out cross validation (LOOCV) as the resampling process, in order to estimate our algorithms' ability without further dividing the already small sampled datasets to training and testing ones.

5.5.1 *Support Vector Machine*

The first classification algorithm that we will use, which is widely used for cancer classification in microarray data, is the support vector machines (SVM). It typically follows these steps:

1. A hyperplane that seperates the data in two classes is found.
2. The algorithm runs recursively in order to maximize the margin between the data and the hyperplane.
3. The mapping of the input data to the high-dimensional feature space is performed by a kernel function.
4. The kernel function is tuned by kernel parameters.
5. The tuning parameters are optimized by the process of cross-validation.
6. After the optimal tuning parameters are derived, class predictions are made for all samples.

7. Total accuracy level is estimated by computing the average classification rate from all repetitions.

We consider the RBF kernel as our kernel function. The optimal parameters C and σ are found through cross validation. We use a grid search, where a set of default values are predefined for the two variables. Pairs of (C, σ) are tried in each step, and the optimal one with the best cross-validation accuracy is selected. Typical values for C and σ are $C = (0.75, 0.9, 1, 1.1, 1.25\ )$ and $\sigma = (0.01, 0.015, 0.2\ )$. The grid of parameters C and σ is passed to the function *train*, along this the training dataset, the training method (in our case SVM with RBF) and the train control method which in our case is the LOOCV. Class predictors are built through a repetitive process, where the optimal set of parameters is selected and the final accuracy level is estimated by averaging the classification rates from all repetitions of the cross validation process.

Class predictors are built, which leads us to the testing part, where we evaluate our model's classification ability. Function *predict* extracts the predictions and class probabilities from the trained model, on a testing dataset. We first validate our model on the training data, and then on the testing tissue data (GSE71989). *Confusion matrix* gives us the classification results, by calculating a cross-tabulation of observed and predicted classes with associated statistics, containing information about various metrics that evaluate our models. The results of all the classification methods for the different training datasets, are presented and discussed in chapter 6.

Subsequently, before we proceed with peripheral blood datasets classification, we make a second attempt with pdac tissue datasets. This time, we shuffle the training dataset, and we keep out the 10% of the samples for testing, while the other 90% is used for training. This is a confirmation step in order to further evaluate our model's classification ability.

We continue with the pdac peripheral blood datasets. *GSE49515, GSE60601* and *GSE15932* are combined to create the peripheral blood training dataset. In the gcrma analysis we will only test on the training dataset, since the two testing datasets have run on the not supported by gcrma, GPL6244 platform. The exact same process is followed in both shuffled tissue and pdac peripheral blood datasets, as in pdac tissue SVM classification, with the results being presented in chapter 6.

5.5.2 *K Nearest Neighbours*

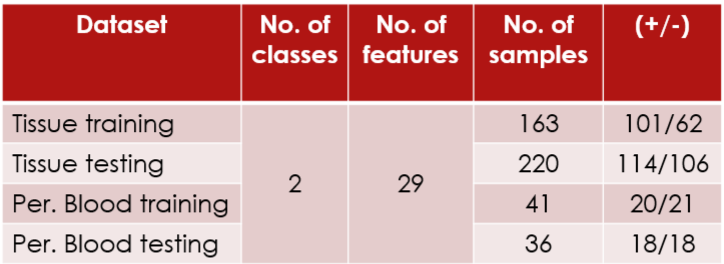
The second classification algorithm that we will use, which is also widely used for cancer classification in microarray data, is the k nearest neighbours (KNN). It typically follows these steps:

1. Find k samples where their distance from a query point are minimized (i.e its neighbours).
2. The k parameter is set by the user, and it represents the number of the neighbours that will be examined.
3. The k parameter is decided by a cross validation method, among a user defined set of values, since it strongly affects the classification ability of the algorithm.
4. Distance weights are calculated for every neighbour, where w refers to the relative closeness of the sample with respect to the query point.
5. The maximun weight is set to the output value of the query point x.
6. The outcome of the query point is predicted by averaging the outcomes of the k nearest neighbors.

We use the Euclidean distance as the distance metric between each query point and its neighbours. The goal of this classification method is to find the appropriate number of neighbours, and predict the outcome of each sample by averaging the outcomes of its neighbours. Thus, it is crucial to find the correct neighbours, which give the best classification accuracy. This process is performed by using the LOOCV method, which is defined in the *trainControl* function, similar to the SVM method. We then fit the knn model by using the *train* function, where we set the method to knn and the train control to LOOCV. The last attribute that needs to be passed in the *train* function, is the k parameter, which will tune the algorithm. There are two ways to tune an algorithm in the caret R package. The first one is by defining a tuneGrid, as we did in the SVM tuning, and the second one is by allowing the method to tune it automatically. This can be done by setting the *tuneLength* variable to a number, which indicates the number of different values to try for the k parameter. We set a relatively large number of *tuneLength* = 20, so as to let the algorithm try 20 different values for k, in each repetition of the cross validation process. *TuneLength* makes a guess of what values to try, by using random selection to find the optimal model parameters. The model is trained, with the cross validation method, and the final classification accuracy is estimated by averaging the class prediction rates of each model run.

Similarly to the process of the SVM classification, we validate our model via the functions *predict* and *confusionMatrix*. KNN classification was performed both for pdac tissue and peripheral blood datasets. The model's classification ability is discussed in chapter 6.

The exact same process of SVM and KNN classification, was followed for the oligo analysis. The only difference is the examined datasets, since this analysis also supports (among others) the GPL6244 Affymetrix GeneChip®. Thus, three more datasets have been added in this analysis, which will be used for testing. We also decided not to use GSE71989, which was used for tissue testing in the gcrma analysis, and only use the two new datasets for tissue testing. The training and testing datasets for the oligo analysis are described in figure 12.



Dataset	No. of classes	No. of features	No. of samples	(+/-)
Tissue training	2	29	163	101/62
Tissue testing			220	114/106
Per. Blood training			41	20/21
Per. Blood testing			36	18/18

Figure 12: Training and testing datasets for the oligo analysis. The positive class refers to patient subjects

The two classification methods (SVM and KNN) were tuned as in the gcrma analysis, while LOOCV was used as a cross validation method as well. All the oligo classification results, along with their comparison to the gcrma analysis, are discussed in chapter 6.

5.6 FEATURE SELECTION

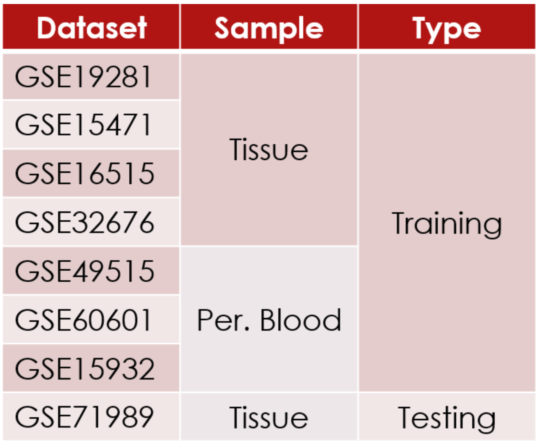
Feature selection is the final step of our analysis. From a clinical perspective, the examination of redundant gene expression levels, may not improve clinical decisions, but result to larger medical examination costs needlessly. Feature extraction and feature selection methods aim on deriving a gene signature from a minimum number of genes, which are highly related with the examined disease. These methods also result to higher accuracy in classification and prediction algorithms [25]. Thus, we conclude our analysis by implementing a feature selection method, from which we expect to derive a gene signature with the least possible number of genes.

We perform the feature selection by using the algorithm of support vector machines - recursive feature elimination (SVM-RFE), which was specifically designed for microarray data. SVM-RFE belongs to the category of the embedded feature selection methods. It incorporates the feature selection (recursive feature elimination) as a part of the training process (support vector machines). More specifically, it aims on determining a smaller than the input dataset, of equally informative/significant features. SVM-RFE also uses the weight magnitude as a ranking criterion. It typically follows these steps:

1. Start with the full input dataset.
2. Train an SVM classifier based on the full dataset features and assign ranking weights to all features.
3. In every run, remove heuristically a specific number of features (set by the user), which have the smaller weights.
4. Recursively repeat the process with the subset of features.
5. Cross validate the classifiers by repeating the experiment k-times.
6. Select the optimal subset of features, which results to the best classification accuracy.

We implement the SVM-RFE process only for the gcrma analysis. We begin by creating the training datasets, for the pdac tissue and peripheral blood. The datasets used for the feature selection process are described in figure 13. The binary variable *Patient* is also added to these datasets, which will divide the samples in two classes. SVM is used in combination with the RBF kernel, and the tuning of the kernel is set by a tuning grid, similarly to the SVM model used in the classification section. We work with 10-fold cross validation for this algorithm, and we set each experiment to be repeated 5 times in order to eliminate statistical variations, since the LOOCV increases significantly the time complexity of our algorithm.

Another parameter that has to be set beforehand, is the number of redundant features that we want to be removed in each iteration. We set that number equal to 5, which means that in every repetition of each fold the features with the worse weight vector will be recursively reduced by 5, until all features are removed. Then the optimal subset is selected for each repetition. The average classification rate of the 5 repetitions is computed, which produces the v-fold classification rate. In each run, the dataset is partitioned in 10 partitions, where the 9 are used for



Dataset	Sample	Type
GSE19281	Tissue	Training
GSE15471		
GSE16515		
GSE32676		
GSE49515	Per. Blood	
GSE60601		
GSE15932		
GSE71989	Tissue	Testing

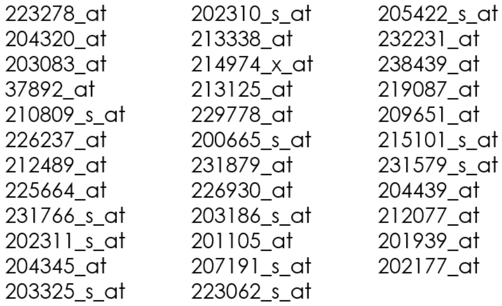
Figure 13: Datasets used for SVM-RFE training and testing

training and 1 for testing (10-fold cross validation). A feature weight vector is learned, based on the training dataset, and the top-ranked features are fed into SVM, while recording the classification accuracy. After the 10-fold cross validation process, the final classification accuracy rate for the model is computed from the average rates of all 10 folds. The last parameter that needs to be set is the *rfeControl*. This parameter defines the way that the feature elimination will be performed. We set this parameter to *caretFuncs*, a package of helper functions that take over the backwards feature selection process.

The process of SVM-RFE is performed by the caret function *rfe*. This function simultaneously calculates the SVM classifiers and proceeds with the recursive feature elimination, according to a cross validation method. In our case, the input parameters for the *rfe* function are the tissue training dataset, the trainControl (10-fold CV), the redundant sizes (5 in each step), the rfeControl (the caretFuncs that will perform the backwards feature elimination), the training method (SVM with RBF kernel), and the SVM tuning grid that is responsible for the tuning of the SVM parameters C and σ.

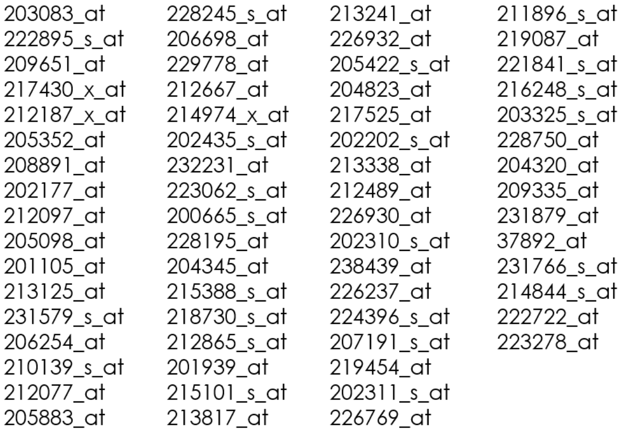
After the training is complete, we result with an optimal subset of features, that also achieve high classification accuracy. The optimal variables found for pdac tissue are 35 genes, extracted from the initial dataset of 74 genes. The 35 genes are listed in figure 14. The next step is to test our SVM classifiers after the RFE process, namely the 35 optimal genes. We test the classification ability of our predictors on both the training dataset and on GSE71989 which is used for testing. The procedure followed for the testing is identical to the one used in the SVM classification section.

The same steps are followed for the SVM-RFE process on the pdac peripheral blood datasets. The training dataset is also used for testing in this case, since we do not have more independent testing datasets in the gcrma analysis. The feature selection process results to 65 optimal variables, presented in figure 15, a subset slightly smaller than the original set of 74 features. The results of both pdac tissue and peripheral blood SVM-RFE models are presented and discussed in chapter 6.



223278_at
204320_at
203083_at
37892_at
210809_s_at
226237_at
212489_at
225664_at
231766_s_at
202311_s_at
204345_at
203325_s_at
202310_s_at
213338_at
214974_x_at
213125_at
229778_at
200665_s_at
231879_at
226930_at
203186_s_at
201105_at
207191_s_at
223062_s_at
205422_s_at
232231_at
238439_at
219087_at
209651_at
215101_s_at
231579_s_at
204439_at
212077_at
201939_at
202177_at

Figure 14: SVM-RFE optimal feature selection subset on pdac tissue datasets (35 genes) (see also figure 57, 65, 66 of Appendix A.3).



203083_at
222895_s_at
209651_at
217430_x_at
212187_x_at
205352_at
208891_at
202177_at
212097_at
205098_at
201105_at
213125_at
231579_s_at
206254_at
210139_s_at
212077_at
205883_at
228245_s_at
206698_at
229778_at
212667_at
214974_x_at
202435_s_at
232231_at
223062_s_at
200665_s_at
228195_at
204345_at
215388_s_at
218730_s_at
212865_s_at
201939_at
215101_s_at
213817_at
213241_at
226932_at
205422_s_at
204823_at
217525_at
202202_s_at
213338_at
212489_at
226930_at
202310_s_at
238439_at
226237_at
224396_s_at
207191_s_at
219454_at
202311_s_at
226769_at
211896_s_at
219087_at
221841_s_at
216248_s_at
203325_s_at
228750_at
204320_at
209335_at
231879_at
37892_at
231766_s_at
214844_s_at
222722_at
223278_at

Figure 15: SVM-RFE optimal feature selection subset on pdac peripheral blood datasets (65 genes) (see also figure 58, 67, 68 of Appendix A.3).

5.7 IDENTIFICATION OF POTENTIAL PANCREATIC CANCER BIOMARKERS

Pancreatic cancer is one of the most dangerous cancer types, and accounts for many deaths every year. The only curative option is the complete surgical resection, a difficult surgical procedure that only 15% of the patients can undergo. Thus, scientific research is focused on the early diagnosis and prognosis of pancreatic cancer (pdac), a critical step in impoving the survival rates. Yet, early diagnosis is difficult and currently inadequate, since patients remain asymptomic until the cancer has reached advanced stages. Furthermore, specific symptoms that are associated with pdac are not yet discovered [44].

In an attempt on improving diagnosis and prognosis of pancreatic cancer, scientists study it in a molecular level, where potential biomolecular markers are examined that could indicate the presence of the disease. These markers contain information that could be useful for early cancer detection, since they distinguish different tumour types from normal cells. Potential tumour biomarkers are extracted from analysis of gene expression data. Systematic analysis of gene expression levels of tumour data can reveal novel tumour markers and associate them with different tumour types. Suboptimal markers can also be combined in order to yield higher sensitivity and specificity.

Several potential proteins and markers have been identified as pancreatic cancer markers, using gene expression array analysis. Their combination can lead to adequate sensitivity and specificity levels for cancer classification and diagnosis. Yet, in order to reach safe conclusions about the validity of these markers, further validation is applied in large scale studies. [44]

In our approach, we examined pdac tissue datasets, from where we extracted some potential pancreatic cancer biomarkers. Tumour markers have contributed to pancreatic cancer treatment, as they are used to monitor the disease progression during chemotherapy or reccurance after surgery. However, they are not effective in early disease detection, since the elevated tumour marker levels indicate the high concentration of cancer cells. In addition, the resection of pancreatic tissue is not an easy procedure, and can not be performed in a regular basis for disease prognosis. Thus, it is crucial to find non-invasive, fast and cost-effective methods, that will contribute in early prognosis and diagnosis of pancreatic cancer.

Fortunatly, a large number of biomarkers can be also found in the serum, making gene expression analysis on blood datasets an attractive field of study. Serum tumour biomarkers are substances produced by tumour cells, which are released into the bloodstream. The measurement of these markers is relatively simple and inexpensive to perform, compared to the invasive methods of pancreatic tissue resection, since it only requires blood extraction. Blood extraction is a non-invasive process, fast, safe to collect, containing widely readable information and can be applied to large populations. Transcriptomic or metabolomic biomarkers, which are collected from the serum (blood or saliva samples), are used for disease diagnosis and prognosis [2]. CEA3 is used as a prognostic marker for various cancer types, though it lacks the required sensitivity and specificity for a presymptomatic marker. CA19-9 is considered the best pancreatic cancer marker found in the serum, despite the fact that it also has limited sensiticity and specificity levels [45].

Thus, we considered the examination of peripheral blood datasets a crucial step in our analysis, in order to suggest some potential biomarkers, that could be used for prognosis and diagnosis of pancreatic cancer. The biomarkers and their application in the clinical setting is summarized in figure 16.

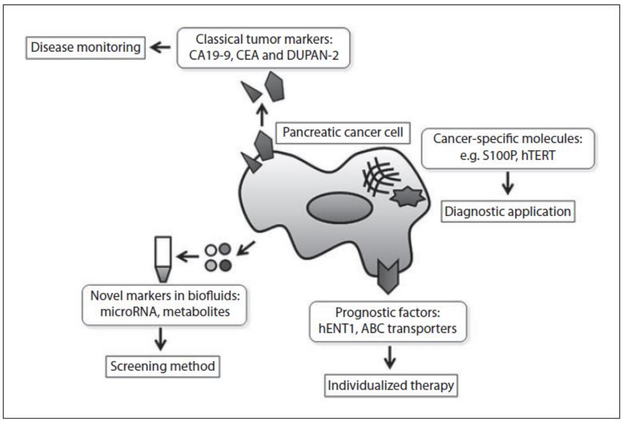

Disease monitoring
Classical tumor markers: CA19-9, CEA and DUPAN-2
Pancreatic cancer cell
Cancer-specific molecules: e.g. S100P, hTERT
Diagnostic application
Novel markers in biofluids: microRNA, metabolites
Screening method
Prognostic factors: hENT1, ABC transporters
Individualized therapy

Figure 16: Biomarkers of pancreatic cancer and clinical applications [2]

In order to result to reliable potential biomarkers, we cross-examined the extracted markers from both pdac tissue and peripheral blood. The initial list of the extracted DEGs (for the gcrma and the oligo analysis) was compared to the optimal subsets of tissue and peripheral blood from the process of SVM-RFE, and their intersection was considered. The two following VENN diagrams contain the mentioned genes.

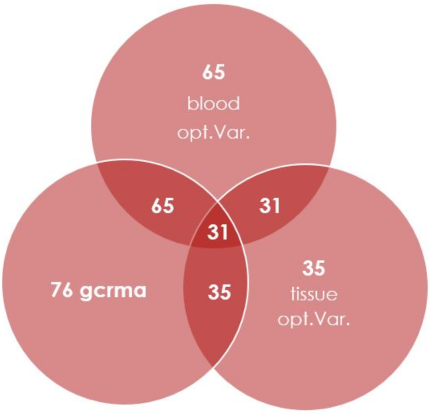


Figure 17: VENN diagram for extracted DEGs in gcrma analysis, compared to the optimal subsets from SVM-RFE

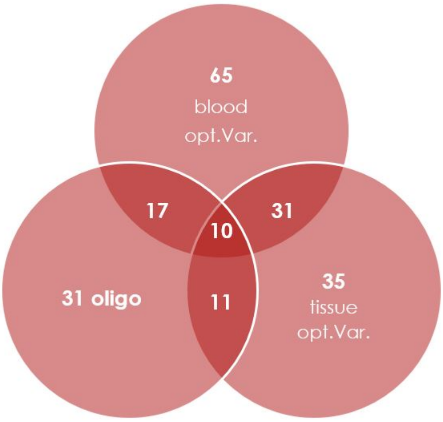


Figure 18: VENN diagram for extracted DEGs in oligo analysis, compared to the optimal subsets from SVM-RFE

- 76-DEGs gcrma list ∩ pdac tissue optimal subset variables ∩ pdac per. blood optimal subset variables : 31 DEGs
- 31-DEGs oligo list ∩ pdac tissue optimal subset variables ∩ pdac per. blood optimal subset variables : 10 DEGs

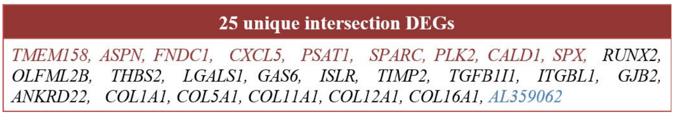
We conclude in 41 DEGs that could be potential pancreatic cancer biomarkers. These markers can both serve as classification, prognosis and diagnosis markers, since they are extracted from both pdac tissue and peripheral blood datasets. Their biological content is discussed in the following section.

5.8 BIOLOGICAL CONTENT

As depicted in the Venn diagrams of figures 17 and 18, the core circles represent the 31 DEGs and the 10 DEGs resulted from the intersection of gcrma-analysis or oligo-analysis and the SVM-RFE optimal subsets respectively.

