TECHNICAL UNIVERSITY OF CRETE

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

DIGITAL SIGNAL & IMAGE PROCESSING LAB

Combined analysis of phenotype and genotype in lung cancer using Radiogenomics framework

Συνδυαστική ανάλυση φαινότυπου και γονότυπου στον καρκίνο του πνεύμονα στο πλαίσιο της Ραδιο-γονιδιωματικής

DIPLOMA THESIS

Author

Dovrou Aikaterini

Advisor

Professor Zervakis Michalis

Committee members

Professor Zervakis Michalis

Professor Karystinos George

Associate Professor Marias Konstantinos (Hellenic Mediterranean University)

CHANIA, AUGUST 2020

Acknowledgements

With the completion of the present work, I would like to thank some people that contributed significantly, academically and emotionally. Firstly, I would like to thank my advisor, Professor Michalis Zervakis, for trusting me with this project and guiding me though the last year of my studies. His advice helped me evolve and better understand specific aspects of my field.

I would also like to thank the members of my committee, Professor George Karystinos and Konstantinos Marias, for their valuable advice during my thesis project.

I owe a huge thank you to Dr. Katerina Bei and Dr. Stelio Sfakianakis for the tremendous help and advice they offered me during the project. Their availability and valuable knowledge were crucial for the completion of the thesis.

I would like to thank Mr Eleftherio Trivizakis, for kindly providing us the radiomic features that were used in our analysis. I would also like to thank Professor Daphne Manoussaki for her valuable advice and collaboration though my studies.

A huge thank you to all my friends for being there for me and supporting me through my studies.

Nothing would be feasible without the tremendous support and help of my family. I would like to thank my parents, Sofia and Nikos, for being always available and supporting me in every possible way in order to achieve my goals, academically, athletically and personally! A huge thank you to my sister, Eleni, for supporting me and giving me advice to overcome any obstacle! Thank you for always being there for me and encouraging me to reach my goals, no matter how difficult they might seem at the time!

Abstract

During the last years, there is an increased interest in the development of models that intend to link cancer imaging features to the tumor genetic profile (Radiogenomics), in order to contribute in the diagnosis, evaluation, treatment planning and prognosis of lung cancer. Imaging features are extracted from the medical standard-of-care images and reflect the tumor phenotype. The tumor phenotype is formed by the rearrangement and the alterations of the genetic information. The gene mutations lead to cell proliferation and thus to cancer spread, which defines the cancer stage. There is an emerging need of valid diagnosis tools for lung cancer staging in order to define the proper treatment planning.

This study aims at investigating correlations among the most significant imaging features and genes in lung cancer and their potential to detect the stage of the patients with Non-Small Cell Lung Cancer (NSCLC). The proposed analysis includes the identification of the differentially expressed genes between cancer and healthy population by the application of the Significance Analysis of Microarrays (SAM) algorithm and the 2-fold change technique. Subsequently, correlation of these genes with the Computed Tomography (CT) imaging-derived features was conducted through the Spearman rank correlation test, SAM for quantitative problems and False Discovery Rate (FDR) methods, revealing 78 significant genes correlated to imaging features. These genes were validated for their diagnostic character through classification and clustering techniques followed by the formation of clusters of co-expressed imaging features (metafeatures). From these two procedures, 77 homogeneous metafeatures and 73 significant genes were identified. These genes were analyzed with least absolute shrinkage and selection operation (LASSO) regression for their ability to predict the metafeatures accurately. Through the analysis, 51 metafeatures that are correlated and can be predicted with the genes, were identified. The last step was comprised of the examination of the predictive ability of the remaining significant genes and metafeatures in lung cancer staging through various classification tests using linear Support Vector Machines (SVM) algorithms. This study concluded that staging cancer could be predicted from a) genes, with an accuracy of 75.00% - 94.11%, b) metafeatures, with an accuracy of 70.83% - 95.00% and c) the combination of metafeatures and genes, with an accuracy of 85.24% - 100.00%. Additionally, artificial imaging features were produced from the linear combination of the genes that could replace the actual metafeatures and predict cancer staging with an accuracy of 76.47% - 83.60%. Finally, signaling and metabolism pathways as well as miRNA targets were revealed during the enrichment analysis of the derived gene signatures.

Περίληψη

Τα τελευταία χρόνια έχει παρατηρηθεί αυξημένο επιστημονικό ενδιαφέρον, για την ανάπτυξη μοντέλων, τα οποία στοχεύουν στην συσχέτιση απεικονιστικών χαρακτηριστικών του καρκίνου με το γενετικό του προφίλ (Ραδιο-γονιδιωματική), ώστε να συμβάλουν στην διάγνωση, αξιολόγηση, θεραπεία και πρόγνωση του καρκίνου του πνεύμονα. Τα απεικονιστικά χαρακτηριστικά εξάγονται από ιατρικές standard-of-care εικόνες και αντιπροσωπεύουν τον καρκινικό φαινότυπο. Ο καρκινικός φαινότυπος δημιουργείται από την αναδιάταξη και τις αλλοιώσεις τις γενετικής πληροφορίας. Η μετάλλαξη των γονιδίων οδηγεί στον κυτταρικό πολλαπλασιασμό και κατά συνέπεια, στην εξάπλωση του καρκίνου, η οποία χαρακτηρίζει το καρκινικό στάδιο. Έγκυρα διαγνωστικά εργαλεία για την αναγνώριση του καρκινικού σταδίου είναι αναγκαία, ώστε να επιλεγεί η κατάλληλη θεραπεία.

Η παρούσα έρευνα έχει ως στόχο την εξερεύνηση συσχετίσεων μεταξύ των πιο σημαντικών απεικονιστικών χαρακτηριστικών και γονιδίων του καρκίνου του πνεύμονα και της δυνατότητάς τους να ανιχνεύσουν το καρκινικό στάδιο ασθενών με μη-μικροκυτταρικό καρκίνο του πνεύμονα (ΜΜΚΠ). Η παρούσα ανάλυση περιλαμβάνει την αναγνώριση των διαφορετικά εκφραζόμενων γονιδίων μεταξύ πληθυσμών που έχουν προσβληθεί από καρκίνο και υγιών πληθυσμών, μέσω της εφαρμογής του αλγορίθμου Significance Analysis of Microarrays (SAM) και της τεχνικής 2-fold change. Εν συνεχεία, υλοποιήθηκαν συσχετίσεις των γονιδίων με παραγόμενα απεικονιστικά χαρακτηριστικά αξονικής τομογραφίας, μέσω των μεθόδων Spearman rank correlation test, SAM για ποσοτικά προβλήματα και False Discovery Rate (FDR), αποκαλύπτοντας 78 σημαντικά γονίδια συσχετιζόμενα με απεικονιστικά χαρακτηριστικά. Τα γονίδια αυτά, αξιολογήθηκαν ως προς την εγκυρότητά τους για τον διαγνωστικό τους χαρακτήρα μέσω τεχνικών ταξινόμησης και clustering. Ακολούθησε ο σχηματισμός clusters από συνεκφραζόμενα απεικονιστικά χαρακτηριστικά (metafeatures). Από αυτές τις δυο διαδικασίες , 77 ομογενή metafeatures και 73 σημαντικά γονίδια αναγνωρίστηκαν. Τα γονίδια αναλύθηκαν μέσω του αλγορίθμου Least Absolute Shrinkage and Selection Operation (LASSO) regression, για να διερευνηθεί η δυνατότητά τους να προβλέψουν με ακρίβεια τα metafeatures. Μέσω της ανάλυσης, 51 metafeatures, τα οποία είναι συσχετιζόμενα και μπορούν να προβλεφθούν μέσω των γονιδίων, αναγνωρίστηκαν. Το τελευταίο στάδιο περιλάμβανε την εξέταση της προβλεπτικής ικανότητας των εναπομεινάντων σημαντικών γονιδίων και metafeatures, του καρκίνου του πνεύμονα, μέσω ποικίλων τεστ ταξινόμησης χρησιμοποιώντας Linear Support Vector Machines (SVM) αλγορίθμους. Η παρούσα έρευνα είχε ως βασικό συμπέρασμα ότι, το καρκινικό στάδιο μπορεί να προβλεφθεί μέσω a) γονιδίων, με ακρίβεια 75.00%-95.11%, b) metafeatures, με ακρίβεια 70.83%-95.00%, και c) συνδυασμού metafeatures και γονιδίων, με ακρίβεια 85.24%-100.00%. Επιπλέον, τεχνητά απεικονιστικά χαρακτηριστικά παράχθηκαν μέσω γραμμικού συνδυασμού γονιδίων, τα οποία δείχνουν ότι μπορούν να αντικαταστήσουν τα πραγματικά metafeatures και να προβλέψουν το καρκινικό στάδιο με ακρίβεια 76.47%-83.60%. Τέλος, ανακαλύφθηκαν σηματοδοτικά και μεταβολικά μονοπάτια καθώς και miRNA targets μέσω της ανάλυσης εμπλουτισμού των παραγόμενων γονιδιακών υπογραφών.

Contents

List of Figures

List of Tables

1. Introduction

Lung cancer is a common and aggressive type of cancer in both men and women. The majority of the affected population is of age 65 or over, while a small proportion of people diagnosed with lung cancer are younger than 45. [1] There are two types of lung cancer: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC). The NSCLC is the most common form, accounting for more than 85% of lung cancer cases and constitutes the leading cause of cancer-related deaths. [2] The main cause of lung cancer is smoking; both active and passive smokers, i.e. people who are non-smokers but are being exposed at secondhand smoke, can ail. Family history for lung cancer, ionizing radiation and exposure to other carcinogens, such as arsenic, chromium and nickel, are some other risk factors for lung cancer, which can act independently or in combination with smoking. [3] Cigarette smoke consists of many cancer-causing substances (carcinogens), such as nickel, carbon monoxide etc, whose inhalation can damage the cells in the lung tissue. The repeated exposure of people in these substances leads to abnormal function of cells and oxidative stress, which increases the risk of developing lung cancer. However, this risk can be decreased after several years of quitting smoking. Therefore, people who smoke for a short period of time, have lower risk of developing lung cancer. [4] The symptoms of lung cancer are not noticeable in early stages. The basic signs occur when the cancer starts to spread (metastasizes) through the lung or to other parts of the human body. When metastasis has progressed, the disease is considered advanced. The treatment and the survival of people with lung cancer depends on the stage of the cancer when it is diagnosed. Early diagnosis, and therefore detection, of the cancer in its earlier stage, when the tumor is confined in a small area, increases the probabilities of curing and survival.

Screening at-risk populations, which are suspected for lung cancer, is suggested by the doctors to detect the disease at an early stage, when the treatment has more probabilities to succeed. The most popular screening tests are Computed Tomography (CT) scan, Magnetic Resonance Imaging (MRI) scan, Positron Emission Tomography (PET) scan and PET/CT scans. CT scans, and specifically low dose CT scans, are widely used as recommended screening test for lung cancer. [5] These medical images illustrate the development and the progression of the cancer, providing valuable information for the clinical diagnosis and the treatment planning. Within the last few years, an evolving field, Radiomics, targets to the extraction of quantitative features from the medical images in order to support decisions for the cancer diagnosis and treatment. [6] This tool evaluates the tumor heterogeneity using medical images, reflecting the tumor phenotype. An emerging branch of Radiomics is Radiogenomics, which investigates the linkage between these imaging-derived features and the tumor genomic profiles.[7] A high-throughput amount of gene expression profiles can be derived from the widely used DNA microarray technology. The Radiogenomics analysis targets to provide noninvasive and comprehensive information about the tumor and its peripheral morphology, contributing significantly in the field of medicine. [8]

1.1. Goal of the study

The goal of this study is the combined analysis of gene expressions data and imaging features in order to investigate their diagnostic potential in cancer staging. Specifically, this study aimed at:

- Investigating genes that have diagnostic character in lung cancer and simultaneously predictive ability of radiomic features.
- Generating reliable mappings for cancer associations between genes and imaging features.
- Developing of potential non-invasive tumor imaging biomarkers.
- Identificating genes and radiomic features that are important in diagnosis of lung cancer staging.

1.2. Thesis outline

This thesis is divided in 7 chapters. Chapter 1 is a brief introduction of the basic goals and concepts. Chapter 2 includes additional information about studies that have been conducted in the field of Radiomics and Radiogenomics. Chapter 3 presents the theoretical background that is necessary for a deeper understanding of the processes in which imaging features and gene expression microarray data are extracted. Chapter 4 includes the technical background about the statistical tests and the machine learning algorithms that were used in our analysis. Chapter 5 describes the proposed methodology and the final results of this study. Chapter 6 summarizes this work and includes a discussion about the proposed analysis. Chapter 7 concludes the findings alongside with proposed future work.

2. Related work

The extraction of imaging features from the tumor region and their connection with the characterization of tumor aggression constitute significant tools for the non-invasive diagnosis, prognosis and evaluation of the disease. Andersen et al. [9] employed the technique of Computed Tomography Texture Analysis (CTTA) in order to differentiate between benign and malignant lymph nodes in the mediastinum. The CTTA technique for each lymph node was performed in two stages, including: a) filtering of images and b) quantification of texture. The first stage implemented by the application of a Laplacian of Gaussian bandpass and spatial filter to highlight fine, medium and coarse textures from the region-of-interest (ROI) of the image. This filter contribute to the extraction and the enhancement of imaging features of different texture degree. The different texture degrees of the image correspond to different values in diameter of the spatial filter. The second stage, which is the quantification of image texture, was performed by histogram analysis, in which the mean gray-level intensity for all filter sizes was calculated from the whole tumor of each lymph node. This mean image intensity was calculated for both group of benign and malignant tumors of filtered and unfiltered images. Furthermore, the mean short axis diameter and the mean long axis diameter was calculated for these two groups. The independent t-test between the malignant and benign groups was performed for each of the three previous features in order to investigate statistically significant difference between these two independent groups. The results showed that only the mean image intensity of the unfiltered images presents statistically significant difference, P = 0.001, between the group of malignant and benign lymph nodes. Specifically, the mean image intensity of the lymph nodes in the malignant group was substantially higher than in the benign group. This feature was subsequently used in a binary logistic regression model to assess the method. The applied CTTA method showed high enough performance by classifying 82.6% of the cases correctly, proving that the texture analysis of CT scans has the ability to help and distinguish differences between malignant and benign lymph nodes from the mediastinum for patients that are suspected for lung cancer. However, with the rapid development of the field of machine learning many researchers focused on the use of deep learning networks to extract features from medical images and classify, predict and evaluate a disease.

Zeju Li et al. [10] constructed a convolutional neural network (CNN) to predict the mutation of gene isocitrate dehydrogenase 1 (IDH1) and compare this deep learning method with the traditional method of extraction radiomic features from medical images. CNN was used to segment tumors from MRI images of patients with low grade glioma and with validated mutation status of IDH1 gene, which could be mutation or wild type. Their approach included the extraction of imaging features from the information of the last convolutional layers of CNN, where more and deeper information about the intensity, the shape and the texture of the tumor exist. The use of feature maps of CNN as imaging features was the main difference of the proposed approach from the standard radiomic approaches, in which imaging features are

calculated directly from the segmented tumors. The feature maps was encoded by a Fisher vector to normalize these features from image slices of different sizes. For each encoded feature map, the first and the second order statistics were calculated; therefore a one-dimensional high-throughput feature vector was produced for each patient and constituted the CNN features that characterize each tumor. Student t-test and F-scores were used to define the most important CNN features, which have statistically significant difference between the two mutation statuses of the IDH1 gene and thus, they are related to the type of the gene. A Support Vector Machine (SVM) with linear kernel was performed for the classification of mutation status of IDH1 gene using the selected significant CNN features and for the assessment of the model. The proposal method, which employ the CNN features, was evaluated by its ability to predict correctly the mutation status of IDH1 gene, which is a significant molecular biomarker. The deep learning-based radiomics method showed better performance compared to the traditional method of extracting radiomic features from the initial medical images.

In addition, the work of Bibault et al. [11] confirms the evolving invasion of deep learning in radiomics. In this work, the calculation of radiomic features was combined with the creation of a deep neural network (DNN) to predict complete response (pCR) of patients with rectal adenocarcinoma after neo-adjuvant chemoradiation. The method implemented the extraction of some clinical, biological and pathological features from patients, of which only the value of T stage showed significant correlation with pCR after performing the Chi-squared test. Furthermore, a large number of radiomic features were extracted from the segmented tumor of the initial CT scans for each patient, which were obtained before the chemoradiation. The features were associated with the following categories: shape, intensity, gray-level co-occurrence matrix 2D and 3D, neighbor intensity difference and Gray Level Run Length matrix. The intra-class correlation coefficient (ICC) was calculated to estimate the robustness of the features and the Wilcoxon test was performed to select features significantly correlated with pCR of rectal cancer. Only 28 features, from the categories of intensity and of gray-level co-occurrence matrix 2D and 3D, were selected for further analysis by satisfying the criterion of ICC to be greater of 0.8 and by applying the Wilcoxon test. Subsequently, the T stage and these 28 robust features were used as input to the DNN and as predictors to another machine learning network, the Support Vector Machines (SVM). Furthermore, the T stage was used separately as predictor to a linear regression model. The performance of these three different models was assessed and the researchers concluded that the DNN has the higher performance (80%) in precise prediction of the pCR of patients with rectal adenocarcinoma after neo-adjuvant chemoradiation.

An advanced technique of deep learning networks is the 3D CNN that present improved performance in the study of medical data than the 2D CNN. Trivizakis et al. [12] created a 3D CNN to classify the liver tumor in primary or metastatic stage using MRI images. The data are used without pre-processing, such as segmentation or definition of ROI, for the network's training. The significant advantage of this model is

that it is constructed from a huge number of trainable parameters and thus data pre-processing is not required. Furthermore, the oncogenic feature maps, which are derived from the interactions of the neurons, lead to richer representation of inner and overall structure of the tumor's environment and of the tumor itself.

Within the last few years, an increased number of studies have been conducted in order to link the phenotype and genotype of the examined disease. Emphasis has been given to the investigation of the correlation between the derived imaging features from clinical images and the genomic data of the disease. These correlations could demonstrate the underlying biology of the imaging features and of genes in order to enhance the accuracy and validity of the clinical results. Zhou et al. [13] created a radiogenomics map to link the derived imaging features from CT scans with the gene expression profiles analyzed by RNA sequencing for patients with NSCLC. More precisely, 87 semantic image features were extracted from the medical images to describe the lung characteristics by a thoracic radiologist using the open-source e-PAD platform. However, only the 35 of the 87 imaging features were used for further statistical analysis due to the criterion of occurrence rate $\geq$ 10% in the initial study dataset. These 35 features captured lung characteristics, such as nodule location, margins, attenuation, ground-glass composition and presence of emphysema. The high-throughput gene expression profiles were grouped together to metagenes as clusters of co-expressed genes, which means that genes with relevant expression profiles were assigned to the same metagene. The grouping of genes yielded to the reduction of their dimensionality. The homogeneity score of each metagene was calculated for the study cohort and for five other validation cohorts to validate the co-expression of genes within each metagene in different datasets. 10 metagenes with the higher homogeneity score were used for further analysis, while their molecular functions were annotated by using public accessible molecular databases, which link the genes with known biological pathways. Furthermore, the metagenes were associated with the survival by using the PRECOG tool, which contains publicly available gene expression profiles and their corresponding survival data for lung adenocarcinoma and lung squamous cell carcinoma. The correlation of the metagenes with survival was assessed by the univariate Cox proportional hazards regression. The final step was the creation of the radiogenomics map by performing t-statictic and Spearman correlation metric to evaluate significant correlations between the top 10 metagenes and the 35 semantic image features. To enhance the validity of these correlations, the False Discovery Rate (FDR) was used to correct for multiple testing. The results showed that 32 significant correlations between CT imaging features and metagenes were produced. For instance, the metagene that is related to late cell cycle was correlated with nodule attenuation and nodule margins and the metagene that is related to the activity of the Epidermal Growth Factor (EGF) pathway was associated with nodule margins and ground-glass opacity. Therefore, this study illustrates a method in which specific imaging features could be linked with specific metagenes that describe molecular properties and activate or deactivate specific molecular pathways. Thus, a comprehensive detection of genetic changes and a noninvasive

identification of molecular properties through imaging features could be achieved for patients that are diagnosed with NSCLC.

Prior to the aforementioned study, similar investigations for the discovery of prognostic imaging biomarkers for the NSCLC, deploying medical images and simultaneously gene expression profiles from the patients, were implemented from Gevaert et al. [14] and Nair et al. [15]. The research of Gevaert et al. [14] investigated PET and CT images from patients with NSCLC, while the research of Nair et al. [15] was focused on the study of PET images from the same patients to extract imaging features that can be used as imaging biomarkers to the evaluation and the prognosis of the disease progression. Gevaert et al. used a) a study cohort with 25 patients with NSCLC, which had PET scans and genomics data, but no follow-up data, b) an external cohort with data from a previous analysis of 63 patients with lung adenocarcinoma, which provided genomics data that were linked with the survival of the patients, while Nair et al. [15] used additionally c) a validation cohort, which provided PET scans of 84 patients with NSCLC and known clinical outcome of the patients. The method that conducted in the framework of radiogenomics comprised of the extraction of the imaging features from the DICOM images of the study cohort by using a suitable program which called RT_image. Regarding to the PET images, the imaging features were related to the measurement of some standard uptake value (SUV) of ^{18}F-2-fluoro-2-deoxyglucose (FDG). The Principal Component Analysis (PCA) technique was performed on the uptake features from PET scans and the first three principal components were used as three new FDG uptake features. The genes were grouped together to form the metagenes by performing an iterative k-means clustering algorithm. The quality of the metagenes was tested by calculating the homogeneity score of each metagene in the study and the external cohort. The correlation between the imaging features and the genes profiles was assessed by using the Spearman rank correlation test, the Significance Analysis of Microarrays (SAM) and the FDR for correction from multiple comparisons. Furthermore, two predictive models were implemented by using generalized linear regression with Lasso Regularization in order to investigate the way to predict the metagenes in terms of image features and the image features in terms of metagenes, which are called predicted image features. The metagenes from the study cohort were mapped to the publicly available gene expressions data with survival of the external cohort. Thus, the imaging features, which correlate with the metagenes, were associated with survival by leveraging public gene expression data. Moreover, the Kaplan-Meier (KM) curves and the Cox-proportional hazards (CPH) testing were used to assess the prognostic ability of the predicted uptake features and of single genes highly associated with FDG features. These tests were performed to define whether a predicted image feature provide independent information with the presence other clinical data, such as the age, gender, smoking status, size and stage of tumor. Additionally, a multivariate survival model based on the predicted image features was constructed by using generalized linear regression with lasso regularization.

In the work of Nair et al. [15] the FDG uptake image features were calculated from the PET images of the validation cohort, similar to the study cohort, in order to define the association between the actual imaging features of the study cohort and the survival. Their prognostic significance was evaluated with the same techniques that were used for the predicted imaging features. The study concluded that only two FDG uptake imaging features, and especially the SUV_{max} and the multivariate-SUV model, remain significantly associated with survival in the validation cohort. The final step was the gene enrichment analysis to correlate these two important imaging features that were expressed in terms of genes, with known molecular pathways of these genes. This analysis was performed by leveraging publicly available molecular databases, which link genes with known molecular pathways.

Radiogenomics constitutes an important tool for precision diagnosis, prognosis and treatment planning in oncology. Many studies for different types of malignancies have been conducted in order to investigate correlations between imaging features with genes and thus with their molecular pathways. [2] Liao et al. [16] conducted a radiogenomic study in patients with glioblastoma multiforme (GBM), which is a brain tumor with high mortality. Radiomic features were extracted from the tumor region of MRI images by using a python software package called Pyradiomics. The patients were divided into two groups according to their survival rate; patients with survival rate shorter than 1 year were grouped together and similarly patients with survival rate longer than 1 year. Feature selection were performed on the extracted radiomic features in order to select the most representative features to construct the model and predict the survival rate of the patients. The Gradient Boosting Decision Tree (GBDT) with these selected features achieved accuracy of 81% for distinguishing patients with short and long survival, which was the highest accuracy among logistic regression model, SVM and K-Nearest Neighbor (KNN). The most relevant genes were simultaneously selected by investigating the genes with statistically significant difference between the two survival groups. The R package DESeq2, the fold change and the t-test were used for the genes selection. The Pearson correlation coefficient between the selected imaging features and the differentially expressed genes in the two groups was calculated in order to investigate associations between them. The results revealed that some radiomic features had the ability to predict the clinical outcome of patients with GBM and simultaneously were associated to significant differentially expressed genes. For example, the textural features showed great ability of predicting the clinical outcome and at the same time were significantly correlated with three relevant genes. Thus, the imaging and molecular data were correlated to provide a precise prognosis and detailed information of the disease.

3. Theoretical background

3.1. Radiomics

Radiomics is an advanced technique of extracting quantitative information from the tumor region of clinical images, in order to provide a comprehensive characterization of the image phenotypes of the tumor [17]. This innovative field has evolved significantly during the past few years and is widely used for tumor evaluation in clinical oncology. It is a novel tool for discovering new imaging biomarkers, by extracting a significant amount of features from medical images, and identifying novel imaging signatures that help to improve diagnosis, prognosis and treatment planning strategies in medical applications. Moreover, radiomics can be applied to any type of standard-of-care clinical images such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). It is an emerging advanced texture analysis technique, which targets the identification of the linkage between tumor imaging biomarkers and the underlying genetic heterogeneity of the tumor or the available data on treatment outcomes, such as survival.[18]
According to Lambin et al. [19] a typical radiomics study can be structured through the following five phases (Figure 1):

a) Data selection, in which the following procedures are determined: i) imaging protocol, ii) the sub-regions or the region of the tumor that will be examined and iii) the expected goal of the study.
b) Medical imaging, which includes the determination of the way of segmentation of the initial medical image (automated, semi-automated, manually) and the phantom studies to gauge the uncertainties (such as organ motion or different imaging protocols) and reduce the risk in cases where patients' images are generated from different scanners.
c) Feature extraction, in which the high-throughput quantitative image features are extracted, that describe the medical image and characterize the region-of-interest (ROI).
d) Exploratory analysis–Feature selection, which enables the investigation of the relationship between features, in order to reduce the dimensionality of the feature vector. This is achieved by clustering radiomic features that are highly correlated and distinguish the ROI. Feature selection can also be implemented using another dimensionality reduction technique called Principal Component Analysis (PCA). In this way, overfitting is more likely to be avoided, because features that are redundant and lack robustness are eliminated whereas only the features that have not similar information are used for further processing. Furthermore, clinical variables, e.g. age, stage of the tumor, smoking status etc. can be included during the exploratory analysis to examine if they have an important role in tumor aggression.

e) Modeling, in which the methods and the models are defined. These methods will use the radiomic features, in order to achieve the goal of the study for prediction or diagnosis or evaluation of a disease. The most usual method for this purpose is machine learning algorithms. However, it is essential to assess model performance in order to identify whether the model is predictive for the target patient population or just for a particular subset of samples. For assessing model performance, validation techniques are used, such as the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC): methods that quantify the sensitivity and the specificity of the model.