In order to gain insight into the biological role of these DEGs, we needed to perform a mapping of probe set identifiers to HUGO Gene Nomenclature (HGNC) symbols and a functional enrichment analysis (figures 59, 60, 69, 70, 71, 72 of Appendix A.3).

First of all, the probe set identifiers from these 41 candidates were mapped to HGNC symbols by using the WebGestalt platform (a popular tool for the interpretation of gene lists derived from high-dimensional data analysis), as shown in figures 59 and 60 of Appendix A.3 [46]. More specifically, in figure 59, it is illustrated that 30 probe set identifiers correspond to 25 unique gene symbols, whereas one probe set identifier could not be mapped to any known gene symbol. Also, in figure 60, it is illustrated that 10 probe set identifiers correspond to 9 unique gene symbols. As one can observe, there is an overlap of the probe set identifiers and their mapped HGNC symbols of the 10 DEGs of the gcrma-intersection with the 31 DEGs of the oligo-intersection, i.e. the 10 common DEGs is a subset of the 31 common DEGs. Thus, the mapping of the probe set identifiers and the HGNC symbol assignment reduced the list of 41 potential cancer biomarkers obtained in the previous step to twenty five known genes, as shown in figure 19.

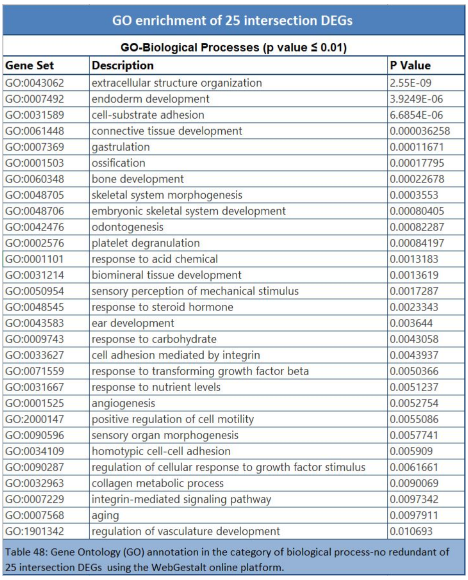


25 unique intersection DEGs

TMEM158, ASPN, FNDC1, CXCL5, PSAT1, SPARC, PLK2, CALD1, SPX, RUNX2, OLFML2B, THBS2, LGALS1, GAS6, ISLR, TIMP2, TGFBI11, ITGBL1, GJB2, ANKRD22, COL1A1, COL5A1, COL11A1, COL12A1, COL16A1, AL359062

Figure 19: The list of 25 unique known genes described by their HGNC symbols. The nine overlapping genes between gcrma-intersection and oligo-intersection are red highlighted. The unknown gene is blue highlighted.

Then, as a second step, to identify the biological processes (BP) and pathways in which these intersection DEGs were involved, we further analyzed the list of 25 genes at the functional level, by performing Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome pathway enrichment analyses using the WebGestalt online tool [47]. A p-value ≤ 0.01 was considered significant. Gene overrepresentation analysis (ORA) resulted in enriched GO-biological processes and pathways as illustrated in figures 20 and 21 respectively.



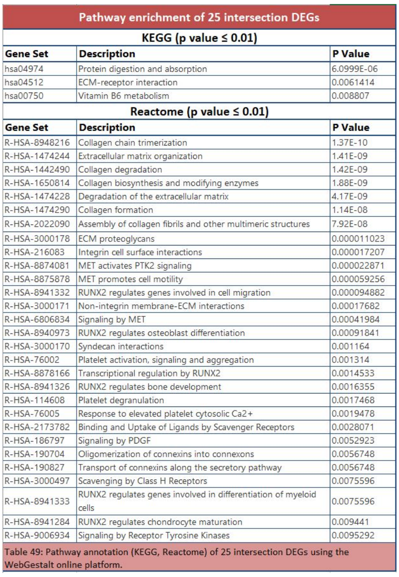
GO enrichment of 25 intersection DEGs

GO-Biological Processes (p value ≤ 0.01)

Gene Set	Description	P Value
GO:0043062	extracellular structure organization	2.55E-09
GO:0007492	endoderm development	3.9249E-06
GO:0031589	cell-substrate adhesion	6.6854E-06
GO:0061448	connective tissue development	0.000036258
GO:0007369	gastrulation	0.00011671
GO:0001503	ossification	0.00017795
GO:0060348	bone development	0.00022678
GO:0048705	skeletal system morphogenesis	0.0003553
GO:0048706	embryonic skeletal system development	0.00080405
GO:0042476	odontogenesis	0.00082287
GO:0002576	platelet degranulation	0.00084197
GO:0001101	response to acid chemical	0.0013183
GO:0031214	biomineral tissue development	0.0013619
GO:0050954	sensory perception of mechanical stimulus	0.0017287
GO:0048545	response to steroid hormone	0.0023343
GO:0043583	ear development	0.003644
GO:0009743	response to carbohydrate	0.0043058
GO:0033627	cell adhesion mediated by integrin	0.0043937
GO:0071559	response to transforming growth factor beta	0.0050366
GO:0031667	response to nutrient levels	0.0051237
GO:0001525	angiogenesis	0.0052754
GO:2000147	positive regulation of cell motility	0.0055086
GO:0090596	sensory organ morphogenesis	0.0057741
GO:0034109	homotypic cell-cell adhesion	0.005909
GO:0090287	regulation of cellular response to growth factor stimulus	0.0061661
GO:0032963	collagen metabolic process	0.0090069
GO:0007229	integrin-mediated signaling pathway	0.0097342
GO:0007568	aging	0.0097911
GO:1901342	regulation of vasculature development	0.010693

Table 48: Gene Ontology (GO) annotation in the category of biological process-no redundant of 25 intersection DEGs using the WebGestalt online platform.

Figure 20: GO enrichment analysis in the category of biological process-no redundant of 25 intersection DEGs using the WebGestalt online platform.



Pathway enrichment of 25 intersection DEGs

KEGG (p value ≤ 0.01)

Gene Set	Description	P Value
hsa04974	Protein digestion and absorption	6.0999E-06
hsa04512	ECM-receptor interaction	0.0061414
hsa00750	Vitamin B6 metabolism	0.008807

Reactome (p value ≤ 0.01)

Gene Set	Description	P Value
R-HSA-8948216	Collagen chain trimerization	1.37E-10
R-HSA-1474244	Extracellular matrix organization	1.41E-09
R-HSA-1442490	Collagen degradation	1.42E-09
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	1.88E-09
R-HSA-1474228	Degradation of the extracellular matrix	4.17E-09
R-HSA-1474290	Collagen formation	1.14E-08
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	7.92E-08
R-HSA-3000178	ECM proteoglycans	0.000011023
R-HSA-216083	Integrin cell surface interactions	0.000017207
R-HSA-8874081	MET activates PTK2 signaling	0.000022871
R-HSA-8875878	MET promotes cell motility	0.000059256
R-HSA-8941332	RUNX2 regulates genes involved in cell migration	0.000094882
R-HSA-3000171	Non-integrin membrane-ECM interactions	0.00017682
R-HSA-6806834	Signaling by MET	0.00041984
R-HSA-8940973	RUNX2 regulates osteoblast differentiation	0.00091841
R-HSA-3000170	Syndecan interactions	0.001164
R-HSA-76002	Platelet activation, signaling and aggregation	0.001314
R-HSA-8878166	Transcriptional regulation by RUNX2	0.0014533
R-HSA-8941326	RUNX2 regulates bone development	0.0016355
R-HSA-114608	Platelet degranulation	0.0017468
R-HSA-76005	Response to elevated platelet cytosolic Ca2+	0.0019478
R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.0028071
R-HSA-186797	Signaling by PDGF	0.0052923
R-HSA-190704	Oligomerization of connexins into connexons	0.0056748
R-HSA-190827	Transport of connexins along the secretory pathway	0.0056748
R-HSA-3000497	Scavenging by Class H Receptors	0.0075596
R-HSA-8941333	RUNX2 regulates genes involved in differentiation of myeloid cells	0.0075596
R-HSA-8941284	RUNX2 regulates chondrocyte maturation	0.009441
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	0.0095292

Table 49: Pathway annotation (KEGG, Reactome) of 25 intersection DEGs using the WebGestalt online platform.

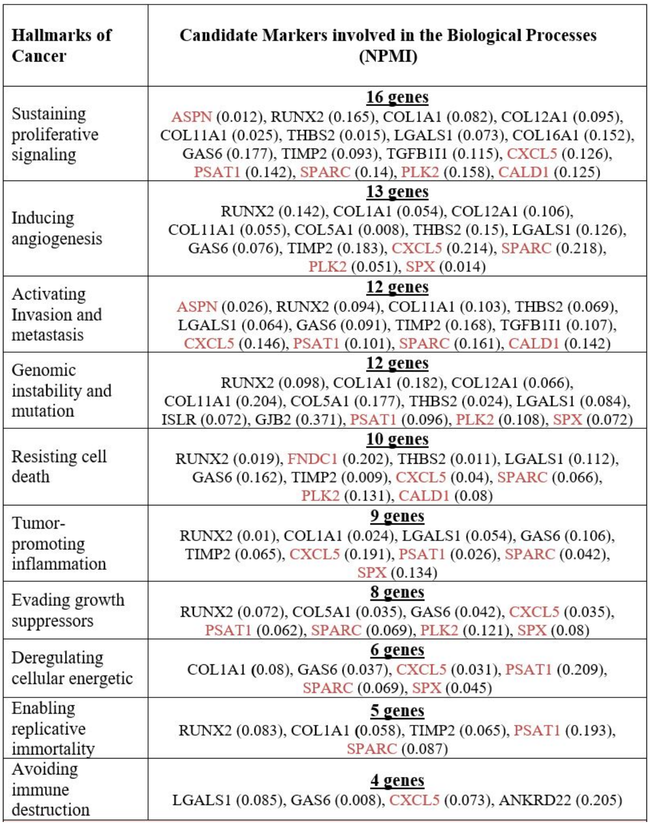
Figure 21: Pathway (KEGG, Reactome) enrichment analysis of 25 intersection DEGs using the WebGestalt online platform.

Our analysis revealed core processes and signaling pathways that are altered in pancreatic cancer subjects compared to healthy subjects (figures 20, 21), some of which may be critical in human embryogenesis [48], [49], [50]. As shown in figure 20, enriched BP functions were integrin-mediated processes, angiogenesis and developmental processes such as endoderm development. Furthermore, as demonstrated in figure 21, pathway enrichment results showed that the 25 intersection DEGs were significantly enriched in ECM and immune system (e.g. scavenger receptors) associated pathways as well as signal transduction (e.g. signaling by receptor tyrosine kinases, signaling by PDGF) and metabolic pathways (e.g. Vitamin B6 metabolism).

To facilitate comprehension of the enrichment results in the context of cancer, we conducted further analysis of the 25 intersection DEGs by utilizing the Cancer Hallmarks Analytics Tool (CHAT) [51]. CHAT was employed to reveal the involvement of the 25 intersection DEGs in the biological processes leading to pancreatic cancer according to the following ten current hallmarks of cancer: Sustaining proliferative signaling, Evading growth suppressors, Avoiding immune destruction, Enabling replicative immortality, Tumor-promoting inflammation, Activating Invasion and metastasis, Inducing angiogenesis, Genomic instability and mutation, Resisting cell death, Deregulating cellular energetic.

By using each of the 25 intersection DEGs as search term, CHAT yielded one or several hallmarks to 22 of these genes, as presented in table 22.

Moreover, we used CHAT to analyze PubMed literature on pancreatic cancer in relation to each gene of the 25 intersection DEGs. CHAT automatic literature analysis revealed that TIMP2, GAS6, CXCL5, and SPARC studies have a hallmark profile more similar to that of pancreatic cancer, as illustrated in figures 23-26.



Hallmarks of Cancer	**Candidate Markers involved in the Biological Processes (NPMI)**
Sustaining proliferative signaling	**16 genes** ASPN (0.012), RUNX2 (0.165), COL1A1 (0.082), COL12A1 (0.095), COL11A1 (0.025), THBS2 (0.015), LGALS1 (0.073), COL16A1 (0.152), GAS6 (0.177), TIMP2 (0.093), TGFB1I1 (0.115), CXCL5 (0.126), PSAT1 (0.142), SPARC (0.14), PLK2 (0.158), CALD1 (0.125)
Inducing angiogenesis	**13 genes** RUNX2 (0.142), COL1A1 (0.054), COL12A1 (0.106), COL11A1 (0.055), COL5A1 (0.008), THBS2 (0.15), LGALS1 (0.126), GAS6 (0.076), TIMP2 (0.183), CXCL5 (0.214), SPARC (0.218), PLK2 (0.051), SPX (0.014)
Activating Invasion and metastasis	**12 genes** ASPN (0.026), RUNX2 (0.094), COL11A1 (0.103), THBS2 (0.069), LGALS1 (0.064), GAS6 (0.091), TIMP2 (0.168), TGFB1I1 (0.107), CXCL5 (0.146), PSAT1 (0.101), SPARC (0.161), CALD1 (0.142)
Genomic instability and mutation	**12 genes** RUNX2 (0.098), COL1A1 (0.182), COL12A1 (0.066), COL11A1 (0.204), COL5A1 (0.177), THBS2 (0.024), LGALS1 (0.084), ISLR (0.072), GJB2 (0.371), PSAT1 (0.096), PLK2 (0.108), SPX (0.072)
Resisting cell death	**10 genes** RUNX2 (0.019), FNDC1 (0.202), THBS2 (0.011), LGALS1 (0.112), GAS6 (0.162), TIMP2 (0.009), CXCL5 (0.04), SPARC (0.066), PLK2 (0.131), CALD1 (0.08)
Tumor-promoting inflammation	**9 genes** RUNX2 (0.01), COL1A1 (0.024), LGALS1 (0.054), GAS6 (0.106), TIMP2 (0.065), CXCL5 (0.191), PSAT1 (0.026), SPARC (0.042), SPX (0.134)
Evading growth suppressors	**8 genes** RUNX2 (0.072), COL5A1 (0.035), GAS6 (0.042), CXCL5 (0.035), PSAT1 (0.062), SPARC (0.069), PLK2 (0.121), SPX (0.08)
Deregulating cellular energetic	**6 genes** COL1A1 (0.08), GAS6 (0.037), CXCL5 (0.031), PSAT1 (0.209), SPARC (0.069), SPX (0.045)
Enabling replicative immortality	**5 genes** RUNX2 (0.083), COL1A1 (0.058), TIMP2 (0.065), PSAT1 (0.193), SPARC (0.087)
Avoiding immune destruction	**4 genes** LGALS1 (0.085), GAS6 (0.008), CXCL5 (0.073), ANKRD22 (0.205)

Figure 22: Classification of 25 putative markers according to the hallmarks of cancer (data shown as NPMI; normalized pointwise mutual information). The nine overlapping genes between gcrma-intersection and oligo-intersection are red highlighted.

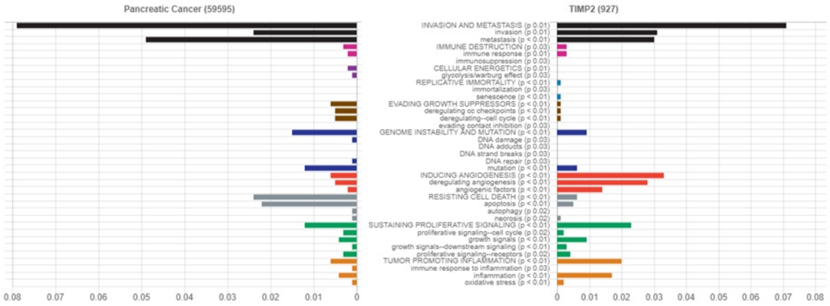


Figure 23: Pancreatic cancer and TIMP2 (data shown as CPROB; conditional probability).

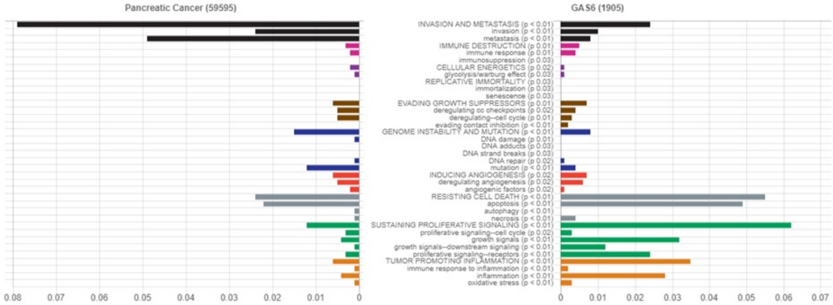


Figure 24: Pancreatic cancer and GAS6 (data shown as CPROB; conditional probability).

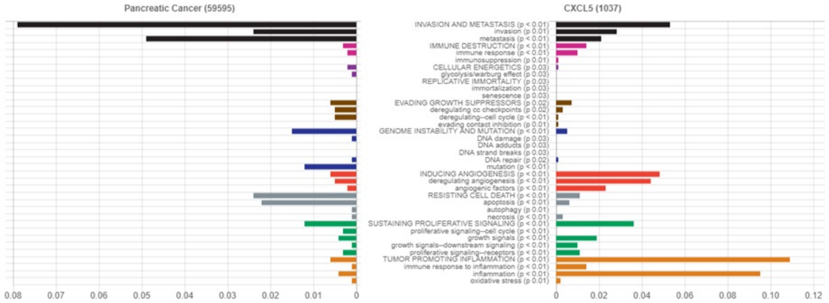


Figure 25: Pancreatic cancer and CXCL5 (data shown as CPROB; conditional probability).

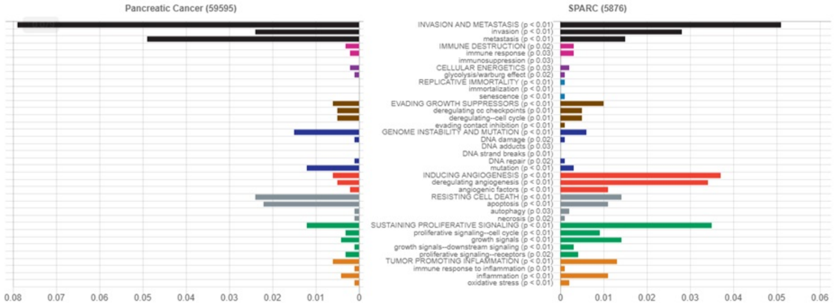


Figure 26: Pancreatic cancer and SPARC (data shown as CPROB; conditional probability).

Even with poor knowledge about the directly interrelation of the 25 intersection DEGs with pancreatic cancer as well as with embryogenesis, we can consider them as potential classifiers of PDAC based on:

- the significant BP processes and pathways highlighted here [52],
- the Cancer Hallmarks and associated molecular pathways underlying the mechanisms involved [53], and
- the abundance of collagens in the 25 intersection DEGs and the role of the ECM in PDAC [54].

The 25 intersection DEGs represent a list of putative biomarkers, whereas TIMP2, GAS6, CXCL5, and SPARC can be considered as the most promising biomarkers that could be easily validated experimentally in peripheral blood.

Part III

FINDINGS

6

RESULTS

In this chapter we present and discuss the results from the steps of feature extraction, cancer classification and feature selection.

6.1 FEATURE EXTRACTION

Microarray gene expression datasets, suffer from the "curse of dimensionality", as already discussed in chapter 2. Thus, feature extraction methods are applied, in order to avoid the issues that a high dimensional feature space introduces. The differentially expressed genes (DEGs) are examined in 4 human embryos, pdac tissue and pdac peripheral blood datasets. We consider the DEGs as the genes with expression values of log_2FoldChange > 2 and FDR < 1%. The lists of the DEGs extracted by each of the 4 datasets, are cross-examined and combined to result to our final feature space.

In order to evaluate the feature extraction method and confirm that the extracted DEGs are indeed significantly differentially expressed, we visualize our data by creating a heatmap with gene clustering, of the extracted DEGs. We present the gcrma analysis heatmaps in this section, while the oligo analysis corresponding heatmaps are presented in Appendix A.1. The heatmaps of the 4 examined datasets are listed and discussed below.

Heatmap 27 describes the DEGs on the human embryos dataset. The genes are divided in clusters, which are represented by the side left colors. We can distinguish two major clusters, where the genes with common expression level are divided. The upper

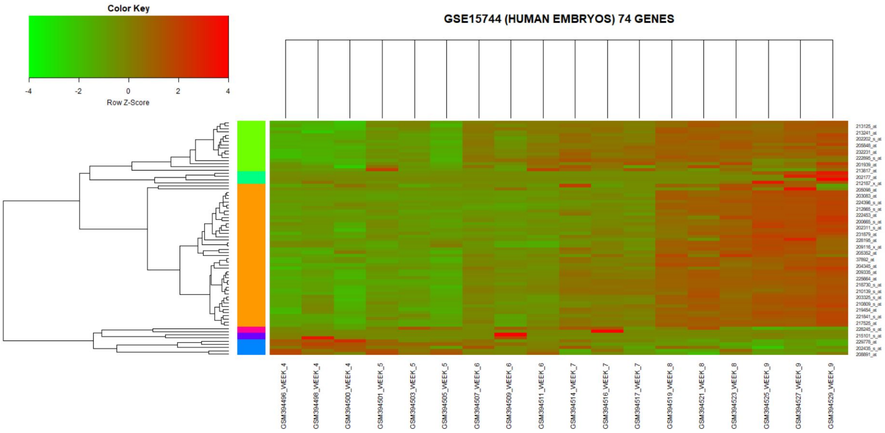


Figure 27: K-means clustering with pearson distances heatmap of DEGs screened on the basis of log_2fold change > 2 and FDR < 0.01.
Notes: GSE15744 human embryos data with 74 DEGs vs 18 samples for 6 weeks of embryonic development (3 samples/week). Red indicates that the expression of genes is relatively upregulated, green indicates that the expression of genes is relatively downregulated.
Abbreviation: DEGs, differentially expressed genes

cluster contains genes that are downregulated in early weeks (week 4 and 5) and upregulated in later weeks (week 8 and 9), while the opposite happens in the lower cluster of genes. We observe a progression of gene expression levels over the weeks, which is an expected characteristic, as the embryo develops. The heatmap also confirms the validatity of the extracted DEGs for this dataset.

Heatmap 28 contains the DEGs on the first of the two pdac tissue examined datasets (GSE32676). This dataset contains samples from early stage pancreatic cancer patients, were the gene expression levels are not significantly differentiated. The information extracted from this heatmap is quite ambiguous, however this is an expected result and does not imply false validity of the extracted DEGs.

Heatmap 29 on the other hand, contains the DEGs on the second of the two pdac tissue examined datasets (GSE71989).

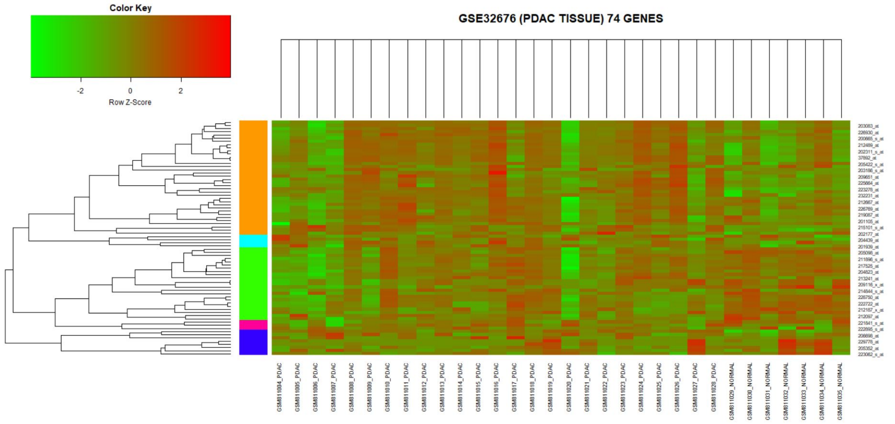


Figure 28: K-means clustering with pearson distances heatmap of DEGs screened on the basis of log_2fold change > 2 and FDR < 0.01. Notes: GSE32676 pdac tissue data with 74 DEGs vs 32 pdac patient and healthy samples. Red indicates that the expression of genes is relatively upregulated, green indicates that the expression of genes is relatively downregulated. Abbreviation: DEGs, differentially expressed genes

Things are quite clear in this dataset, and useful information can be extracted from the heatmap. We observe 2 major clusters of genes, which are significantly differentiated between the 2 classes (pdac-normal). This heatmap confirms the quality of the input data, since GSE71989 refers to advanced stage pdac tissue samples, where we expect more significant differences in gene expression levels.

The last of the 4 datasets used in the feature extraction process is described by heatmap 30, which contains the DEGs of the peripheral blood dataset (GSE49515). We can confirm the differences in the expression levels between the two classes in this dataset as well. GSE49515 contains similar expression levels for some genes. The Pearson distances are close to zero for these genes and they cannot be displayed in the heatmap, thus they are left empty. The 5 produced clusters give us useful insights for the DEGs and confirm their validity.

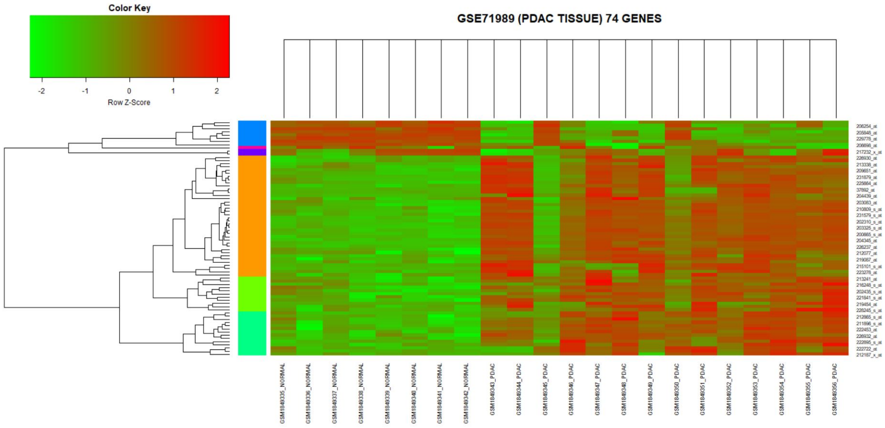


Figure 29: K-means clustering with pearson distances heatmap of DEGs screened on the basis of log_2fold change > 2 and FDR < 0.01. Notes: GSE71989 pdac tissue data with 74 DEGs vs 22 pdac patient and healthy samples. Red indicates that the expression of genes is relatively upregulated, green indicates that the expression of genes is relatively downregulated. Abbreviation: DEGs, differentially expressed genes

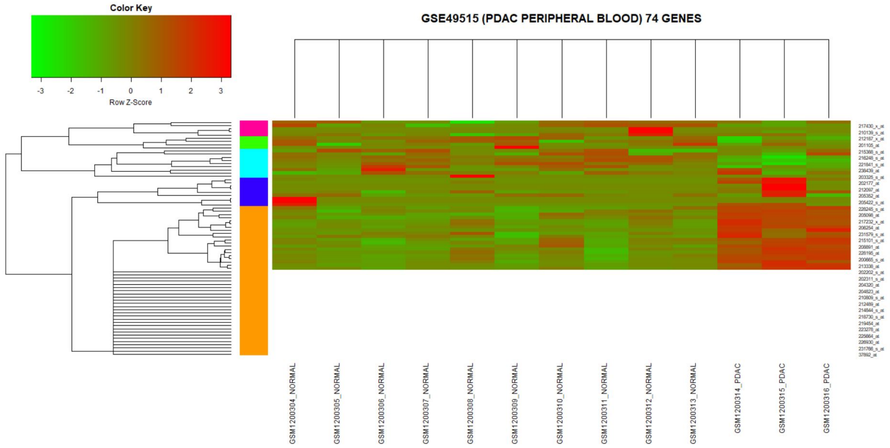


Figure 30: K-means clustering with pearson distances heatmap of DEGs screened on the basis of log_2fold change > 2 and FDR < 0.01. Notes: GSE49515 pdac peripheral blood data with 74 DEGs vs 13 pdac patient and healthy samples. Red indicates that the expression of genes is relatively upregulated, green indicates that the expression of genes is relatively downregulated. Abbreviation: DEGs, differentially expressed genes

6.2 CLASSIFICATION

The cancer classification results are presented in this section. After confirming the validity of the 74 extracted DEGs, we use them as classifiers for our machine learning classification methods. Firstly, heatmaps of the training and testing datasets are presented. Subsequently, some validation metrics for our classification models are examined, and the results are presented and discussed in tables and charts.

6.2.1 *Heatmaps*

The feature extraction process for the gcrma analysis, concluded in 74 DEGs, which synthesize the classification models' feature space. We visualize these genes with heatmaps and clustering for the training and testing datasets of the classification process. Examining these heatmaps is an non-trivial procedure, since it can provide us with useful information about the quality of our classifiers.

Heatmap 31 contains the DEGs on the training pdac tissue dataset used for SVM and KNN. This dataset is created from the combination of 4 different datasets, with different expression levels. Thus, we donnot expect to observe significant expression level differences. On the contrary, these smoothly spread intensities can result to better training of the models, since they will have better discrimination ability for tough classification decisions. The genes are divided in 4 clusters, where we can observe similar, non-significant differences in expression levels.

Heatmap 32 on the other hand, contains the DEGs on GSE71989, which is used as the testing pdac tissue dataset for our models. This dataset contains advanced stage pdac tissue samples, where the differences in expression levels are significant. We use this dataset for testing, since the classes are more distinct, which can result to better classification predictions. Therefore, we expect high classification accuracy on the testing dataset, due to the large class distances. The genes are divided in 2 major clusters, where we can observe the significant differences between the two classes.

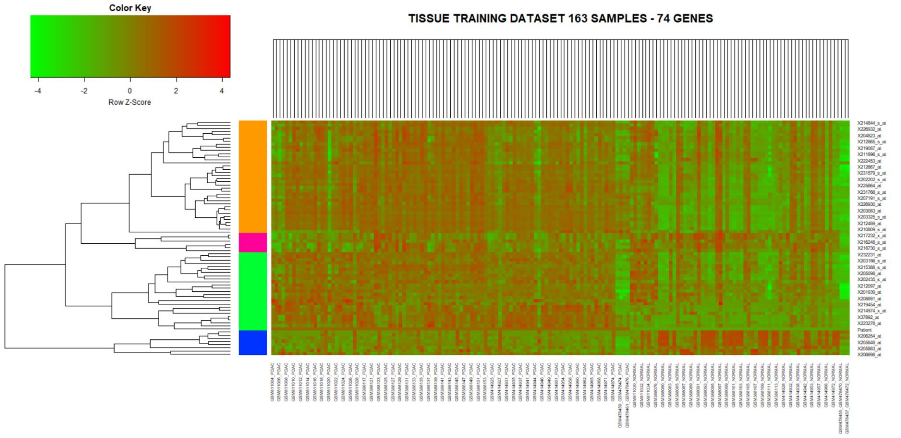


Figure 31: Pdac tissue training dataset heatmap containing data with 74 DEGs vs 163 pdac patient and healthy samples

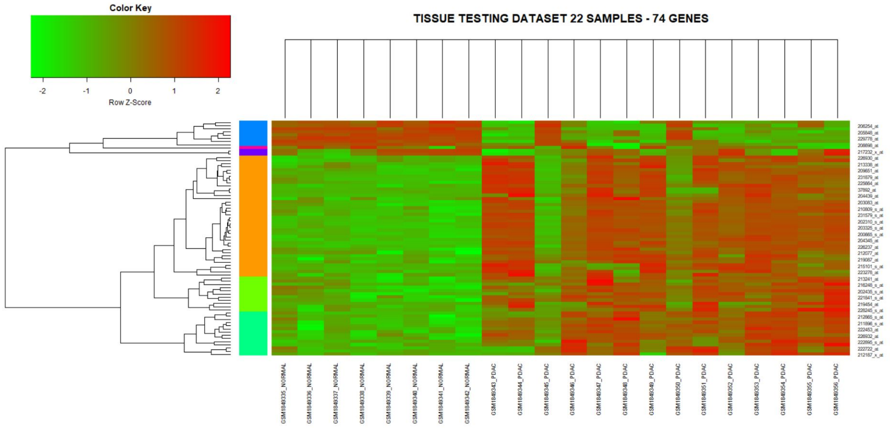


Figure 32: Pdac tissue testing dataset heatmap containing data with 74 DEGs vs 22 pdac patient and healthy samples

Finally, we examine heatmap 33, which contains the DEGs on the pdac peripheral blood training dataset. The training dataset is created by the combination of 3 different datasets, which different genes expression levels. That results to an ambiguous heatmap, since we cannot classify the samples into two classes by clustering them. However, this is not necessarily bad, because the short distance between the two classes, can lead us to models with better classification ability. The genes are divided in 4 clusters, which expose the difference between the expression levels of different datasets, rather than the difference between the two classes.

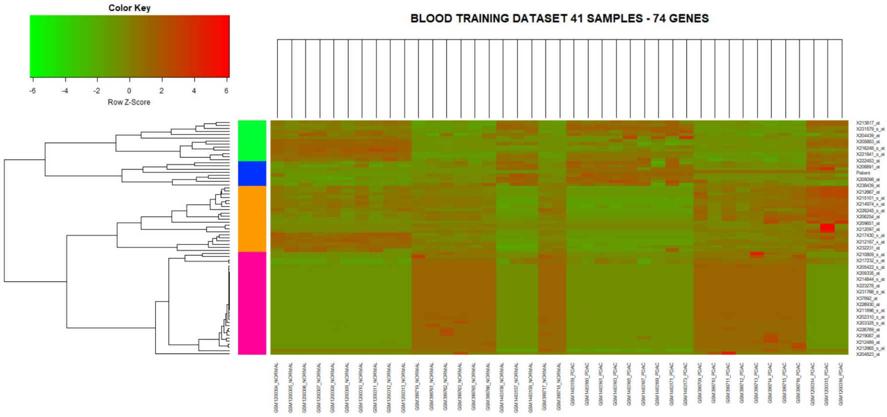
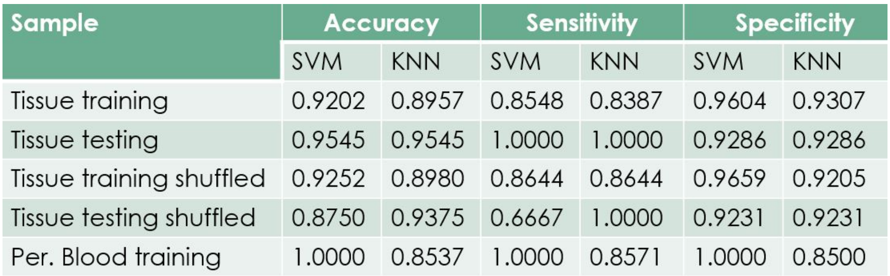


Figure 33: Pdac peripheral blood training dataset heatmap containing data with 74 DEGs vs 41 pdac patient and healthy samples

6.2.2 *Cancer classification rates*

We have used 11 pdac datasets for the cancer classification section. Pdac tissue and peripheral blood datasets were examined seperately. The datasets where divided in training and testing, and leave one out cross validation was used. The training datasets were fed into two classification algorithms, support vector machines (SVM) and k-nearest neighbours (KNN). After the training stage, the models were tested on both training and testing datasets. Three significant metrics were used for the evaluation of our models: accuracy, sensitivity and specificity. Accuracy refers to the ability of the models to classify correctly the samples. Sensitivity, or true positive rate, is a metric that expresses the ratio of the correctly predicted positive samples, with respect to all positive samples. Specificity on the other hand, or false positive rate, describes the ratio of the falsely predicted positive samples (which are negative), with respect to all negative samples. These three metrics are estimated on all classification attempts, and they are presented in table 34. A comparison of the two algorithms is also presented in plot 35.



Sample	Accuracy		Sensitivity		Specificity	
	SVM	KNN	SVM	KNN	SVM	KNN
Tissue training	0.9202	0.8957	0.8548	0.8387	0.9604	0.9307
Tissue testing	0.9545	0.9545	1.0000	1.0000	0.9286	0.9286
Tissue training shuffled	0.9252	0.8980	0.8644	0.8644	0.9659	0.9205
Tissue testing shuffled	0.8750	0.9375	0.6667	1.0000	0.9231	0.9231
Per. Blood training	1.0000	0.8537	1.0000	0.8571	1.0000	0.8500

Figure 34: SVM and KNN Classification metrics on various cancer classification attempts. 74 DEGs were used as classifiers on binary class data

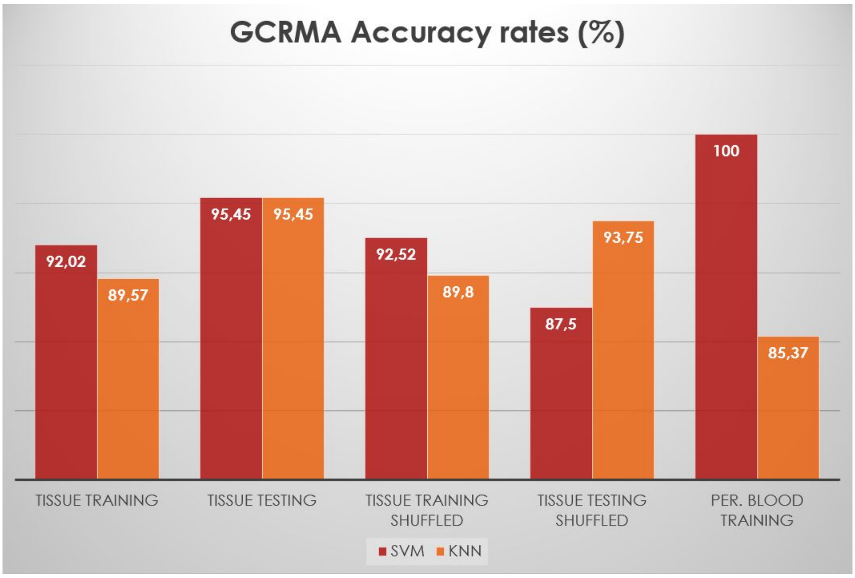


Figure 35: SVM and KNN comparison of accuracy rates on cancer classification attempts.74 DEGs were used as classifiers on binary class data

From the evaluation of the experimental results, we end up with the following observations:

- Both SVM and KNN perform exceptionally on pdac tissue and peripheral blood data. Accuracy rates of more than 90% are achieved, which indicates the high classification ability of our models. Good cancer classification is achieved, based on the 74 DEGs which are used as classifiers.

- Both SVM and KNN perform even better on independent pdac tissue testing data, where the accuracy rates exceeded 95%. This happens due to the fitness of our testing data, which have large class distances, as they are obtained from advanced stage pdac patients. That is also the reason why relatively lower accuracy rates are achieved in the training data, where the class distances are smaller (see also heatmap 31).

- SVM generally performs better than the KNN in all experiments. Especially in pdac peripheral blood training dataset, SVM manages to classify all samples correctly, while KNN achieves 85% accuracy rate.

- Slightly lower accuracy levels are observed in the shuffled and split tissue training dataset, where KNN also performs better on testing data. The lower accuracy rates are independent of the shuffling of the dataset. Worse classification happens in this dataset, due to the smaller amount of classifiers, since only 90% of the training dataset was used for training and 10% was used for testing. The shuffling of the dataset would not affect the classification process eitherway, since the data are cross validated with the leave one out cross validation method (LOOCV), where all samples are used as testing data, independently of their position.

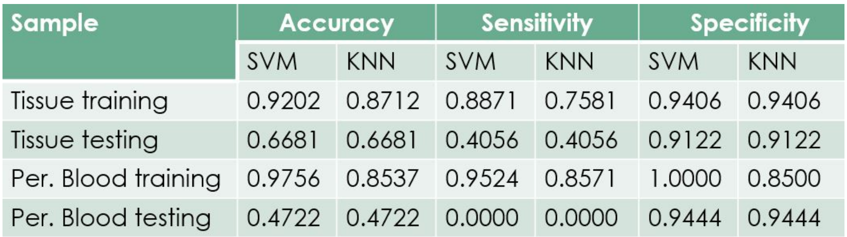
- The computational time of both algorithms was significantly low. SVM run for about 20 seconds, while KNN needed about 12 seconds to finish. Despite the fact that

LOOCV was used (which increases the time complexity as it cross validates the training process for N times, where N is the number of features), the running time was kept low. That happens due to the grid on parameter tuning, where specific values were tried, and the length of function *tuneLength* which was set to small values. The small number of samples and features also contributed in the low time complexity of each method.

- Sensitivity and specificity are two good metrics that show us on which data class the algorithms suffer more and score worse classification rates. Sensitivity is lower in most cases, which indicates that our methods manage to better classify the data of the negative class (patient subjects) as patients, but achieve worse classification results when classifying positive class data (healthy subjects) as healthy.

- Lastly, no overfitting was observed, since our methods achieved high classification rates. This happens due to the feature extraction process, where the redundant and irrelevant to the classification features that introduce data overfitting were successfully removed.

We also present the classification results for the oligo analysis. We decided to proceed with this analysis, in order to integrate 2 more pdac tissue and 1 peripheral blood datasets, which are used for testing. The same steps were followed, with the same parameter tuning and cross validation method. The step of tissue shuffling was skipped. The corresponding heatmaps are listed in Appendix A.2. The results are presented in table 36, and a comparative plot of the two classification methods follows 37. The results are discussed afterwards.