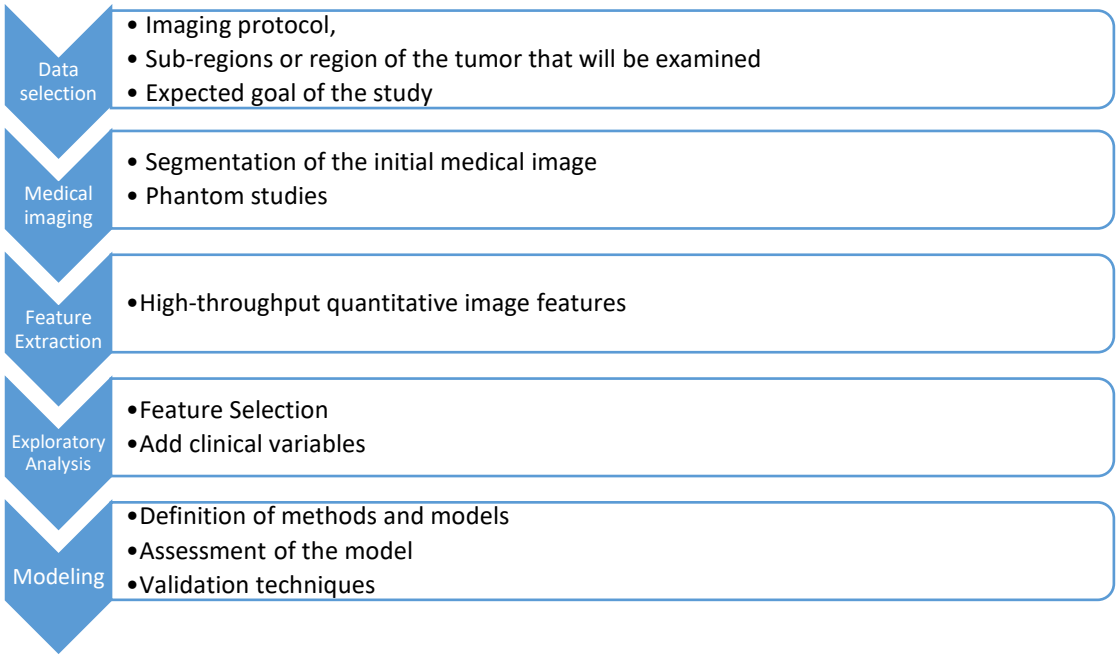


Figure 1. Five Stages of a Radiomics Study (Lambin, 2017)

Radiomics focuses on the processing of imaging data prior to the treatment. A specific field of radiomics is Delta-Radiomics [19] in which quantitative features are acquired from the medical images over the course of a treatment. The purpose of this technique is to determine whether radiomic features change during therapy by measuring the value of these features after the desired time of treatment. The result assist in understanding the way that the human organism responds to treatment and the way prognosis, diagnosis, prediction, monitoring and assessment of therapeutic response are improved.

3.1.1. Tumor imaging biomarkers – Radiomic features

Tumor imaging biomarkers are extracted from medical images. They are used to quantify the tumor burden describing the macroscopic and microscopic structures of

a tumor.[18] Macroscopic structures refer to the shape and the size of a tumor, which are extracted using specific criteria, e.g. WHO, RECIST, CHOI, mRECIST,[20] while microscopic structures refer to biological or pathological characteristics within a tumor, such as the hemodynamic parameters and the local image textural patterns (e.g. signal intensity, heterogeneity, histogram analysis, wavelet transformations).
The imaging evaluation of the tumor response to treatment is based on measuring and comparing the values of these imaging biomarkers before and after the treatment. The tumor response is then classified into four categories: a) complete response (CR), b) partial response (PR), c) stable disease (SD) and d) progressive disease (PD).
Radiomic features are an alternative definition of these tumor imaging biomarkers, which can be extracted from images to comprehensively characterize the tumor phenotypes and can be used in the radiomic analysis. They have the potential to uncover disease characteristics that fail to be estimated with the naked eye.

3.1.1.1. Texture Analysis

The texture analysis methods quantify the tumor heterogeneity and characterize the biological or pathological changes of micro-structures within the tumor in order to evaluate the tumor's response [18]. The texture analysis of images comprises of the calculation of several imaging features, such as histogram statistics features, run-length (RL features) texture features to encode the coarseness of the tumor [21], gray level co-occurrence matrix (GLCM) and shape features to describe the spatial shape of the tumor.

There are two principal aspects of texture analysis, that can be used independently or in combination, to quantify the tumor heterogeneity [22]:

- Image transformation for the extraction of imaging features that highlight the texture properties. This method uses filters which transfer the imaging features to larger scales in order to reduce the effect of photon noise and to enhance the evaluation and the quantification of the heterogeneity. There are two transformations that are widely used: a) Fourier transform, which describes the image in terms of frequencies that are defined by the shapes and the size of the image characteristics and b) Wavelet transform, which provide information for the spatial location of the imaging features in addition to their frequency characteristics. A specific category of wavelet transformations is the non-orthogonal wavelet filters which have the potential to highlight precisely the imaging features of a particular size. The Laplacian of band-pass Gaussian filter (LoG spatial filter) belongs to the category of the non-orthogonal wavelet filters. It is mainly used for the extraction and the enhancement of imaging features of fine, medium and coarse textures.
- Image quantification for the characterization of the texture, equivalent to definition of the category of the tumor region (normal vs abnormal, less or

more aggressive disease). The techniques for the calculation of the parameters that quantify the image texture are divided in the following categories:

- *Structural approaches*, which are used when there is no information about the discrete shape of the object's boundary and the probability of the specific object to be in a particular location.
- *Model-based approaches,* in which mathematical models, such as fractal and stochastic models, are used to interpret the image texture by comparing images that are generated by models. For example, fractal dimension can be used as an indicator of surface-texture and shows the similarity between shapes of different scales.
- *Statistical approaches,* which are based on representations of the texture by using the properties that are being derived from the distribution and the relationship among the gray-level intensity values of image. First order, second order and higher order statistics were calculated by using the histogram due to their property to differ significantly in the description of the gray-level distribution of the image. Specifically, the following features can be calculated from the histogram:
  - *First order statistics* which are based on the probability distributions of the gray-level pixel values, such as the mean value and the entropy.
  - *Second order statistics* which are based on the jointly probability distributions of pair of pixels, such as variance, standard deviation of the histogram, correlation, gray level run length (GLRL) or gray level co-occurrence matrices (GLCM). The GLRL matrix quantify the size of consecutive pixels with the same gray-level intensity in a fixed direction and thus provide the size of homogeneous runs along specific axis for each gray level [23]. The GLCM matrix contains the jointly probability occurrence of pairs of gray values along fixed axes within the image.
  - *Third order statistics,* such as skewness for the histogram's asymmetry.
  - *Fourth order statistics,* such as kurtosis for the measurement of the tail-heaviness of the distribution.

The texture analysis technique is used for the evaluation of the tumor response in treatment, the detection and characterization of the lesions due to the differences that are noticed between the texture features of the tumor and the surrounding tissue as well as between different diseases.

3.2. Radiogenomics

Radiogenomics is a field of science that aims to associate imaging features with their gene profiles in order to non-invasively predict tumor genomic alterations identifying specific imaging features.[7] These imaging-derived features (phenotypes) can be linked with genomic data, in order to understand their biological underpinnings, such as their molecular pathways. Furthermore, this linkage may improve the prediction accuracy of clinical outcomes.[17] What differentiates radiomics from radiogenomics is that radiomics refers to a general branch of study in which imaging features from patient scans are converted into quantitative data while radiogenomics is a specific application where imaging features are linked to genomic profiles. This field aims to bypass both the issue of invasiveness and the sampling bias which is used in biopsies by using non-invasive radiological images to analyze the full tumor burden.

From the biological aspect, genes carry the necessary information for the functioning of cells and the synthesis of functional structures such as proteins. [24] The genotype of the cell determines the amount of protein that is present in that cell. The synthesis of these functional structures is implemented through the information transfer of messages that are formed from genes. These processes are critical due to the fact that they assist to the formation of the characteristics features or phenotypes of the cells, such as normal and cancer cells.[24] Hence, the phenotype of the cells and as a result the organism, is "determined" by the genes in combination with the environmental factors. Thus, radiogenomics targets to investigate this underlying biological meaning of association of phenotype and genotype in an organism's cells. To conclude, the flow of the linkage from genes to organism phenotype is illustrated in the following Figure 2:

Figure 2. Flow of genes to organism phenotype

Tumor genetic profile knowledge, which can be gained non-invasively with radiogenomics, is highly important to clinicians and assists the precision of medicine. Genomic alterations are a hallmark of cancer and the accumulation of genetic mutations results in unchecked cell proliferation. Gene profiles can be used as a prognostic biomarker to predict survival or as a predictive biomarker to predict treatment response, helping to plan clinical decisions and especially treatment selection [7]. Precision therapy takes advantage of tumor-specific biology to inhibit the action of tumor-associated proteins or enzymes, mutated receptors or other oncogenic molecular vulnerabilities. The detailed knowledge of genetic structure of the tumor, for instance driver mutations, can provide much-needed guidance on the prediction of the disease progression and the selection of an efficient therapy.

3.3. DNA Microarray technology

DNA

Deoxyribonucleic acid (DNA) [25] is the hereditary material in humans and all mammalian-organisms. The same DNA exists in all the cells of the organism and controls their activity. Cells are fundamental structures of each living organism. Most DNA is located in the cell nucleus (nuclear DNA) and a small amount of DNA can also be found in the mitochondria (mitochondrial DNA or mtDNA). Mitochondria are structures within a cell that convert the energy from the food consumed into a useful form; thus they are considered the "powerhouse" of the cell.

DNA is a molecule composed of two polynucleotide chains that are held together primarily by hydrogen bounds and coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all organisms. This molecule consists of four fundamental molecular units called nucleotides which they are arranged in two long strands that form the double helix. Each nucleotide contains a phosphate group, a (deoxyribose) sugar and a nitrogen base. The four types of nucleotides are distinguished by their distinct nitrogen base: adenine (A), cytosine (C), guanine (G) and thymine (T). Each of the four nucleotide DNA bases is linked in pairs to form units called base pairs. Hence, base A links to T, reversely base T links to base A and similarly base C links to G and G to C. [24], [25], [26]

The specific pairing of DNA bases (A-T and C-G) is called base-sequence complementarity. A DNA sequence is a specific type of ordering base pairs in DNA strands. [24], [26]

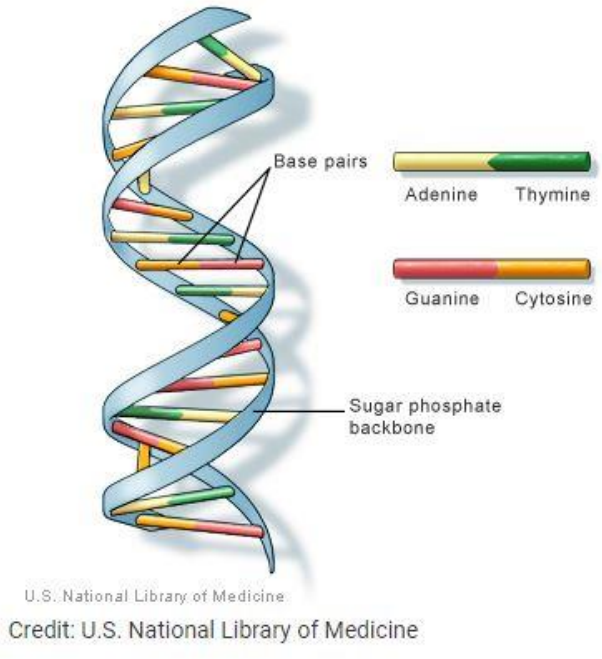


Figure 3. DNA double helix and the complementarity of the DNA bases (https://ghr.nlm.nih.gov/primer/basics/dna)

Genes

A gene is the basic physical and functional unit of heredity and they consist of DNA. Some of them act as instructions to form other molecules called proteins (coding DNA). However, most of the genes do not code for proteins (non-coding DNA), but they are fundamental for cells function, particularly the control of gene activity. In the human organism, genes vary in size from a few hundred DNA bases to more than 2 million bases. The "Human Genome Project" estimated that humans have approximately 20,500 genes. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different among people which results from small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features. Scientists keep track of genes by giving them unique names or symbols. Thus, genome of one living organism is considered the whole DNA sequence of the organism, including all of its genes. Therefore, a genome contains all the essential information which are needed to build and maintain an organism. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all cells that have a nucleus. [24], [25]

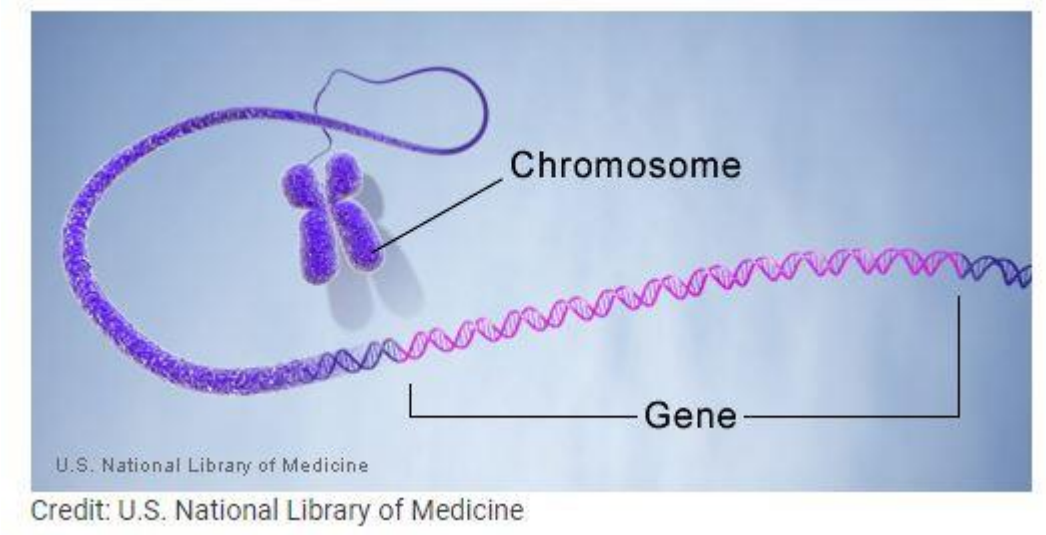


Figure 4. Gene representation (https://ghr.nlm.nih.gov/primer/basics/gene)

Gene expression profile

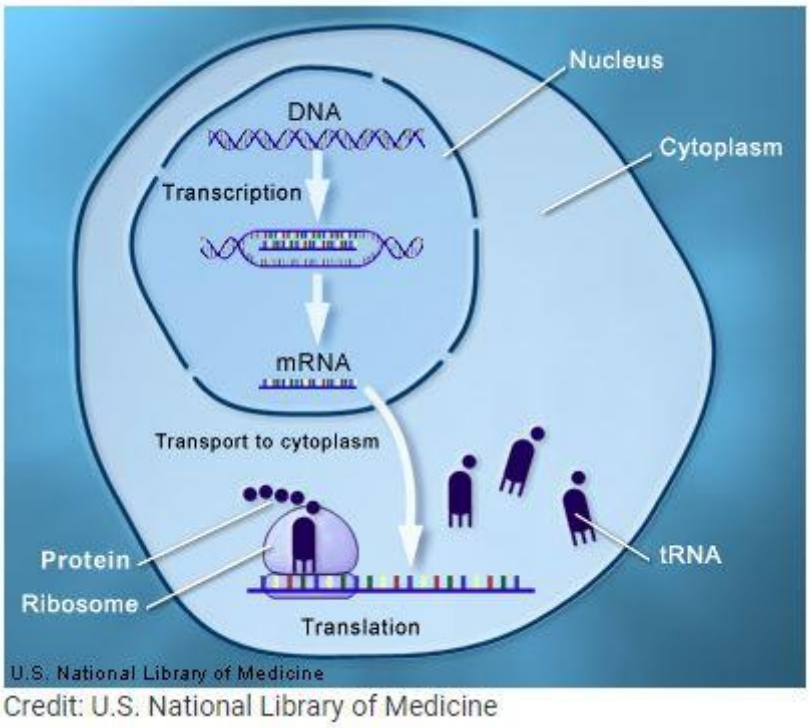
Genes can be responsible for the production of proteins' molecules. The flow of information from genes and consequently from DNA to proteins is achieved via two major processes: transcription and translation. [24], [26], [25]

1. During the process of transcription, the gene's DNA sequence is transcribed into mRNA (messenger RNA). The name of messenger RNA comes from its property to convey the information (or message) from the DNA out of nucleus into the cytoplasm. Both RNA and DNA are nucleic acids with the differences

that RNA is a single stranded rather than double helix found in DNA, the sugar in its nucleotide is ribose rather than deoxyribose and has the base Uracil (U) instead of the base thymine (T) that exists in DNA. That means that in RNA base Uracil is complementary to Adenine forming a hydrogen bond between these two bases. The other two bases, Cytosine and Guanine, remain complementary to each other. Hence, the synthesis of RNA chain is implemented by adding nucleotides with base A, C, G and U where a T,G,C and A base is found in DNA template strand with respect to the bases' complementarity.

2. The other process, the translation, takes place in the cytoplasm of the cell. During this procedure mRNA translates into amino acid sequence of proteins, which conduct different cell functions.

The activity of proteins provides the genetic information that is contained in the DNA. Thus, the transcription and translation, in which a gene's DNA sequence is initially transcribed into mRNA and then into a protein, are called gene expression. To be more specific, the level of a gene expression shows the approximate number of produced copies of RNA from that gene in one cell and is related to the amount of derived proteins. [26]



Credit: U.S. National Library of Medicine

Figure 5. Graphical representation of the processes and the flow information for protein's formation (https://ghr.nlm.nih.gov/primer/howgeneswork/makingprotein)

DNA Microarray technology

DNA Microarray technology is widely used to explore and measure the gene expression profiles with which scientists have the ability to identify gene functions and contribute to cancer diagnosis. The advantage of this technology is that it has the potential to simultaneously measure the relative expression level of thousands of genes within a cell or a tissue within a short period of time. This technology is based on the property of complementarity of the four nucleotide DNA or RNA bases. The two major types of microarray experiments depending on the DNA probes that are used in them, are complementary DNA (cDNA) microarray and oligonucleotide arrays (abbreviated oligo chips). [26]

A DNA microarray (Figure 6) measures the amount of mRNA expression levels of a gene, which is the gene expression at the transcription level [24]. For this reason, the cellular mRNA of the cell is extracted in order to be measured. Two crucial procedures occur during this measurement process: a) reverse transcription and b) hybridization. During the process of reverse transcription, the mRNA of a gene, which is experimentally isolated from a cell, is reverse-transcribed into a complementary DNA copy called cDNA, which is double-stranded. In specific cases, this double-stranded cDNA is feasible to be reverse-transcribed into a complementary RNA copy called cRNA. The second procedure, hybridization, is the process in which the molecules of nucleic acid recognize and link in pairs to molecules with a complementary sequence. In this case, the two single strands of DNA or RNA are being base paired. The two strands of DNA are separated by heating in a characteristic melting temperature, above 65°C. In following step the reduction of the temperature results in the re-binding of the two single stranded, which originate from a DNA or/and an RNA molecule, on the principle of base pairing (complementarity). It is important to note that when hybridization between a DNA and an RNA molecule occurs, a single stranded DNA which has been produced from a melted DNA molecule, binds strongly to its complementary RNA in a way that prevents the two single DNA strands of re-coupling with each other. [25]

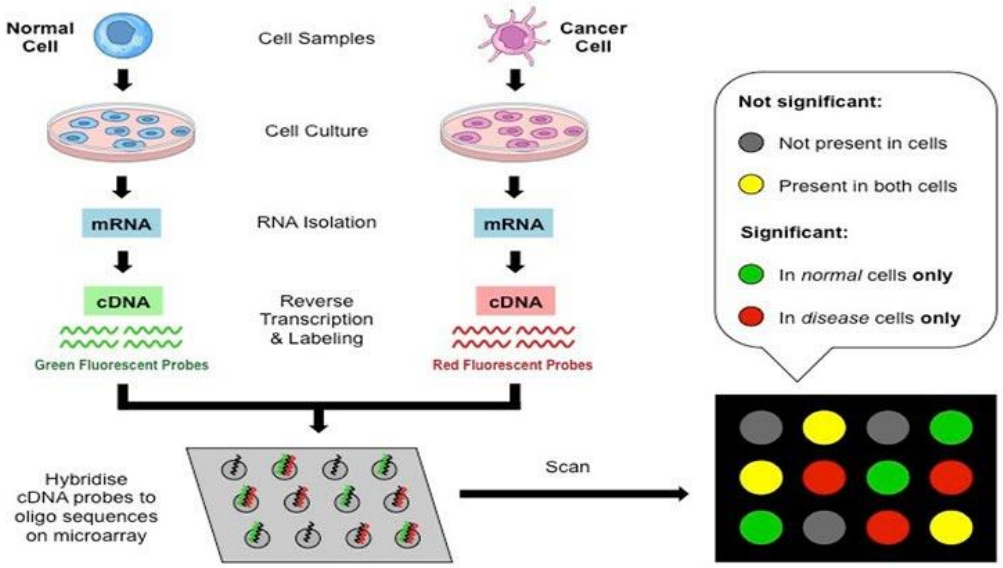


Figure 6. The cancer and the normal tissues are RNA isolated, reverse transcribed and labelled with fluorescent dyes. The probes of the DNA microarray technology is hybridized and the spots are dyed in the appropriate color. (Image from: https://microbenotes.com/dna-microarray/)

A microarray is a small chip made of chemically coated glass, nylon membrane or silicon.[27] In most cases, it is a microscopic glass slide on which molecules of complementary DNA (cDNA) are placed to certain locations called spots. There is a fundamental difference between cDNA microarrays and oligo chips. The cDNA microarrays consist of long sequences of cDNA, while oligo chips consist of short sequences of single-stranded cDNA called synthetic oligonucleotides (abbreviated to oligos).[26] The tethered cDNA sequences or oligos are called probes and they represent known genes or segments of known DNA sequences. The reverse-transcribed form of the extracted cellular mRNA is referred to as the target.[25] Therefore, the construction of the chip is the first process of the basic procedure that DNA microarrays follow. Following the first process, the mRNA of the cell is extracted and reverse transcribed into cDNA, for the case of cDNA microarrays, and into cRNA, for the case of oligo chips. These molecules constitute the target in each type of the corresponding microarray. The target is labeled with fluorescent dye and then hybridized to the probes on the surface of the chip. With the hybridization, each single strand of target cDNA or cRNA is bound with probes (double-stranded or single stranded cDNA respectively) by finding and linking complementary nucleotide base pairs with hydrogen bonds. Once hybridization has completed, the glass side (i.e. microarray) is washed to be cleaned from non-hybridized molecules and scanned with a laser scanner to obtain images. The signal intensity of the labeled and hybridized targets is determined from the scanning of these images. The more intense fluorescent dyes correspond to higher amount of cDNA or cRNA that is hybridized to each probe. Thus, DNA microarrays succeed to measure the gene expression profile levels by measuring the relative mRNA abundance of the gene. This abundance is estimated by the measurement of the intensity of the fluorescent dyes that is emitted from the hybridization of the probes with the targets.[25],[27]

Additional differences exist during the process of target labelling with fluorescent dyes and the structure of spots that will be used for the hybridization of genes, depending on the types of DNA microarrays used. In cDNA microarrays the mRNA molecules are extracted from two samples, the one is the control sample (i.e. cells of normal tissue) and the other is the test sample (i.e. cells of tested tissue). These two mRNA samples are reverse-transcribed into cDNA and are labelled with 2 different fluorescent dyes. The fluorescent dye, Cye-3 (green), labels the cDNA molecules that correspond to the cells of the control sample and the fluorescent dye, Cye-5 (red), labels the cDNA that correspond to the cells of the test sample. This target mixture of the two dyes is hybridized to the probes on the glass slides of the microarray. Then, the dye of each spot determines which of the two populations (i.e. control or test sample) has greater amount of cDNA molecules. To be more specific, a spot is dyed in red color, if the amount of cDNA is greater in the test sample, while it is dyed in green color, if the cDNA is in higher amount in the control sample. Hence, the relative mRNA abundance (i.e. gene expression) of the gene in the cell can be measured by fluorescence intensity of each of the two dyes in each spot. The log ratio between the two intensities of these two dyes represents the gene expression profile. Each spot corresponds to a gene and the color of the spot indicates whether the gene is expressed (colored) or not and the relative level of gene expression in the two samples.[26], [25], [24], [28].

For the case of oligo chips, the mRNA molecules are extracted from the test sample and then reverse-transcribed into cDNA, which is double-stranded and then converted into cRNA. After reverse transcription occurs, this target is fluorescently labeled with a single dye. The main difference of this technology is the usage of probe redundancy. Probe redundancy is used to identify a gene and measure the relative gene expression of a set of well-chosen small segments of cDNA unique to the DNA of the gene and not only a spot (like in the case of cDNA microarrays). Specifically, a gene in oligo chips is represented by a set of probe pairs, which a probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe. The MM probe is identical to the corresponding PM probe except from the central base of the nucleotide, which is replaced with its complementary base. Therefore, the PM probe is complementary to the target gene sequence and thus, detect uniquely the gene by the hybridization. The MM probe works as “control”. Hence, high intensity for the PM probe and low intensity for the MM probe are expected for a specific gene in the cell sample, under ideal circumstances. Thus, the use of the probe pairs and the fact that a set of probe pairs are used simultaneously and are scattered across the microarray to identify the same gene, contribute to decrease the chance of cross-hybridization and reduce the noise present in the signal.[24],[25] Cross-hybridization is a situation where fragments of the reverse – transcribed mRNA of target hybridize to similar complementary probes but not to the real complementary probe and thus, false detections can be caused.[25]

In conclusion, both cDNA microarrays and oligo chips measure the relative expression levels of each gene; however, the calculation of the ratio of signal intensity between the test sample and the control sample follows a different process in each case. Thus,

these two types of microarray technology are based on similar technical concepts behind the measurement's procedure and hence the produced gene values from both methods share the same biological semantics. [27]

4. Technical Background

4.1. Significance Analysis of Microarrays (SAM)

Significance Analysis of Microarrays (SAM) is a statistical method for genomic expression data mining. Tusher et al. [29] proposed this statistical algorithm, which targets to discover the most significant genes in a set of microarray experiments.