Sample	Accuracy		Sensitivity		Specificity	
	SVM	KNN	SVM	KNN	SVM	KNN
Tissue training	0.9202	0.8712	0.8871	0.7581	0.9406	0.9406
Tissue testing	0.6681	0.6681	0.4056	0.4056	0.9122	0.9122
Per. Blood training	0.9756	0.8537	0.9524	0.8571	1.0000	0.8500
Per. Blood testing	0.4722	0.4722	0.0000	0.0000	0.9444	0.9444

Figure 36: SVM and KNN Classification metrics on various cancer classification attempts (oligo analysis). 29 DEGs were used as classifiers on binary class data

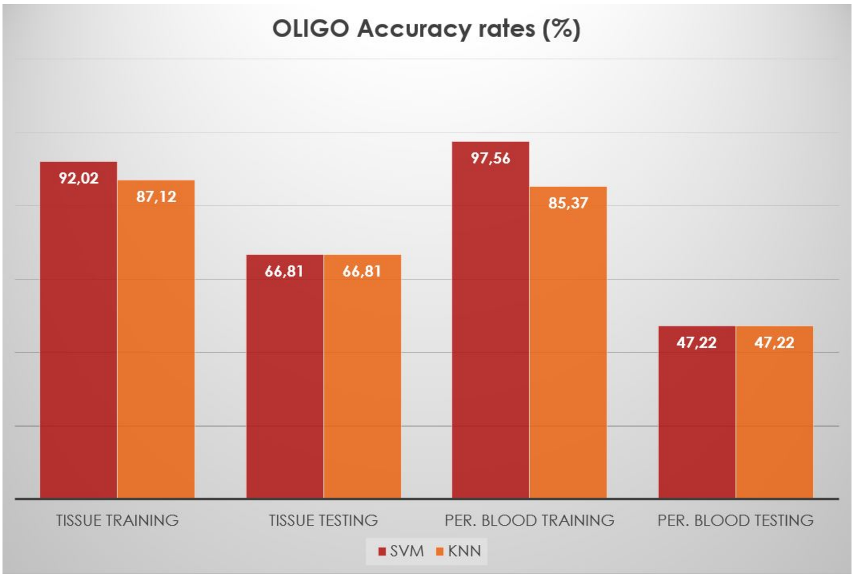


Figure 37: SVM and KNN comparison of accuracy rates on cancer classification attempts (oligo analysis). 29 DEGs were used as classifiers on binary class data

From the evaluation of the experimental results, we observe that our methods performed exceptionally on both pdac tissue and peripheral blood training datasets. High classification rates were expected, since the datasets were the same as in the gcrma analysis, with the difference that the classifiers were 29 DEGs instead of 74. The negative conclusion that arises from this analysis, is the non-fitness of the testing datasets. Both in pdac tissue and peripheral blood testing datasets, low accuracy rates were observed. Especially in the pdac peripheral blood testing dataset, the classification rate was below 50%, which can lead to two conclusions: Either our models are unsuccessful, or the testing data are not suitable for testing with these models. Having tested our models in various experiments, where accuracy rates of over 90% were achieved, we conclude that the testing data from the platform GPL6244 are not suitable for our models, thus we donnot proceed with further analysis or discussion for the oligo implementation.

6.3 FEATURE SELECTION

The feature selection results are presented in this section. After performing the cancer classification with SVM and KNN with 74 DEGs as classifiers, we attempt to extract a smaller gene signature, by implementing a widely used feature selection method, support vector machines - recursive feature elimination (SVM-RFE). Firstly, heatmaps of the training and testing datasets are presented. Subsequently, we evaluate our models by examining some validation metrics, which are presented and discussed in tables and charts.

6.3.1 *Heatmaps*

The feature selection process is performed on the 74 DEGs pdac tissue and peripheral blood training datasets. The SVM-RFE process provides us with the optimal variable subsets for pdac tissue and peripheral blood. The 3 following heatmaps describe these datasets, which are in fact subsets of the heatmaps presented in the classification step. Thus, the comments on these heatmaps will not be repeated in this section, and they can be found under section 6.2.1. Heatmap 38 refers to the pdac tissue training dataset, heatmap 39 refers to the pdac tissue testing dataset and heatmap 40 refers to the pdac peripheral blood training dataset.

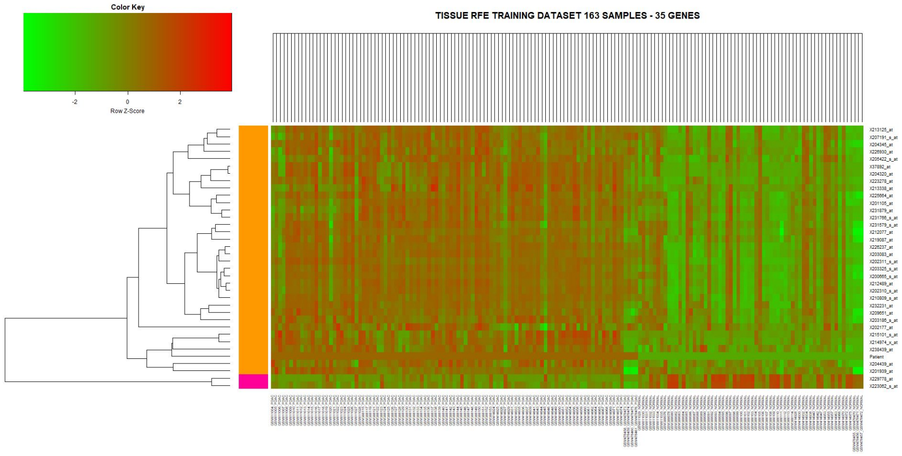


Figure 38: Pdac tissue training dataset used for SVM-RFE heatmap containing data with 35 DEGs (optimal subset) vs 163 pdac patient and healthy samples

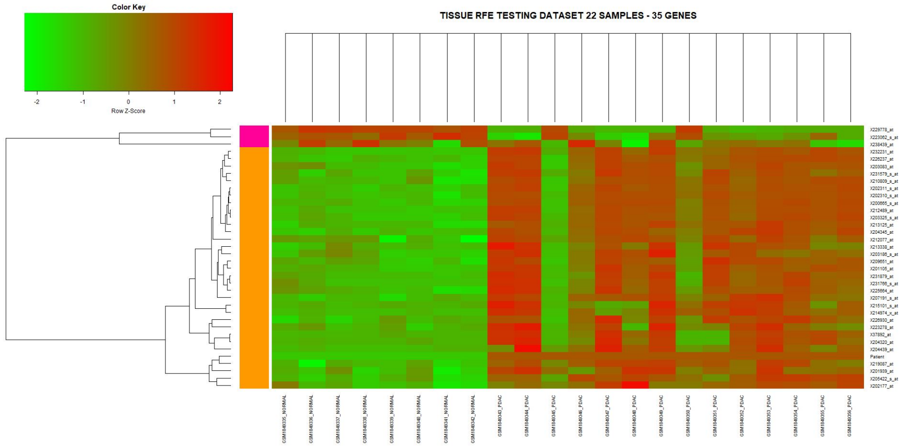


Figure 39: Pdac tissue testing dataset used for SVM-RFE heatmap containing data with 35 DEGs (optimal subset) vs 22 pdac patient and healthy samples

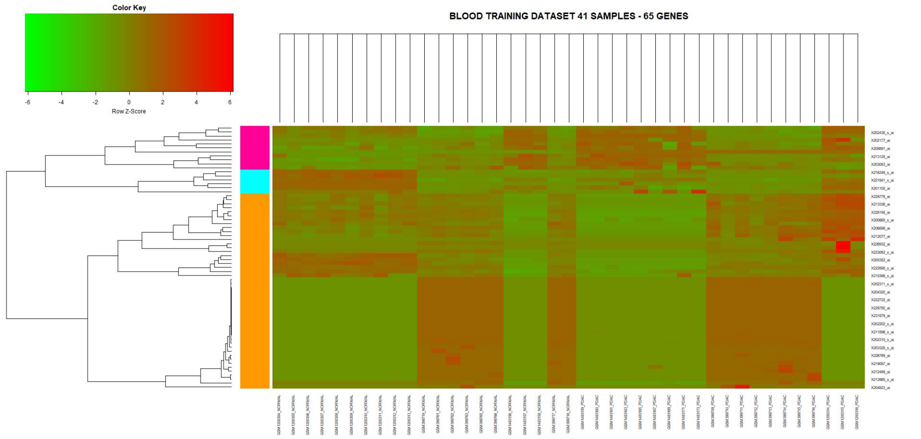
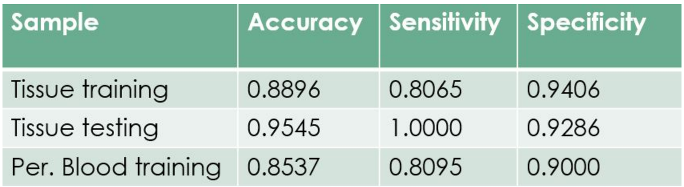


Figure 40: Pdac peripheral blood training dataset used for SVM-RFE heatmap containing data with 65 DEGs (optimal subset) vs 41 pdac patient and healthy samples

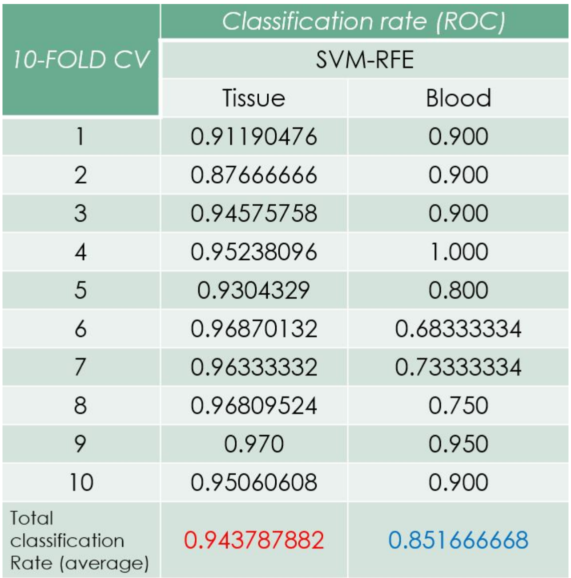
6.3.2 *Cancer classification rates*

We have used 8 pdac datasets for the feature selection. Pdac tissue and peripheral blood datasets were examined seperately. The training datasets are the same as in the step of cancer classification. We only used 1 dataset for pdac tissue testing. 10 fold cross validation was used as a validation method, instead of LOOCV. SVM-RFE is a time-consuming process, with higher time complexity than SVM or KNN. That happens because the recursive feature elimination works with a predefined number of predictors which will be removed in each step. In order to acquire the optimal subset, the SVM classifiers are trained and cross validated in each iteration of the RFE, until all features are eliminated. This process can result to high computational costs, which led us to use 10-fold CV instead of LOOCV. Accuracy, sensitivity and specificity are the metrics that will evaluate the performance of our models, and they are presented in table 41. Moreover, the classification rates from the folds of 10-fold CV are presented in table 42. Furthermore, the accuracy rates of all the examined subsets which led to the optimal subsets for pdac tissue and peripheral blood datasets, are presented in table 43, along with chart 44 and chart 45. Finally, a comparison of the accuracy rates of SVM, KNN and SVM-RFE methods is presented in plot 46.



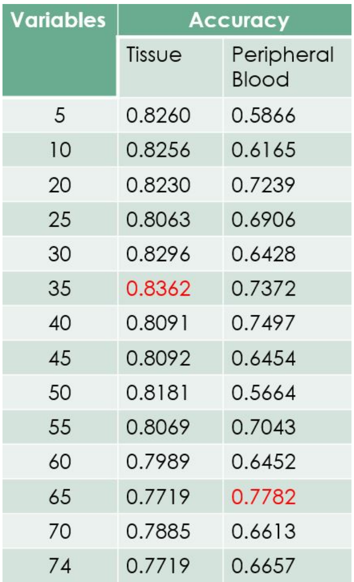
Sample	Accuracy	Sensitivity	Specificity
Tissue training	0.8896	0.8065	0.9406
Tissue testing	0.9545	1.0000	0.9286
Per. Blood training	0.8537	0.8095	0.9000

Figure 41: SVM RFE metrics on various feature selection attempts. 35 DEGs for pdac tissue and 65 DEGs for pdac peripheral blood datasets were used as classifiers on binary class data



10-FOLD CV	Classification rate (ROC)	
	SVM-RFE	
	Tissue	Blood
1	0.91190476	0.900
2	0.87666666	0.900
3	0.94575758	0.900
4	0.95238096	1.000
5	0.9304329	0.800
6	0.96870132	0.68333334
7	0.96333332	0.73333334
8	0.96809524	0.750
9	0.970	0.950
10	0.95060608	0.900
Total classification Rate (average)	0.943787882	0.851666668

Figure 42: 10-Fold CV classification rates for the optimal features subsets (35 DEGs for pdac tissue and 65 DEGs for pdac peripheral blood datasets).



Variables	Accuracy	
	Tissue	Peripheral Blood
5	0.8260	0.5866
10	0.8256	0.6165
20	0.8230	0.7239
25	0.8063	0.6906
30	0.8296	0.6428
35	0.8362	0.7372
40	0.8091	0.7497
45	0.8092	0.6454
50	0.8181	0.5664
55	0.8069	0.7043
60	0.7989	0.6452
65	0.7719	0.7782
70	0.7885	0.6613
74	0.7719	0.6657

Figure 43: SVM-RFE feature subset accuracy levels

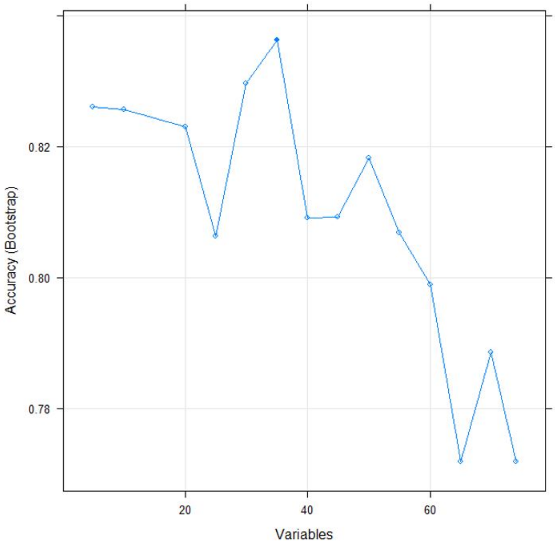


Figure 44: SVM-RFE optimal variables subset of 35 DEGs for pdac tissue datasets

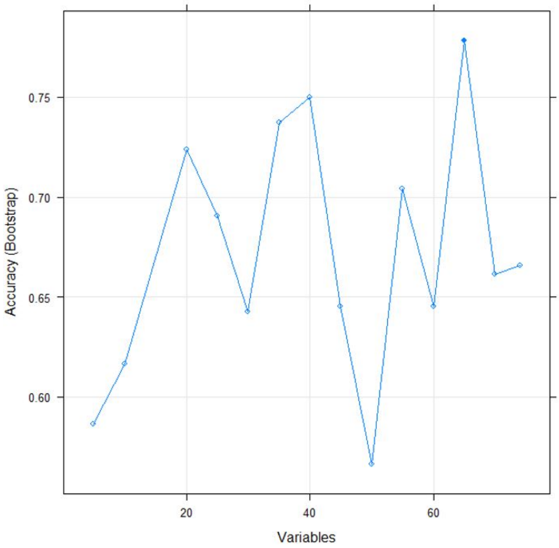


Figure 45: SVM-RFE optimal variables subset of 65 DEGs for pdac peripheral blood datasets

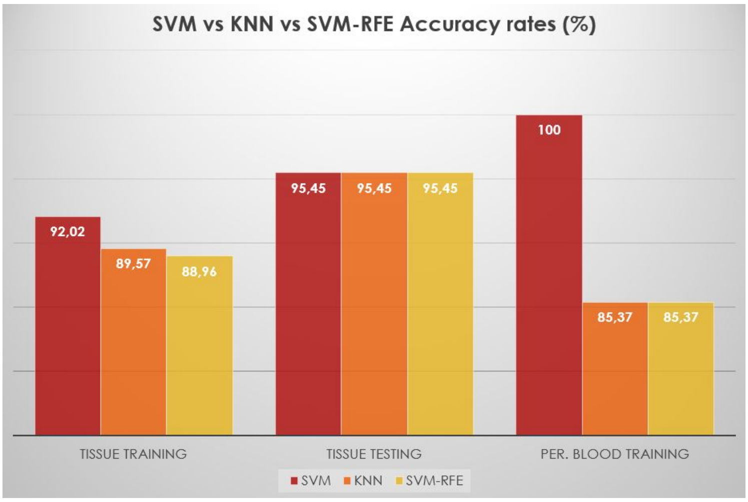


Figure 46: SVM, KNN and SVM-RFE comparison of accuracy rates on cancer classification attempts. Different number of DEGs were used as classifiers on binary class data

From the evaluation of the experimental results and the presented plots, we result to the following observations:

- The SVM-RFE achieves great classification rates (>85%). That means that the SVM classifiers, built on the subsets of the 35 and 65 DEGs for pdac tissue and peripheral blood respectively, have high classification ability. Thus, we can conclude that the feature selection process is successful, since it manages to filter out the redundant classifiers, without undermining the models' accuracy levels.

- Sensitivity is slightly worse than specificity, which indicates that our models tend to classify the data of the positive class (healthy subjects) as healthy, with worse accuracy.

- Regarding the 10-fold CV classification rates, ROC was used instead of accuracy. The 10-fold CV rates are regarding the optimal subset of DEGs, 35 for pdac tissue and 65 for pdac peripheral blood, where the average classification rate is also calculated. The average classification rate differs from the extracted classification rate for the training dataset, since the first one refers to 1 fold each time, while the second refers to the whole dataset. Each fold was repeated 5 times, where each fold value is the average of the 5 repeats.

- Table 43 contains the accuracy levels of each subset, which leads us to the select the optimal one for pdac tissue and peripheral blood. Here we mention that the number of reduced features in each iteration is predifined, and it is independent of their feature weights. Thus, the optimal subset is selected between some predefined values that we set beforehand, and not based on the exact number of the best classifiers.

- SVM-RFE is a time demanding process, since it has greater time complexity than a simple classification algorithm, demanding some minutes to run. That is an expected observation, and it is also the reason why we have chosen 10-fold CV over LOOCV.

- Finally, the comparative plot between SVM, KNN and SVM-RFE shows that our feature selection method performed similarly the two classification methods, achieving high accuracy rates on both pdac tissue and peripheral blood datasets. Thus, we conclude that SVM-RFE filtered out the correct redundant features, and overall was a successful process that provided us with a smaller gene signature.

7

CONCLUSIONS

Microarray gene expression analysis is a useful tool in bioinformatics and is widely used for gene expression profiling, especially in human cancer studies. According to recent studies [5], there is a correlation of some signaling pathways that participate in pancreatic cancer tumorigenesis, with the ones activated during the process of embryogenesis.

In our analysis, the significant genes activated in both pancreatic cancer (pdac) and human embryos datasets are cross-examined and a list of the significantly involved in these pathways genes is extracted. These genes are later used for pancreatic cancer classification. This thesis aims to propose a gene signature for genes involved in both processes, and to contribute some machine learning models that will be used for pancreatic cancer classification. Our goal is to confirm this existing theory that associates pancreatic cancer with embryogenesis, by suggesting the involved genes, and to add some pancreatic cancer classification methods to the literature, in an attempt to improve cancer identification and classification in a molecular level.

Based on the results of our analysis, we can conclude the following:

- The correlation of differentially expressed genes (DEGs) between pdac and human embryos datasets is confirmed. The extracted DEGs were cross-validated in pdac tissue, pdac peripheral blood and human embryos datasets and heatmaps of those genes were plotted. These heatmaps of the 74 extracted DEGs highlighted both the significance of the genes in each dataset, and the correlation in the expression levels progression between the examined datasets.

- The feature extraction process, where we implemented statistical methods to extract the DEGs, resulted in good discrimination of the significant features. The extracted genes that were characterized as significant, were confirmed as DEGs by the heatmaps on each dataset. This process of data mining gives us useful insights on which genes are activated during pancreatic cancer and embryogenesis. A gene signature is extracted, which can be interpreted biologically to identify the examined biomolecular signaling pathways.

- Furthermore, the feature extraction process was a prerequisite for the cancer classification methods, since it filtered out all the redundant and irrelevant to the analysis features. A lower feature space with only features associated with the analysis, results to higher and more accurate classification rates.

- Cancer classification methods used to classify the two class data into patient or healthy subjects, also performed well. They managed to classify independent testing data of pdac and healthy subjects with high accuracy rates, based on the extracted DEGs which were used as classifiers. The proposed models of SVM and KNN algorithms, can be used on more independent testing data in order to classify future subjects as patients or healthy. The extracted features can be added to the literature as pancreatic cancer classifiers or predictors, in the attempt of improving cancer classification and prediction.

- The feature selection method (SVM-RFE) also achieved high classification accuracy and managed to filter out more redundant features. A smaller gene signature was extracted, which is desirable, since the lower dimensional feature space can improve the accuracy and computational time of future cancer classification/prediction methods.

- Finally, we emphasize on the significance of a good preprocessing analysis for microarray data. The preprocessing performed by *oligo* was not as efficient as the one by *gcrma*. It identified less genes as differentially expressed, and was unsuccessful on cancer classification for independent testing data.

Concerning the biological evaluation, the enriched biological processes and pathways that assigned to the 25 intersection genes are important with respect to different aspects of pancreatic carcinogenesis and to some crucial events of embryogenesis. Moreover, we support the notion that our "25 gene signature" in its entirety can play a classification role in discriminating patients with pancreatic cancer from healthy controls, and we emphasize the role of TIMP2, GAS6, CXCL5, and SPARC as potent predictors.

Overall, we conclude this thesis by adding the gene signature of the significantly involved genes in the common signaling pathways between pancreatic cancer and embryogenesis to the literature. We also propose two cancer classification models, hoping to contribute to the efforts made in achieving better classification and prediction of the incurable disease of pancreatic cancer.

8

FUTURE WORK

A significant amount of work can be done in cancer classification at a molecular level using microarray gene expression analysis. The three independent major steps that were followed in this thesis, can be implemented with different approaches.

The process of feature extraction can be implemented by various data mining techniques proposed in the literature. It would be a good practice for future studies to focus on extracting the differentially expressed genes with different data mining methods, on different datasets, and cross-validate their findings with the ones proposed in this thesis. This would strongly suggest the validity of these markers, which could be used as a reliable criterion for cancer classification and prediction. Feature selection could also concern future researchers, since various feature selection methods are proposed in the literature, which could result to better subsetting of the extracted features.

Different classification methods can also be a subject of future studies. Widely used algorithms on cancer classification can be trained on the proposed features, in order to possibly achieve better classification rates and evaluate the quality of the features. Artificial neural networks, random forest and deep learning are related with cancer classification and prediction, and their evaluation on the proposed features could be a future subject of study. Different datasets can also be tested as another evaluation metric for the quality of extracted genes.

Finally, this study can be extended by using the proposed classifiers as predictors in machine learning methods used for cancer prediction. Independent pancreas tissue and peripheral blood samples can be used as testing datasets for cancer prediction methods with the described predictors. Future research is needed to achieve better pancreatic cancer prediction, and this study could be a major step towards this goal.

Part IV

APPENDIX

A

APPENDIX CHAPTER

A.1 FEATURE EXTRACTION HEATMAPS

The heatmaps of each of the 4 datasets used for the feature extraction process are presented in this section. These heatmaps contain only the extracted genes of each dataset instead of all 76 extracted features.

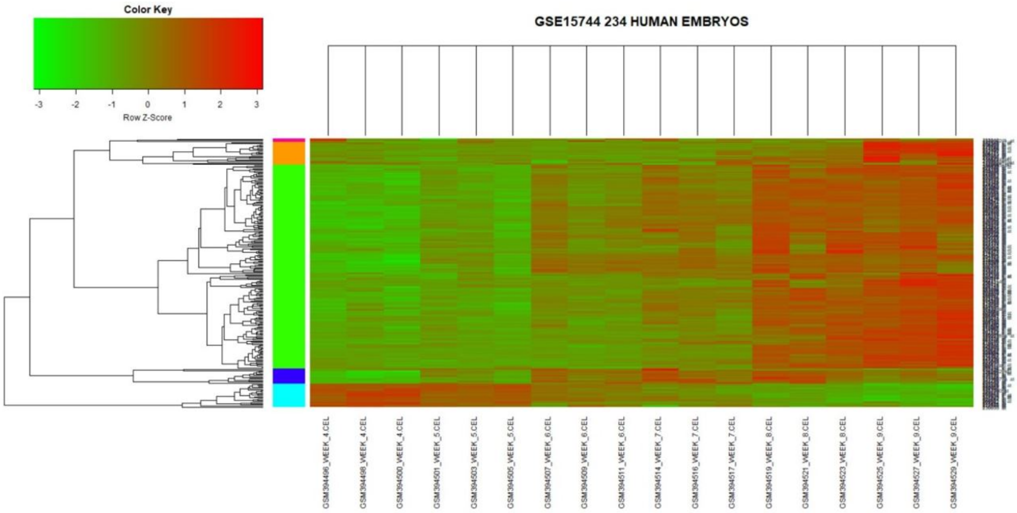


Figure 47: Human embryos heatmap (GSE15744) containing data with 234 own extracted DEGs vs 18 human embryos samples

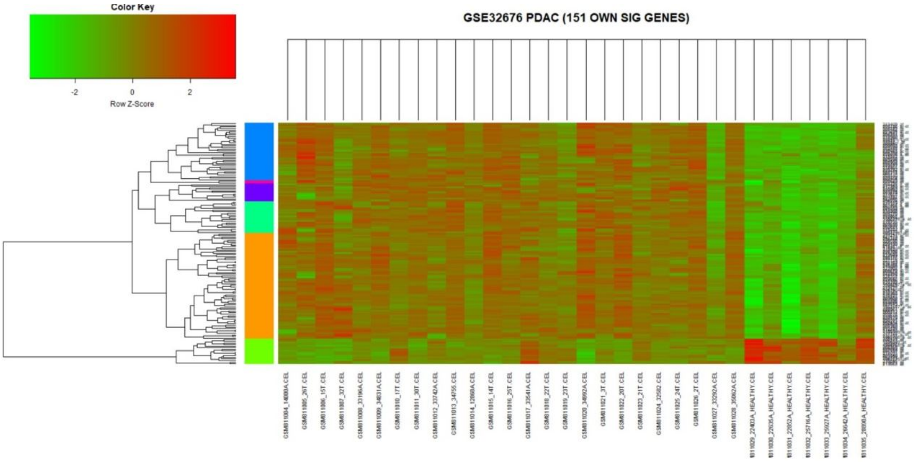


Figure 48: Pdac tissue heatmap (GSE32676) containing data with 151 own extracted DEGs vs 32 pdac patient and healthy samples

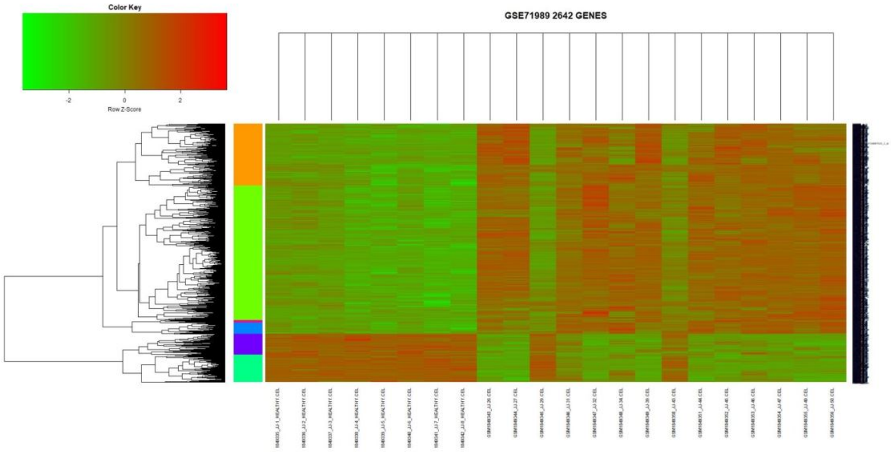


Figure 49: Pdac tissue heatmap (GSE71989) containing data with 2642 own extracted DEGs vs 22 pdac patient and healthy samples

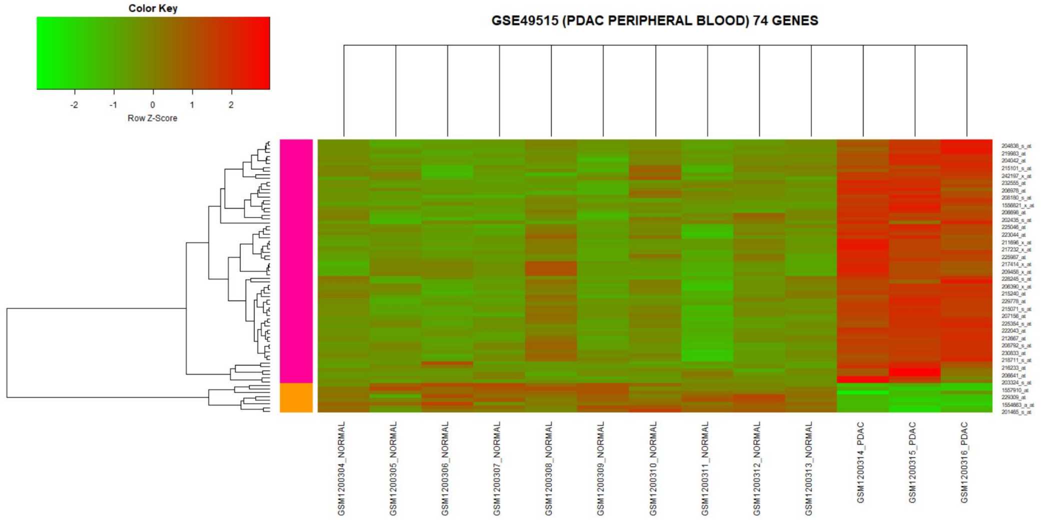


Figure 50: Pdac peripheral blood heatmap (GSE49515) containing data with 74 own extracted DEGs vs 13 pdac patient and healthy samples

A.2 CLASSIFICATION HEATMAPS ON OLIGO ANALYSIS

The training and testing datasets for SVM and KNN classification are presented in this section.

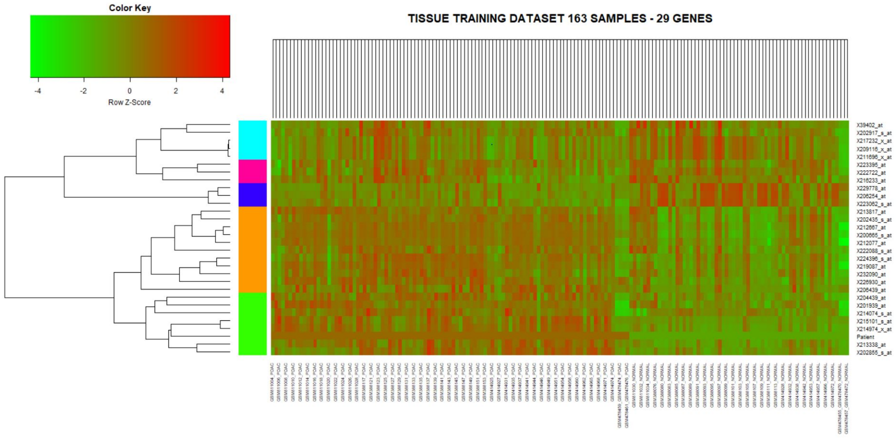


Figure 51: Pdac tissue training dataset heatmap containing data with 29 DEGs vs 163 pdac patient and healthy samples (oligo analysis)

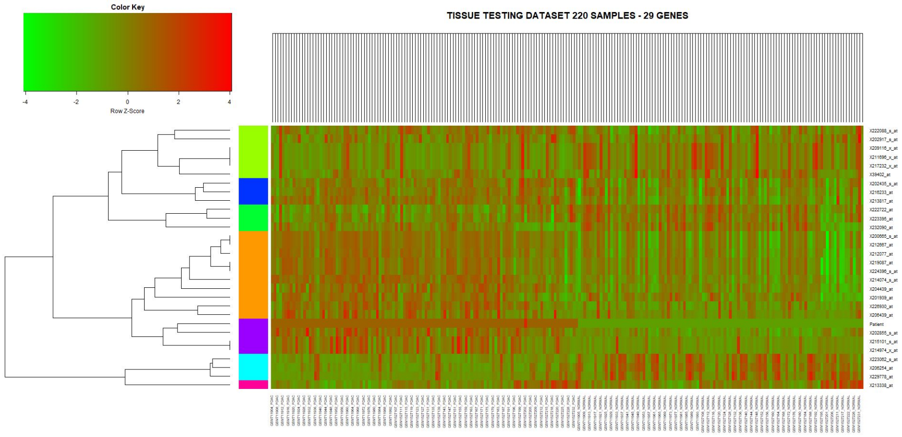


Figure 52: Pdac tissue testing dataset heatmap containing data with 29 DEGs vs 220 pdac patient and healthy samples (oligo analysis)

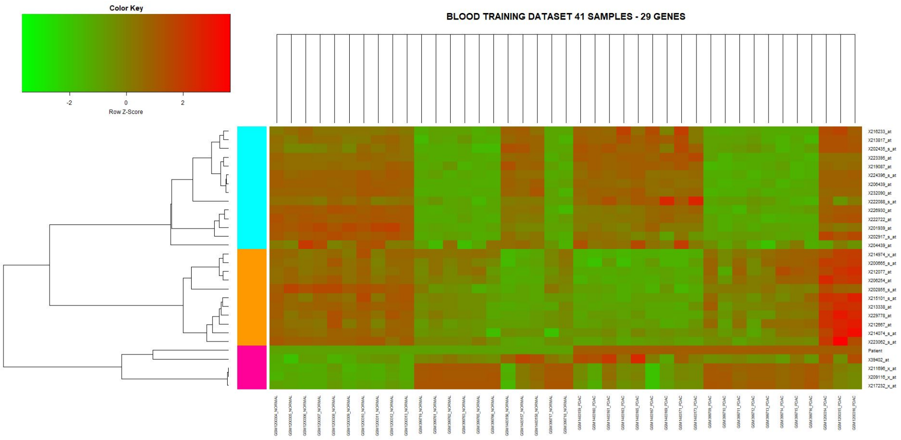


Figure 53: Pdac peripheral blood training dataset heatmap containing data with 29 DEGs vs 41 pdac patient and healthy samples (oligo analysis)

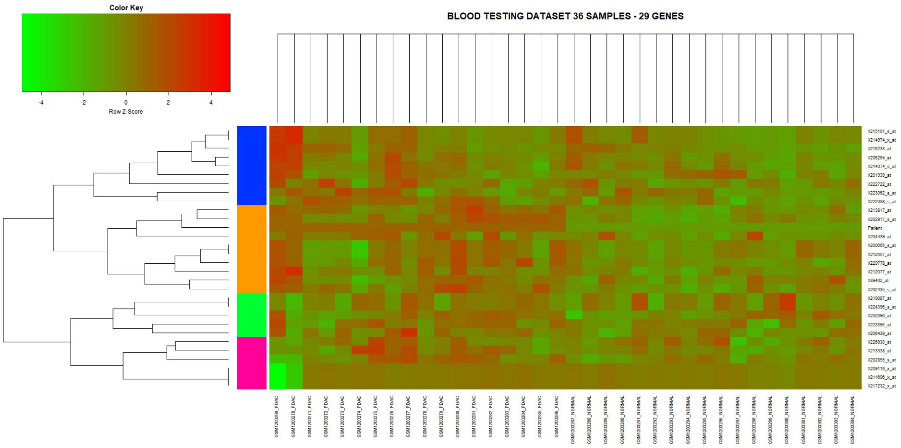
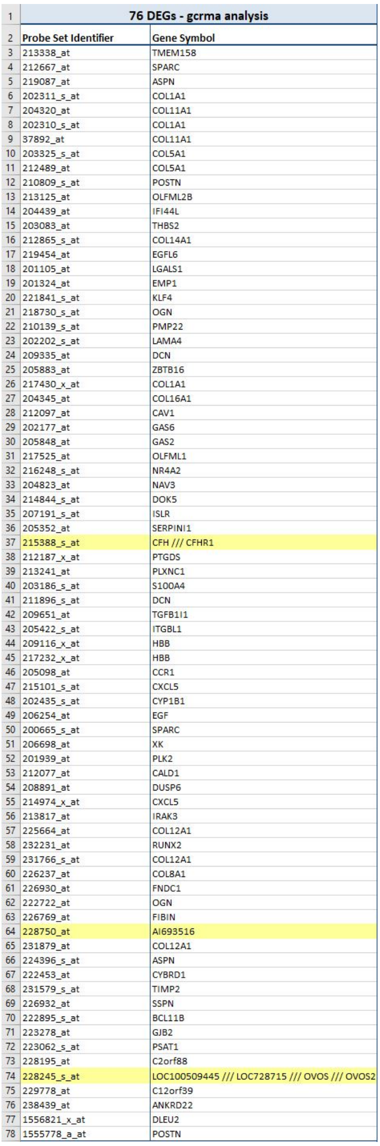


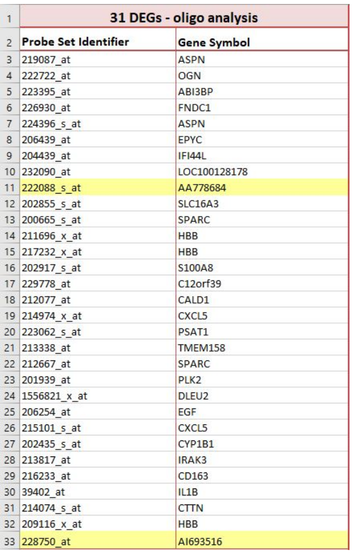
Figure 54: Pdac peripheral blood testing dataset heatmap containing data with 29 DEGs vs 36 pdac patient and healthy samples (oligo analysis)

A.3 GENE LISTS AND THEIR ANNOTATION



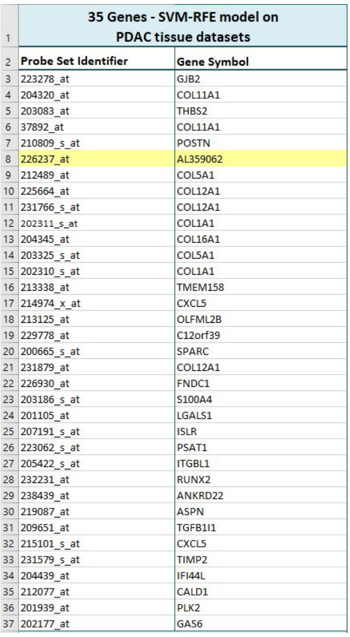
1	76 DEGs - gcrma analysis	
2	**Probe Set Identifier**	**Gene Symbol**
3	213338_at	TMEM158
4	212667_at	SPARC
5	219087_at	ASPN
6	202311_s_at	COL1A1
7	204320_at	COL11A1
8	202310_s_at	COL1A1
9	37892_at	COL11A1
10	203325_s_at	COL5A1
11	212489_at	COL5A1
12	210809_s_at	POSTN
13	213125_at	OLFML2B
14	204439_at	IFI44L
15	203083_at	THBS2
16	212865_s_at	COL14A1
17	219454_at	EGFL6
18	201105_at	LGALS1
19	201324_at	EMP1
20	221841_s_at	KLF4
21	218730_s_at	OGN
22	210139_s_at	PMP22
23	202202_s_at	LAMA4
24	209335_at	DCN
25	205883_at	ZBTB16
26	217430_x_at	COL1A1
27	204345_at	COL16A1
28	212097_at	CAV1
29	202177_at	GAS6
30	205848_at	GAS2
31	217525_at	OLFML1
32	216248_s_at	NR4A2
33	204823_at	NAV3
34	214844_s_at	DOK5
35	207191_s_at	ISLR
36	205352_at	SERPINI1
37	215388_s_at	CFH /// CFHR1
38	212187_x_at	PTGDS
39	213241_at	PLXNC1
40	203186_s_at	S100A4
41	211896_s_at	DCN
42	209651_at	TGFB1I1
43	205422_s_at	ITGBL1
44	209116_x_at	HBB
45	217232_x_at	HBB
46	205098_at	CCR1
47	215101_s_at	CXCL5
48	202435_s_at	CYP1B1
49	206254_at	EGF
50	200665_s_at	SPARC
51	206698_at	XK
52	201939_at	PLK2
53	212077_at	CALD1
54	208891_at	DUSP6
55	214974_x_at	CXCL5
56	213817_at	IRAK3
57	225664_at	COL12A1
58	232231_at	RUNX2
59	231766_s_at	COL12A1
60	226237_at	COL8A1
61	226930_at	FNDC1
62	222722_at	OGN
63	226769_at	FIBIN
64	228750_at	AI693516
65	231879_at	COL12A1
66	224396_s_at	ASPN
67	222453_at	CYBRD1
68	231579_s_at	TIMP2
69	226932_at	SSPN
70	222895_s_at	BCL11B
71	223278_at	GJB2
72	223062_s_at	PSAT1
73	228195_at	C2orf88
74	228245_s_at	LOC100509445 /// LOC728715 /// OVOS /// OVOS2
75	229778_at	C12orf39
76	238439_at	ANKRD22
77	1556821_x_at	DLEU2
78	1555778_a_at	POSTN

Figure 55: 76 DEGs as extracted from the gcrma analysis. The genes are described by their gene symbols using WebGestalt 2013. Identifiers in yellow background were mapped to multiple gene symbols or could not be mapped to any gene symbol



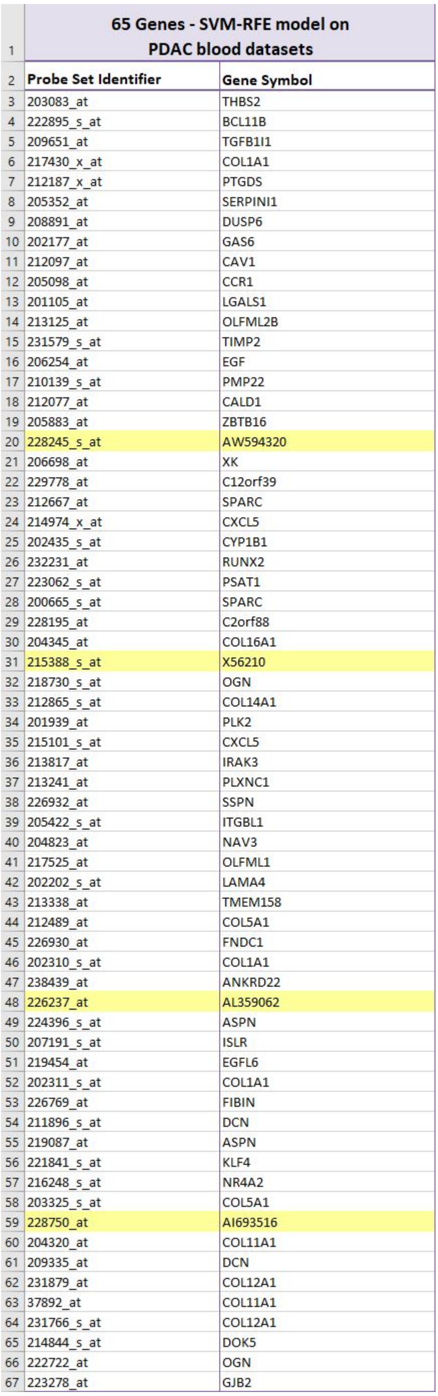
1	31 DEGs - oligo analysis	
2	**Probe Set Identifier**	**Gene Symbol**
3	219087_at	ASPN
4	222722_at	OGN
5	223395_at	ABI3BP
6	226930_at	FNDC1
7	224396_s_at	ASPN
8	206439_at	EPYC
9	204439_at	IFI44L
10	232090_at	LOC100128178
11	222088_s_at	AA778684
12	202855_s_at	SLC16A3
13	200665_s_at	SPARC
14	211696_x_at	HBB
15	217232_x_at	HBB
16	202917_s_at	S100A8
17	229778_at	C12orf39
18	212077_at	CALD1
19	214974_x_at	CXCL5
20	223062_s_at	PSAT1
21	213338_at	TMEM158
22	212667_at	SPARC
23	201939_at	PLK2
24	1556821_x_at	DLEU2
25	206254_at	EGF
26	215101_s_at	CXCL5
27	202435_s_at	CYP1B1
28	213817_at	IRAK3
29	216233_at	CD163
30	39402_at	IL1B
31	214074_s_at	CTTN
32	209116_x_at	HBB
33	228750_at	AI693516