The gene expression measurements from microarray experiments are imported as input to SAM as well as the response variable from each experiment. The response variable determines the class in which each sample belongs for each gene. It is defined according to the response type of the examined problem. Some examples of response types are:

A) *Quantitative*, in which the response variable is real-valued, such as blood pressure, values of an imaging feature etc.

B) *Two-class unpaired,* in which the response variable is expressed by an integer (1 or 2) and refers to the class of each measurement of different experimental units. For example, the two classes could be cancer and normal samples from different patients.

C) *Multiclass,* in which similarly the response variable is expressed by an integer (1,2,…) and refers to the class of each measurement. The only difference with the two-class problem is that the number of classes is greater than 2.

D) *Paired,* in which there are groups of two sets of measurements with the same experimental unit in each group. The response variable indicates the group of each measurement (i.e. 1,2,…), while the sign of the response variable (i.e. positive or negative) shows the set of each group that the measurement belongs. For instance, a two class-paired problem is a group of samples from the same patients that are measured before and after the treatment.

E) *One-class,* in which the response variable is equal to 1 for all the measurements due to the fact that they belong to the same class. The goal of this problem is, usually, to test whether the mean of each gene expression differs from zero.

SAM computes a score statistic d_i for each gene i in order to identify genes with expression that vary with statistical significance across the response variable. This statistic d_i is called "relative difference" in gene expression and is similar to the t-statistic. It expresses the change in gene expression across the response variable

relative to the standard deviation of the gene. Specifically, the statistic d_i is calculated according to the following equation:

$$d_i = \frac{r_i}{s_i + s_0}$$

where index i = 1,2,….p indicates the number of gene.
For the following definitions the index j is supposed to be the indicator of the number of sample/measurement, y_j is the value of the response variable of each sample j and x_{ij} is the value of the gene i for the sample j.

The calculation of the numerator r_i and the s_i is based on the response type of the problem. The equations for these two terms of statistic d_i are shown for the quantitative and the two-class problem.

Two class problem

The numerator r_i is calculated as the difference in the mean values of the gene expressions between the two classes (i.e. normal measurements belong to the class 1 and cancer measurements belong to the class 2). The equation of r_i is:

$$r_i = \overline{x_1}(i) - \overline{x_2}(i)$$

where $\overline{x_1}(i)$ is the average expression value of class 1 and $\overline{x_2}(i)$ is the average expression value of class 2 for each gene i.

The s_i is the pooled standard deviation [30] of the gene i, that is the mean standard deviation of repeated measurements under the assumption that the standard deviation of each class C_1 and C_2 remains the same. The equation of s_i is:

$$s_i = \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \frac{\sum_{x \in C_1}(x_i - \overline{x_1}(i))^2 + \sum_{x \in C_2}(x_i - \overline{x_2}(i))^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 are the number of measurements in class C_1 and C_2 respectively.

Quantitative problem

The numerator r_i is calculated as the linear regression coefficient of each gene i on the response variable y. For each gene the r_i is calculated in order to find the solution to the problem $y = r_i x_i$. The equation of r_i is:

$$r_i = \frac{\sum_j y_j (x_{ij} - \overline{x_i})}{\sum_j (y_j - \overline{y_j})^2}$$

where $\overline{x_i}$ is the mean value of gene i and $\overline{y_j}$ is the mean value of the response variable of sample j.

The s_i is the standard error of r_i and its equation is:

$$s_i = \frac{\widehat{\sigma_i}}{[\sum_j (y_j - \overline{y_j})^2]^{1/2}}$$

where $\widehat{\sigma_i}$ is the square root of residual error and it is equal to:

$$\widehat{\sigma_i} = [\frac{\sum_j (x_{ij} - \hat{x}_{ij})^2}{n-2}]^{1/2}$$

where n is the number of measurements and:

$$\hat{x}_{ij} = \hat{\beta}_{i0} + r_i y_j$$

$$\hat{\beta}_{i0} = \bar{x}_j - r_i \bar{y}_j$$

The factor s_0 is called exchangeability factor and is constant across all the genes. It is expressed as a percentile of the standard deviation values of all genes. The role of the factor s_0 is to include a "penalty" term next to the standard deviation s_i in order to protect genes with expressions close to 0. These genes have small standard deviation and the factor s_0 protect them from having large scores of statistic d_i. [31] Hence, the value of s_0 restricts the effect of the fluctuations in genes' variance, ensuring that the distribution of d_i from all genes is independent of the gene expression levels. [29]

SAM initially computes the statistic d_i for each gene from the input values. The d_i value is the *observed score* and after it is calculated, SAM ranks all the observed scores d_i in ascending order. Subsequently, it performs repeated permutations on the data by randomly changing the class of each sample in order to assess the statistical significance of each gene related to the response. For each permutation, b, the statistic d_i^b for each gene is re-calculated and these statistic values are also ranked in ascending order $d_{(1)}^b \leq d_{(2)}^b \leq \cdots \leq d_{(p)}^b$. The following step is the calculation of the *expected score* in each place (1,2,….p) as the average value $E[d_{(i)}]$ of the statistic values across all the permutations. The equation of the expected score $E[d_{(i)}]$ is $\mathrm{E[d(i)]} = \frac{1}{B}\sum_b d_{(i)}^b$.

To identify significant genes in expression, SAM defines the delta (Δ) value, which is a threshold for significance. Specifically, SAM calculates the difference in value between the *observed score* d_i and the *expected score* $E[d_{(i)}]$ for each gene. Consequently, SAM identifies the first gene that satisfies the criterion $d_i - E[d_{(i)}] > \Delta$. The value d_i of this first gene "k" that satisfies this criterion constitutes the upper cut-point: $cut_{up} = d_k$. All genes that have d_i values > cut_{up} are considered as "positive significant".
Similarly, SAM calculates for each gene the difference $E[d_{(i)}] - d_i$. Then, it recognizes the first gene that satisfies the reverse criterion $E[d_{(i)}] - d_i > \Delta$. The value d_i of the first gene "m" that satisfies this criterion defines the lower cut-point $cut_{low} = d_m$. All genes that have d_i values < cut_{low} are considered as "negative significant".

Hence, genes with score higher than the threshold are deemed significant. The threshold for significance is determined by the tuning parameter Δ, which is defined by the user and is depended on the number of false positives that are acceptable. False positives or *falsely called* are the genes that are identified as significant, while in fact they are not. To estimate the proportion of these genes, which are identified as significant by chance, the FDR is calculated. Firstly, SAM calculates the total number

or the median number of *falsely called* genes. For each permutation the number of falsely significant genes is equal to the number of genes that exceed the horizontal cutoffs, meaning that genes have d_i value > cut_{up} or d_i value < cut_{low}. The total number of *falsely called* genes is calculated as the average number of *falsely called* genes from all permutations. Similarly, the median number of *falsely called* genes is calculated as the median number of falsely called genes among all permutations. The FDR is computed as the ratio between the number of the *falsely called* genes and the number of the genes called significant.
Therefore, as the Δ value decreases, the cutoffs decrease and thus, the number of genes called significant increases with a cost of an increasing FDR.

Additionally, SAM computes the q-value for each gene, which is the lower FDR that can be achieved over all rejection regions from all Δ, at which the gene is called significant. [30] The q-value measures the significance of each gene: as $|d_i|$ increases, the corresponding q-value decreases. Hence, the lower the q-value of a gene is, the higher the significance of this gene.

4.1.1. Fold Change

Fold change measures the change of a quantity between two measurements A and B. It is defined as the ratio between these two measurements, i.e. $FC = \frac{A}{B}$ is the fold change of A with respect to B.

Fold change is a common method in analysis of gene expression data, where genes with change in their expression level between different experimental measurements are deemed significant. Moreover, a particular cutoff in fold change is defined in order to specify the acceptable and necessary amount of change in values to characterize a gene as significant. For this reason, the term 'X-fold change' is used to describe an increase of multiple X in expression levels of a quantity in a population compared to its expression levels in another population. For example, 2-fold change increase between A and B means that A is twice as big as B or alternatively A is "2 times" larger than B (in other words A is 200% of B). Due to the widely use of fold change in gene expression analysis, SAM has the extra option of setting a non-zero fold-change parameter as a more stringent criterion in the exploratory procedure analysis of significant genes.

4.2. Correlation tests – Statistical Significance

Correlation coefficient is a simple statistical measure of the strength of the relationship between two variables and the direction of this association. [32] The values of the correlation coefficient vary between -1 and +1. The extreme values of ± 1 indicate a perfect association between the two variables, while the values that are closer to 0 indicate weaker relationship between them. Thus, higher absolute value of the correlation coefficient corresponds to stronger association. The direction of the relationship is defined by the sign (+ or -) of the correlation coefficient. A positive sign (+) shows positive correlation, which means that both variables move in the same direction. Positive correlation exists when one variable increases as the other one increases as well, or one variable decreases while the other one, similarly, decreases. Inversely, a negative sign (-) indicates negative correlation, meaning that the two variables move in the opposite direction. Negative correlation indicates that one variable increases when the other one decreases and vice versa.

In statistics, there are several correlation tests that depend on different statistical hypothesis in order to measure the relationship between two variables. The most well-known correlation tests are:

- ***Pearson correlation coefficient***

The Pearson correlation coefficient measures the strength and the direction of the linear relationship between two variables.

This test assumes that both variables should be normally distributed and they should be interval or ratio variables. Furthermore, it assumes linearity, which means that a straight line relationship between the variables is expected to be formed.

The formula of the Pearson r coefficient is:

$$r_{Pearson} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where X and Y are the two variables,

cov(X,Y) is the covariance matrix of X and Y,

μ_X and μ_Y are the mean values of X and Y respectively,

$\sigma_X = \sqrt{E[X^2] - (E[X])^2}$ is the standard deviation of X,

$\sigma_Y = \sqrt{E[Y^2] - (E[Y])^2}$ is the standard deviation of Y

- ***Spearman's rank correlation test***

The Spearman's rank correlation test is a non-parametric test that measures the strength of the relationship between two variables and specifically, the degree to which this relationship is monotonic. It determines whether there is an arbitrary monotonic function, which describes the relationship between two variables.

This test assumes that any assumptions about the distribution of the variables are not required. Moreover, the condition of having linear relationship between the two variables and using interval or ratio variables are not necessary.

The two variables X and Y are converted into ranks rg_X and rg_Y in order to compute the Spearman's rank correlation coefficient. The Spearman's rank correlation coefficient, which has the abbreviation of Greek letter ρ (rho), is equal to the Pearson correlation coefficient between the rank variables. The general formula of the Spearman's rank correlation coefficient is:

$$r_{Spearman} = \rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

where $cov(rg_X, rg_Y)$ is the covariance matrix of the rank variables,
σ_{rg_X} is the standard deviation of the rank variable X,
σ_{rg_Y} is the standard deviation of the rank variable Y

If there are no tied ranks, meaning that all ranks are distinct integers, the ρ coefficient can be calculated using the following simpler formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of observations and
$d_i = rg_X(i) - rg_Y(i)$ is the difference between the ranks of each i observation

- ***Kendall's tau correlation test***

The Kendall's tau correlation test is a non-parametric test that measures the relationship between two ordinal variables.

The test assumes that there is no need of knowledge about the distribution of the variables, similarly to the Spearman's rank correlation test. Furthermore, the variables must be at least at an ordinal scale.

The Kendall's tau correlation coefficient is used as an alternative correlation test to the Pearson correlation, when the assumptions of this test are not fulfilled. Additionally, it constitutes an alternative correlation test to the Spearman rank correlation test when there is a small sample size with many tied ranks.

The Kendall's tau correlation coefficient has high value when the ranks of the observations of two variables are similar, while it has low value when they are dissimilar.

The formula of the Kendall's tau correlation coefficient, which is abbreviated by the Greek letter τ (tau), is equal to:

$$r_{Kendall} = \tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}$$

where n_c is the number of concordant pairs, i.e. pairs ordered in the same way,

n_d is the number of discordant pairs, i.e. pairs ordered differently,
the denominator represents the binomial coefficient for the number of ways to select two items from n items.

Statistical significance

In order to assess the statistical significance of each correlation, the corresponding p-value is calculated. The p-value is the probability that the obtaining results are "extreme" or "more extreme" compared to the observed results of a statistical hypothesis test, under the assumption that the null hypothesis is true. [33] Hence, the p-value evaluates how well the data reject the null-hypothesis, which is defined as the statement that there is no relationship between the variables.

The significance level (α) is the probability that the null hypothesis is rejected, while the null hypothesis is actually true. Thus, it represents the proportion of obtaining false positives. For example, a significance level of 5% means that 5% of all tests may result in false positives.

A small p-value indicates that there is not enough evidence to accept the null hypothesis. Specifically, a small p-value, less than the significance level α, indicates that there is statistically significant correlation between the two variables.

The p-value is obtained by a sampling distribution, which is generated by re-sampling the values of the two variables. Specifically, the correlation coefficient (i.e. statistic) between the two variables is calculated for each random sampling of the variables' values resulting in the sampling distribution of this statistic which expresses the distribution of that statistic.
The original value of the statistic, which is computed from the initial values of the two variables, is checked against this distribution. To assess the statistical significance of this statistic, there are two alternative ways: a) the one-tailed test and b) the two-tailed test. The one-tailed test is a statistical test in which the critical area is one-sided. To be more specific, in the one-tailed test the statistic needs to satisfy one direction, i.e. it can either be greater than or less than a specific value, but not both (Figure 7). The critical area is the region of the distribution in which, if the estimated statistic falls into the area, the alternative hypothesis will be accepted. In contrast, the two-tailed test is a statistical test in which the critical area is two-sided so that the statistic can be greater than or less than a specific value. To be more specific, half of the α is used to test the statistical significance in one direction and half of the alpha to test the statistical significance in the other direction (Figure 7). Thus, the statistic is considered significant when the original value of this statistic is checked against the sampling distribution and falls into the α% critical area. The p-value is calculated using the sampling distribution of the statistic under the null hypothesis and the type of test (one-sided or two-sided). For the lower-sided test, the p-value is the cumulative distribution function of this statistic. [34]

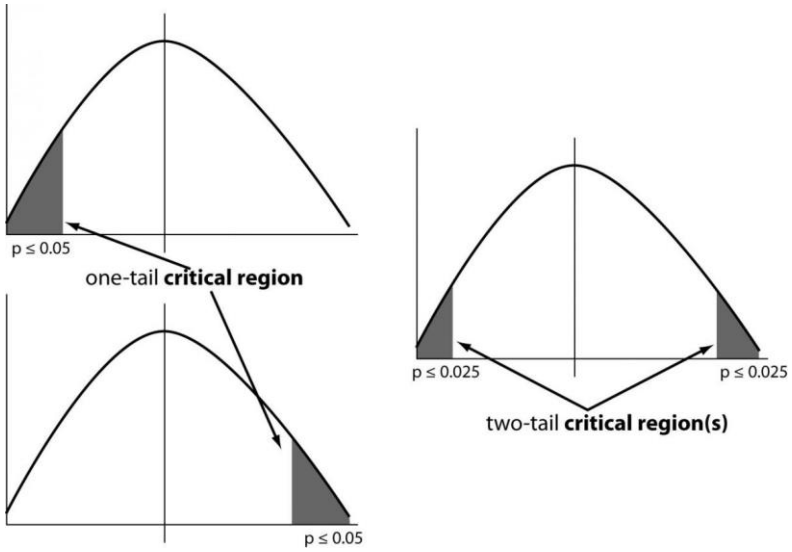


Figure 7. Graphical representation of the critical area of the one-tailed test (left) and two-tailed test (right) for a significance level of 5% (https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce)

4.2.1. Multiple Comparisons

Multiple statistical comparisons may emerge when a set of statistical tests are performed simultaneously. In this case, there is an increased risk of type I errors to occur. [35] Type I error is a result that indicates that the null hypothesis is rejected incorrectly or in other words that a condition exists when it actually does not; thus, it is ~~a~~ false positive. This could mean that the likelihood of obtaining significant results by chance is increased. To solve this problem, there are two widely used procedures for correction due to such multiple comparisons: a) Bonferroni correction and b) adjusting the false discovery rate using Benjamini – Hochberg procedure [36].

Bonferroni correction

Bonferroni correction is a conservative test which controls the family-wise error rate (FWER). The FWER is the probability of making at least one type I error in an entire set of tests. Thus, Bonferroni correction controls the FWER, guarding against the chance of making one or more type I errors (i.e. false positives). Supposing that the familywise error rate is defined to be equal to 0.05, Bonferroni correction secures that if the null hypothesis is true, the probability that the family of tests includes one or more false positives is equal or less than 0.05.
Bonferroni works finding the critical value for an individual test by dividing the familywise error rate by the total number of the family of tests. The p-value of the result of a test must be less than the critical value, in order to be statistically significant.

Bonferroni is a strict criterion and is used when a single false positive in a family of tests would be a problem. It is appropriate when a small number of tests are performed and only few results are expected to be significant. In a large number of tests Bonferroni correction may be too strict to lead in incorrectly acceptance of the null hypothesis (i.e. false negatives).

Benjamini – Hochberg procedure

Benjamini – Hochberg procedure controls the false discovery rate. False Discovery Rate (FDR) is the expected proportion of "discoveries" (i.e. significant results) that are actually false positives. Thus, this procedure controls the low proportion of false positives.
Benjamini – Hochberg works by sorting the p-values in ascending order and ranking them. Then, the critical value of each individual test is calculated with the equation $\frac{i}{m}Q$, where i is the rank, m is the total number of tests and Q is the desired false discovery rate. The largest p-value that satisfies P < $\frac{i}{m}Q$ is significant as well as all the p-values that are smaller than it, even if they are not lower than their critical values.
This statistical approach is less strict and sensitive than Bonferroni. Thus, it is preferred when there is a large number of tests.

4.3. Classification Methods

4.3.1. Support Vector Machine Classifier

Support Vector Machine (SVM) [37] is a machine learning algorithm and supervised learning model that analyzes data for classification and regression analysis. SVM targets to find an optimal hyperplane in an N-dimensional space, where N is the number of features, which distinctly classifies the data points. SVM is a binary classification, meaning that it aims to find the hyperplane that separates the data points of two classes.

The SVM tries to solve the classification problem based on two concepts: a) large-margin separation and b) kernel functions. Thus, SVM targets to find the hyperplane that maximizes the margin between the two classes in the space (Figure 8). The margin is the distance between the hyperplane and the support vectors. The support vectors are the points of each class closer to the hyperplane. SVM algorithm uses a set of mathematical functions which are defined as the kernel. The kernel function takes as input the data and transforms it in the required form. Thus, the kernel function defines the dimension of the feature space in which the training data will be classified. The basic kernels are the linear, the Gaussian and the polynomial whereas the selection of the appropriate kernel function is crucial for the classifier's performance.

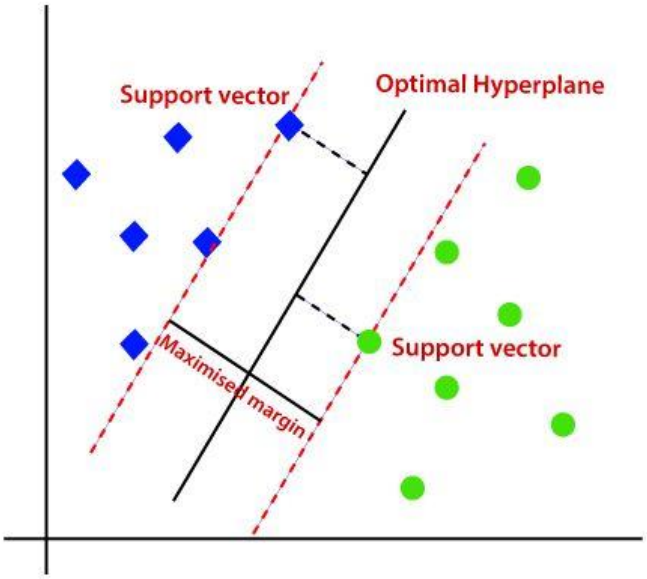


Figure 8. The black line represents the optimal hyperplane that separates the two classes and is chosen by the SVM classifier. (https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm)

4.3.2. K – Nearest Neighbors algorithm

The K – Nearest Neighbors algorithm [38] is a supervised machine learning algorithm that is used for classification and regression problems. The KNN algorithm makes the assumption that similar data points are close to each other. When a new sample has to be classified, the distances between this new sample and the training data points are calculated. The KNN selects the k-nearest points according to their distances and assigns the new sample to the dominant sign. The sign indicates the class label; if there are two classes, the samples of the positive class are signed with "+" while the samples of the negative class are signed with "-".
The main drawback of the KNN is that there is ambiguity in the selection of the initial "K", which is the number of the nearest "neighbors" that the algorithm should include into the process. Additionally, it has high computational cost due to the fact that it needs to compute distances between each query and all the training samples.

4.3.3. Multiclass classification

Multiclass classification is the classification task that consists of data driven from more than two classes. Most of the machine learning classifiers, such as SVM and KNN, are by nature binary. However, there are two popular techniques to solve the problem of multiclass classification by forming multiple binary classifiers, the one-vs-all and one-vs-one technique. [39]

One-vs-all (OVA)

The one-vs-rest approach reduces the multiclass problem by creating K binary classifiers, where K is the number of different classes. Each of the K binary classifiers

is trained with the samples of the k^{th} class as positive samples and all the other samples as negatives. When an unknown sample is tested, it is assigned to the class of the classifier that produces the maximum score.

One-vs-one (OVO)

The one-vs-one approach reduces the multiclass problem by comparing each class to each other class. A binary classifier is built using the samples of a pair of classes from the initial training set and learning to discriminate the two classes, while discarding the samples of the rest classes. Thus, $\frac{K(K-1)}{2}$ binary classifiers are required in order to combine all the possible pairs of classes. When an unknown sample is tested, the class that gets the more votes is selected.

4.4. Clustering methods

Clustering is an unsupervised machine learning technique that aims to group together relative data points. The data points that are grouped together in the same cluster are expected to have similar properties and features, while data points of different clusters should have significantly dissimilar behavior. There are two popular clustering algorithms: a) K-means clustering and b) Hierarchical clustering. [40]

K-means clustering

K-means clustering is a widely used clustering algorithm. The method initially selects a predetermined number of clusters, which is symbolized by K and randomly initializes their respective cluster centroids. For each data point the distance between the point and each group center is calculated. The data point is assigned to the cluster whose center is closest to it. The cluster centroids are re-calculated by averaging all the vectors of the points that belong in the group. This procedure is repeated iteratively until the cluster centroids do not change significantly between the iterations. Thus, the algorithm groups together data points without having any knowledge about their labels.

Hierarchical clustering

Hierarchical clustering is an agglomerative algorithm, which results in a tree-like structure which is called dendrogram (Figure 9). The approach starts by considering each data point as a separate cluster. Consequently, it identifies the two clusters that are closest together based on a distance measure and merges them to one cluster. This procedure is repeated iteratively until all the clusters are merged together in one. The height in which the tree cutoff is set, determines the number of the derived clusters of the algorithm.

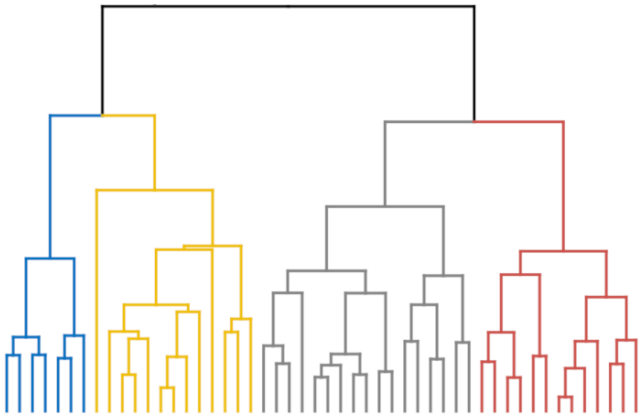


Figure 9. Structure of a dendrogram (https://www.datanovia.com/en/lessons/divisive-hierarchical-clustering/)

4.5. Regression Methods

Regression is a statistical method that examines and estimates the relationship between a dependent variable and one or more predictors. It is a reliable method for identifying the impact of the predictors on the dependent variable. The dependent variable is the main variable that the regression model aims to predict. Predictors (or independent variables) are the variables that are hypothesized to influence and are used to predict the dependent variable. When the number of predictors is more than one, the process is called multiple regression.

Linear regression consists of finding the best-fitting straight line through the observed data points. The best-fitting line represents the regression line. Thus, linear regression attempts to model the linear relationship between the dependent variable Y and the predictors X. The model is expressed in matrix form as: Y = X*w + ε, where w is the vector of regression coefficients that needs to be estimated and ε is the error term.
The most common method for fitting a regression line is the least-squares method. This method calculates the best-fitting straight line for the observed data by minimizing the sum of squares of the vertical deviations from each data point to the line. This vertical deviation is equal to 0, when the point lies on the fitted line exactly. Thus, the cost function, which actually represents the difference between the estimated values and the actual data we are trying to fit, is equal to:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{P} w_j x_{ij}\right)^2$$

where N and P are the number of observations and predictors, respectively
y_i is the actual value of the dependent variable for the i-th observation
$\hat{y}_i$ is the estimated value from the regression model for the i-th observation
w_j is the vector with the weights (or regression coefficients) for each predictor
x_{ij} is the value of the "j" predictor for the i-th observation

However, when the number of predictors is greater than the number of observations, the regression model may result to overfitting, providing inaccurate results. The shrinkage – regularization methods can be used in order to solve the model complexity and prevent overfitting. There are two well-known regularization methods [41]: a) least absolute shrinkage and selection operation (LASSO) regression [42] and b) ridge regression.

4.5.1. Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is a type of linear regression analysis that uses a shrinkage procedure. Shrinkage is a statistical process to shrink data values towards a specific point, such as zero. Lasso regression uses L1 regularization, meaning that it adds a "penalty" term of the absolute value of the magnitude of the regression coefficients in the cost function. Specifically, the cost function of Lasso is equal to:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{P} |w_j|$$

The hyperparameter λ is a tuning parameter that expresses the amount of the strinkage. As the value of λ increases, more and more coefficients are set to zero and the corresponding independent variables are eliminated. On the contrast, when λ = 0, no regularization is performed and thus no parameters are eliminated.

This type of regularization can lead to zero coefficients and thus some predictors are completely neglected for the evaluation of the dependent variable. Hence, lasso selects the most important predictors which are considered to predict adequately the outcome. Lasso regression performs feature selection and regularization, reducing the chance of overfitting and enhancing the prediction accuracy and the significance of the selected predictors.