Figure 56: 31 DEGs as extracted from the oligo analysis. The genes are described by their gene symbols using WebGestalt 2013. Identifiers in yellow background could not be mapped to any gene symbol



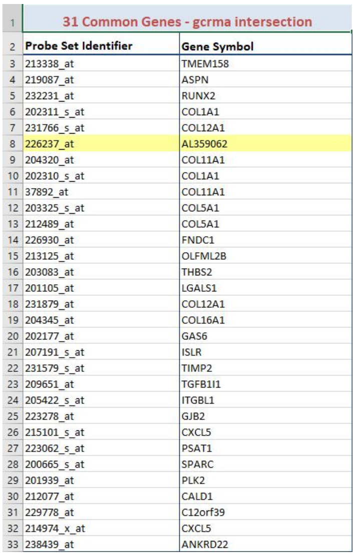
1	35 Genes - SVM-RFE model on PDAC tissue datasets	
2	Probe Set Identifier	Gene Symbol
3	223278_at	GJB2
4	204320_at	COL11A1
5	203083_at	THBS2
6	37892_at	COL11A1
7	210809_s_at	POSTN
8	226237_at	AL359062
9	212489_at	COL5A1
10	225664_at	COL12A1
11	231766_s_at	COL12A1
12	202311_s_at	COL1A1
13	204345_at	COL16A1
14	203325_s_at	COL5A1
15	202310_s_at	COL1A1
16	213338_at	TMEM158
17	214974_x_at	CXCL5
18	213125_at	OLFML2B
19	229778_at	C12orf39
20	200665_s_at	SPARC
21	231879_at	COL12A1
22	226930_at	FNDC1
23	203186_s_at	S100A4
24	201105_at	LGALS1
25	207191_s_at	ISLR
26	223062_s_at	PSAT1
27	205422_s_at	ITGBL1
28	232231_at	RUNX2
29	238439_at	ANKRD22
30	219087_at	ASPN
31	209651_at	TGFB1I1
32	215101_s_at	CXCL5
33	231579_s_at	TIMP2
34	204439_at	IFI44L
35	212077_at	CALD1
36	201939_at	PLK2
37	202177_at	GAS6

Figure 57: 35 genes as selected from SVM-RFE optimal feature selection subset on pdac tissue datasets. The genes are described by their gene symbols using WebGestalt 2013. Identifiers in yellow background could not be mapped to any gene symbol



1	65 Genes - SVM-RFE model on PDAC blood datasets	
2	Probe Set Identifier	Gene Symbol
3	203083_at	THBS2
4	222895_s_at	BCL11B
5	209651_at	TGFB1I1
6	217430_x_at	COL1A1
7	212187_x_at	PTGDS
8	205352_at	SERPINI1
9	208891_at	DUSP6
10	202177_at	GAS6
11	212097_at	CAV1
12	205098_at	CCR1
13	201105_at	LGALS1
14	213125_at	OLFML2B
15	231579_s_at	TIMP2
16	206254_at	EGF
17	210139_s_at	PMP22
18	212077_at	CALD1
19	205883_at	ZBTB16
20	228245_s_at	AW594320
21	206698_at	XK
22	229778_at	C12orf39
23	212667_at	SPARC
24	214974_x_at	CXCL5
25	202435_s_at	CYP1B1
26	232231_at	RUNX2
27	223062_s_at	PSAT1
28	200665_s_at	SPARC
29	228195_at	C2orf88
30	204345_at	COL16A1
31	215388_s_at	X56210
32	218730_s_at	OGN
33	212865_s_at	COL14A1
34	201939_at	PLK2
35	215101_s_at	CXCL5
36	213817_at	IRAK3
37	213241_at	PLXNC1
38	226932_at	SSPN
39	205422_s_at	ITGBL1
40	204823_at	NAV3
41	217525_at	OLFML1
42	202202_s_at	LAMA4
43	213338_at	TMEM158
44	212489_at	COL5A1
45	226930_at	FNDC1
46	202310_s_at	COL1A1
47	238439_at	ANKRD22
48	226237_at	AL359062
49	224396_s_at	ASPN
50	207191_s_at	ISLR
51	219454_at	EGFL6
52	202311_s_at	COL1A1
53	226769_at	FIBIN
54	211896_s_at	DCN
55	219087_at	ASPN
56	221841_s_at	KLF4
57	216248_s_at	NR4A2
58	203325_s_at	COL5A1
59	228750_at	AI693516
60	204320_at	COL11A1
61	209335_at	DCN
62	231879_at	COL12A1
63	37892_at	COL11A1
64	231766_s_at	COL12A1
65	214844_s_at	DOK5
66	222722_at	OGN
67	223278_at	GJB2

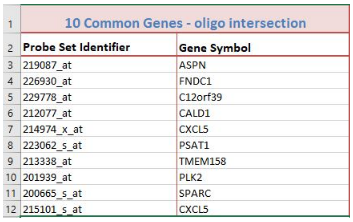
Figure 58: 65 genes as selected from SVM-RFE optimal feature selection subset on pdac blood datasets. The genes are described by their gene symbols using WebGestalt 2013. Identifiers in yellow background could not be mapped to any gene symbol



31 Common Genes - gcrma intersection

Probe Set Identifier	Gene Symbol
213338_at	TMEM158
219087_at	ASPN
232231_at	RUNX2
202311_s_at	COL1A1
231766_s_at	COL12A1
226237_at	AL359062
204320_at	COL11A1
202310_s_at	COL1A1
37892_at	COL11A1
203325_s_at	COL5A1
212489_at	COL5A1
226930_at	FNDC1
213125_at	OLFML2B
203083_at	THBS2
201105_at	LGALS1
231879_at	COL12A1
204345_at	COL16A1
202177_at	GAS6
207191_s_at	ISLR
231579_s_at	TIMP2
209651_at	TGFB1I1
205422_s_at	ITGBL1
223278_at	GJB2
215101_s_at	CXCL5
223062_s_at	PSAT1
200665_s_at	SPARC
201939_at	PLK2
212077_at	CALD1
229778_at	C12orf39
214974_x_at	CXCL5
238439_at	ANKRD22

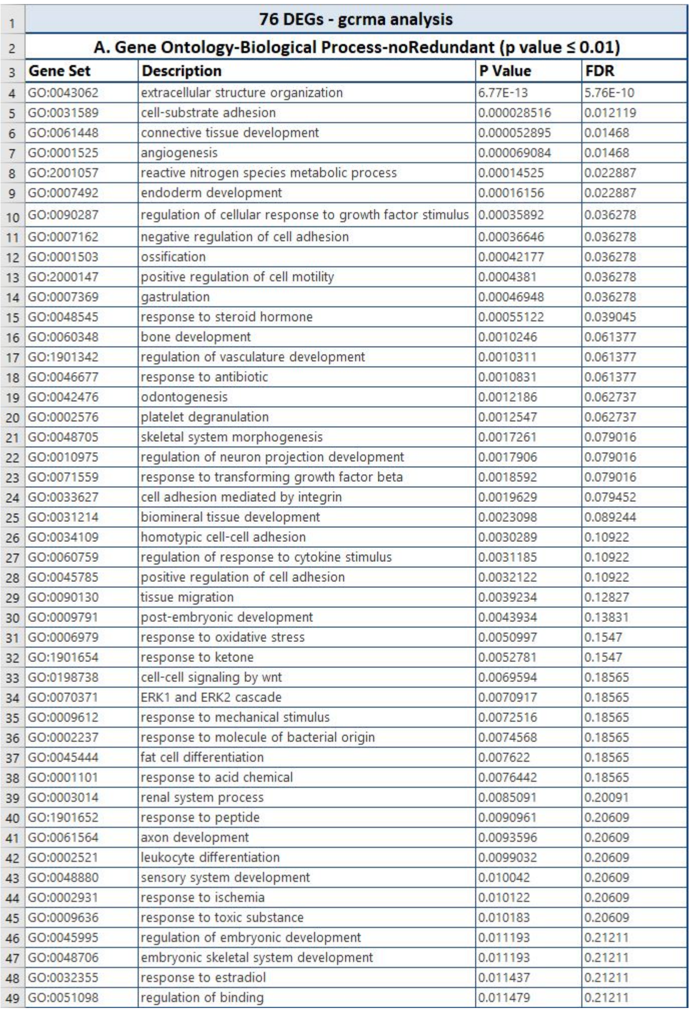
Figure 59: 31 genes as extracted from the intersection of the subset of gcrma analysis, and the optimal subsets from SVM-RFE (tissue and blood). The genes are described by their gene symbols using WebGestalt 2013. Identifier in yellow background could not be mapped to any gene symbol



10 Common Genes - oligo intersection

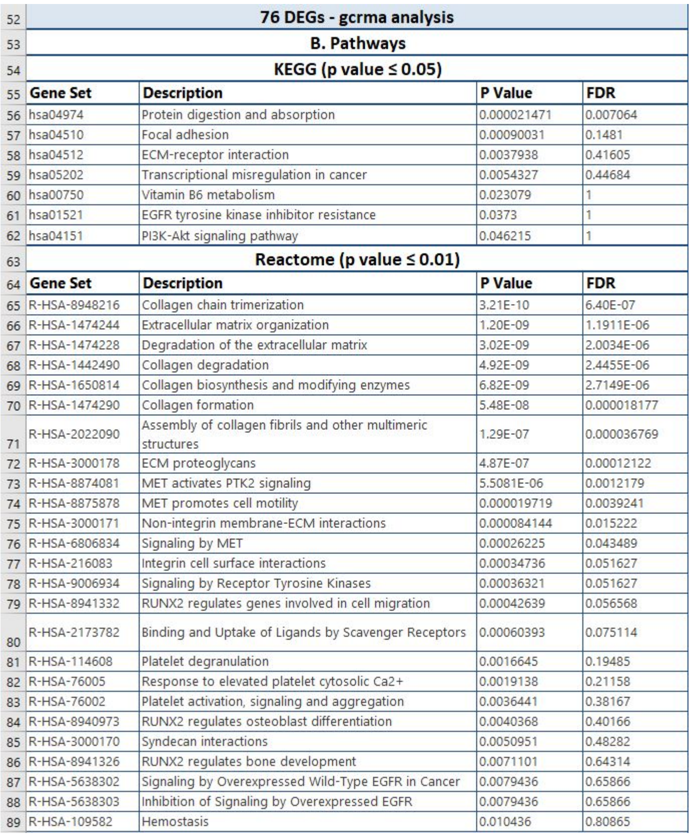
Probe Set Identifier	Gene Symbol
219087_at	ASPN
226930_at	FNDC1
229778_at	C12orf39
212077_at	CALD1
214974_x_at	CXCL5
223062_s_at	PSAT1
213338_at	TMEM158
201939_at	PLK2
200665_s_at	SPARC
215101_s_at	CXCL5

Figure 60: 10 genes as extracted from the intersection of the subset of oligo analysis, and the optimal subsets from SVM-RFE (tissue and blood). The genes are described by their gene symbols using WebGestalt 2013



1	**76 DEGs - gcrma analysis**			
2	**A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)**			
3	**Gene Set**	**Description**	**P Value**	**FDR**
4	GO:0043062	extracellular structure organization	6.77E-13	5.76E-10
5	GO:0031589	cell-substrate adhesion	0.000028516	0.012119
6	GO:0061448	connective tissue development	0.000052895	0.01468
7	GO:0001525	angiogenesis	0.000069084	0.01468
8	GO:2001057	reactive nitrogen species metabolic process	0.00014525	0.022887
9	GO:0007492	endoderm development	0.00016156	0.022887
10	GO:0090287	regulation of cellular response to growth factor stimulus	0.00035892	0.036278
11	GO:0007162	negative regulation of cell adhesion	0.00036646	0.036278
12	GO:0001503	ossification	0.00042177	0.036278
13	GO:2000147	positive regulation of cell motility	0.0004381	0.036278
14	GO:0007369	gastrulation	0.00046948	0.036278
15	GO:0048545	response to steroid hormone	0.00055122	0.039045
16	GO:0060348	bone development	0.0010246	0.061377
17	GO:1901342	regulation of vasculature development	0.0010311	0.061377
18	GO:0046677	response to antibiotic	0.0010831	0.061377
19	GO:0042476	odontogenesis	0.0012186	0.062737
20	GO:0002576	platelet degranulation	0.0012547	0.062737
21	GO:0048705	skeletal system morphogenesis	0.0017261	0.079016
22	GO:0010975	regulation of neuron projection development	0.0017906	0.079016
23	GO:0071559	response to transforming growth factor beta	0.0018592	0.079016
24	GO:0033627	cell adhesion mediated by integrin	0.0019629	0.079452
25	GO:0031214	biomineral tissue development	0.0023098	0.089244
26	GO:0034109	homotypic cell-cell adhesion	0.0030289	0.10922
27	GO:0060759	regulation of response to cytokine stimulus	0.0031185	0.10922
28	GO:0045785	positive regulation of cell adhesion	0.0032122	0.10922
29	GO:0090130	tissue migration	0.0039234	0.12827
30	GO:0009791	post-embryonic development	0.0043934	0.13831
31	GO:0006979	response to oxidative stress	0.0050997	0.1547
32	GO:1901654	response to ketone	0.0052781	0.1547
33	GO:0198738	cell-cell signaling by wnt	0.0069594	0.18565
34	GO:0070371	ERK1 and ERK2 cascade	0.0070917	0.18565
35	GO:0009612	response to mechanical stimulus	0.0072516	0.18565
36	GO:0002237	response to molecule of bacterial origin	0.0074568	0.18565
37	GO:0045444	fat cell differentiation	0.007622	0.18565
38	GO:0001101	response to acid chemical	0.0076442	0.18565
39	GO:0003014	renal system process	0.0085091	0.20091
40	GO:1901652	response to peptide	0.0090961	0.20609
41	GO:0061564	axon development	0.0093596	0.20609
42	GO:0002521	leukocyte differentiation	0.0099032	0.20609
43	GO:0048880	sensory system development	0.010042	0.20609
44	GO:0002931	response to ischemia	0.010122	0.20609
45	GO:0009636	response to toxic substance	0.010183	0.20609
46	GO:0045995	regulation of embryonic development	0.011193	0.21211
47	GO:0048706	embryonic skeletal system development	0.011193	0.21211
48	GO:0032355	response to estradiol	0.011437	0.21211
49	GO:0051098	regulation of binding	0.011479	0.21211

Figure 61: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 76 DEGs as extracted from the gcrma analysis



76 DEGs - gcrma analysis

B. Pathways

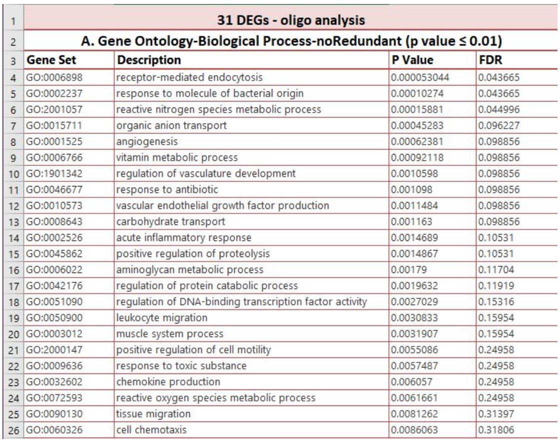
KEGG (p value ≤ 0.05)

Gene Set	Description	P Value	FDR
hsa04974	Protein digestion and absorption	0.000021471	0.007064
hsa04510	Focal adhesion	0.00090031	0.1481
hsa04512	ECM-receptor interaction	0.0037938	0.41605
hsa05202	Transcriptional misregulation in cancer	0.0054327	0.44684
hsa00750	Vitamin B6 metabolism	0.023079	1
hsa01521	EGFR tyrosine kinase inhibitor resistance	0.0373	1
hsa04151	PI3K-Akt signaling pathway	0.046215	1

Reactome (p value ≤ 0.01)

Gene Set	Description	P Value	FDR
R-HSA-8948216	Collagen chain trimerization	3.21E-10	6.40E-07
R-HSA-1474244	Extracellular matrix organization	1.20E-09	1.1911E-06
R-HSA-1474228	Degradation of the extracellular matrix	3.02E-09	2.0034E-06
R-HSA-1442490	Collagen degradation	4.92E-09	2.4455E-06
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	6.82E-09	2.7149E-06
R-HSA-1474290	Collagen formation	5.48E-08	0.000018177
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	1.29E-07	0.000036769
R-HSA-3000178	ECM proteoglycans	4.87E-07	0.00012122
R-HSA-8874081	MET activates PTK2 signaling	5.5081E-06	0.0012179
R-HSA-8875878	MET promotes cell motility	0.000019719	0.0039241
R-HSA-3000171	Non-integrin membrane-ECM interactions	0.000084144	0.015222
R-HSA-6806834	Signaling by MET	0.00026225	0.043489
R-HSA-216083	Integrin cell surface interactions	0.00034736	0.051627
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	0.00036321	0.051627
R-HSA-8941332	RUNX2 regulates genes involved in cell migration	0.00042639	0.056568
R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.00060393	0.075114
R-HSA-114608	Platelet degranulation	0.0016645	0.19485
R-HSA-76005	Response to elevated platelet cytosolic Ca2+	0.0019138	0.21158
R-HSA-76002	Platelet activation, signaling and aggregation	0.0036441	0.38167
R-HSA-8940973	RUNX2 regulates osteoblast differentiation	0.0040368	0.40166
R-HSA-3000170	Syndecan interactions	0.0050951	0.48282
R-HSA-8941326	RUNX2 regulates bone development	0.0071101	0.64314
R-HSA-5638302	Signaling by Overexpressed Wild-Type EGFR in Cancer	0.0079436	0.65866
R-HSA-5638303	Inhibition of Signaling by Overexpressed EGFR	0.0079436	0.65866
R-HSA-109582	Hemostasis	0.010436	0.80865

Figure 62: b) Pathway annotation (KEGG, Reactome) of 76 DEGs as extracted from the gcrma analysis

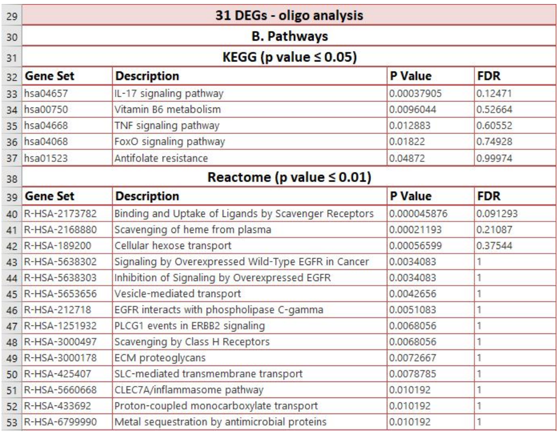


31 DEGs - oligo analysis

A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)

Gene Set	Description	P Value	FDR
GO:0006898	receptor-mediated endocytosis	0.000053044	0.043665
GO:0002237	response to molecule of bacterial origin	0.00010274	0.043665
GO:2001057	reactive nitrogen species metabolic process	0.00015881	0.044996
GO:0015711	organic anion transport	0.00045283	0.096227
GO:0001525	angiogenesis	0.00062381	0.098856
GO:0006766	vitamin metabolic process	0.00092118	0.098856
GO:1901342	regulation of vasculature development	0.0010598	0.098856
GO:0046677	response to antibiotic	0.001098	0.098856
GO:0010573	vascular endothelial growth factor production	0.0011484	0.098856
GO:0008643	carbohydrate transport	0.001163	0.098856
GO:0002526	acute inflammatory response	0.0014689	0.10531
GO:0045862	positive regulation of proteolysis	0.0014867	0.10531
GO:0006022	aminoglycan metabolic process	0.00179	0.11704
GO:0042176	regulation of protein catabolic process	0.0019632	0.11919
GO:0051090	regulation of DNA-binding transcription factor activity	0.0027029	0.15316
GO:0050900	leukocyte migration	0.0030833	0.15954
GO:0003012	muscle system process	0.0031907	0.15954
GO:2000147	positive regulation of cell motility	0.0055086	0.24958
GO:0009636	response to toxic substance	0.0057487	0.24958
GO:0032602	chemokine production	0.006057	0.24958
GO:0072593	reactive oxygen species metabolic process	0.0061661	0.24958
GO:0090130	tissue migration	0.0081262	0.31397
GO:0060326	cell chemotaxis	0.0086063	0.31806

Figure 63: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 31 DEGs as extracted from the oligo analysis