4.5.2. Ridge Regression

Ridge regression is a linear regression that uses shrinkage without leading to the elimination of predictors. Ridge regression uses L2 regularization, meaning that it adds a "penalty" term equivalent to the square of the magnitude of the regression coefficients. The cost function of Ridge regression is:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{P} {w_j}^2$$

Similarly to LASSO, the hyperparameter λ is the same tuning parameter that controls the strength of the penalty term. When λ = 0 ridge regression performs the least-squares method. When λ = ∞, all regression coefficients are shrunk to zero.
Ridge regression shrinks the coefficients, but it does not eliminated them. Thus, the method reduces the model complexity and multi-collinearity, but it does not perform feature selection.

5. Methodology and Results

5.1. Methodology Overview

This project demonstrates a combined analysis of gene expression microarray data and radiomic features data acquired from CT medical images from patients with NSCLC, in order to contribute to valid diagnosis and prognosis of lung cancer. Statistical methods and tests were used to examine significant correlations between gene expressions and imaging features and identify their potential to the lung cancer diagnosis. Furthermore, machine learning methods were used to classify and cluster problems as well as to evaluate them, in order to investigate deeper associations and enhance the significance and the diagnostic ability of radiomic and genomic data.

The steps of the proposed analysis are briefly mentioned below (Figure 10). Each step is described in detail in its dedicated section.

1. Data acquisition.
2. Identification of Differentially Expressed Genes using SAM and 2-fold change.
3. Investigation of correlations between genes and radiomic features using Spearman rank correlation test and quantitative SAM.
4. Extra validation of the extracted significant genes.
5. Clustering of radiomic features.
6. Construction of predictive models of radiomic features in terms of genes.
7. Prediction of lung cancer's staging using genomic and imaging feature data.
8. Gene Enrichment Analysis

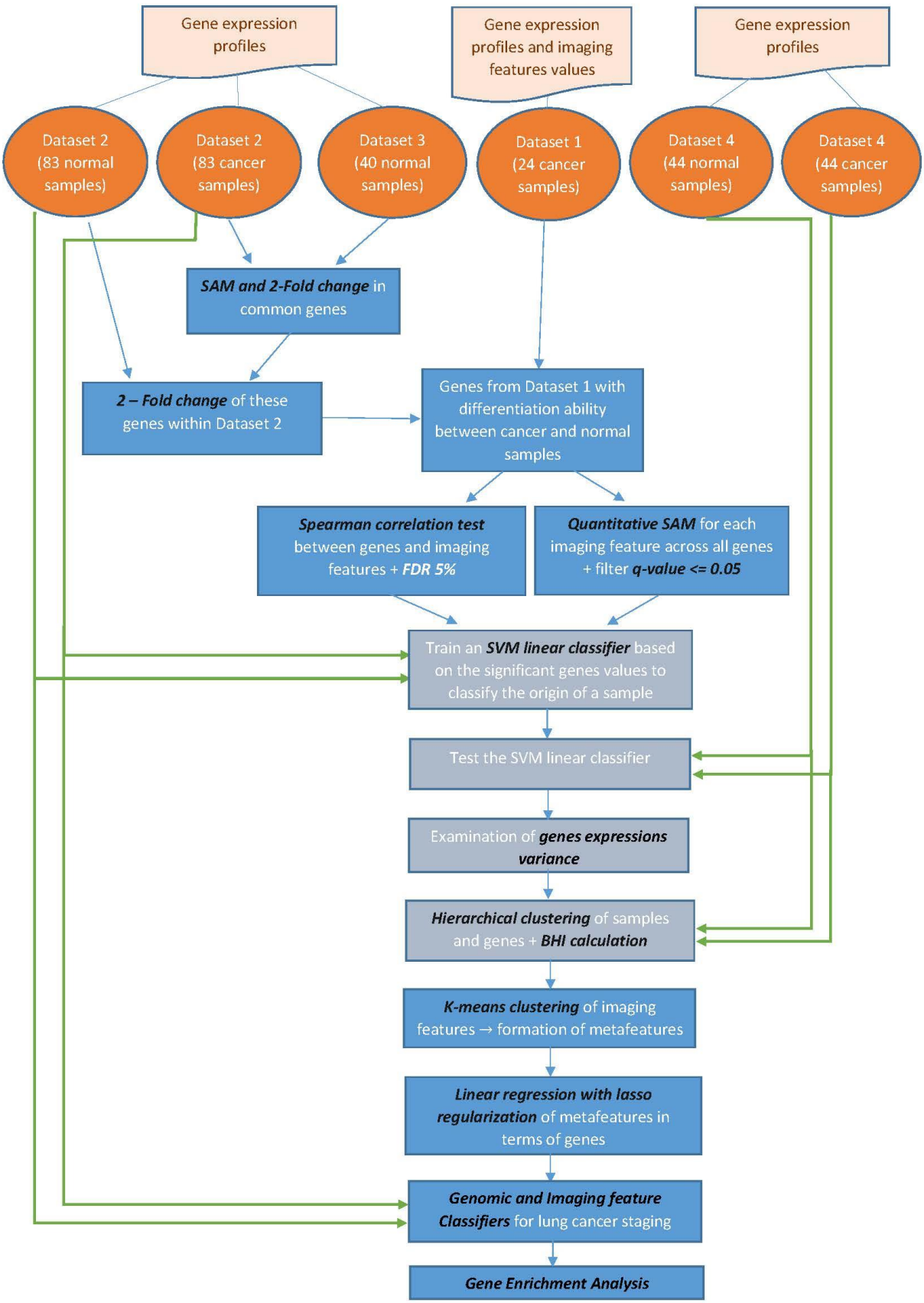


Figure 10. Flowchart of the proposed analysis.

5.2. Description of Datasets

Four different datasets were used during the initial procedure analysis. The gene expression microarray data of all datasets were obtained from the publicly available Gene Expression Omnibus (GEO) database. In each dataset the probes were coded into their corresponding Entrez Gene ID according to the Illumina platform. Due to the fact that one Entrez Gene ID may map to more than one probes, the probe with the higher gene expression value was used to express the corresponding Entrez Gene ID. The gene expression values of all probes of all datasets have been preprocessed with the same method: quantile normalization and log2 transformation.

- The *main dataset* is **dataset GSE28827** ([14], [15]) (abbreviated Dataset1), which contains gene expression microarray data and CT radiomic feature values for 26 patients with NSCLC. This dataset contains gene expression profiles for 24371 different genes, regarding to the Entrez Gene terminology, for each patient. Furthermore, it contains CT scans for each patient, which are obtained from the publicly available Cancer Imaging Archive (TCIA) database. The radiomic features were provided by Eleftherios Trivizakis (University of Crete) and were extracted using the open-source python package pyradiomics [43]. This package requires the 3D-ROI of the scan for each patient. Scans with ROI < 10 pixels were excluded; thus, 24 patients were used for further analysis. Hence, 749 CT radiomic features were extracted for the 24 patients. According to pyradiomics, features were computed on the original images as well as on derived filtered images. In order to efficiently process the images, several filters were applied on the original image, such as the Laplacian of Gaussian, Wavelet, Square, Square Root, Logarithm, Exponential and Gradient. Consequently, imaging features were calculated for each filtered and unfiltered image related to the following categories [44]:
  - ***First order statistics***, such as energy, entropy, the minimum, the maximum and the mean gray level intensity, standard deviation, skewness, kurtosis etc. The first order statistics features describe the distribution of voxel intensities within the image ROI.
  - ***Shape 3D features***, such as maximum 3D diameter, surface area etc. The shape 3D features include descriptors of the three-dimensional (3D) size and shape of the ROI. They are independent from the gray level intensity distribution and thus they have the same values for all the original and the filtered images.
  - ***Gray Level Co-occurrence Matrix (GLCM),*** such as autocorrelation, cluster tendency, contrast etc. The GLCM describes the second-order joint probability function of the image region defined by the mask.
  - ***Gray Level Size Zone Matrix (GLSZM),*** such as gray level Non-Uniformity, Gray Level Variance etc. The GLSZM quantifies gray level zones in the image, where the number of connected voxels that share the same gray level intensity constitutes a gray level zone.

- ***Gray Level Run Length Matrix (GLRLM),*** such as High Gray Level Run Emphasis, Long Run Emphasis etc. The GLRLM quantifies gray level runs, which are the length of consecutive pixels that have the same gray level intensity.
- ***Neighboring Gray Tone Difference Matrix (NGTDM),*** such as coarseness, complexity, contrast etc. The NGTDM quantifies the difference between a gray level intensity and the average gray level intensity of its neighbors.
- ***Gray Level Dependence Matrix (GLDM),*** such as dependence entropy, dependence variance etc. The GLDM quantifies gray level dependencies in an image.

Additionally, dataset GSE28827 includes the information for the cancer staging of each patient.

- **Dataset GSE75037** (abbreviated Dataset2) contains only gene expression microarray data for both cancer and normal samples. There are 83 patients for each population (cancerous or normal) in this dataset, resulting in 166 patients in total. It includes gene expression values for 19227 different genes for each patient. The cancer staging of each cancer sample is also provided.

- A control dataset, which is the **Dataset GSE76925** (abbreviated Dataset3), was used during the proposed method. From this dataset, only the gene expression microarray data from the normal samples were deployed in the analysis. It contains gene expression values for 17130 different genes for 40 samples.

- A new dataset, which is the **Dataset GSE18842** (abbreviated Dataset4), was only used during the validation procedure. Hence, this dataset is defined as the *validation dataset*. It includes gene expression microarray data for both cancer and normal samples. There are 44 patients for each population (cancerous or normal) in this dataset, resulting in 88 patients in total.

Table 1 represents an overview of the used datasets.

Table 1. Overview of datasets

Authors	Dataset	GSE	Genes	Cancer	Normal	Radiomic features
Nair et al. (2012)	1	28827	24371	24 (samples)	- (samples)	749
Girard et al. (2016)	2	75037	19227	83	83	-
Morrow et al. (2017)	3	76925	17130	-	40	-
Sanchez-Palencia et al. (2011)	4	18842	-	44	44	-

5.3. Differentially Expressed Genes Analysis

The first step of the analysis was the identification of genes that have differentiation ability between cancer and normal samples. These differentially expressed genes have the potential to distinguish cancer from normal tissues and subsequently constitute diagnostic biomarkers for lung cancer.

The gene values of cancer samples of Dataset2 and the gene values of normal samples of Dataset3 were used to examine which genes are significant and have differentiation ability in a set of microarray experiments. The cancer samples and the normal samples of two different datasets were used in order to generalize the results. Dataset2 has 19227 different genes and Dataset3 has 17130 different genes. The common genes, which are the genes that exist in both datasets, are 16252. The expression profiles of these 16252 genes were used for further analysis.

However, due to the fact that the cancer and normal samples of each gene are derived from two different datasets, cross normalization is required as a pre-processing step. More precisely, the gene values of the cancer and the normal samples have been derived from different staff members, different platforms and under different laboratory conditions, which are known as “batch effects”. To restrict the batch effects and make the two different datasets comparable, cross-normalization is essential. Hence, mean-centering normalization is applied in both datasets independently as a pre-processing step. The mean-centering normalization is performed by calculating the mean value of each dataset independently and subtracting this value from each value within the dataset. Thus, the 0 point of each dataset is redefined as its mean value.

SAM implementation

After pre-processing was completed, SAM was used to identify genes that differ significantly between the two sets of microarray experiments. The mean-centered values of the 16252 common genes from Dataset2 and Dataset3 were imported as input to SAM. Apart from gene expression profiles of microarray experiments, the response variable for each experiment should also be imported as input to SAM. The response variable determines the class in which each sample of each gene belongs to. In this case, there are two classes: a) the normal samples of Dataset3, which constitute the first class and are marked with label ‘1’ and similarly, b) the cancer samples of Dataset2 form the second class and are marked with label ‘2’. Hence, there are 40 samples from class ’1’ and 83 samples from class ‘2’ for each gene. They correspond to normal and cancer groups with samples from different patients; thus, SAM was performed for the *two-class unpaired* problem. Two parameters of SAM, Δ value and the number of permutations, need to be defined in order to run SAM. The parameter

Δ was chosen with respect to the criterion of minimum FDR. In this case, the minimum FDR was equal to 0 and thus, the Δ value that corresponds to this FDR was chosen as the desired one. The value of Δ was 3.94. For the number of permutations, it is considered that the higher the number of repetitions, the better the results. According to Damle et al. [45] ,who showed that as the number of permutation increases, the FDR decreases. However, after some number of permutations, FDR has no change or very slight change. At this stage, permutations can be terminated. Hence, the number of permutations was set equal to 1000, which is an enough big number to provide precise results.
Furthermore, SAM has the extra option of setting a non-zero Minimum Fold Change. This is a more stringent criterion, because genes must satisfy the extra criterion of changing at least a pre-specified amount, in order to be called 'significant'. For this reason, SAM was performed with the aforementioned options and the additional setting of 2-fold change. Hence, positive and negative significant genes were derived after performing SAM with 2-fold change.

SAM with 2-fold change identified 7014 significant genes with a q-value equal to 0%. The 5260 of these 7014 genes were declared as positive significant, which indicates that their expression profiles are higher in cancer samples (group 2) than in normal samples (group 1). Conversely, 1754 of the 7014 significant genes were identified as negative significant; thus, their expression profiles are higher in normal samples (group 1) than in cancer samples (group 2).

2-Fold Change within Dataset2

These 7014 derived significant genes were expected to have the potential to diagnose normal from cancer samples. However, the differentiation ability of these genes had to be examined more accurately, due to the fact that the gene analysis was performed on two completely different datasets. The batch effects were restricted, but it is uncertain that they are eliminated. To enhance the significance of these genes and to provide more precise diagnosis, 2-fold change between cancer and normal samples of Dataset2 was performed. The examination of the expressions of these genes between the 2 states (cancer and normal) of the same dataset can provide more accurate gene expression analysis. The genes that continue to have "2 times" bigger expressions in one state instead of the other state within Dataset2, remain to be deemed significant in lung cancer diagnosis.
Fold change is calculated initially by performing the anti-log of all values for each gene, i.e. 2^(value), and finding the average value for each group (normal – group 1 and cancer – group 2 of Dataset2). These average expression levels of each gene under each of two states within Dataset2 are marked as $\overline{x_{1i}}$ and $\overline{x_{2i}}$ respectively. The fold change is calculated as the ratio between these two average values.
Hence, a positive gene must satisfy the extra demand of

$$2\_FC = \frac{\overline{x_{2i}}}{\overline{x_{1i}}} \geq 2$$

in order to maintain to its positive significance.
Respectively, a negative gene must satisfy the corresponding extra demand of

$$2\_FC = \frac{\overline{x_{2i}}}{\overline{x_{1i}}} \leq \frac{1}{2}$$

in order to preserve its negative significance.

The result was that 2415 of the 7014 genes remain significant according to the 2-fold change within Dataset2. 1573 genes from them were positive significant, while the rest 842 were negative significant. Hence, a fewer number of genes was proved to have the ability to discriminate cancer from normal samples after the extra examination of their diagnostic potential.

Identification of significant genes in Dataset1

Dataset1 is the only dataset that consists of gene expression data and radiomic features data. Thus, it constitutes the main dataset of this study in order to investigate the relationship between gene expression profiles and imaging features and their contribution to the prediction of lung cancer staging. Consequently, these 2415 significant genes were identified in Dataset1. 2370 of the 2415 significant genes existed in patients of Dataset1. The 1540 were positive significant, while the 830 were negative significant.

This procedure analysis for the identification of genes with high differentiation ability between cancer and normal samples will be considered as *step A*. In Figure 11 is depicted the flowchart of this procedure analysis and the corresponding results.

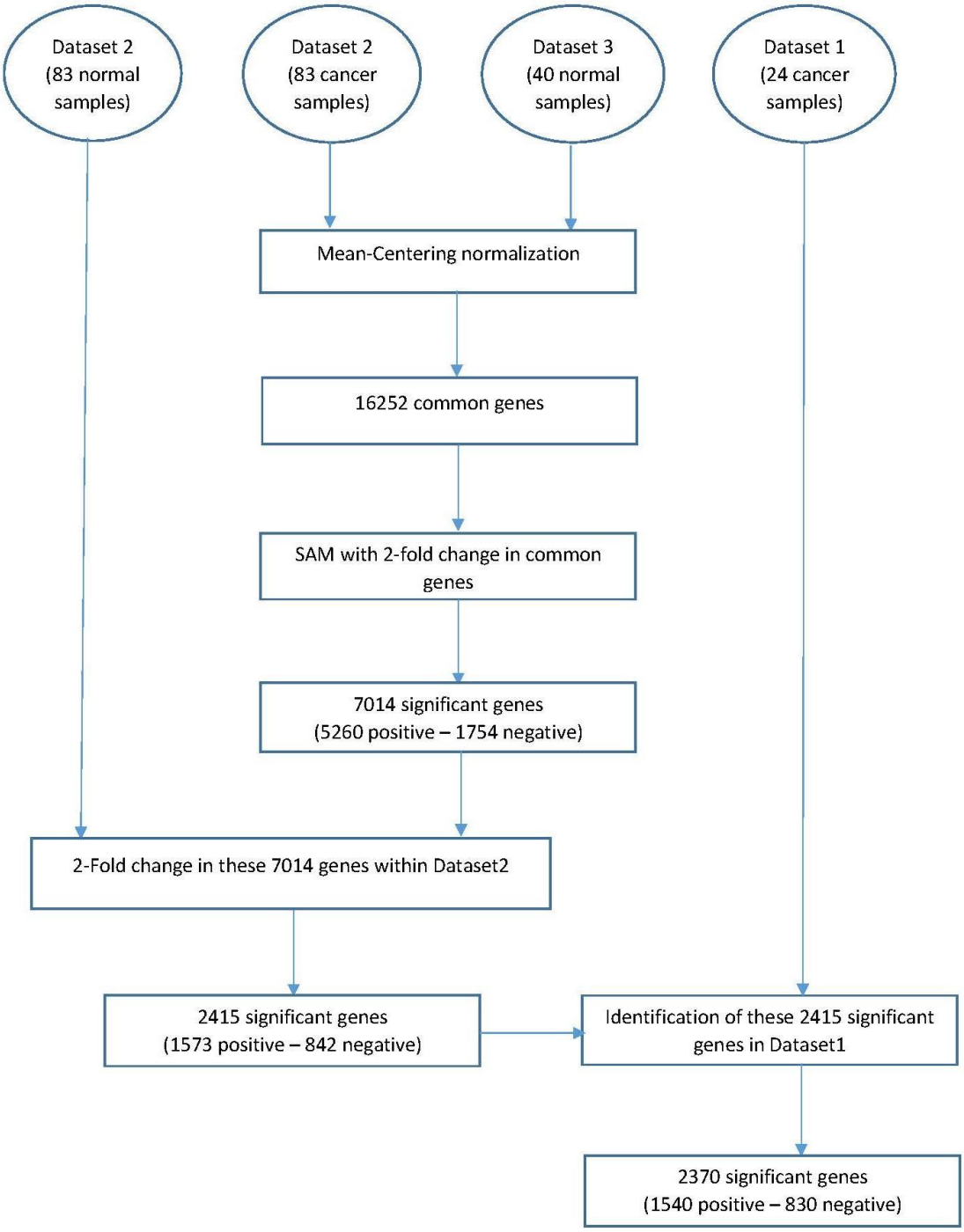


Figure 11. Flowchart of Differentially Expressed Genes Analysis (step A)

5.4. Correlation of genes with radiomic features

The genes that were identified to demonstrate diagnostic character, from the previous step analysis, were used to examine their correlation with the radiomic features. These correlations are important for lung cancer diagnosis in order to investigate the underlying biology and connection of the genotype and the phenotype of lung cancer. Radiomic features reflect the tumor heterogeneity, which is caused by the mutations of the genes in lung cancer.

In this step, the gene expression values of the 2370 significant genes and the 749 CT radiomic features from Dataset1 were used for further analysis. Each gene and each radiomic feature have values for 24 samples (patients). However, 42 imaging features were observed to have the same value across all samples, meaning that the variable does not change. When a variable does not change, its variance and thus its standard deviation is equal to 0. Hence, these imaging features are considered to have no correlation with any gene. These 42 radiomic features were excluded and the remaining 707 imaging features were used to reveal possible correlations with the genes. To investigate these associations, two statistical methods were implemented. The whole step, including both statistical methods, is defined as *step B*.

5.4.1. Spearman rank correlation test

First method: Spearman rank correlation test *(step B1)*

The first method used was the Spearman rank correlation test and is marked as *step B1*. Specifically, the Spearman rank correlation test was performed between every gene and every imaging feature. This statistical test was selected due to the fact that it is a non-parametric test, meaning that there is no assumption about the distribution of the variables. The test aims to find a monotonic relationship between the two variables and thus is more general than the Pearson correlation test, which measures the specific linear relationship. Furthermore, the values of the imaging features have no tied ranks; thus, the Kendall's tau correlation test is not required.

The rho (ρ) values and the corresponding p-values of the Spearman rank correlation test were produced for each pair of genes and imaging features. This test is implemented using the build-in function *corr* of Matlab. According to this function, the p-values are computed using the Student's t-distribution for two-tailed test by doubling the more significant value of the two (right or left) one-tailed p-values.

The total number of correlation tests were equal to 2370 (genes) x 707 (imaging features) = 1675590 tests. This is the number of all possible pairs of genes and imaging features. Due to the extremely large number of tests, we applied FDR-controlling as a statistical approach to correct multiple comparisons and enhance the statistical

significance of the derived correlations. The Benjamini-Hochberg procedure was used to correct the p-values using the FDR correction. This approach is a more robust method and is preferred when the number of tests is large. FDR aims to control the proportion of "discoveries" (significant results) that are actually false positives and is appropriate when seeking for "discoveries", because it is a less strict criterion than Bonferroni method. Bonferroni is rather preferred with small amounts of multiple comparisons or significant results, not in a quest for many "discoveries". When there are large numbers of tests, as in this case, Bonferroni could be so strict that it may produce false negatives; thus, it may discard desired significant correlations. Consequently, we used the Benjamini-Hochberg method to correct the p-values of the correlation due to its increased power in set of many tests However, such an extreme large number of tests (1675590) reduces the robustness of FDR Hence, due to the fact that during this step we aimed to investigate correlations between individual genes and each imaging feature, FDR across each gene was applied to correct multiple comparisons. In other words, the correlations between a gene and all imaging features were corrected for each gene separately. A vector with the p-values of the correlations between each gene and all imaging features was imported as input to FDR. The length of the vector was equal to 707, which is the number of the radiomic features. Thus, 2370 independent FDR tests (equal to the number of genes) were performed to correct multiple comparisons. FDR equal to 5% was applied in order to guarantee that only 5% of the significant correlations may be false positives.

This statistical analysis resulted in 6883 statistically significant correlations among genes and radiomic features. These statistically significant correlations refer to 95 different genes, as some genes are correlated with more than one imaging feature. To be more specific 95 from the initial 2370 different genes were correlated with at least one imaging feature.

5.4.2. Quantitative SAM

Second method: Quantitative SAM *(step B2)*

The second method to investigate significant correlations between genes and imaging features was the SAM using the *quantitative problem*. This approach is defined as *step B2* of the whole procedure analysis. The SAM investigates significant genes with respect to their response variable, meaning that it identifies genes significantly correlated with the response variable. Thus, the gene values of the 2370 differentially expressed genes, which have been derived from *step A*, was imported as input to the quantitative SAM and the continuous-valued imaging feature as the response variable. The SAM for quantitative problems method was used due to the fact that the desired response variable is a continuous-valued variable. Thus, a SAM with all genes values for each imaging feature independently was implemented. Therefore, 707 different SAM tests were implemented; one for each imaging feature. In each SAM, correlations between the particular imaging feature and the genes were investigated.

However, SAM for quantitative problems uses the "linear regression coefficient" as the numerator r_i of the statistic d_i by regressing the expression of each gene "i" on the response variable "y". In this thesis, Spearman's rank correlation coefficient is preferred to be used as statistic d_i , which defines the significant or not significant correlation of the gene "i" with the response variable (i.e. imaging feature) by the implementation of the SAM permutation-based algorithm. Hence, *step B2* is comparable to *step B1*, while they are using the same correlation coefficient (rho value).

For this reason, the input of each quantitative SAM (i.e. genes and imaging features values) was transformed before importing to SAM. The relation between a linear regression coefficient "b" and the Pearson correlation coefficient r_{xy} is:

$$b = r_{xy}\frac{s_y}{s_x}$$

Hence, the linear regression coefficient "b" is equal to Pearson correlation coefficient r_{xy} , when the standard deviation of the response variable "y" and the gene "x" are equal to 1. In this case, SAM will use the Pearson correlation coefficient r_{xy} as a statistic score in the computations. However, the Spearman rank correlation coefficient between two variables is equal to the Pearson correlation coefficient of the ranked variables.

Thus, the transformation, that is required, is to transform the values of each gene "x" and each response variable "y" into ranks and then scale-divide them with their standard deviations independently, forming $x' = \frac{x}{s_x}$ and $y' = \frac{y}{s_y}$ respectively, in order to have $s_{x'} = s_{y'} = 1$.

Proof of $s_{x'} = s_{y'} = 1$:

$$var(x') = var\left(\frac{x}{s_x}\right) = E\left(\frac{x}{s_x} - E\left(\frac{x}{s_x}\right)\right)^2 = E\left(\frac{1}{s_x}\big(x - E(x)\big)\right)^2 = \frac{1}{s_x^2}E(x - E(x))^2 = \frac{1}{s_x^2}\,var(x)$$

$$\text{Thus, } var(x') = \frac{1}{s_x^2}\,var(x) = \frac{var(x)}{var(x)} = 1 \text{ and } s_{x'} = \sqrt{var(x')} = \sqrt{1} = 1$$

Similarly, it can be proved that $s_{y'}$ is equal to 1.

Additionally, to prove that Spearman rank correlation coefficient is equal to the Pearson correlation coefficient of these ranked-scaled variables x' and y', it is needed to be shown that Pearson correlation is invariant to scaling transforms i.e. $r_{x'y'} = r_{xy}$.

Proof of $r_{x'y'} = r_{xy}$:

$$r_{x'y'} = \frac{cov(x', y')}{s_{x'}\,s_{y'}} = cov(x', y') = E\big[\big(x' - E(x')\big)\big(y' - E(y')\big)\big] \Leftrightarrow$$

$$r_{x'y'} = E\left[\left(\frac{x}{s_x} - E\left(\frac{x}{s_x}\right)\right)\left(\frac{y}{s_y} - E\left(\frac{y}{s_y}\right)\right)\right] = E\left[\frac{1}{s_x}\big(x - E(x)\big)\frac{1}{s_y}\big(y - E(y)\big)\right] \Leftrightarrow$$

$$r_{x'y'} = \frac{1}{s_x s_y}E[(x - E(x))(y - E(y))] = \frac{1}{s_x s_y}cov(x, y) = r_{xy}$$

In conclusion, SAM will actually use the Spearman rank correlation coefficient of the original data in its computations with the transformation of the values of each gene and ~~of~~ imaging feature into ranks and their deviation with their standard deviations. The numerator r_i of the statistic d_i, that SAM uses in its computations, expresses the ρ correlation coefficient between gene "i" and each response variable.