31 DEGs - oligo analysis

B. Pathways

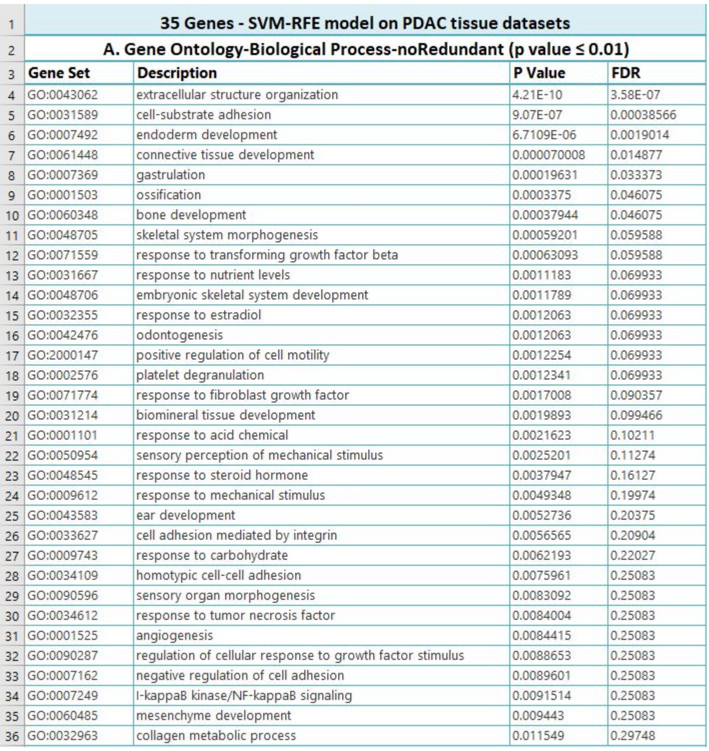
KEGG (p value ≤ 0.05)

Gene Set	Description	P Value	FDR
hsa04657	IL-17 signaling pathway	0.00037905	0.12471
hsa00750	Vitamin B6 metabolism	0.0096044	0.52664
hsa04668	TNF signaling pathway	0.012883	0.60552
hsa04068	FoxO signaling pathway	0.01822	0.74928
hsa01523	Antifolate resistance	0.04872	0.99974

Reactome (p value ≤ 0.01)

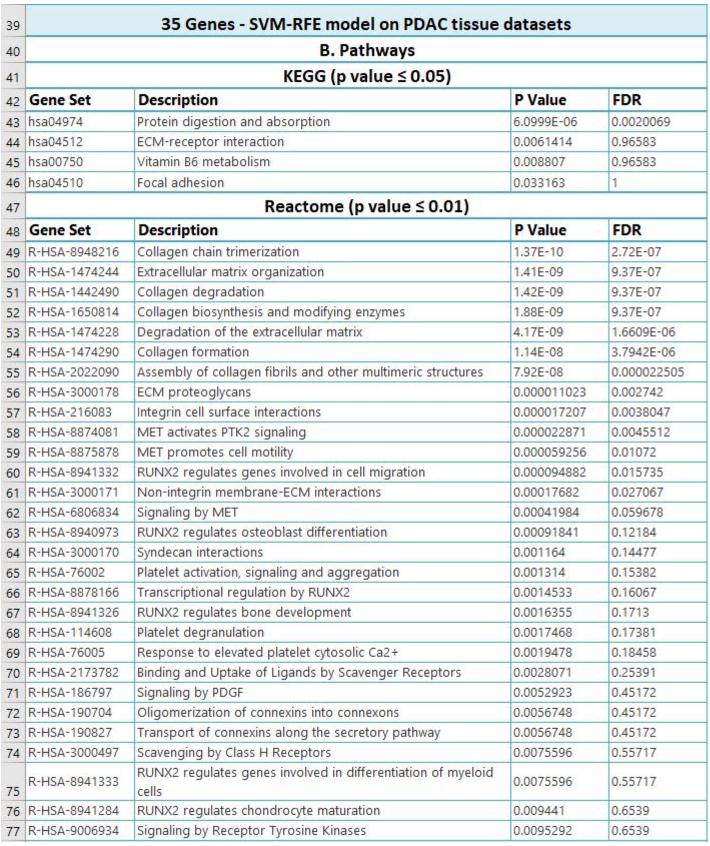
Gene Set	Description	P Value	FDR
R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.000045876	0.091293
R-HSA-2168880	Scavenging of heme from plasma	0.00021193	0.21087
R-HSA-189200	Cellular hexose transport	0.00056599	0.37544
R-HSA-5638302	Signaling by Overexpressed Wild-Type EGFR in Cancer	0.0034083	1
R-HSA-5638303	Inhibition of Signaling by Overexpressed EGFR	0.0034083	1
R-HSA-5653656	Vesicle-mediated transport	0.0042656	1
R-HSA-212718	EGFR interacts with phospholipase C-gamma	0.0051083	1
R-HSA-1251932	PLCG1 events in ERBB2 signaling	0.0068056	1
R-HSA-3000497	Scavenging by Class H Receptors	0.0068056	1
R-HSA-3000178	ECM proteoglycans	0.0072667	1
R-HSA-425407	SLC-mediated transmembrane transport	0.0078785	1
R-HSA-5660668	CLEC7A/inflammasome pathway	0.010192	1
R-HSA-433692	Proton-coupled monocarboxylate transport	0.010192	1
R-HSA-6799990	Metal sequestration by antimicrobial proteins	0.010192	1

Figure 64: b) Pathway annotation (KEGG, Reactome) of 31 DEGs as extracted from the oligo analysis



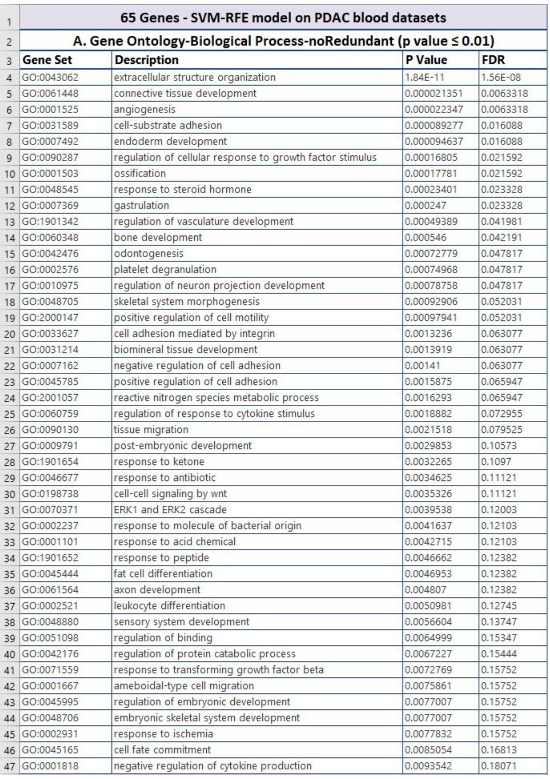
1	35 Genes - SVM-RFE model on PDAC tissue datasets			
2	A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)			
3	**Gene Set**	**Description**	**P Value**	**FDR**
4	GO:0043062	extracellular structure organization	4.21E-10	3.58E-07
5	GO:0031589	cell-substrate adhesion	9.07E-07	0.00038566
6	GO:0007492	endoderm development	6.7109E-06	0.0019014
7	GO:0061448	connective tissue development	0.000070008	0.014877
8	GO:0007369	gastrulation	0.00019631	0.033373
9	GO:0001503	ossification	0.0003375	0.046075
10	GO:0060348	bone development	0.00037944	0.046075
11	GO:0048705	skeletal system morphogenesis	0.00059201	0.059588
12	GO:0071559	response to transforming growth factor beta	0.00063093	0.059588
13	GO:0031667	response to nutrient levels	0.0011183	0.069933
14	GO:0048706	embryonic skeletal system development	0.0011789	0.069933
15	GO:0032355	response to estradiol	0.0012063	0.069933
16	GO:0042476	odontogenesis	0.0012063	0.069933
17	GO:2000147	positive regulation of cell motility	0.0012254	0.069933
18	GO:0002576	platelet degranulation	0.0012341	0.069933
19	GO:0071774	response to fibroblast growth factor	0.0017008	0.090357
20	GO:0031214	biomineral tissue development	0.0019893	0.099466
21	GO:0001101	response to acid chemical	0.0021623	0.10211
22	GO:0050954	sensory perception of mechanical stimulus	0.0025201	0.11274
23	GO:0048545	response to steroid hormone	0.0037947	0.16127
24	GO:0009612	response to mechanical stimulus	0.0049348	0.19974
25	GO:0043583	ear development	0.0052736	0.20375
26	GO:0033627	cell adhesion mediated by integrin	0.0056565	0.20904
27	GO:0009743	response to carbohydrate	0.0062193	0.22027
28	GO:0034109	homotypic cell-cell adhesion	0.0075961	0.25083
29	GO:0090596	sensory organ morphogenesis	0.0083092	0.25083
30	GO:0034612	response to tumor necrosis factor	0.0084004	0.25083
31	GO:0001525	angiogenesis	0.0084415	0.25083
32	GO:0090287	regulation of cellular response to growth factor stimulus	0.0088653	0.25083
33	GO:0007162	negative regulation of cell adhesion	0.0089601	0.25083
34	GO:0007249	I-kappaB kinase/NF-kappaB signaling	0.0091514	0.25083
35	GO:0060485	mesenchyme development	0.009443	0.25083
36	GO:0032963	collagen metabolic process	0.011549	0.29748

Figure 65: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 35 gene list as selected from SVM-RFE optimal feature selection subset on pdac tissue datasets



39	35 Genes - SVM-RFE model on PDAC tissue datasets			
40	B. Pathways			
41	KEGG (p value ≤ 0.05)			
42	**Gene Set**	**Description**	**P Value**	**FDR**
43	hsa04974	Protein digestion and absorption	6.0999E-06	0.0020069
44	hsa04512	ECM-receptor interaction	0.0061414	0.96583
45	hsa00750	Vitamin B6 metabolism	0.008807	0.96583
46	hsa04510	Focal adhesion	0.033163	1
47	Reactome (p value ≤ 0.01)			
48	**Gene Set**	**Description**	**P Value**	**FDR**
49	R-HSA-8948216	Collagen chain trimerization	1.37E-10	2.72E-07
50	R-HSA-1474244	Extracellular matrix organization	1.41E-09	9.37E-07
51	R-HSA-1442490	Collagen degradation	1.42E-09	9.37E-07
52	R-HSA-1650814	Collagen biosynthesis and modifying enzymes	1.88E-09	9.37E-07
53	R-HSA-1474228	Degradation of the extracellular matrix	4.17E-09	1.6609E-06
54	R-HSA-1474290	Collagen formation	1.14E-08	3.7942E-06
55	R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	7.92E-08	0.000022505
56	R-HSA-3000178	ECM proteoglycans	0.000011023	0.002742
57	R-HSA-216083	Integrin cell surface interactions	0.000017207	0.0038047
58	R-HSA-8874081	MET activates PTK2 signaling	0.000022871	0.0045512
59	R-HSA-8875878	MET promotes cell motility	0.000059256	0.01072
60	R-HSA-8941332	RUNX2 regulates genes involved in cell migration	0.000094882	0.015735
61	R-HSA-3000171	Non-integrin membrane-ECM interactions	0.00017682	0.027067
62	R-HSA-6806834	Signaling by MET	0.00041984	0.059678
63	R-HSA-8940973	RUNX2 regulates osteoblast differentiation	0.00091841	0.12184
64	R-HSA-3000170	Syndecan interactions	0.001164	0.14477
65	R-HSA-76002	Platelet activation, signaling and aggregation	0.001314	0.15382
66	R-HSA-8878166	Transcriptional regulation by RUNX2	0.0014533	0.16067
67	R-HSA-8941326	RUNX2 regulates bone development	0.0016355	0.1713
68	R-HSA-114608	Platelet degranulation	0.0017468	0.17381
69	R-HSA-76005	Response to elevated platelet cytosolic Ca2+	0.0019478	0.18458
70	R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.0028071	0.25391
71	R-HSA-186797	Signaling by PDGF	0.0052923	0.45172
72	R-HSA-190704	Oligomerization of connexins into connexons	0.0056748	0.45172
73	R-HSA-190827	Transport of connexins along the secretory pathway	0.0056748	0.45172
74	R-HSA-3000497	Scavenging by Class H Receptors	0.0075596	0.55717
75	R-HSA-8941333	RUNX2 regulates genes involved in differentiation of myeloid cells	0.0075596	0.55717
76	R-HSA-8941284	RUNX2 regulates chondrocyte maturation	0.009441	0.6539
77	R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	0.0095292	0.6539

Figure 66: b) Pathway annotation (KEGG, Reactome) of 35 gene list as selected from SVM-RFE optimal feature selection subset on pdac tissue datasets

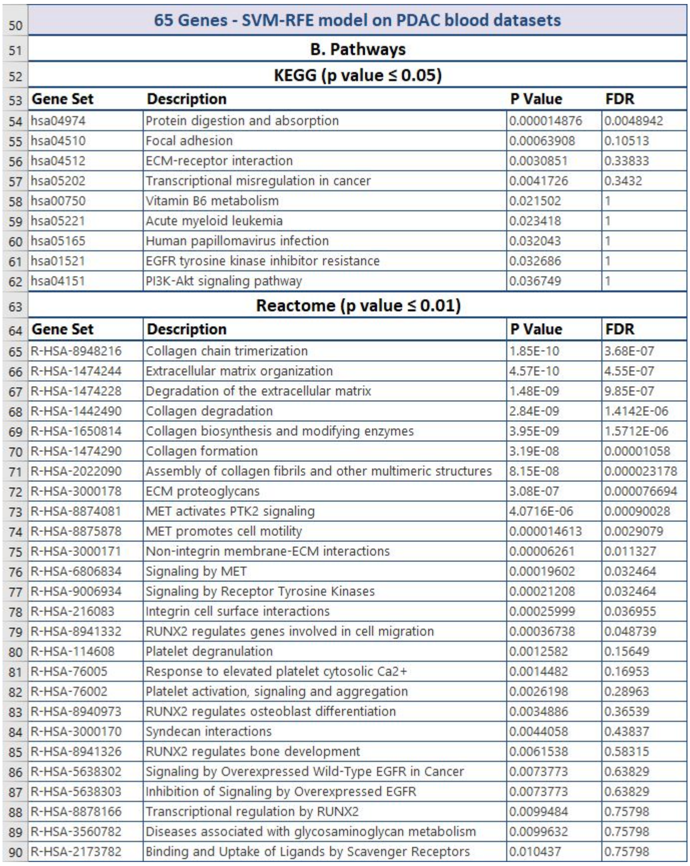


65 Genes - SVM-RFE model on PDAC blood datasets

A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)

Gene Set	Description	P Value	FDR
GO:0043062	extracellular structure organization	1.84E-11	1.56E-08
GO:0061448	connective tissue development	0.000021351	0.0063318
GO:0001525	angiogenesis	0.000022347	0.0063318
GO:0031589	cell-substrate adhesion	0.000089277	0.016088
GO:0007492	endoderm development	0.000094637	0.016088
GO:0090287	regulation of cellular response to growth factor stimulus	0.00016805	0.021592
GO:0001503	ossification	0.00017781	0.021592
GO:0048545	response to steroid hormone	0.00023401	0.023328
GO:0007369	gastrulation	0.000247	0.023328
GO:1901342	regulation of vasculature development	0.00049389	0.041981
GO:0060348	bone development	0.000546	0.042191
GO:0042476	odontogenesis	0.00072779	0.047817
GO:0002576	platelet degranulation	0.00074968	0.047817
GO:0010975	regulation of neuron projection development	0.00078758	0.047817
GO:0048705	skeletal system morphogenesis	0.00092906	0.052031
GO:2000147	positive regulation of cell motility	0.00097941	0.052031
GO:0033627	cell adhesion mediated by integrin	0.0013236	0.063077
GO:0031214	biomineral tissue development	0.0013919	0.063077
GO:0007162	negative regulation of cell adhesion	0.00141	0.063077
GO:0045785	positive regulation of cell adhesion	0.0015875	0.065947
GO:2001057	reactive nitrogen species metabolic process	0.0016293	0.065947
GO:0060759	regulation of response to cytokine stimulus	0.0018882	0.072955
GO:0090130	tissue migration	0.0021518	0.079525
GO:0009791	post-embryonic development	0.0029853	0.10573
GO:1901654	response to ketone	0.0032265	0.1097
GO:0046677	response to antibiotic	0.0034625	0.11121
GO:0198738	cell-cell signaling by wnt	0.0035326	0.11121
GO:0070371	ERK1 and ERK2 cascade	0.0039538	0.12003
GO:0002237	response to molecule of bacterial origin	0.0041637	0.12103
GO:0001101	response to acid chemical	0.0042715	0.12103
GO:1901652	response to peptide	0.0046662	0.12382
GO:0045444	fat cell differentiation	0.0046953	0.12382
GO:0061564	axon development	0.004807	0.12382
GO:0002521	leukocyte differentiation	0.0050981	0.12745
GO:0048880	sensory system development	0.0056604	0.13747
GO:0051098	regulation of binding	0.0064999	0.15347
GO:0042176	regulation of protein catabolic process	0.0067227	0.15444
GO:0071559	response to transforming growth factor beta	0.0072769	0.15752
GO:0001667	ameboidal-type cell migration	0.0075861	0.15752
GO:0045995	regulation of embryonic development	0.0077007	0.15752
GO:0048706	embryonic skeletal system development	0.0077007	0.15752
GO:0002931	response to ischemia	0.0077832	0.15752
GO:0045165	cell fate commitment	0.0085054	0.16813
GO:0001818	negative regulation of cytokine production	0.0093542	0.18071

Figure 67: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 65 gene list as selected from SVM-RFE optimal feature selection subset on pdac blood datasets



65 Genes - SVM-RFE model on PDAC blood datasets

B. Pathways

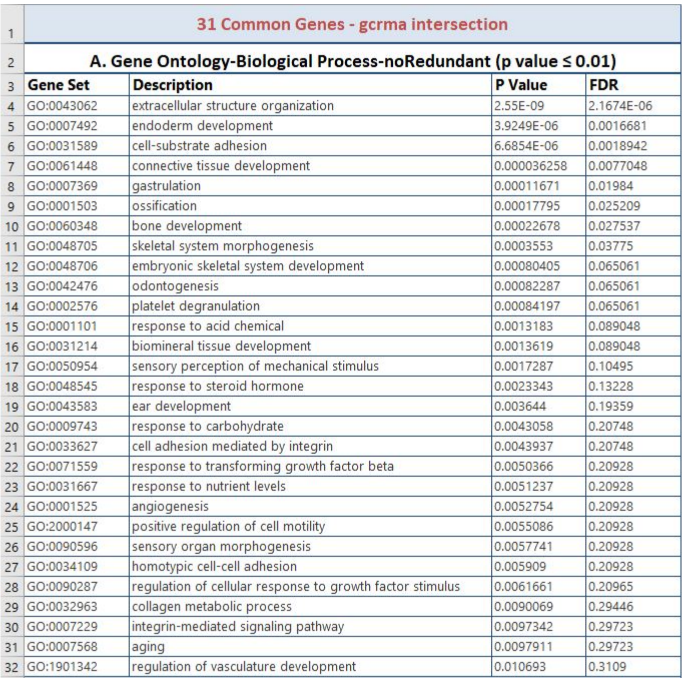
KEGG (p value ≤ 0.05)

Gene Set	Description	P Value	FDR
hsa04974	Protein digestion and absorption	0.000014876	0.0048942
hsa04510	Focal adhesion	0.00063908	0.10513
hsa04512	ECM-receptor interaction	0.0030851	0.33833
hsa05202	Transcriptional misregulation in cancer	0.0041726	0.3432
hsa00750	Vitamin B6 metabolism	0.021502	1
hsa05221	Acute myeloid leukemia	0.023418	1
hsa05165	Human papillomavirus infection	0.032043	1
hsa01521	EGFR tyrosine kinase inhibitor resistance	0.032686	1
hsa04151	PI3K-Akt signaling pathway	0.036749	1

Reactome (p value ≤ 0.01)

Gene Set	Description	P Value	FDR
R-HSA-8948216	Collagen chain trimerization	1.85E-10	3.68E-07
R-HSA-1474244	Extracellular matrix organization	4.57E-10	4.55E-07
R-HSA-1474228	Degradation of the extracellular matrix	1.48E-09	9.85E-07
R-HSA-1442490	Collagen degradation	2.84E-09	1.4142E-06
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	3.95E-09	1.5712E-06
R-HSA-1474290	Collagen formation	3.19E-08	0.00001058
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	8.15E-08	0.000023178
R-HSA-3000178	ECM proteoglycans	3.08E-07	0.000076694
R-HSA-8874081	MET activates PTK2 signaling	4.0716E-06	0.00090028
R-HSA-8875878	MET promotes cell motility	0.000014613	0.0029079
R-HSA-3000171	Non-integrin membrane-ECM interactions	0.00006261	0.011327
R-HSA-6806834	Signaling by MET	0.00019602	0.032464
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	0.00021208	0.032464
R-HSA-216083	Integrin cell surface interactions	0.00025999	0.036955
R-HSA-8941332	RUNX2 regulates genes involved in cell migration	0.00036738	0.048739
R-HSA-114608	Platelet degranulation	0.0012582	0.15649
R-HSA-76005	Response to elevated platelet cytosolic Ca2+	0.0014482	0.16953
R-HSA-76002	Platelet activation, signaling and aggregation	0.0026198	0.28963
R-HSA-8940973	RUNX2 regulates osteoblast differentiation	0.0034886	0.36539
R-HSA-3000170	Syndecan interactions	0.0044058	0.43837
R-HSA-8941326	RUNX2 regulates bone development	0.0061538	0.58315
R-HSA-5638302	Signaling by Overexpressed Wild-Type EGFR in Cancer	0.0073773	0.63829
R-HSA-5638303	Inhibition of Signaling by Overexpressed EGFR	0.0073773	0.63829
R-HSA-8878166	Transcriptional regulation by RUNX2	0.0099484	0.75798
R-HSA-3560782	Diseases associated with glycosaminoglycan metabolism	0.0099632	0.75798
R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.010437	0.75798

Figure 68: b) Pathway annotation (KEGG, Reactome) of 65 gene list as selected from SVM-RFE optimal feature selection subset on pdac blood datasets

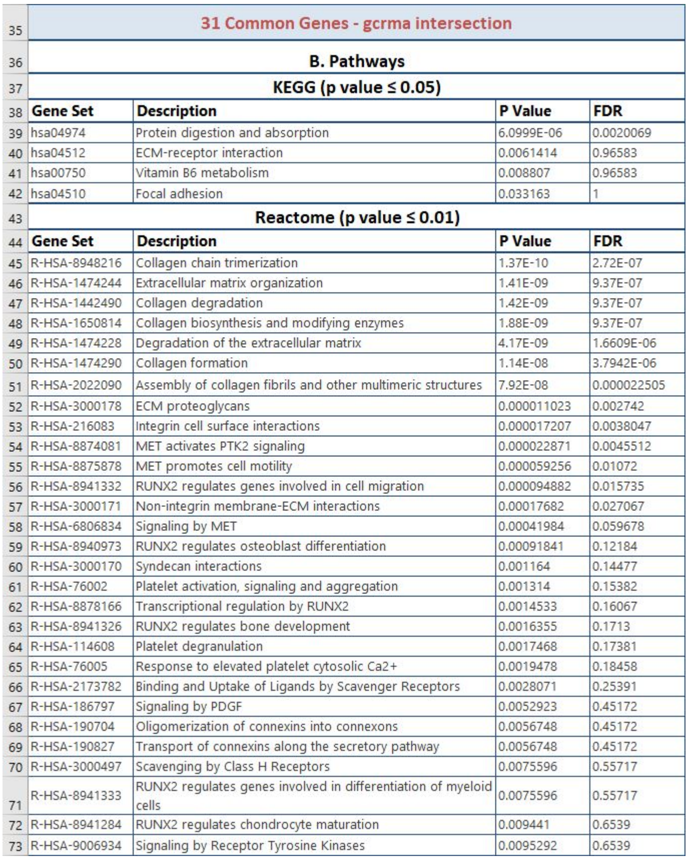


31 Common Genes - gcrma intersection

A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)

Gene Set	Description	P Value	FDR
GO:0043062	extracellular structure organization	2.55E-09	2.1674E-06
GO:0007492	endoderm development	3.9249E-06	0.0016681
GO:0031589	cell-substrate adhesion	6.6854E-06	0.0018942
GO:0061448	connective tissue development	0.000036258	0.0077048
GO:0007369	gastrulation	0.00011671	0.01984
GO:0001503	ossification	0.00017795	0.025209
GO:0060348	bone development	0.00022678	0.027537
GO:0048705	skeletal system morphogenesis	0.0003553	0.03775
GO:0048706	embryonic skeletal system development	0.00080405	0.065061
GO:0042476	odontogenesis	0.00082287	0.065061
GO:0002576	platelet degranulation	0.00084197	0.065061
GO:0001101	response to acid chemical	0.0013183	0.089048
GO:0031214	biomineral tissue development	0.0013619	0.089048
GO:0050954	sensory perception of mechanical stimulus	0.0017287	0.10495
GO:0048545	response to steroid hormone	0.0023343	0.13228
GO:0043583	ear development	0.003644	0.19359
GO:0009743	response to carbohydrate	0.0043058	0.20748
GO:0033627	cell adhesion mediated by integrin	0.0043937	0.20748
GO:0071559	response to transforming growth factor beta	0.0050366	0.20928
GO:0031667	response to nutrient levels	0.0051237	0.20928
GO:0001525	angiogenesis	0.0052754	0.20928
GO:2000147	positive regulation of cell motility	0.0055086	0.20928
GO:0090596	sensory organ morphogenesis	0.0057741	0.20928
GO:0034109	homotypic cell-cell adhesion	0.005909	0.20928
GO:0090287	regulation of cellular response to growth factor stimulus	0.0061661	0.20965
GO:0032963	collagen metabolic process	0.0090069	0.29446
GO:0007229	integrin-mediated signaling pathway	0.0097342	0.29723
GO:0007568	aging	0.0097911	0.29723
GO:1901342	regulation of vasculature development	0.010693	0.3109

Figure 69: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 31 common gene list as extracted from the intersection of the gcrma analysis subset, and the optimal SVM-RFE (tissue and blood) subsets



35 31 Common Genes - gcrma intersection

36 B. Pathways

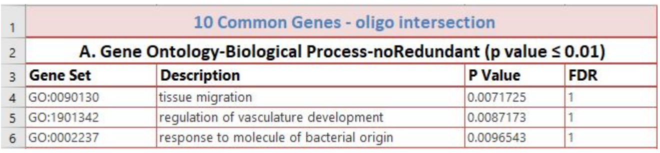
37 KEGG (p value ≤ 0.05)

	Gene Set	Description	P Value	FDR
39	hsa04974	Protein digestion and absorption	6.0999E-06	0.0020069
40	hsa04512	ECM-receptor interaction	0.0061414	0.96583
41	hsa00750	Vitamin B6 metabolism	0.008807	0.96583
42	hsa04510	Focal adhesion	0.033163	1

43 Reactome (p value ≤ 0.01)

	Gene Set	Description	P Value	FDR
45	R-HSA-8948216	Collagen chain trimerization	1.37E-10	2.72E-07
46	R-HSA-1474244	Extracellular matrix organization	1.41E-09	9.37E-07
47	R-HSA-1442490	Collagen degradation	1.42E-09	9.37E-07
48	R-HSA-1650814	Collagen biosynthesis and modifying enzymes	1.88E-09	9.37E-07
49	R-HSA-1474228	Degradation of the extracellular matrix	4.17E-09	1.6609E-06
50	R-HSA-1474290	Collagen formation	1.14E-08	3.7942E-06
51	R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	7.92E-08	0.000022505
52	R-HSA-3000178	ECM proteoglycans	0.000011023	0.002742
53	R-HSA-216083	Integrin cell surface interactions	0.000017207	0.0038047
54	R-HSA-8874081	MET activates PTK2 signaling	0.000022871	0.0045512
55	R-HSA-8875878	MET promotes cell motility	0.000059256	0.01072
56	R-HSA-8941332	RUNX2 regulates genes involved in cell migration	0.000094882	0.015735
57	R-HSA-3000171	Non-integrin membrane-ECM interactions	0.00017682	0.027067
58	R-HSA-6806834	Signaling by MET	0.00041984	0.059678
59	R-HSA-8940973	RUNX2 regulates osteoblast differentiation	0.00091841	0.12184
60	R-HSA-3000170	Syndecan interactions	0.001164	0.14477
61	R-HSA-76002	Platelet activation, signaling and aggregation	0.001314	0.15382
62	R-HSA-8878166	Transcriptional regulation by RUNX2	0.0014533	0.16067
63	R-HSA-8941326	RUNX2 regulates bone development	0.0016355	0.1713
64	R-HSA-114608	Platelet degranulation	0.0017468	0.17381
65	R-HSA-76005	Response to elevated platelet cytosolic Ca2+	0.0019478	0.18458
66	R-HSA-2173782	Binding and Uptake of Ligands by Scavenger Receptors	0.0028071	0.25391
67	R-HSA-186797	Signaling by PDGF	0.0052923	0.45172
68	R-HSA-190704	Oligomerization of connexins into connexons	0.0056748	0.45172
69	R-HSA-190827	Transport of connexins along the secretory pathway	0.0056748	0.45172
70	R-HSA-3000497	Scavenging by Class H Receptors	0.0075596	0.55717
71	R-HSA-8941333	RUNX2 regulates genes involved in differentiation of myeloid cells	0.0075596	0.55717
72	R-HSA-8941284	RUNX2 regulates chondrocyte maturation	0.009441	0.6539
73	R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	0.0095292	0.6539

Figure 70: b) Pathway annotation (KEGG, Reactome) of 31 common gene list as extracted from the intersection of the gcrma analysis subset, and the optimal SVM-RFE (tissue and blood) subsets

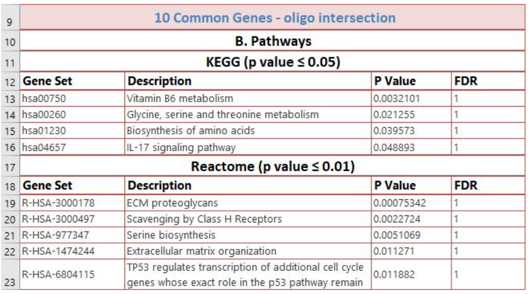


1 10 Common Genes - oligo intersection

2 A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.01)

	Gene Set	Description	P Value	FDR
4	GO:0090130	tissue migration	0.0071725	1
5	GO:1901342	regulation of vasculature development	0.0087173	1
6	GO:0002237	response to molecule of bacterial origin	0.0096543	1

Figure 71: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 10 common gene list as extracted from the intersection of the oligo analysis subset, and the SVM-RFE (tissue and blood) optimal subsets



9	10 Common Genes - oligo intersection			
10	B. Pathways			
11	KEGG (p value ≤ 0.05)			
12	Gene Set	Description	P Value	FDR
13	hsa00750	Vitamin B6 metabolism	0.0032101	1
14	hsa00260	Glycine, serine and threonine metabolism	0.021255	1
15	hsa01230	Biosynthesis of amino acids	0.039573	1
16	hsa04657	IL-17 signaling pathway	0.048893	1
17	Reactome (p value ≤ 0.01)			
18	Gene Set	Description	P Value	FDR
19	R-HSA-3000178	ECM proteoglycans	0.00075342	1
20	R-HSA-3000497	Scavenging by Class H Receptors	0.0022724	1
21	R-HSA-977347	Serine biosynthesis	0.0051069	1
22	R-HSA-1474244	Extracellular matrix organization	0.011271	1
23	R-HSA-6804115	TP53 regulates transcription of additional cell cycle genes whose exact role in the p53 pathway remain	0.011882	1

Figure 72: b) Pathway annotation (KEGG, Reactome) of 10 common gene list as extracted from the intersection of the oligo analysis subset, and the SVM-RFE (tissue and blood) optimal subsets

A.4 SAMPLE R SCRIPTS

Some sample R scripts are included in this section, describing the main steps we followed in our analysis.

Listing A.1: Feature extraction on pdac dataset

```
cl = ifelse(grepl("NORMAL", colnames(input), fixed=TRUE
    ), 'Normal', 'PDAC')
f = factor(cl, levels=c("PDAC","Normal"))
design = model.matrix(~0+f)
colnames(design) = c("PDAC","Normal")
data.fit = lmFit(input,design)
contrast.matrix = makeContrasts(PDAC—Normal,levels=
    design)
data.fit.con = contrasts.fit(data.fit, contrast.matrix)
data.fit.eb = eBayes(data.fit.con)
tab = topTable(data.fit.eb, adjust.method="BH", sort.by =
    "logFC", p.value=0.01, number=Inf, lfc=2)
```

Listing A.2: SVM Training

```
trctrl <— trainControl(method = 'LOOCV',classProbs =
    TRUE, verboseIter = TRUE,summaryFunction =
    twoClassSummary)
#SVM RADIAL
```

```
 svmRadialgrid <- expand.grid(sigma = c(.01, .015, 0.2) ,C
      = c (0.75, 0.9, 1, 1.1, 1.25) )
 svmRadial <- train( Patient ~ . ,data = training ,method
      = 'svmRadial',metric = "ROC",trControl = trctrl ,
      tuneGrid=svmRadialgrid,verbose = FALSE)
#SVM RADIAL PREDICTION
 results <- predict(svmRadial, newdata=training)
svm_training_prediction <- confusionMatrix( results ,
     training $Patient )
```

Listing A.3: KNN Training

```
#K NEAREST NEIGHBORS
knn_fit <- train( Patient ~ . ,data = training ,method = '
     knn',metric = "ROC",trControl = trctrl ,tuneLength =
     20)
#K NEAREST NEIGHBORS PREDICTION
 results <- predict(knn_fit , newdata=training)
knn_training_prediction <- confusionMatrix( results ,
     training $Patient )
```

Listing A.4: Feature Selection process (SVM-RFE)

```
svmRadialgrid <- expand.grid(sigma = c(.01, .015, 0.2) ,C
     = c (0.75, 0.9, 1, 1.1, 1.25) )
 trctrl <- trainControl(method = 'repeatedcv' ,number=10,
     repeats=5,classProbs = TRUE,verboseIter = TRUE,
     summaryFunction = twoClassSummary)
control <- rfeControl( functions =caretFuncs ,number=5)
 prediction <- rfe( training [,1:74], training [,75],
     trControl = trctrl , sizes =c
        (5,10,20,25,30,35,40,45,50,55,60,65,70) , rfeControl =
     control,method = "svmRadial" ,tuneGrid = svmRadialgrid ,
     verbose = TRUE)
 results <- predict(prediction , newdata=training)
svm_training_prediction <- confusionMatrix( results $pred,
     training $Patient )
```

Listing A.5: Sample Heatmap construction

```
input<−read.table("example.txt" , header = TRUE, sep = " "
    , dec = " . ")
hr <− hclust(as. dist (1−cor(t(input) , method="pearson")),
    method="complete")
mycl <− cutree(hr , h=max(hr$height)/1.8)
mycolhc <− rainbow(length(unique(mycl)), start=0.1, end
    =0.9)
mycolhc<− mycolhc[as.vector(mycl)]
mycol <− colorpanel(512, "green" , "red" )
hm<−heatmap.2(as.matrix(input), Rowv=as.dendrogram(hr),
    Colv="FALSE", col=mycol,
scale="row", density. info ="none", trace="none",
    RowSideColors=mycolhc,
margins =c (12,9) , main="Sample Heatmap with Clustering")
```

BIBLIOGRAPHY

[1] Peng Liang and B. Arthur Pardee. Analysing differential gene expression in cancer. *Nature Reviews Cancer*, 3(11):869–876, 2003.

[2] S. Hamada and T. Shimosegawa. Biomarkers of pancreatic cancer. *Pancreatology*, 2:14–19, 2011.

[3] Li Donghui, Xie Keping, Wolff Robert, and L. Abbruzzese James. Pancreatic cancer. *The Lancet*, 363(9414):1049–1057, 2004.

[4] J.P. Morris, S.C. Wang, and M. Hebrok. Kras, hedgehog, wnt and the twisted developmental biology of pancreatic ductal adenocarcinoma. *Nature Reviews Cancer*, 10(10):683–695, 2010.

[5] Xiao Yang, Shaoming Zhu, Li Li, Li Zhang, Shu Xian, Yanqing Wang, and Yanxiang Cheng. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. *OncoTargets and Therapy*, Volume 11:1457–1474, 2018.

[6] Rabia Aziz, , C.k. Verma, and Namita Srivastava. Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, 4(1):179–197, 2017.

[7] U.S. National Library of Medicine. Home - geo - ncbi. https://www.ncbi.nlm.nih.gov/geo/.

[8] Gene expression atlas for human embryogenesis. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15744.

[9] Integrative survival-based molecular profiling of human pancreatic cancer [mrna]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32676.

[10] The gene expression of normal pancreatic and pdac tissues. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71989.

[11] S100p is a metastasis-associated gene that facilitates transendothelial migration of pancreatic cancer cells. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19281.

[12] Whole-tissue gene expression study of pancreatic ductal adenocarcinoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471.

[13] Expression data from mayo clinic pancreatic tumor and normal samples. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16515.

[14] Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28735.

[15] Microarray gene-expression profiles of 69 pancreatic tumors and 61 adjacent non-tumor tissue from patients with pancreatic ductal adenocarcinoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62452.

[16] Expression profiling of pbmc from patients with hepatocellular carcinoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49515.

[17] Gene expression data from cd14++ cd16- classical monocytes from healthy volunteers and patients with pancreatic ductal adenocarcinoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60601.

[18] Blood biomarkers of pancreatic cancer associated diabetes identified by peripheral blood-based gene expression profiles. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15932.

[19] Expression data from peripheral blood in pancreatic ductal adenocarcinoma (pdac) patients. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49641.

[20] Rabia Musheer, C.K. Verma, and Namita Srivastava. t-independent component analysis for svm classification of dna- microarray data. *International Journal of Bioinformatics Research*, 6:305–312, 03 2015.

[21] Hassan Tariq, Elf Eldridge, and Ian Welch. An efficient approach for feature construction of high-dimensional microarray data by random projections. *Plos One*, 13(4), 2018.

[22] Affymetrix home page. http://www.affymetrix.com/site/mainPage.affx.

[23] R. Irizarry. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[24] Z. Wu and R. Irizarry. Description of gcrma package. *Bioconductor*, 2011.

[25] Yijun Sun, Sinisa Todorovic, and Steve Goodison. Local-learning-based feature selection for high-dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1610–1626, 2010.

[26] I. Guyon, J. Weston, and S. Barnhill. Gene selection for cancer classification using support vector machines. *Kluwer Academic Publishers*, 46:389–422, 2002.

[27] Mehdi Pirooznia, Jack Y Yang, Mary Qu Yang, and Youping Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(Suppl 1), 2008.

[28] K. Jung, T. Friede, and T. Beißbarth. Reporting fdr analogous confidence intervals for the log fold change of differentially expressed genes. *BMC Bioinformatics*, 12(1), 2011.

[29] Shilin Zhao, Yan Guo, Quanhu Sheng, and Yu Shyr. Advanced heat map and clustering analysis using heatmap3. *BioMed Research International*, 2014:1–6, 2014.

[30] S. Van Dongen and A. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations. 1:1–5, 2012.

[31] M. Zekić-Sušac, S. Pfeifer, and N Šarlija. A comparison of machine learning methods in a high-dimensional classification problem. *Business Systems Research Journal*, 5(3):82–96, 2014.

[32] C. Bishop. Neural networks for pattern recognition. *University Press*, 1995.

[33] Muni S. Srivastava and Tatsuya Kubokawa. Comparison of discrimination methods for high dimensional data. *Journal Of The Japan Statistical Society*, 37(1):123–134, 2007.

[34] S. Hussein, P. Kandel, C. Bolan, M. Wallace, and U. Bagci. Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches. *IEEE Transactions on Medical Imaging*, 38(8):1777–1787, 2019.

[35] E. Schadt, C. Li, B. Ellis, and W. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 84(S37):120–125, 2001.

[36] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[37] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, 2005.

[38] R-project. https://www.r-project.org/about.html.

[39] Bioconductor gcrma package. https://www.bioconductor.org/packages/release/bioc/html/gcrma.html.

[40] Bioconductor oligo package. https://www.bioconductor.org/packages/release/bioc/html/oligo.html.

[41] Bioconductor limma package. https://bioconductor.org/packages/release/bioc/html/limma.html.

[42] Gplots package. https://cran.r-project.org/web/packages/gplots/index.html.

[43] Caret package. https://topepo.github.io/caret/index.html.

[44] J. Liang, E. Kimchi, K. Staveley-O'Carroll, and D. Tan. Diagnostic and prognostic biomarkers in pancreatic carcinoma. *International journal of clinical and experimental pathology*, 2:1–10, 2009.

[45] W. Zhou, L. Sokoll, D. Bruzek, L. Zhang, V. Velculescu, S. Goldin, R. Hruban, S. Kern, S. Hamilton, D. Chan, B. Vogelstein, and K. Kinzler. Identifying markers for pancreatic cancer by gene expression analysis. *Cancer Epidemiology and Prevention Biomarkers*, 7:109–112, 1998.

[46] J. Wang, D. Duncan, Z. Shi, and B. Zhang. Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic Acids Research*, pages 77–83, 2013.

[47] Y. Liao, J. Wang, Z. Shi, and B. Zhang. Webgestalt 2019: gene set analysis toolkit with revamped uis and apis. *Nucleic Acids Research*, 47:199–205, 2019.

[48] U. Gormus, S. Gulecyilmaz, E.M. Altinkilic, and T. Isbir. The relationship of embryogenesis, carcinogenesis and angiogenesis. *Tumor Angiogenesis*, 1:2–19, 2016.

[49] N.M. Aiello and B.Z. Stanger. Echoes of the embryo: using the developmental biology toolkit to study cancer. *Disease Models & Mechanisms*, 9:105–114, 2016.

[50] M. Reichert, K. Blume, A. Kleger, D. Hartmann, and G. Von Figura. Developmental pathways direct pancreatic cancer initiation from its cellular origin. *Stem Cells International*, pages 1–8, 2016.

[51] S. Baker, I. Ali, I. Silins, S. Pyysalo, Y. Guo, J. Högberg, U. Stenius, and A. Korhonen. Cancer hallmarks analytics tool (chat): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, 33:3973–3981, 2017.

[52] F.C. Kelleher, D. Fennely, and M. Rafferty. Common critical pathways in embryogenesis and cancer. *Acta Oncologica*, 45:375–388, 2006.

[53] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144:646–674, 2011.

[54] M. Weniger, K.C. Honselmann, and A.S. Liss. The extracellular matrix and pancreatic cancer: A complex relationship. *Cancers (Basel)*, 10:316, 2018.