After transforming the input, the 707 SAM tests were performed to identify significant correlations between genes and each imaging feature. The number of permutations was set equal to 1000 and the Δ value was chosen in each SAM so that the minimum FDR is achieved. The minimum FDR is equal to 0, but in some cases the SAM execution failed to achieve FDR = 0. In these cases, SAM used the Δ value that corresponds to minimum FDR, which may reach high and unacceptable values, such as FDR = 58%. To exclude the genes that correspond to these unacceptable values of FDR and simultaneously were identified as significant from SAM, a filter of q-value <= 0.05 was applied. Hence, significant genes and consequently significant correlations were considered only the genes that correspond to q-value <= 0.05.

This second statistical analysis resulted in 651 statistically significant correlations. Similarly with the *step B1*, some genes had significant correlations with more than one imaging feature. Thus, these 651 correlations referred to 137 different significant genes.

5.4.3. Combination of the two statistical methods

Both the Spearman rank correlation method and the quantitative SAM method investigate possible statistically significant associations between important genes with high differentiation ability and radiomic features. The statistical significance of these correlations is secured by applying the extra criterion of FDR 5% in both cases. There was a remarkable reduction in the number of significant genes in both methods, after searching for genes that, apart from their differentiation ability, satisfy the extra demand of being correlated with radiomic features. The significant genes after *step A* were 2370, while after *step B1* and *step B2* were reduced to 95 and 137 respectively.

The common genes of these two methods (*step B1* and *step B2*) were used for further analysis to enhance their significance. The number of these genes was equal to 78, meaning that these 78 genes are significantly correlated with imaging features in both *B1* and *B2* steps. They are genes that have high discrimination ability and simultaneously significant correlations with radiomic features after the application of the two different methods. Thus, they seem to have a high impact on the detection and the diagnosis of lung cancer.

The flow of the statistical tests and their results that are conducted during the whole analysis of *step B* is depicted in Figure 12.

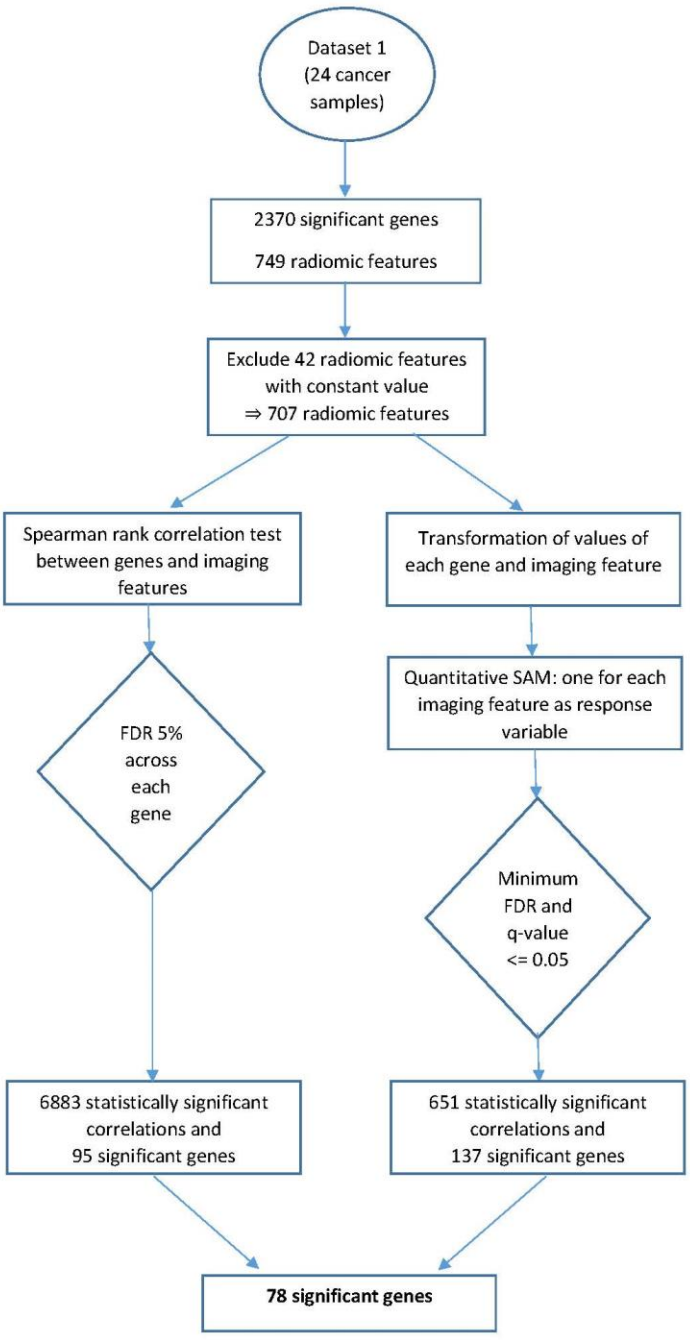


Figure 12. Workflow for the investigation of correlations between genes and imaging features (step B)

5.5. Data visualization with Heatmaps

The expression profiles of the 78 significant genes, that have been derived so far, were expected to differ significantly between cancer and normal tissues. Heatmap is a data visualization technique which describes the magnitude of the values by coloring them from a predefined color spectrum. It is widely used for the visualization and the interpretation of gene expression data.

Two heatmaps were created for the data visualization of the 78 significant genes in order to visualize the difference in their gene expression profiles between the cancer and the normal samples. The heatmaps were made in R programing language using the function *heatmap.2* with row scaling.

1st Heatmap:

The first heatmap was constructed using normal and cancer samples from different datasets. Specifically, 40 normal samples from Dataset3 and 107 cancer samples (83 cancer samples from Dataset2 and 24 cancer samples from Dataset1) were used for making the heatmap as shown in the following Figure 13:

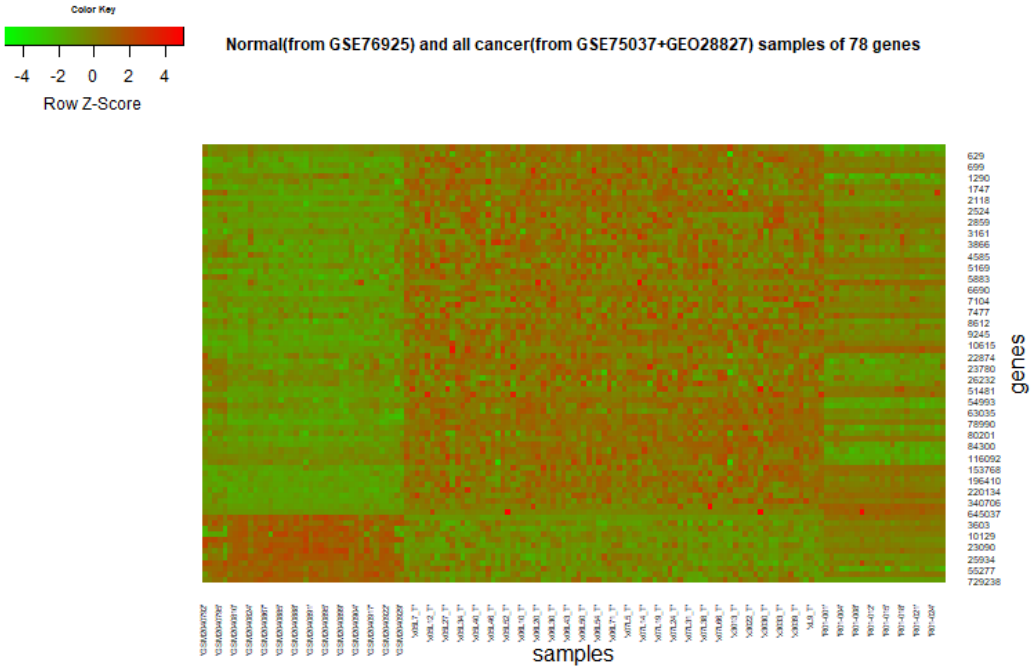


Figure 13. Heatmap for 78 genes using 40 normal samples from Dataset3 and 107 cancer samples from Dataset2 and Dataset1

The samples are represented in the columns and the genes in the rows. Looking at the figure 13 from top to bottom, the positive significant genes are displayed first, followed by the negative significant genes. Furthermore, looking at the figure 13 from left to right, the normal samples are displayed first, followed by the cancer samples. Red color indicates higher values and green color indicates lower, as indicated in the 'colorkey' bar.

The positive significant genes are the genes that have higher values in cancer samples than in normal. On the contrary, negative significant genes are the genes that have higher values on normal samples than in cancer. We can discriminate these two groups (positive or negative) of genes from the first heatmap (Figure 13), because the cancer samples for the positive genes are labelled mainly with red colors while the normal samples of these genes are labelled with more shades of green color. Similarly, the cancer samples for the negative genes are labelled with green color while the normal samples are labelled with mainly red color. However, the cancer samples of the Dataset1 (right samples on the figure 13) do not depict clearly this difference between the normal and the cancer samples for all genes. The fact that the gene values originate from different datasets can explain the presence of batch effects.

2nd Heatmap:

The second heatmap was constructed using normal and cancer samples from the same dataset. Hence, 83 normal and 83 cancer samples from Dataset2 were used to create the heatmap as shown in the following Figure 14:

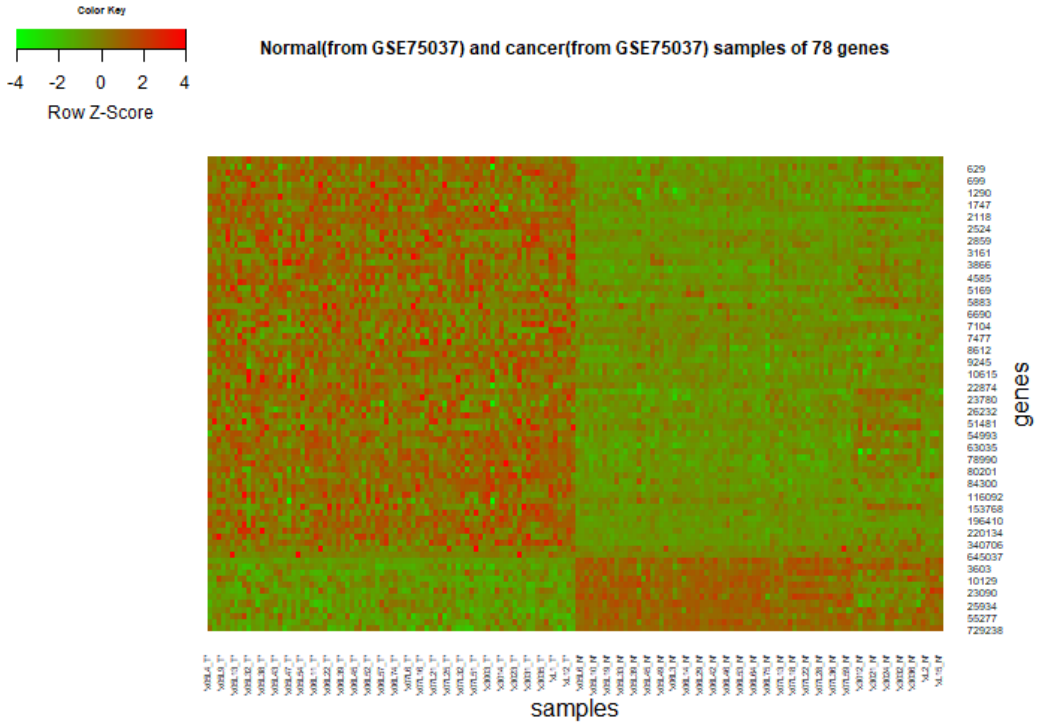


Figure 14. Heatmap for 78 genes using 83 normal and 83 cancer samples from Dataset2

The only difference with the previous heatmap (Figure 13), is that the cancer samples in this case have been displayed in the left side of the figure 14 and the normal samples in the right side.

This heatmap has the potential to differentiate more accurately the two groups of genes (positive and negative) and the two groups of samples (cancer and normal). The higher values of the positive genes for the cancer samples are represented quite clearly with the red shades and the lower values for the normal samples with the green

color. On the opposite, the higher values of the negative genes for the normal samples and the lower values for the cancer samples are expressed with the red color in normal samples and the green color in cancer samples, respectively.

Therefore, both heatmaps show the difference in the expression profiles of these genes between cancer and normal samples. The heatmaps confirm the discrimination and diagnostic potential of these 78 genes in lung cancer.

5.6. Extra validation of genes

The 78 genes have shown a noticeable behavior for lung cancer diagnosis and its underlying biological behavior detection related to the phenotype of the tumor. They tend to have the potential to be used as significant biomarkers for NSCLC. However, it is crucial to validate further their significance and role in the detection of lung cancer in order to provide reliable findings for precise medicine. Thus, several steps have been conducted in order to validate the significance of these genes. All steps that are used during this section constitute the *step C* of the whole analysis.

5.6.1. Examination of genes' predictive ability in classification

This step analysis, which constitutes *step C1*, aims at determining the ability of the genes to predict if a sample belongs to cancer or normal tissues. To empower the outcome of this classification, a new Dataset, Dataset4, which has not yet been used, is introduced as a testing set. However, the new Dataset4 does not contain 5 of the 78 significant genes. Thus, these 5 genes were excluded from the analysis.

A SVM linear classifier was used for the classification of the samples. The SVMs have many features that make them attractive for classification using gene expression data. Specifically, SVM has the advantage to deal with data that has unknown distribution and in general, there is not much information about it, due to the use of different kernels. Furthermore, it shows good performance, when the dimension of the feature space is greater than the number of observations. Additionally, it has the ability to identify outliers and reduce their influence in finding the separating hyperplane, resulting in better classification scores. [37] For all these reasons, SVM is adopted in this step for classification using gene expression data. Moreover, all kernels of SVM were tried (linear, gaussian, polynomial) in order to assess its performance. The kernel with the better performance was the linear and thus it was selected for the classifier. The other two kernels fitted with higher efficiency the training dataset that resulted in overfitting; thus, they could not predict the new unseen dataset.

The feature vector of the SVM linear classifier is composed of the 73 different genes. The classifier was trained at the 83 cancer and 83 normal samples of Dataset2. Thus,

the training set was equal to 166 samples. The test set was the 88 samples of the new unseen Dataset4, which contained 44 cancer and 44 normal samples (Figure 15).

Mean-centering as cross-normalization technique was performed on the two datasets due to the fact that the gene expression profiles have been produced from different staff and under different experimental conditions. The prediction of the classifier is one of the two classes: 'tumor' or 'control'. The class 'tumor' is considered as the positive class and the class 'control' is considered as the negative class.

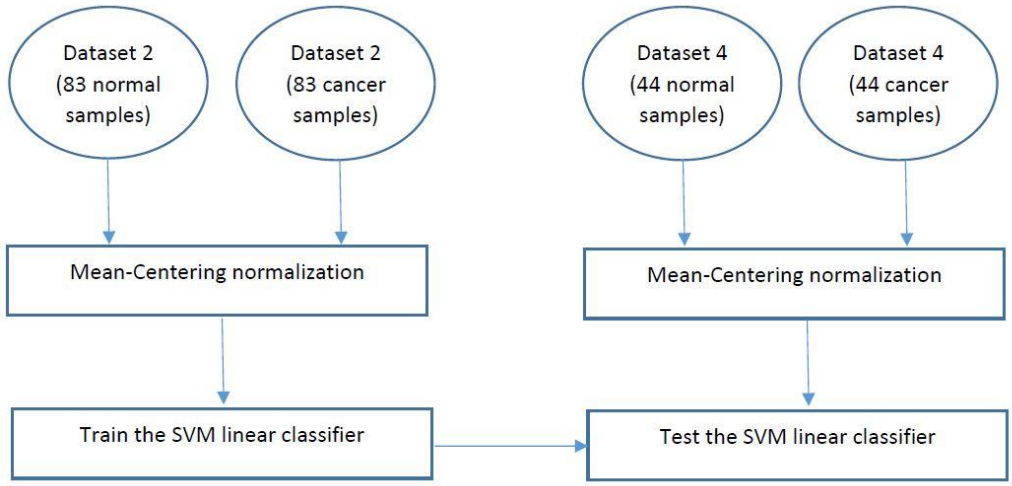


Figure 15. Workflow of the examination of predictive ability of genes in tissue classification

The classifier was evaluated by examining its ability to predict if a sample is cancerous or normal. The classifier showed great performance achieving accuracy equal to 92.05% (Table 3). Thus, the classifier has the ability to predict correctly, in high percentage, if a sample is cancerous or normal. Furthermore, the confusion matrix (Table 2) was produced in order to show the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In this case, the definition of these measurements are:

- True Positives (TP) is the number of cancer samples that were classified correctly as cancer sample.
- True Negatives (TN) is the number of normal samples that were classified correctly as normal sample.
- False Positives (FP) is the number of normal samples that were misclassified as cancer sample.
- False Negatives (FN) is the number of cancer samples that were misclassified as normal sample.

Furthermore, two additional validity metrics (Table 3) were used to evaluate the performance of the classifier using the confusion matrix. The first one is the Sensitivity, which expresses the ability of the classifier to correctly classify a sample as 'tumor'. In other words, it gives the probability (as percentage) that a sample is 'tumor' given that

the sample is cancerous. The second one is the Specificity, which expresses the ability of the classifier to correctly classify a sample as 'control'. Thus, it gives the probability (as percentage) that a sample is 'control' given that the sample is actually normal.

Equations:

$$Sensitivity = \frac{Number\ of\ true\ positives}{Total\ number\ of\ cancerous\ samples} = \frac{TP}{TP + FN}$$

$$Specificity = \frac{Number\ of\ true\ negatives}{Total\ number\ of\ normal\ samples} = \frac{TN}{TN + FP}$$

The scores of Sensitivity and Specificity were also quite high. Sensitivity was equal to 84.09%, meaning that a huge proportion of cancer samples were correctly determined as cancer. Specificity was equal to 100%, meaning that all normal samples were identified by the classifier as normal. Hence, the classifier does not produce false alarms, while none of the normal samples was identified as cancerous.

Table 2. Confusion Matrix of the SVM classifier using 73 different genes for tissue classification

		Predicted	
		Cancer	Normal
Actual	Cancer	TP = 37	FN = 7
	Normal	FP = 0	TN = 44

Table 3. Validity metrics for evaluation performance of the SVM classifier using 73 different genes for tissue classification

Accuracy	92.05%
Sensitivity	84.09%
Specificity	100%

The high performance of this classifier, assessing all the validity metrics, indicates that these 73 genes have the potential to predict accurately if a sample belongs to cancer or normal sample. Hence, the significance of these genes is enhanced, while their

diagnostic ability in lung cancer was examined with another machine learning method and was tested in a new dataset.

5.6.2. Examination of genes expression variance

The expression profiles of these 73 genes were analyzed to investigate the genes' values variance. Specifically, three measurements were calculated:

- The variance of each of these 73 genes
- The mean variance from all genes for each state (cancer or normal) within each dataset
- The difference of each gene's variance from the mean variance for each population in each dataset

The previous measurements were computed on the datasets that have been already used during the analysis, apart from the *validation dataset* (Dataset4). Thus, the expression profiles of the 73 genes in Dataset1, cancer population in Dataset2, normal population in Dataset2 and Dataset3 were deployed for this step. This step analysis is defined as *step C2*. Therefore, four mean variance values were calculated; one for each of the aforementioned datasets and their distinct populations.

After calculating and examining the differences of each gene's variance from the mean variance in each dataset, it was concluded that most of the genes present small differences. The small difference indicates that the values of these 73 genes have not huge fluctuations with respect to the mean variance of all genes within the dataset. Thus, the range of the values of the gene does not differ significantly from the mean range of the values of all genes within the dataset. Hence, these 73 genes seem to be compact in all datasets.

5.6.3. Calculation of Biological Homogeneity Index (BHI)

The Biological Homogeneity Index was used as an extra criterion to comprehensively evaluate and increase the significance of the 73 genes in lung cancer prediction. Biological Homogeneity Index (BHI) [46] is a measure that assesses how biologically homogeneous the clusters are. Biologically homogeneous clusters are the clusters that their samples belong to the same biological class.

The calculation of the BHI concludes the following procedure:

- Hierarchical clustering with one minus the Pearson's correlation coefficient as a measure of dissimilarity.
- Determination of the number of the derived clusters from clustering.
- Calculation of the BHI to assess the homogeneity of the derived clusters.

The equation for the calculation of BHI is:

$$BHI = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in\mathcal{D}_j} I(C(x)=C(y))$$

where k is the number of statistical clusters

n_j is the number of samples within cluster j

D_j express the cluster j

C(x) and C(y) are the biological class of sample x and y respectively

Should we assume that x and y are two samples that have been assigned to the same statistical class, the indicator function I(C(x) = C(y)) will take the value 1 if C(x) and C(y) are referred to the same biological class.

As the samples x and y are assigned to the same statistical class, it is expected to originate from the same biological class. Thus, the maximum and better value of BHI is equal to 1, meaning that all the samples within each statistical cluster belong simultaneously to the same biological class, producing completely homogeneous clusters. Therefore, BHI measures the proportion of sample pairs with same biological classes that are grouped together to the same statistical class.

The BHI was calculated for two different clustering cases using the expression profiles of the 73 significant genes of Dataset4. This step analysis, which is denoted as *step C3*, deploys the gene values of cancer and normal samples from the *validation dataset* (Dataset4) in order to examine further their significance using a new independent dataset (Figure 16).

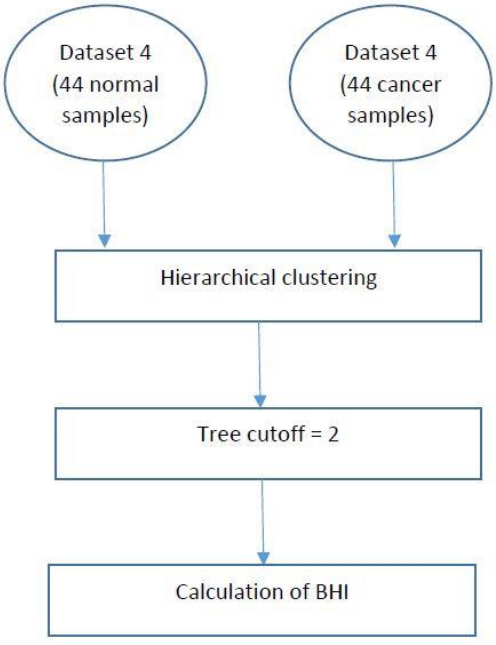


Figure 16. Flowchart for the calculation of BHI

The two different clustering cases are the following:

- First case: clustering of samples based on the gene expression profiles and calculation of BHI for assessment of the biological homogeneity of the clusters.
- Second case: clustering of genes followed by the calculation of BHI for the evaluation of the biological homogeneity of the clusters.

The two cases are described in detail separately.

First case: Clustering of samples based on the gene values

This first case refers to the clustering of samples based on the values of the 73 significant genes from the Dataset4. The BHI aims to validate if these 73 significant genes have the ability to produce biologically homogeneous clusters. Hence, it investigates the potential of the genes to group together samples with similar biological behavior forming biological classes.

The biological classes of the samples were "cancer" and "normal". Each sample was represented by 73 coordinates, one for each of the 73 genes. The proposed hierarchical clustering was performed in order to create clusters of samples. The number of clusters that were derived from the hierarchical clustering, was set equal to 2. Thus, the cutoff of the derived tree was set equal to 2 due to the two classes of the samples (cancer or normal). To be more specific, the goal of this step was to investigate if the genes could produce two biologically homogeneous clusters.

This procedure analysis resulted in a BHI = 0.8563135, a quite high value, meaning that the two clusters are adequately biologically homogeneous with respect to the class. Hence, the larger proportion of sample pairs with same biological class were grouped together to the same statistical class based on the expression profiles of the 73 significant genes. These genes seem to have the potential to group together biologically similar samples and thus derive biologically homogeneous clusters related to the functional class.

Second case: Clustering of genes

The second case refers to the clustering of these 73 significant genes using their expression profiles from the 88 samples of Dataset4, followed by the calculation of BHI. In this case, the BHI is calculated in order to examine if the genes of the same biological class are grouped together in the same statistical cluster, producing homogeneous clusters.

The biological classes of the genes are "positive" and "negative". This characterization has been produced during the first step (*step A*) of differentially expressed genes analysis and identifies the genes for the whole procedure analysis. Each gene was represented by 88 coordinates, one for each sample of Dataset4. Similarly to the first case of the clustered samples, the hierarchical clustering tree cutoff was set equal to two producing two clusters. The value of cutoff was determined by the number of the biological classes of the genes, which was two (positive and negative). The goal was to examine if the genes of the two biological classes could be grouped together in two statistical clusters with respect to their biological class.

The outcome of the analysis in this case derived a BHI = 0.8929619, concluding to a larger amount of gene pairs with same biological class that were grouped together to the same statistical class based on their expression profiles across all samples. Therefore, the larger proportion of the positive genes of these 73 genes were grouped together and the same was observed for the negative genes, resulting to biologically homogeneous clusters.

In conclusion, in both cases the value of BHI was high, meaning that the genes have the potential to group together tissue samples of the same biological class. Simultaneously genes of the same biological class are grouped together in the same statistical class. Thus, the ability of the 73 significant genes to discriminate cancer indications on a human's organism was examined and validated by two different scopes in a validation dataset in this *step C3*, leading to promising results.

5.7. Clustering of radiomic features

The significance of genes, which have been identified to contribute to lung cancer diagnosis, was validated during the process of *step C*. These genes, apart from their contribution to lung cancer diagnosis, revealed significant correlations with radiomic features, showing a combination of genetic and imaging information. According to the proposed model the next step was to examine further the potential associations between the radiomic features in order to explore groupings of relative imaging features. During this step analysis *(step D)*, we aimed at conducting the clustering of the radiomic features (Figure 17).

The only dataset that contains radiomic features data is the *main Dataset*, Dataset1, with features of 24 patients. From the aforementioned analysis of *step B*, 42 imaging features have been excluded due to its constant value across all samples; thus, the 707 imaging features was used for further analysis. These features which have been extracted from the tumor region, quantify the tumor shape and heterogeneity. They reflect the tumor characteristics, providing useful insight for the tumor morphology. The clustering of the radiomic features is essential to identify groups of relatively similar imaging features and simultaneously reduce the number of significant imaging features in lung cancer.

The values of the 707 different radiomic features are measured in different scales therefore they will not contribute equally to the analysis and the formation of the clusters. In order to overcome this problem, standardization is applied as a pre-processing step of the data. The standardization procedure encompassed the calculation of the mean value and the standard deviation of each radiomic feature. The mean value was then subtracted from the initial value of the radiomic feature and divided with the standard deviation. Thus, all radiomic features were comparable to each other while they have mean value equal to 0 and standard deviation equal to 1.

The iterative K-means algorithm was implemented to group together similar imaging features. K-means algorithm is a widely used clustering algorithm, as it is simple, quite fast and easy to understand and implement. The algorithm aims to achieve local convergence via calculating iteratively the within-cluster sum of squared distances, modifying group membership of each point to reduce the within-cluster sum of squared distances, and computing new cluster centers. [47] Additionally, two important advantages of K-means are that it works well with unlabeled datasets, meaning that there is no evidence to guide the way that the dataset should be grouped into clusters and the data should not be exclusively linearly separable.

The number of iterations of K-means algorithm was set equal to 200 in order to achieve convergence of the algorithm. To avoid finding local minimum, the number of replicates was set equal to 10, meaning that for each iteration the algorithm will start from a different set of initial starting points for 10 times. The number of K was chosen from testing 100 different values for initial K (the range was from 1 to 100). The best value of K was chosen according to its better score in *Silhouette* [48] and *Davies – Bouldin* criterion ([26], [49]). A brief explanation of these two criteria is the following:

- **Silhouette score** shows how an element is similar to the other elements of the same cluster in range from -1 to 1. Values closer to 1 mean that the sample is far away from the neighboring clusters, those closer to 0 that it is very close to the neighboring clusters whereas closer to -1 that it is assigned to the wrong clusters.
- **Davies Bouldin index criterion** is expressed as the average similarity measure of each cluster with its most similar cluster. Similarity is the maximum ratio (R_{ij}) of within-cluster distances to between-cluster distances for each pairwise of clusters. Thus, clusters which are farther apart and less dispersed will result in a better score. The minimum score is zero. Smaller values are preferred.

However, the value of K was chosen arbitrary large in order to produce enough clusters with a homogeneity score > 0.75. The Homogeneity score ([13], [14], [15]) was calculated by averaging all pair-wise Spearman correlation coefficients within each cluster.

Thus, the value of K that satisfies all the aforementioned criteria was equal to 95. Additionally, for K = 95 the Inertia and the Calinski – Harabasz criterion [50] were evaluated in order to assess the performance of clustering.

- **Inertia** is defined as the sum of the squared distance between each member of a cluster and its cluster centroid. It expresses the intra cluster distance and it is expected to be small, because the distance between the points within a cluster should be as low as possible, leading to compactness of the cluster.
- **Calinski – Harabasz** criterion is defined as the between-cluster dispersion (BCD) and the within-cluster dispersion (WCD) ratio. This is also known as the variance ratio criterion. Larger values are preferred.

The values of these validity metrics are depicted in Table 4. These measurements refer to clustering results applying K = 95. Thus, 95 clusters of relatively similar imaging features were produced.

Table 4. Results of the validity metrics for the evaluation of K-means clustering algorithm on radiomic features

Criterion	*Results*
Silhouette score	0.4148
Davies – Bouldin index	1.0074
Inertia	1407.99
Calinski - Harabasz	63.7785

The homogeneity score of each of the 95 derived clusters was computed in order to assess the homogeneity of each cluster. The criterion was that the clusters that had homogeneity score > 0.75 were considered adequately homogeneous. 77 of the 95 clusters satisfied this criterion and thus they were used for further analysis. Each of the 77 homogeneous clusters entangles related features in the compact form of a 'metafeature'. Thus, we will refer to each of these homogeneous clusters with the compact notion of ~~a~~ 'metafeatures'. Each metafeature consisted of a different number of imaging features that constitute the cluster. Hence, we represent each metafeature by the nearest imaging feature to its cluster centroid; this nearest imaging feature constitutes the principal component of the metafeature. For the cases that more than one imaging features had the same minimum distance from the cluster centroid, the first imaging feature with the minimum distance was chosen as the principal component of the metafeature.

In conclusion, this *step D* produced clusters of adequately co-expressed imaging features that are important in lung cancer diagnosis and stage identification. Furthermore, the initial number of imaging features was reduced by identifying radiomic features that can be grouped together in homogeneous clusters. The compact form of the metafeature is used in order to express a group of co-expressed imaging features by one representative imaging feature from the group.

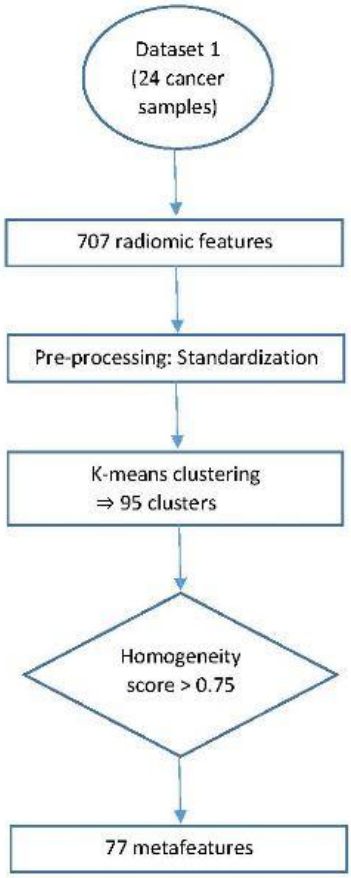


Figure 17. Flowchart of the procedure for clustering radiomic features (step D)

5.8. Predictive model of radiomic features in terms of genes

At this point, the analysis has revealed 73 genes and 77 metafeatures that are important for the diagnosis and the evaluation of lung cancer. The genes indicate a diagnostic character and are correlated with radiomic features. It is now important to investigate if they have also the ability to predict the values of metafeatures, which are actually radiomic features. To be more specific, it is crucial to explore if they have the potential to produce the radiomic features, providing a robust tool for the formation of artificial imaging features from genes.

The goal of this step analysis (step E) is to implement prognostic models of metafeatures in terms of genes. Thus, we investigate combination of genes that have the ability to predict the metafeatures (Figure 18).

Multiple linear regression is used in order to model the relationship between the dependent variable, which is the metafeature, and the independent variables, which are the genes. The 73 genes constitute the predictors of the model and each of the 77 metafeature represents the dependent variable of the regression analysis. Hence, 77 models using linear regression were performed, in order to construct a model between the genes and the metafeature.

The gene expression data and the radiomic features data from the main dataset, (Dataset1) were used for the regression analysis, as it is the only dataset that contains both genomic and imaging data. Dataset1 is comprised of data from 24 patients. The linear regression model for each metafeature has N = 24 observations (sample size) and P = 73 predictors (genes). Hence, the number of predictors is greater than the number of observations (P > N); therefore, the linear regression approach is problematic and the derived regression coefficients are unreliable. The least squared method is unreliable when there are too many predictors, due to the infinite number of solutions for a given problem. Moreover, when the number of predictors is greater than the number of observations, the learned hypothesis may fit the training set very well, yet fail to generalize in new samples, leading to overfitting. Overfitting is not desirable and may lead to inaccurate results. To solve this problem, LASSO regularization is used. LASSO regularization includes an extra "penalty" term in the cost function that tries to minimize it, in order to enhance the prediction ability of the regression model. This penalty term shrinks some of the regression coefficients of the predictors to 0, leading to feature selection and reducing variance. Thus, these predictors with coefficients equal to 0 are completely neglected for the evaluation of the dependent variable. On the other hand, ridge regression reduces the regression coefficient of some predictors that are considered to be less important and gives weight to the more important predictors. Thus, it gives different importance weights to the predictors without eliminating the unimportant variables. Therefore, LASSO regularization was selected in order to avoid overfitting and multicollinearity by reducing the number of predictors and selecting only the important ones.

The hyperparameter lambda (λ), is a shrinkage parameter of LASSO algorithm which controls the amount of shrinkage imposed on the coefficients and has to be tuned. Thus, the leave one-out cross validation (LOOCV) technique is performed to assess the model's performance and define the proper lambda value. The lambda value that minimizes the cross validated Mean Squared Error (MSE) is chosen to subsequently select its corresponding regression coefficients. The LOOCV procedure operates by dividing the samples into K subsets randomly, where K is the initial total number of samples. Then, the K-1 subsets were used in order to train the model. The remaining K^{th} sample is used to test the model. This procedure is repeated for K times in order to give the opportunity to each sample to be used as test set.

The cross validated MSE is the average value of the MSE from all the cross validation tests. The MSE is the average of the squares of the errors, meaning that is the average squared difference between the actual and the predicted value. It shows how close are the predicted values from the actual ones, reflecting how well they fit on the predictive model.

Equation: $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

where y_i is the actual value of the i-th observation and

$\hat{y}_i$ is the estimated value of the i-th observation

After selecting the λ value, the regression coefficients that correspond to its value are used to predict the dependent variable. In order to assess the model's performance, the R-squared is calculated according to the following equation:

$$R^2 = 1 - \frac{\sum_{k;observations}(Y_{actual} - Y_{predicted})^2}{\sum_{k;observations}(Y_{actual} - Y_{mean})^2}$$

R-squared, which is also known as coefficient of determination, represents the proportion of the variance for the dependent variable that is explained by the predictors in the regression model. The maximum and best value of the R-squared coefficient is equal to 1, when the estimated variable is identical to the actual; thus there are no errors from the predictive model. Hence, the accuracy of the model was calculated with the R-squared.

The criterion of R-squared > 0.70 (out of a max measure of 1) was applied, in order to identify the models that adequately predict the metafeatures. The result was that 53 of the 77 models satisfy this criterion, meaning that 53 metafeatures can be predicted from genes with an accuracy 70% and greater. Each one of these 53 predicted metafeatures is called "pMetafeature", using the prefix "p" to identify the predicted metafeatures. Furthermore, each pMetafeature was predicted from a subset of genes that have non-zero regression coefficients. The subset of the genes that predict the pMetafeature is called "signature of pMetafeature" and differ among the pMetafeatures.

Additionally, some extra validity metrics were used in order to assess the performance of the predictive models. [41] Specifically, for each of the 53 metafeatures the following metrics were calculated:

- **Normalized Root Mean Squared error (Normalized RMSE)**:

$$\frac{\sqrt{\frac{\sum_{n;observations}(Y_{actual} - Y_{predicted})^2}{N}}}{Y_{max} - Y_{min}}$$

  where Y_{max} is the maximum value of the Y_{actual} and
  Y_{min} is the minimum value of the Y_{actual}

  Similarly to MSE, the normalized root mean squared error measures the goodness of fit of the predictive model. Values closer to 0 correspond to better predictive ability of the model.

- **Pearson Correlation between the predicted Metafeature and the actual Metafeature** in order to assess the relationship between the estimated and the actual values. Values closer to 1 mean stronger association between the two variables.

- **Cross Validated Normalized RMSE:**

$$\frac{\sqrt{MSE}}{Y_{max} - Y_{min}}$$

Similarly to MSE, the cross validated normalized root mean squared error measures how well the data is fitted. Values closer to 0 mean better predictive performance of the model.

The range of the derived values of the extra validity metrics for the 53 metafeatures are depicted in Table 5. The values of the normalized RMSE as well as the values of the cross validated normalized RMSE for all the 53 metafeatures were acceptably small, indicating that the estimated values were close to the actual. Furthermore, the Pearson correlation coefficient between the pMetafeature and the actual metafeature for all the examined metafeatures was greater than 0.88, showing desired strong relationships between them. Simultaneously all the corresponding p-values of Pearson correlation coefficients were significantly lower than 1%, indicating strong evidence for rejecting the null hypothesis.

Table 5. Range of the values of the validity metrics for the predictive models of metafeatures in terms of genes

VALIDITY METRIC	MIN VALUE	MAX VALUE
NORMALIZED RMSE	0.007066477	0.146459931
PEARSON CORRELATION (PMETAFEATURE + ACTUAL METAFEATURE)	0.881558633	0.999754279
CROSS NORMALIZED RMSE	0.121271463	0.361192079

However, the metafeatures that can be predicted with high accuracy from the genes needed to be statistically correlated to them according to *step B*. To be more specific, the metafeatures are important to have predictive and simultaneously statistical correlation with genes in order to enhance their significance in lung cancer. After examining the correlations between the 53 metafeatures and the genes according to *step B*, 2 of the 53 metafeatures resulted not to be statistically correlated with genes. Thus, these two metafeatures were excluded from the further analysis, resulting to 51 significant metafeatures.

Thus, we reduced the number of metafeatures and consequently the number of imaging features that are important in characterization of cancer. Furthermore, more

reliable and specific mappings for cancer associations between genes and imaging features have been investigated through step B and E. However, all the 73 genes were participated in at least one predictive model; thus, no further reduction of the number of genes was found.

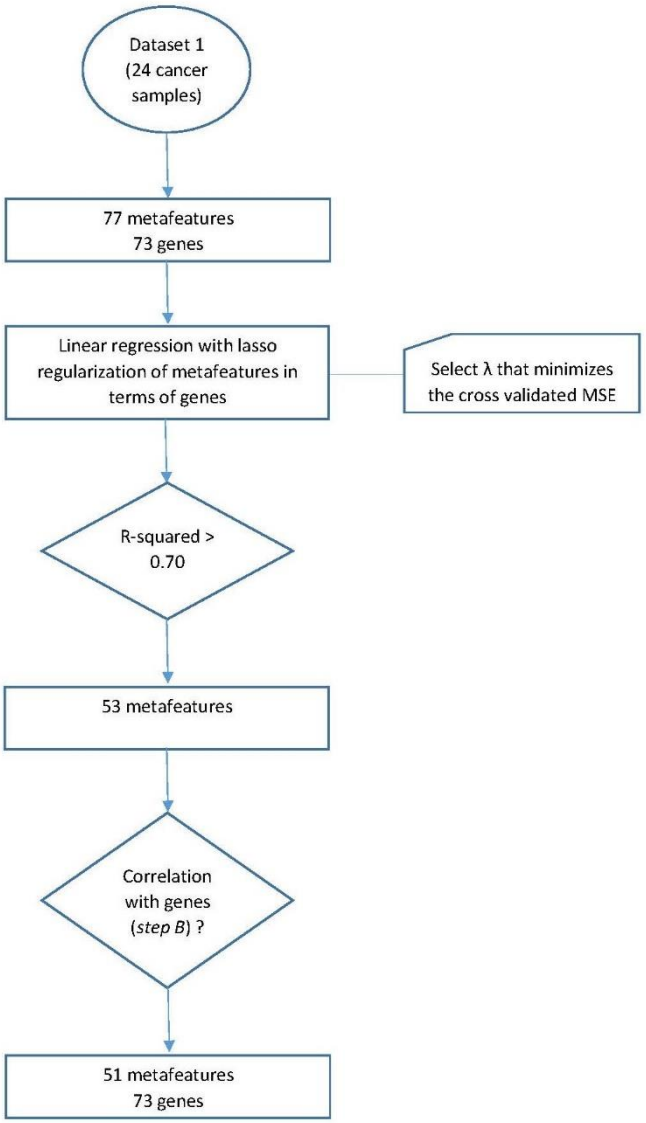


Figure 18. Workflow of regression analysis of metafeatures in terms of genes (step E)

5.9. Cancer Staging Classification

Early and valid diagnosis of lung cancer staging is crucial for the treatment planning and consequently for the patients' outcome. Lung cancer diagnosis at its early stages increase the probabilities for successful treatment and treatment. Hence, the impact of the 73 significant genes and the 51 metafeatures on predicting lung cancer staging has to be examined. Through the procedure analysis, these genomic and imaging features have shown significantly statistical and predictive relationship between them as well as detective potential for lung cancer. In this final step (*step F*), we investigate their diagnostic ability in lung cancer staging in order to reveal their underlying biological significance.

We performed two sets of classification tests in order to comprehensively examine the contribution of the genes and metafeatures in lung cancer staging prediction. In both sets of classification tests, the SVM linear classifier was used due to its simplicity and advantages. Furthermore, the SVM linear classifier had better (in some cases slight better) performance than KNN, Naïve Bayes, Decision Tree and SVMs with Gaussian or polynomial kernels. Thus, SVM linear classifier was selected to perform all the classification tests due to its robustness.
The first set of classification tests were performed using Dataset1, which is the only dataset with gene expressions and imaging data. It contains data from 24 patients, which is a small sample size. The second set of classification tests were performed using Dataset1 and Dataset2 as training and testing respectively and vice versa. Dataset2 contains data from 84 patients with lung cancer, which is also a small sample size.
Thus, the small sample size and the large number of feature space in each classifier may lead to underfitting of the model and thus poor performance. To overcome this problem, we apply a feature selection method to simplify the model by reducing the number of features which constitute the feature vector of the classifier. The feature selection methods reduce data complexity while they select the subset of the most important and relevant features, leading, in most cases, to more efficient performance. The Support Vector Machine - Recursive Feature Elimination (SVM - RFE) method was used as feature selection method. This method was selected due to the fact that SVM-RFE is an SVM-based feature selection algorithm and was created for gene expression data by Guyon et al. [51], being appropriate for our classification tests. It promises reducing computational time and higher classification accuracy rates. [51],[52] The SVM-RFE is a wrapper-type feature selection algorithm, meaning that RFE "wraps" the machine learning algorithm (i.e. SVM) to compute the scores of the features and select the more important ones. More precisely, it aims to recognize the most relevant features according to the ranking weights. It begins by computing the importance scores for all features and then removes the least important predictors. The model is re-built and the importance scores of the remaining features is calculated again. The process terminates when the optimal subset of important features is selected. Briefly, the main steps are the following:

1. Initialize with the full input dataset.
2. Train the SVM model and calculate the ranking weight for each feature.
3. Find a specific number of features (set by the user) with the lower weights and remove them.
4. Recursively repeat steps 2 and 3 with the subset of features.
5. Terminate when the optimal subset of features is defined.

5.9.1. Classification tests based on Dataset1

The first set of classification tests (*step F1*) were performed on Dataset1. Specifically, Dataset1 is the only dataset that contains expression profiles for the 73 significant genes and imaging values for the 51 metafeatures. Additionally, it includes the cancer staging for all 24 patients. In terms of cancer staging classification the information refers to T, N and M staging. The T (tumor) stage describes the tumor size and location, the N (nodal) stage indicates the spread of lung cancer to the lymph nodes around the lung and the M (metastasis) stage refers to the metastasis of cancer to other organs. [53] The combination of the status of the three descriptors determines the final lung cancer stage, which ranges from zero to four, expressed by Roman numerals (0-IV). The lower the stage, the less cancer has been spread. More precisely, stages 0, I and II are more premature stages, while stages III and IV are more advanced. The final cancer stage of each patient was derived from the combination of T, N and M staging according to the American Joint Committee on Cancer (AJCC) TNM system.

Dataset1 contains patients with staging 0, I, II and III (Table 6). The number of patients are not distributed equally to the four cancer stages, forming an imbalanced dataset. For example, 14 from the 24 patients had stage I, meaning that more than half of the dataset consists of patients with stage I. Thus, the imbalanced dataset in combination with the quite small initial sample size of the dataset lead to an exploratory classification analysis of different cancer cases. Each case consists of different number of patients, while combinations of different cancer stages were used.
Specifically, the five different classification cases were:

- ***Case 1.1:*** *4 stages (0,I,II,III) – 24 samples*
- ***Case 1.2:*** *3 stages (I,II,III – except 0) – 20 samples*
- ***Case 1.3:*** *3 stages (0,I,II (stages II and III are combined into 'II' stage)) – 24 samples*
- ***Case 1.4:*** *2 stages (I and II (stages II and III are combined into 'II' stage) – except 0) - 20 samples*
- ***Case 1.5:*** *2 stages (I and III – except 0,II) – 17 samples*

Thus, we investigated the ability of the genes and the metafeatures to predict the cancer staging in different cases, in order to find the cases in which they had the better predictive performance.

Table 6. Overview of lung cancer staging in Dataset1

Stage	Number of patients
0	4
I	14
II	3
III	3

In the cases that the number of cancer stages was greater than two, the approach one-vs-one was used to perform the multi-class classification due to the fact that it is less sensitive to the problems of imbalanced datasets than the one-vs-all technique. Furthermore, standardization was applied as pre-processing of the data, in order to transform the values of the genes and the metafeatures in the same scale. To evaluate the classifiers' performance, the LOOCV technique was used due to the limited sample size of the dataset and the absence of test sets. Additionally, the SVM-RFE technique was directly performed to the classifiers in order to prevent the classifier from a poor performance due to the fact that the dimension of the feature space was greater than the sample size of the dataset. The number of the features that constitute the optimal subset was chosen by applying all the possible numbers of features to select and distinguish the one which leads to the best performance of the classifier.

For each of the aforementioned cases, we implemented and evaluated the following classifiers:

1. SVM-RFE using the **actual metafeatures** as feature vector
2. SVM-RFE using the **genes** as feature vector
3. SVM-RFE using the **pMetafeatures** as feature vector
4. SVM using the **pMetafeatures** of the actual selected (from 1) metafeatures + selected **genes** (from 2) as feature vector
5. SVM-RFE using the **pMetafeatures** of the actual selected (from 1) metafeatures + selected **genes** (from 2) as feature vector

The first and the second classifiers were implemented in order to assess the ability of the actual radiomic features and the genes, respectively, to classify the cancer stage of a sample. The third classifier was performed to evaluate the potential of the metafeatures that have been predicted from genes to estimate the staging of the cancer and thus their ability to replace the actual metafeatures. The fourth classifier combined the predicted metafeatures of the actual metafeatures that have been selected as significant from the SVM-RFE method of the first classifier and the genes that have been selected as significant from the SVM-RFE technique of the second

classifier. Thus, this classifier is constituted from genes, while the predicted metafeatures are a linear combination of a number of genes. Therefore, it assesses the diagnostic potential in cancer staging of a group of genes. The last classifier performed an SVM-RFE algorithm on the feature vector of the fourth classifier in order to evaluate if there is excess information on this feature space, concluding to the most significant and representative features for lung cancer classification.

The accuracy of the aforementioned classifiers for all the examined cases are depicted in Table 7.

Table 7. Accuracy of classification tests with each feature vector for all 5 cases

	Actual metafeatures	Genes	Predicted metafeatures	Predicted metafeatures+genes	Selected pFeatures+genes
1st case	70.83%	75%	70.83%	70.83%	87.5%
2nd case	85%	85%	85%	85%	90%
3rd case	87.5%	91.66%	91.66%	87.5%	95.83%
4th case	95%	90%	95%	90%	100%
5th case	94.11%	94.11%	88.23%	88.23%	100%

The first case (1.1) which involved all the cancer stages had the poorer performance results. We assume that it is rather logical and can be explained by the fact that there were many different classes (i.e. 4 cancer stages) for such a small sample size (i.e. 24 patients), thus, the classifiers do not have an adequate sample size to be trained efficiently. The remaining cases (1.2, 1.3, 1.4, 1.5) showed satisfactory performance results in all classifiers, achieving an accuracy of at least 85%. Hence, we concluded to the assumption that the performance results would improve by reducing the number of the different predictive classes of the classifiers. This could be expected as a bigger amount of samples of each different class, with respect to the total sample size, would be provided for the training of the classifier.

Furthermore, it is important to notice that the accuracy of the classifiers that deployed the predicted metafeatures as feature vector is quite similar to the classifiers with the actual metafeatures in all cases. Thus, the predicted metafeatures are comparable to prediction ability of the actual metafeatures in lung cancer staging, showing promising results that can be used as diagnostic biomarkers. Additionally, the classifiers with the genes as feature vector have equally high accuracy with the classifiers with the actual and predicted metafeatures. Hence, similar to metafeatures, genes seem to have the ability to predict the cancer staging precisely. The accuracy of the fourth classifier shows that the combination of the selected predicted metafeatures and the selected genes results to similar diagnostic ability with the classifiers that use only genes or only metafeatures. However, the last classifier has even better accuracy, which is the

highest over all the classifiers. Thus, the fourth classifier seems to have redundant features which were eliminated from the SVM-RFE, leading to a more precise performance of the classifier. Therefore, the combination of the most relevant metafeatures and genes has the better diagnostic ability on lung cancer staging.
The size of the optimal subset of features after performing SVM-RFE is shown on Table 8. The initial length of the feature vector with genes and metafeatures was 73 and 51 respectively. Thus, the number of the selected genes and the selected metafeatures (either actual or predicted) was reduced in less than the half of the initial. This significant reduction indicates that there were enough redundant genes or metafeatures for predicting the cancer staging, while using the values of Dataset1. Moreover, the last column of Table 8 shows that in the last classifier, which had the better performance, both pMetafeatures and genes were selected after applying SVM-RFE. The presence of pMetafeatures and genes implies that both the combination of genes (i.e. pMetafeatures) and the individual genes are essential for a precise prediction of lung cancer staging.

Table 8. Number of selected features after performing SVM-RFE for classifiers 1, 2, 3 and 5 for all cases.

	Selected Actual metafeatures (1)	Selected Genes (2)	Selected Predicted metafeatures (3)	Selected pMetafeatures+genes (5)
1st case	20	30	13	16 (3 pMetafeatures)
2nd case	10	22	10	22 (2 pMetafeatures)
3rd case	20	30	10	25 (7 pMetafeatures)
4th case	20	42	17	57 (17 pMetafeatures)
5th case	15	30	12	28 (6 pMetafeatures)

5.9.2. Classification tests based on Dataset1 and Dataset2

The small number of patients of Dataset1 makes it essential for a further examination of the diagnostic ability of genes and radiomic features in cancer staging. The enhancement of the significance of the individual genes and the signatures of pMetafeatures was expected to provide more information about their potential on cancer staging prediction.

This set of classification tests (*step F2*) were performed using Dataset1 and Dataset2. Dataset2 contains 83 patients with NSCLC. For each patient the T, N, and M staging is provided. Hence, the final cancer stage of each patient were derived similarly to Dataset1. This dataset has patients with staging I, II, III and IV (Table 9).

Table 9. Overview of lung cancer staging in Dataset2

Stage	Number of patients
I	50
II	20
III	11
IV	2

Dataset2 contains expression profiles for the 73 significant genes, but it does not provide information about any radiomic features. Thus, we produced artificial metafeatures, which represent artificial radiomic features. Each artificial metafeature was produced by using the corresponding regression coefficients of the 73 genes that have been derived during the lasso regression of *step E*. To be more specific, we multiplied the vector of the regression coefficients with the expression profiles of the 73 genes of Dataset2 in order to produce the pMetafeature. This process was performed for all the 51 significant metafeatures using the corresponding vector of regression coefficients for each one of them. Thus, for all the patients of Dataset2 was produced artificial imaging features by using their actual expression profiles of the 73 genes and the derived vector of regression coefficients. The vector of the regression coefficients contains the appropriate weights for each one of the 73 genes. During the examination analysis of *step E*, this vector of each pMetafeature was proven to have the ability, in combination with the values of the 73 genes, to predict accurately the values of the actual metafeatures. This predictive ability will be verified further during the *step F2*.

Patients with stages 0 and IV were excluded for further analysis, as they were included only in Dataset1 and Dataset2, respectively. Dataset1 has few samples and low generalization capacity. Furthermore, stages I and II are quite close, as they constitute premature stages of lung cancer and thus it is more difficult to distinguish them. On the contrary, stage III is a more advanced stage and thus, it is easier to distinguish its difference from stages I and II. Due to the low generalization capacity of the dataset1 and the biological nature of the different cancer stages, we decided to perform classification tests using only the patients with cancer stage I and III. Thus, we selected the most distant stages which have an adequately proportion of samples in both datasets according to their total sample size.

We performed two different classification cases in this *step F2*:

- **Case 2.1: Training at Dataset 1 – Testing at Dataset 2**

2 stages (I and III) – Training set: 17 samples – Testing set: 61 samples

- **Case 2.2: Training at Dataset 2 – Testing at Dataset 1**

2 stages (I and III) – Training set: 61 samples – Testing set: 17 samples

Case 2.1 is the most important, while the *main dataset* of the whole procedure analysis, Dataset1, is used to train the classifier. Thus, the outcome of the classifier

will depend on how well it was trained by Dataset1. Furthermore, testing on Dataset2 could generalize the results by examining the impact of the genes and the radiomic features that have been derived and trained by Dataset1 on another dataset. The fact that the Dataset2 does not provide radiomic features and thus artificial radiomic features were created, is significant for the evaluation of the replacement ability of radiomic features from genes.
Case 2.2 is performed to validate further the importance of the genes in predicting the cancer staging and forming artificial imaging features that can also predict the stage of the cancer. Two important points in this case is that the classifiers are trained in a larger dataset than in case 2.1 and are tested in a dataset that contains actual values of radiomic features.

Two pre-processing techniques were used before the implementation of the classifiers, the mean-centering and the standardization. The mean-centering algorithm is used to cross-normalize the values of genes and imaging features from the two different datasets, targeting to the restriction of batch effects. Additionally, standardization is used as a second pre-processing step in order to make the values of different scales comparable.

For each of the two cases (case 2.1 and case 2.2), the following classifiers were implemented:

1. SVM using **genes** as feature vector
2. SVM-RFE using **genes** as feature vector
3. SVM using **imaging features** as feature vector
4. SVM-RFE using **imaging features** as feature vector
5. SVM using the selected **imaging features** (from 4) + the selected **genes** (from 2) as feature vector
6. SVM-RFE using the selected **imaging features** (from 4) + the selected **genes** (from 2) as feature vector

The first classifier was evaluated in order to assess the diagnostic ability of the 73 significant genes on lung cancer staging. The second classifier was used to remove possible redundant genes, leading to the improvement of the classifier's performance. The third and the fourth classifier were used to assess if the genes can produce artificial radiomic features that can predict the stage of the cancer. To be more specific, they evaluate the potential of genes to replace the predictive ability of imaging features. Finally, the fifth and the sixth classifier were performed to assess if the combination of groups of genes, that express imaging features and individual genes, can lead to better cancer staging classification scores.

The accuracy of the 6 aforementioned classifiers for the two examined cases was calculated (Table 10). In both cases the classifiers (2) and (4), which are the classifiers with a subset of genes and metafeatures, respectively, showed slight better performance compared to classifiers (1) and (3), respectively. Thus, there were some

redundant features in both classifiers with the initial number of genes and metafeatures, which were removed.

The classifier (3) of the case 2.1 was trained in the actual metafeatures of Dataset1 and tested in the artificial metafeatures of Dataset2. Classifier (4), which was derived from classifier (3) after applying SVM-RFE, had the potential to predict the cancer staging with an accuracy of 83.60%, although the values of the radiomic features of this testing set were artificially produced. Thus, this classification test revealed the ability of the group of the 73 genes to produce radiomic features, providing CT imaging information of tumors that have not undergone CT scanning. Furthermore, classifier (4) of the case 2.2 indicated similar results. More precisely, this classifier was trained in the artificial metafeatures of Dataset2 and was tested in the actual metafeatures of Dataset1. The accuracy of the classifier was not reduced significantly compared to classifier (4) of case 1.1, showing that the classifier was trained adequately with the artificial imaging features. Thus, the classifier maintained the ability to predict correctly an acceptable percentage (82.35%) of the cancer staging of the patients based on their actual radiomic features.

In case 2.1, classifier (5) with the combination of metafeatures and genes had one of the highest accuracies among the classifiers whereas in case 2.2 this classifier resulted in the highest accuracy. Hence, both individual genes and combination of genes (i.e. metafeatures) showed more precise detection of lung cancer, similarly to the set of classification tests of *step F1*. It is important to highlight that classifier (6) had slightly better or the same accuracy than classifier (5) in case 2.1 and 2.2, respectively. Thus, the optimal subset of the metafeatures and the genes seemed to have been selected during the construction of classifier (2) and (4).

Table 10. Accuracy of classification tests with each feature vector for the two training cases.

	Genes (1)	Genes after SVM-RFE (2)	Metafeatures (3)	Metafeatures after SVM-RFE (4)	Selected metafeatures+genes (5)	Metafeatures+genes after SVM-RFE (6)
Case 2.1	77.04%	78.68%	81.96%	83.60%	83.60%	85.24%
Case 2.2	70.58%	82.35%	76.47%	82.35%	88.23%	88.23%

The size of the subset of the features after applying SVM-RFE in the classifiers is depicted in Table 11. In case 2.1 there is reduction of the number of important genes and metafeatures in both classifiers (2) and (4), respectively. In case 2.2 a significant reduction on the number of genes in classifier (2) was occurred. By comparing the accuracy of 82.35% of the classifier (2) towards the accuracy of 70.58% of classifier (1), the results indicate that the initial 73 genes in this case were not capable to predict the lung cancer staging; hence, a significant reduction of their number was required. On the contrary, in the last column which contains the number of the selected

metafeatures and genes after applying SVM-RFE, it was confirmed that the optimal subset had already be selected. In case 2.1 the size of the combined list of metafeatures and genes was 50 (genes) + 18 (metafeatures), resulting to 68 features. The size of the final subset after SVM-RFE in this list was equal to 59. Similarly, in case 2.2 the size of combined list was 5 (genes) + 20 (metafeatures), leading to 25 features whereas the size of the final subset of classifier (6) was 23. Thus, a slight reduction of the features of the combined list was performed. Furthermore, it is important to highlight that the final optimal subset of classifier (6), which had the better classification rates, was consisted of both metafeatures and genes. This result complies with the first set of classification tests (*step F1*) that both, individual genes and combination of genes which form the metafeatures, are required for a more accurate prediction of lung cancer staging.

Table 11. Number of selected features after performing SVM-RFE for all classifiers for the two training cases.

	Selected Genes (2)	**Selected Metafeatures (4)**	**Selected Metafeatures+genes (6)**
Case 2.1	50	18	59 (16 pMetafeatures)
Case 2.2	5	20	23 (19 pMetafeatures)

5.10. Enrichment Analysis

The WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) [54] was used to study functional categories in different biological contexts, including biological processes, pathways, and miRNA targets (see Box), that are overrepresented among the signatures' selected genes. WebGestalt is a popular tool for the interpretation of gene lists derived from high-throughput "-omics" studies. The current version of WebGestalt (2019) supports 155175 functional categories from well-known public databases, such as the gene ontology (GO) [55], Kyoto Encyclopedia of Genes and Genomes (KEGG) [56], Protein ANalysis THrough Evolutionary Relationships (Panther) [57], Wikipathway [58], and Molecular Signatures Database (MSigDB) [59], and computational analyses.

To include gene- and miRNA-disease related information, another public resource used was the Integrated Genomic Database of Non-Small Cell Lung Carcinoma (IGDB.NSCLC)[60] as a complement.

Box: Key Terms at a Glance
Gene Ontology [61]**,** [62]**:** *The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable. The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects*: *1.* ***Molecular Function:*** *Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". Molecular-level activities performed by gene products.* ***2. Cellular Component:*** *The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion), or stable macromolecular complexes of which they are parts (e.g., the ribosome).* ***3. Biological Process:*** *The larger processes, or 'biological programs' accomplished by multiple molecular activities. Examples of broad biological process terms are DNA repair or signal transduction. Examples of more specific terms are pyrimidine nucleobase biosynthetic process or glucose transmembrane transport.*
Pathways [63]**:** *A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell. It can trigger the assembly of new molecules, such as a fat or protein, turn genes on and off, or spur a cell to move.*
miRNAs [64]**:** *microRNAs (miRNAs) are small endogenous non-coding RNAs that function as the universal specificity factors in post-transcriptional gene silencing. Discovering miRNAs, identifying their targets and further inferring miRNA functions have been a critical strategy for understanding normal biological processes of miRNAs and their roles in the development of disease.*
KEGG [56]**:** *KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:* ***1.*** *Metabolism,* ***2.*** *Genetic Information Processing,* ***3.*** *Environmental Information Processing****, 4.*** *Cellular Processes,* ***5.*** *Organismal Systems,* ***6.*** *Human Diseases,* ***7.*** *Drug Development.*
Panther [57]: *The PANTHER Pathway ontology uses controlled vocabulary to describe pathways, their components, and the relationships among them. The PANTHER Pathway ontology has four key classes:* ***1.*** *Pathway class,* ***2.*** *Molecule class,* ***3.*** *Reaction class and relationships,* ***4.*** *Cell type or subcellular compartment class.*
Wikipathways Cancer [58]**:** *WikiPathways captures the collective knowledge represented in biological pathways. Wikipathways Cancer is a cancer-related subset of WikiPathways.*
MSigDB [59]: *The Molecular Signatures Database (MSigDB) is one of the most widely used and comprehensive databases of gene sets for performing gene set enrichment analysis.*
WebGestalt [65]**:** *WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) is one of the first software applications that integrate functional enrichment analysis and information visualization for the management, information retrieval, organization, visualization and statistical analysis of large sets of genes.*

Overrepresentation Analysis on Differential Expressed Genes (DEGs)

According to both screening criteria (Spearman & FDR across imaging features, SAM) of differentially expressed genes, 78 common genes showed significant alterations of expression levels in patients with NSCLC compared to normal control patients, including 66 up-regulated and 12 down-regulated genes, respectively.
In order to clarify the potential roles of genes included in this signature, we performed overrepresentation analysis on the 78 common genes in the context of the aforementioned databases, using WebGestalt. The cutoff criterion for enrichment analyses was decided as $p \leq 0.05$ and the false discovery rate (FDR) as < 0.05 for statistically significant terms.

Supporting (S) Table 1a provides several GO biological processes that were enriched but did not reach statistical significance ($p \leq 0.05$, FDR > 0.05), such as acute inflammatory response, nucleoside triphosphate metabolic process, carbohydrate catabolic process, and vasculogenesis. KEGG and Panther analysis results showed that 78 genes were enriched ($p \leq 0.05$, FDR > 0.05) in several metabolism related pathways, including Glycolysis/Gluconeogenesis, Riboflavin metabolism, Starch and sucrose metabolism, Fructose galactose metabolism, and Pentose phosphate pathway process (Table S1b). Wikipathway cancer analysis results showed that 78 genes were enriched ($p \leq 0.05$, FDR > 0.05) in specific signaling pathways, such as ATR Signaling, DNA IR-damage and cellular response via ATR, ATM Signaling Pathway, and DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM. Among these enriched signaling pathways, the later reaches statistical significance ($p = 0.00017$, FDR $= 0.01$). Additionally, Table S1c provides the results of miRNA targets enrichment analysis reporting enriched miRNAs ($p < 0.05$), such as MIR-143, MIR-224, MIR-29A, MIR-29B, MIR-29C, MIR-423, MIR-380-3P, MIR-365, MIR-17-3P that did not reach statistical significance (FDR = 1) .

Overrepresentation Analysis of Genes in the Classification Signatures

In order to characterize the GO biological processes, pathways, and miRNA targets associated with the CT-derived imaging-correlated genes, genes of classification cases 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, and 2.2 were analyzed in the context of several databases, as aforementioned, using the online WebGestalt [66]. The cutoff for enrichment analyses was defined at a $p \leq 0.05$ and the false discovery rate (FDR) was set at < 0.05 for statistically significant terms.

Since these gene sets (classification cases 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, and 2.2) relied on the same 78 gene signature, it is expected to find similar functional categories with some degree of variation in their ranking. Indeed, in part GO analysis results showed that genes were enriched in process of vasculogenesis ($p \leq 0.05$) in all classification signatures (see Tables S2a-S8a), but also in process of acute inflammatory response, carbohydrate catabolic process, snRNA metabolic process, cell cycle checkpoint, double-strand break repair and phospholipase C-activating G protein-coupled receptor signaling pathway ($p \leq 0.05$) in more than one classification signatures (see Tables S2a-S8a).

Similar, but to a greater extent, several metabolism related pathways were enriched in the majority of the classification signatures (see Tables S2b-S8b), including Glycolysis/Gluconeogenesis, Riboflavin metabolism, Starch and sucrose metabolism, Fructose galactose metabolism, and Pentose phosphate pathway ($p \leq 0.05$). Also, several cancer signaling pathways were enriched in one or more classification signatures ($p \leq 0.05$), including Cadherin signaling pathway (Tables S2b, S3b) and HIF-1 signaling pathway (Tables S2b, S4b), as well as DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM in the majority of the classification signatures (see Tables S4b-S8b), which reaches statistical significance in the classification case 2.1 ($p = 0.00017$, FDR = 0.01).

In addition, Tables S2c-S8c provide the results of miRNA targets enrichment analysis reporting enriched miRNAs ($p < 0.05$), such as MIR-143, MIR-29A, MIR-29B, MIR-29C, MIR-224, MIR-423, MIR-380-3P, MIR-365, MIR-17-3P that did not reach statistical significance (FDR = 1).

Based on the above analysis, the **DNA IR-Double Strand Breaks (DSBs) and cellular response via *ATM*** was the most prominent finding in the majority of the classification signatures and was the only statistical significant term among all functional categories in classification signatures (classification case 2.1, $p = 0.00017$, FDR = 0.01). In our study, the genes that involved in the pathway **DNA IR-Double Strand Breaks (DSBs) and cellular response via *ATM*** are: *PARP1* (poly(ADP-ribose) polymerase 1), *FANCD2* (FA complementation group D2), *RAD9A* (RAD9 checkpoint clamp component A), and *NABP2* (nucleic acid binding protein 2), which operate via *ATM* gene (Figure 19) [67].

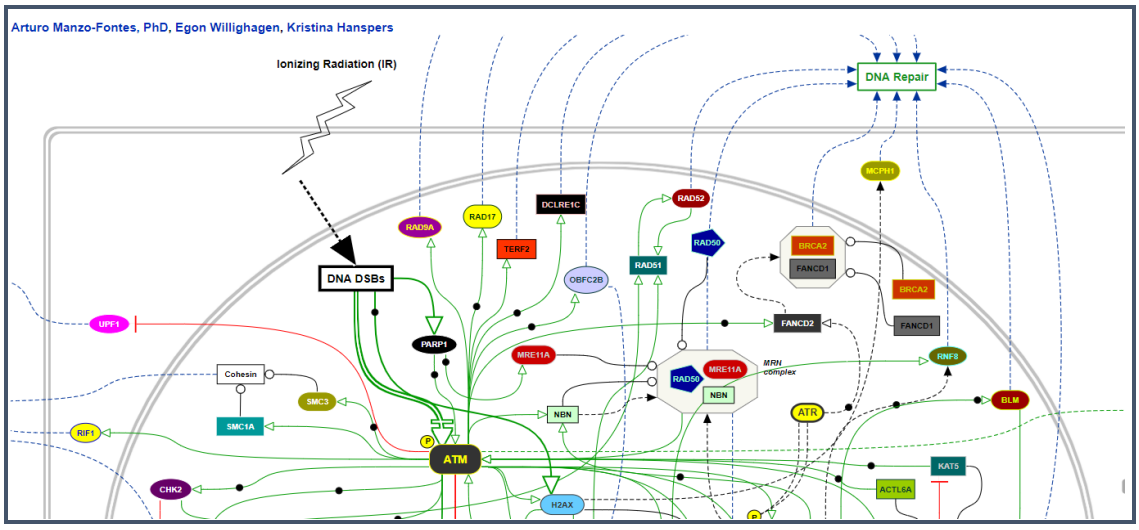


Figure 19. Wikipathway cancer (WP3959): DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM.

Ataxia-telangiectasia mutated (ATM) gene is critical in maintaining genomic integrity and plays a key role in the cellular DNA damage response. In response to DNA double-strand breaks, *ATM* phosphorylates downstream proteins involved in cell-cycle checkpoint arrest, DNA repair, and apoptosis [68]. Petersen et al. (2017) reported that *ATM* loss seems to be an early event in NSCLC carcinogenesis and is an independent prognostic factor associated with worse survival in stage II/III patients. Thus, according

to our analysis, *PARP1*, *FANCD2*, *RAD9A*, *NABP2* could be further experimentally evaluated for their prognostic role in NSCLC.

Meanwhile, most "classification signatures" include *COL5A2* (collagen type V alpha 2 chain) gene, which has already been reported to be involved in the pathological process of colorectal cancer, adenomas, breast cancer, osteosarcoma, bladder cancer etc [69], and was also found significant in five studies of NSCLC, according to IGDB.NSCLC Database [60]. In addition, WebGestalt analysis provided also enriched miRNAs ($p < 0.05$), such as MIR-363, MIR-503, and MIR-22 (Tables S2c-S8c), which were found significant in several NSCLS studies, according to IGDB.NSCLC database. These miRNAs could also be further experimentally evaluated for their roles in the development of NSCLC disease, considering that miRNAs targeting pathways provide potential candidates for therapeutic intervention against various pathological conditions.

To summarize, our analysis supports the notion that the DNA IR-Double Strand Breaks (DSBs) and cellular response via *ATM* seems to be a key signaling pathway in NSCLC. The related classification case 2.1 (see Table S7b), which involved training at dataset 1-GSE28827 (17 samples), testing at dataset 2-GSE75037 (61 samples), and 2 stages (I and III), can be viewed as the most significant "classification signature", that is also consistent with the statistical results. Moreover, our analysis provides indices for the emerging role of *COL5A2, PARP1*, *FANCD2*, *RAD9A*, *NABP2* in NSCLC, and the potential significance of MIR-363, MIR-503, and MIR-22 in NSCLC.

6. Discussion - Summary

Mutations of genes and cell proliferation are hallmarks of cancer. The mutations affect genes in ways that disturb the natural growth and death cycle of cells. These changes lead to abnormal and upregulated cell division forming the tumor. A gene mutation is an alteration in its DNA sequence. Thus, gene mutations cause changes in their expression profiles. We investigated the genes, whose level of expression profiles differ significantly between cancer and normal samples. A comprehensive analysis was conducted during *step A* using SAM and 2-fold change for genomic expression data mining. This analysis resulted to 2370 significant genes with differentiation ability between cancer and normal samples. The larger proportion of them (1540 genes) was positive significant, meaning that they have higher expression profiles in cancer population than in normal. The remaining 830 genes were negative significant with higher expression profiles in normal than in cancer state.

Screening tests are widely used in order to detect and characterize a disease, such as lung cancer. Radiomics deploy textural analysis as a tool for the evaluation of the tumor heterogeneity using medical images. It generates an amount of quantitative imaging features, which reflect the shape, size and texture of the tumor. Liu et al. [6] summarize many studies that have been conducted in the field of Radiomics for the detection, evaluation and prognosis of different type of cancers. In recent years, an extension of Radiomics, which is the Radiogenomics, aim to combine genomic and radiomic data in order to investigate possible correlations between them, leading to increase precision in diagnosis and assessment of cancer. [2] We examine the possible correlations between the 2370 significant genes and the CT radiomic features with two methods, the Spearman rank correlation test + FDR 5% and the quantitative SAM (q-value <= 0.05). The 78 common genes from the 2370 initial genes were identified to reveal statistically significant correlations with both methods. Hence, these 78 genes showed differentiation and diagnostic ability (*step A*) and simultaneously were highly correlated with radiomic features (*step B*). A two-step reduction procedure had been implemented in order to identify the number of genes that are more important in lung cancer.

The genomic data is publicly available from the GEO database, providing open access to the researchers. Thus, we used a new unseen dataset in order to assess further the significance of the genes that had been selected from our analysis on lung cancer diagnosis. The ability of the genes to classify the origin of a sample (cancer or normal) and produce homogeneous clusters with respect to the biological annotation was examined through *step C*. The results of the classification were promising, while the classifier achieved accuracy 92.05%, sensitivity 84.09% and specificity 100%. Furthermore, the BHI of the clustering of samples based on the gene values was 85.63% and of the clustering of genes was 89.29%, showing the potential to group together samples of the same biological class.

Once the significance of the genes had been validated with different approaches, the examination of the potential of the radiomic features was essential. Radiomics, in most cases, extracts a high-throughput amount of imaging features. To reduce the dimensionality of the imaging features and explore associations between them, we

perform K-means clustering algorithm on the radiomic features. Thus, 77 clusters of co-expressed imaging features were produced, forming the metafeatures. Simultaneously, the number of radiomic features was reduced, concluding in the most important ones for the description of the tumor characteristics.

Similarly to Gevaert et al. [14], we investigate the potential of the significant genes to predict the values of the CT radiomic features. Only the metafeatures that can be predicted adequately ($R^2 > 0.70$) and are highly correlated (*step B*) with the genes are considered as significant for further analysis, which are called as pMetafeatures. Thus, we have discovered genes with validated diagnostic character (*step A and C*), which were highly correlated with radiomic features (*step B*) and simultaneously have prediction ability of radiomic features (*step E*). Hence, more reliable and specific associations were produced between the genomic and the imaging data of lung cancer.

The identification of the lung cancer staging is crucial for the treatment planning. We explore the potential of the significant metafeatures and genes in cancer staging prediction. The classification results revealed that the use of both metafeatures and genes as predictors leads to higher accuracy of predicting the cancer stage. Furthermore, the linear combination of the significant genes can produce artificial imaging features that can determine the lung cancer staging of the patients, providing radiomic data in patients who have not undergone screening tests. Thus, there are indications that genes can replace the predictive ability of imaging features in cancer staging.
Furthermore, enrichment analysis reveal the functional processes related to cancer of the selected significant genes. Thus, genes are enriched with biological processes, pathways and miRNA targets.

An example of the combined analysis of genetic and imaging associations that had been derived from our analysis, is illustrated at Figure 20. More precisely, the gene RAD9A was enriched in the significant signaling pathway DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM, which participates in NSCLC carcinogenesis. This gene has shown diagnostic potential among cancer and normal samples through *step A* analysis. Furthermore, through *step B1 and step B2* it is correlated only with the *log_1_original_glcm_InverseVariance* radiomic feature, which constitutes an imaging feature of the Gray Level Co-occurrence Matrix category. This imaging feature belongs to Metafeature 80 (through *step C*), which is represented by the *log_1_original_glrlm_RunPercentage* imaging feature of the Gray Level Run Length Matrix. Both the GLCM and GLRLM describe the texture of the medical image. Simultaneously, the gene RAD9A participates in the prediction of metafeature 80 (through *step E*) and thus can predict the equivalent feature *log_1_original_glcm_InverseVariance*. This gene is involved also in the prediction of the metafeature 87, which is represented by the *log_1_original_glcm_JointEnergy* imaging feature of the GLCM category. Thus, gene RAD9A is correlated and can predict features that describe the texture of the tumor. Moreover, gene RAD9A and both metafeatures 80 and 87 participate in the prediction of lung cancer staging according to the case 2.1 (*step F*) which is the most significant "classification signature". Simultaneously, this "classification signature" has revealed the emerging role of gene RAD9A in NSCLC carcinogenesis. In conclusion, this example (Figure 20) shows the

statistical and predictive relationships between a gene with differentiation ability and some textural radiomic features as well as the involvement of their combination in prediction of lung cancer staging.

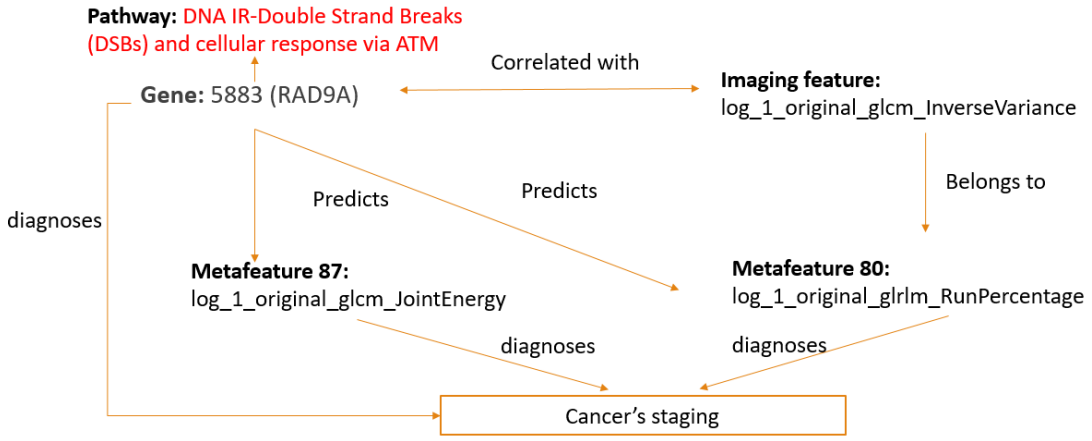


Figure 20. A graphical representation of the associations between the gene RAD9A and the radiomic features, which were derived from our analysis

However, this study has limitations. The main drawback is the small sample size of genomic and imaging data. Specifically, the only dataset that contains both gene expression profiles and radiomic features, is consisted of 24 patients, which is a quite enough small sample size. Moreover, we used cross validation techniques to validate the accuracy of lung cancer classification due to the small sample size. Datasets with larger samples are required in order to enhance the predictive ability of genes and imaging features on lung cancer staging. Furthermore, the absence of publicly available radiomic data restrict the research to investigate and validate further the significance of radiomic features and the potential to be characterized as imaging-based biomarkers. Additionally, PET images measure the metabolic processes in the human body and thus is considered more representative for revealing imaging and molecular associations than CT scans in lung cancer. Generally, the data from patients with cancer are sensitive data; thus, permission in most cases is not guaranteed. Additional datasets and higher-population datasets are needed to evaluate more precisely the diagnostic ability of genes and radiomic features in lung cancer staging.

7. Conclusions – Future Work

This thesis aims to investigate and perfume the combined analysis of gene expression profiles, which constitute the genotype, and CT radiomic features data, which constitute the phenotype, in order to contribute in the precise diagnosis of lung cancer staging. Specifically, we investigated statistically and predictive correlations between differentially expressed genes and radiomic features, using statistical tests and machine learning algorithms. Moreover, we identified the potential of linear combination of significant genes, which can estimate imaging features, to replace the predictive ability of actual radiomic features on lung cancer staging. The combination of these group of significant genes and specific individual important genes revealed the highest performance on cancer staging classification tests in our analysis. Additionally, we indicate that groups of significant genes can produce artificial radiomic features in patients with no radiomic data, showing similar behavior with the actual imaging features on the prediction of lung cancer staging. Finally, the enrichment analysis indicates several signaling and metabolism pathways, several miRNA targets and the emerging role of some genes in NSCLC.

Radiogenomics is an emerging field in which further investigation is needed to examine and validate the importance of the molecular and imaging data in order to achieve a precise diagnosis, prognosis and evaluation of NSCLC. Datasets with larger sample size is an essential prerequisite for radiomics and radiogenomics studies. Furthermore, datasets which contain genomic and imaging data are essential in order to provide adequate information about the genotype and the phenotype of lung cancer. These datasets could be used for further testing the classifiers of cancer staging and providing more accurate results. Simultaneously, predictive models of genes in terms of radiomic features could be examined in order to reveal imaging-based biomarkers, providing genetic non-invasive information in patients which had not undergone biopsy procedure. Moreover, PET images should be used for the extraction of radiomic features, as these images reflect effectively the metabolic processes; thus, more specific associations will be identified between the molecular data and the PET radiomic features. Additionally, the relationship of the imaging features and/or genes with the survival or other clinical data could be examined, when these information is provided. An emerging field of machine learning, the deep learning algorithms, could be used for automatic extraction of radiomic features from medical images in order to explore their ability to provide more informative imaging features. Finally, deep learning algorithms can be used for classification tasks in order to investigate their ability to lead to better classification scores.

References

[1] J. Didkowska, U. Wojciechowska, M. Mańczuk, and J. Lobaszewski, "Lung cancer epidemiology: Contemporary and future challenges worldwide," *Ann. Transl. Med.*, vol. 4, no. 8, pp. 1–11, 2016, doi: 10.21037/atm.2016.03.11.

[2] R. Lo Gullo, I. Daimiel, E. A. Morris, and K. Pinker, "Combining molecular and imaging metrics in cancer: radiogenomics," *Insights Imaging*, vol. 11, no. 1, pp. 1–17, 2020, doi: 10.1186/s13244-019-0795-6.

[3] J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, and P. Boffetta, "Risk factors for lung cancer worldwide," *Eur. Respir. J.*, vol. 48, no. 3, pp. 889–902, 2016, doi: 10.1183/13993003.00359-2016.

[4] H. A. Tindle *et al.*, "Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study," *J. Natl. Cancer Inst.*, vol. 110, no. 11, pp. 1201–1207, 2018, doi: 10.1093/jnci/djy041.

[5] R. Thawani *et al.*, "Radiomics and radiogenomics in lung cancer: A review for the clinician," *Lung Cancer*, vol. 115, 2017, pp. 34–41, 2018, doi: 10.1016/j.lungcan.2017.10.015.

[6] Z. Liu *et al.*, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges," *Theranostics*, vol. 9, no. 5, pp. 1303–1322, 2019, doi: 10.7150/thno.30309.

[7] Z. Bodalal, S. Trebeschi, T. D. L. Nguyen-Kim, W. Schats, and R. Beets-Tan, "Radiogenomics: bridging imaging and genomics," *Abdom. Radiol.*, vol. 44, no. 6, pp. 1960–1984, 2019, doi: 10.1007/s00261-019-02028-w.

[8] T. Wang, H. Sun, Y. Guo, and L. Zou, "18F-FDG PET/CT Quantitative Parameters and Texture Analysis Effectively Differentiate Endometrial Precancerous Lesion and Early-Stage Carcinoma," *Mol. Imaging*, vol. 18, no. 36, 2019, doi: 10.1177/1536012119856965.

[9] M. B. Andersen, S. W. Harders, B. Ganeshan, J. Thygesen, H. H. T. Madsen, and F. Rasmussen, "CT texture analysis can help differentiate between malignant and benign lymph nodes in the mediastinum in patients suspected for lung cancer," *Acta radiol.*, vol. 57, no. 6, pp. 669–676, 2016, doi: 10.1177/0284185115598808.

[10] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, 2017, doi: 10.1038/s41598-017-05848-2.

[11] J. E. Bibault *et al.*, "Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer," *Sci. Rep.*, vol. 8, no. 1, pp. 1–8, 2018, doi: 10.1038/s41598-018-30657-6.

[12] E. Trivizakis *et al.*, "Extending 2-D Convolutional Neural Networks to 3-D for Advancing Deep Learning Cancer Classification with Application to MRI Liver

Tumor Differentiation," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 3, pp. 923–930, 2019, doi: 10.1109/JBHI.2018.2886276.

[13] M. Zhou *et al.*, "Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications," *Radiology*, vol. 286, no. 1, pp. 307–315, 2018, doi: 10.1148/radiol.2017161845.

[14] O. Gevaert *et al.*, "Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data," *Radiology*, vol. 264, no. 2, pp. 387–396, 2012, doi: 10.1148/radiol.12111607/-/DC1.

[15] V. S. Nair *et al.*, "Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer," *Cancer Res.*, vol. 72, no. 15, pp. 3725–3734, 2012, doi: 10.1158/0008-5472.CAN-11-3943.

[16] X. Liao, B. Cai, B. Tian, Y. Luo, W. Song, and Y. Li, "Machine-learning based radiogenomics analysis of MRI features and metagenes in glioblastoma multiforme patients with different survival time," *J. Cell. Mol. Med.*, vol. 23, no. 6, pp. 4375–4385, 2019, doi: 10.1111/jcmm.14328.

[17] J. Wu, K. K. Tha, L. Xing, and R. Li, "Radiomics and radiogenomics for precision radiotherapy," *J. Radiat. Res.*, vol. 59, pp. i25–i31, 2018, doi: 10.1093/jrr/rrx102.

[18] W.-L. Cai and G.-B. Hong, "Quantitative image analysis for evaluation of tumor response in clinical oncology," *Chronic Dis. Transl. Med.*, vol. 4, no. 1, pp. 18–28, 2018, doi: 10.1016/j.cdtm.2018.01.002.

[19] P. Lambin *et al.*, "Radiomics: The bridge between medical imaging and personalized medicine," *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017, doi: 10.1038/nrclinonc.2017.141.

[20] T. Tirkes, M. A. Hollar, M. Tann, M. D. Kohli, F. Akisik, and K. Sandrasegaran, "Response criteria in oncologic imaging: Review of traditional and new criteria," *Radiographics*, vol. 33, no. 5, pp. 1323–1341, 2013, doi: 10.1148/rg.335125214.

[21] D. H. Xu, A. S. Kurani, J. D. Furst, and D. S. Raicu, "Run-length encoding for volumetric texture," *Proc. Fourth IASTED Int. Conf. Vis. Imaging, Image Process.*, no. January 2004, pp. 534–539, 2004.

[22] B. Ganeshan and K. A. Miles, "Quantifying tumour heterogeneity with CT," *Cancer Imaging*, vol. 13, no. 1, pp. 140–149, 2013, doi: 10.1102/1470-7330.2013.0015.

[23] S. Rizzo *et al.*, "Radiomics: the facts and the challenges of image analysis," *Eur. Radiol. Exp.*, vol. 2, no. 1, 2018, doi: 10.1186/s41747-018-0068-z.

[24] D. V. Nguyen, A. B. Arpat, N. Wang, and R. J. Carroll, "DNA microarray experiments: Biological and technological aspects," *Biometrics*, vol. 58, no. 4, pp. 701–717, 2002, doi: 10.1111/j.0006-341X.2002.00701.x.

[25] P. Sebastiani, E. Gussoni, I. S. Kohane, and M. F. Ramoni, “Statistical Challenges in Functional Genomics,” no. May 2014, 2003, doi: 10.1214/ss/1056397486.

[26] B. Ristevski, S. Loshkovska, S. Dzeroski, and I. Slavkov, “A comparison of validation indices for evaluation of clustering results of DNA microarray data,” *2nd Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2008*, pp. 587–591, 2008, doi: 10.1109/ICBBE.2008.143.

[27] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, 2004, doi: 10.1109/TKDE.2004.68.

[28] S.-B. Cho and H.-H. Won, “Machine learning in DNA microarray analysis for cancer classification,” *Proc. First Asia-Pacific Bioinforma. Conf. Bioinforma. 2003-Volume 19*, 2014, pp. 189–198, 2003.

[29] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 9, pp. 5116–5121, 2001, doi: 10.1073/pnas.091062498.

[30] J. D. Storey and R. Tibshirani, “SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays,” pp. 272–290, 2003, doi: 10.1007/0-387-21679-0_12.

[31] J. J. Chen, S. J. Wang, C. A. Tsai, and C. J. Lin, “Selection of differentially expressed genes in microarray data analysis,” *Pharmacogenomics J.*, vol. 7, no. 3, pp. 212–220, 2007, doi: 10.1038/sj.tpj.6500412.

[32] S.-D. Bolboaca and L. Jäntschi, “Pearson versus Spearman, Kendall’s tau correlation analysis on structure-activity relationships of biologic active compounds,” *Leonardo J. Sci.*, vol. 5, no. 9, pp. 179–200, 2006.

[33] M. Jafari and N. Ansari-Pour, “Why, when and how to adjust your P values?,” *Cell J.*, vol. 20, no. 4, pp. 604–607, 2019, doi: 10.22074/cellj.2019.5992.

[34] G. D. Ruxton and M. Neuhäuser, “When should we use one-tailed hypothesis testing?,” *Methods Ecol. Evol.*, vol. 1, no. 2, pp. 114–117, 2010, doi: 10.1111/j.2041-210x.2010.00014.x.

[35] C. Naugler and K. Lesack, “An open-source software program for performing Bonferroni and related corrections for multiple comparisons,” *J. Pathol. Inform.*, vol. 2, no. 1, p. 52, 2011, doi: 10.4103/2153-3539.91130.

[36] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”, vol 57, pp. 289-300, 1995.

[37] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu, “Gene expression data classification using Support Vector Machine and mutual information-based gene selection,” *Procedia Comput. Sci.*, vol. 47, no. C, pp. 13–21, 2014, doi: 10.1016/j.procs.2015.03.178.

[38] M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid-Ahmed, “Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers,”

pp. 1541–1546, 2008, doi: 10.1109/iccit.2007.148.

[39] M. Aly, “Survey on multiclass classification methods,” *Neural Netw*, pp. 1–9, 2005.

[40] T. S. Chen *et al.*, “A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray,” *Proc. 2005 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2005*, vol. 2005, pp. 405–408, 2005, doi: 10.1109/ispacs.2005.1595432.

[41] J. O. Ogutu, T. Schulz-Streeck, and H. P. Piepho, “Genomic selection using regularized linear regression models: ridge regression,” *BMC proceedings. BioMed Cent.*, vol. 6, no. 2, 2012.

[42] R. Tibshirani, "Regression shrinkage and selection via the lasso”, pp. 1–28, 1994.

[43] J. J. M. Van Griethuysen *et al.*, “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[44] G. Pinheiro *et al.*, “Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020, doi: 10.1038/s41598-020-60202-3.

[45] M. T. Damle and M. Kshirsagar, “Role of Permutations in Significance Analysis of Microarray and Clustering of Significant Microarray Gene list,” *Int. J. Comput. Sci. Issues*, vol. 9, no. 2, pp. 342–344, 2012.

[46] S. Datta and S. Datta, “Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes,” *BMC Bioinformatics*, vol. 7, pp. 1–9, 2006, doi: 10.1186/1471-2105-7-397.

[47] H. Wang and M. Song, “Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming,” vol. 3, no. 2, pp. 29–33, 2011.

[48] N. Bolshakova and F. Azuaje, “Cluster validation techniques for genome expression data,” *Signal Processing*, vol. 83, no. 4, pp. 825–833, 2003, doi: 10.1016/S0165-1684(02)00475-9.

[49] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[50] T. Caliński and J. Harabasz, “A Dendrite Method For Cluster Analysis,” *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[51] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene selection for cancer classification using Support Vector Machines”, *Machine Learning*, vol. 46, pp. 389-422, 2002.

[52] M. L. Huang, Y. H. Hung, W. M. Lee, R. K. Li, and B. R. Jiang, “SVM-RFE based feature selection and taguchi parameters optimization for multiclass SVM Classifier,” *Sci. World J.*, 2014, doi: 10.1155/2014/795624.

[53] W. Lim, C. A. Ridge, A. G. Nicholson, and S. Mirsadraee, “The 8th lung cancer TNM classification and clinical staging system: Review of the changes and clinical implications,” *Quant. Imaging Med. Surg.*, vol. 8, no. 7, pp. 709–718, 2018, doi: 10.21037/qims.2018.08.02.

[54] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, “WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W199–W205, 2019, doi: 10.1093/nar/gkz401.

[55] S. Carbon *et al.*, “The Gene Ontology Resource: 20 years and still GOing strong,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, 2019, doi: 10.1093/nar/gky1055.

[56] M. Kanehisa and Subramaniam, “The KEGG database,” *Novartis Found. Symp.*, vol. 247, pp. 91–103, 2002, doi: 10.1002/0470857897.ch8.

[57] H. Mi and P. Thomas, “PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools”, *Protein Networks and Pathways Analysis*, vol. 563,pp. 123-140, 2009, doi: 10.1007/978-1-60761-175-2.

[58] D. N. Slenter *et al.*, “WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D661–D667, 2018, doi: 10.1093/nar/gkx1064.

[59] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011, doi: 10.1093/bioinformatics/btr260.

[60] S. Kao *et al.*, “IGDB.NSCLC: Integrated genomic database of non-small cell lung cancer,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. 972–977, 2012, doi: 10.1093/nar/gkr1183.

[61] http://geneontology.org/docs/ontology-documentation/ . [Accessed: 14-Aug-2020].

[62] http://geneontology.org/. [Accessed: 14-Aug-2020].

[63] https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet . [Accessed: 14-Aug-2020].

[64] B. Liu, J. Li, and M. J. Cairns, “Identifying miRNAs, targets and functions,” *Brief. Bioinform.*, vol. 15, no. 1, pp. 1–19, 2014, doi: 10.1093/bib/bbs075.

[65] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.,” *Nucleic Acids Res.*, vol. 41, pp. 77–83, 2013, doi: 10.1093/nar/gkt439.

[66] http://www.webgestalt.org/. [Accessed: 14-Aug-2020].

[67] https://www.wikipathways.org/index.php/Pathway:WP3959. [Accessed: 14-Aug-2020].

[68] L. F. Petersen *et al.*, “Loss of tumour-specific ATM protein expression is an independent prognostic factor in early resected NSCLC,” *Oncotarget*, vol. 8, no.

24, pp. 38326–38336, 2017, doi: 10.18632/oncotarget.16215.

[69] X. T. Zeng, X. P. Liu, T. Z. Liu, and X. H. Wang, “The clinical significance of COL5A2 in patients with bladder cancer: A retrospective analysis of bladder cancer gene expression data,” *Med. (United States)*, vol. 97, no. 10, pp. 10–13, 2018, doi: 10.1097/MD.0000000000010091.

Appendix

Biological Evaluation - Supporting (S) Tables

S1. Gene Annotation - 78 common Differential Expressed Genes (cDEGs)
S1A. Enriched Gene Ontology (GO) biological processes (BP)
S1B. Enriched Pathways
S1C. Enriched miRNA targets

S2. Gene Annotation - Classification Case 1.1 (47 Genes)
S2A. Enriched Gene Ontology (GO) biological processes (BP)
S2B. Enriched Pathways
S2C. Enriched miRNA targets

S3. Gene Annotation - Classification Case 1.2 (34 Genes)
S3A. Enriched Gene Ontology (GO) biological processes (BP)
S3B. Enriched Pathways
S3C. Enriched miRNA targets

S4. Gene Annotation - Classification Case 1.3 (61 Genes)
S4A. Enriched Gene Ontology (GO) biological processes (BP)
S4B. Enriched Pathways
S4C. Enriched miRNA targets

S5. Gene Annotation - Classification Case 1.4 (71 Genes)
S5A. Enriched Gene Ontology (GO) biological processes (BP)
S5B. Enriched Pathways
S5C. Enriched miRNA targets

S6. Gene Annotation - Classification Case 1.5 (58 Genes)
S6A. Enriched Gene Ontology (GO) biological processes (BP)
S6B. Enriched Pathways
S6C. Enriched miRNA targets

S7. Gene Annotation - Classification Case 2.1 (68 Genes)
S7A. Enriched Gene Ontology (GO) biological processes (BP)
S7B. Enriched Pathways
S7C. Enriched miRNA targets

S8. Gene Annotation - Classification Case 2.2 (70 Genes)
S8A. Enriched Gene Ontology (GO) biological processes (BP)
S8B. Enriched Pathways
S8C. Enriched miRNA targets

S1. Gene Annotation - 78 common Differential Expressed Genes (cDEGs)

78 common Differential Expressed Genes (cDEGs)
Intersection of Spearman & FDR across imaging features and SAM

Entrez Gene ID	Gene Symbol	Gene Name	Entrez Gene ID	Gene Symbol	Gene Name
10882	C1QL1	complement component 1, q subcomponent-like 1	54825	CDHR2	cadherin-related family member 2
5883	RAD9A	RAD9 homolog A (S. pombe)	83933	HDAC10	histone deacetylase 10
443	ASPA	aspartoacylase	196410	METTL7B	methyltransferase like 7B
284185	LINC00482	long intergenic non-protein coding RNA 482	645037	GAGE2B	G antigen 2B
26232	FBXO2	F-box protein 2	8347	HIST1H2BC	histone cluster 1, H2bc
2177	FANCD2	Fanconi anemia, complementation group D2	10615	SPAG5	sperm associated antigen 5
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	3603	IL16	interleukin 16
1261	CNGA3	cyclic nucleotide gated channel alpha 3	5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
2859	GPR35	G protein-coupled receptor 35	10129	FRY	furry homolog (Drosophila)
3866	KRT15	keratin 15	2524	FUT2	fucosyltransferase 2 (secretor status included)
114907	FBXO32	F-box protein 32	2027	ENO3	enolase 3 (beta, muscle)
2705	GJB1	gap junction protein, beta 1, 32kDa	153768	PRELID2	PRELI domain containing 2
4585	MUC4	mucin 4, cell surface associated	27094	KCNMB3	potassium large conductance calcium-activated channel, subfamily M beta member 3
63035	BCORL1	BCL6 corepressor-like 1	79035	NABP2	nucleic acid binding protein 2
55277	FGGY	FGGY carbohydrate kinase domain containing	23534	TNPO3	transportin 3
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)
1741	DLG3	discs, large homolog 3 (Drosophila)	23090	ZNF423	zinc finger protein 423
2118	ETV4	ets variant 4	25894	PLEKHG4	pleckstrin homology domain containing, family G (with RhoGef domain) member 4
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3	3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
55311	ZNF444	zinc finger protein 444	79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
7125	TNNC2	troponin C type 2 (fast)	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
142	PARP1	poly (ADP-ribose) polymerase 1	7104	TM4SF4	transmembrane 4 L six family member 4
1290	COL5A2	collagen, type V, alpha 2	347853	TBX10	T-box 10
1747	DLX3	distal-less homeobox 3	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
8612	PPAP2C	phosphatidic acid phosphatase type 2C	7477	WNT7B	wingless-type MMTV integration site family, member 7B
80201	HKDC1	hexokinase domain containing 1	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
65260	SELRC1	Sel1 repeat containing 1	23414	ZFPM2	zinc finger protein, multitype 2
54993	ZSCAN2	zinc finger and SCAN domain containing 2	3026	HABP2	hyaluronan binding protein 2
9244	CRLF1	cytokine receptor-like factor 1	127845	GOLT1A	golgi transport 1A
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1	220134	SKA1	spindle and kinetochore associated complex subunit 1
4796	TONSL	tonsoku-like, DNA repair protein	3853	KRT6A	keratin 6A
6289	SAA2	serum amyloid A2	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
84300	MNF1	mitochondrial nucleoid factor 1	25934	NIPSNAP3A	nipsnap homolog 3A (C. elegans)
340706	VWA2	von Willebrand factor A domain containing 2	26112	CCDC69	coiled-coil domain containing 69
170487	ACTL10	actin-like 10	78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
10045	SH2D3A	SH2 domain containing 3A	51481	VCX3A	variable charge, X-linked 3A
729238	SFTPA2	surfactant protein A2	629	CFB	complement factor B
638	BIK	BCL2-interacting killer (apoptosis-inducing)	6878	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
202374	STK32A	serine/threonine kinase 32A	23780	APOL2	apolipoprotein L, 2

Table S1: 78 common differentially expressed genes (cDEGs) as extracted from the intersection of the 1st approach (Spearman+FDR across imaging features), and the 2nd approach (SAM). The 73 genes highlighted in blue are used as input in the next steps of the proposed methodology. The genes are described by their gene symbols and gene names using WebGestalt 2013.

S1A. Enriched Gene Ontology (GO) biological processes (BP)

78 common Differential Expressed Genes (cDEGs)
Intersection of Spearman & FDR across imaging features and SAM

A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)

Gene Set	Description	P Value	FDR
GO:0002526	acute inflammatory response	0.0039297	1
GO:0009141	nucleoside triphosphate metabolic process	0.0087898	1
GO:0009123	nucleoside monophosphate metabolic process	0.010953	1
GO:0007586	digestion	0.017864	1
GO:0043954	cellular component maintenance	0.022214	1
GO:0016052	carbohydrate catabolic process	0.027134	1
GO:0060249	anatomical structure homeostasis	0.027795	1
GO:0050918	positive chemotaxis	0.029445	1
GO:0005996	monosaccharide metabolic process	0.029654	1
GO:0090305	nucleic acid phosphodiester bond hydrolysis	0.032409	1
GO:0061458	reproductive system development	0.033196	1
GO:0009314	response to radiation	0.037651	1
GO:0001570	vasculogenesis	0.039365	1
GO:0007059	chromosome segregation	0.041553	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.041842	1

Table S1: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 78 cDEGs as extracted from the intersection of the 1st approach (Spearman+FDR across imaging features), and the 2nd approach (SAM). The enrichment analysis was performed by WebGestalt 2019.

S1B. Enriched Pathways

78 cDEGs - Intersection of Spearman & FDR across imaging features and SAM			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.012163	1
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.023218	1
hsa03460	Fanconi anemia pathway	0.026215	1
hsa00740	Riboflavin metabolism	0.036896	1
hsa00010	Glycolysis / Gluconeogenesis	0.04007	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.02794	1
P02744	Fructose galactose metabolism	0.03824	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	**DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM**	**0.00017467**	**0.01345**
WP4016	DNA IR-damage and cellular response via ATR	0.0094637	0.36435
WP2516	ATM Signaling Pathway	0.020857	0.53533
WP3875	ATR Signaling	0.050732	0.88186

Table S1: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 78 cDEGs as extracted from the intersection of the 1st approach (Spearman+FDR across imaging features), and the 2nd approach (SAM). The enrichment analysis was performed by WebGestalt 2019.

S1C. Enriched miRNA targets

78 DEGs - Intersection of Spearman & FDR across imaging features and SAM		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.010415	1
GTGACTT,MIR-224	0.012003	1
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.019426	1
ACCGAGC,MIR-423	0.024496	1
ATTACAT,MIR-380-3P	0.039115	1
GGGCATT,MIR-365	0.04339	1
ACTGCAG,MIR-17-3P	0.044852	1

Table S1: c) miRNA target annotation (MSigDB) of 78 cDEGs as extracted from the intersection of the 1st approach (Spearman+FDR across imaging features), and the 2nd approach (SAM). The enrichment analysis was performed by WebGestalt 2019.

S2. Gene Annotation - Classification Case 1.1

Classification Case 1.1 - 47 Genes		
Entrez Gene ID	**Gene Symbol**	**Gene Name**
10882	C1QL1	complement component 1, q subcomponent-like 1
443	ASPA	aspartoacylase
284185	LINC00482	long intergenic non-protein coding RNA 482
26232	FBXO2	F-box protein 2
2177	FANCD2	Fanconi anemia, complementation group D2
11142	PKIG	inhibitor gamma
54825	CDHR2	cadherin-related family member 2
83933	HDAC10	histone deacetylase 10
196410	METTL7B	methyltransferase like 7B
8347	HIST1H2BC	histone cluster 1, H2bc
10615	SPAG5	sperm associated antigen 5
2705	GJB1	gap junction protein, beta 1, 32kDa
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
10129	FRY	furry homolog (Drosophila)
2524	FUT2	fucosyltransferase 2 (secretor status included)
4585	MUC4	mucin 4, cell surface associated
63035	BCORL1	BCL6 corepressor-like 1
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
153768	PRELID2	PRELI domain containing 2
2027	ENO3	enolase 3 (beta, muscle)
23534	TNPO3	transportin 3
699	BUB1	budding uninhibited by benzimidazoles 1 homolog
23090	ZNF423	zinc finger protein 423
2118	ETV4	ets variant 4
25894	PLEKHG4	pleckstrin homology domain containing, family G (with RhoGef domain) member 4
5169	ENPP3	pyrophosphatase/phosphodiesterase 3
55311	ZNF444	zinc finger protein 444
79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
1290	COL5A2	collagen, type V, alpha 2
1747	DLX3	distal-less homeobox 3
7104	TM4SF4	transmembrane 4 L six family member 4
347853	TBX10	T-box 10
9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
80201	HKDC1	hexokinase domain containing 1
7477	WNT7B	wingless-type MMTV integration site family, member 7B
54993	ZSCAN2	zinc finger and SCAN domain containing 2
9244	CRLF1	cytokine receptor-like factor 1
23414	ZFPM2	zinc finger protein, multitype 2
3026	HABP2	hyaluronan binding protein 2
3853	KRT6A	keratin 6A
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
26112	CCDC69	coiled-coil domain containing 69
84300	MNF1	mitochondrial nucleoid factor 1
23780	APOL2	apolipoprotein L, 2

Table S2: Based on the 73 gene list (S1), 47 genes were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 4 stages (0, I, II, III). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S2A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 1.1 - 47 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.0039869	1
GO:0009141	nucleoside triphosphate metabolic process	0.0061138	1
GO:0016052	carbohydrate catabolic process	0.0062348	1
GO:0009123	nucleoside monophosphate metabolic process	0.0073894	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.0099717	1
GO:0046434	organophosphate catabolic process	0.013342	1
GO:0001570	vasculogenesis	0.014266	1
GO:0009100	glycoprotein metabolic process	0.01435	1
GO:0061458	reproductive system development	0.01953	1
GO:0046939	nucleotide phosphorylation	0.022709	1
GO:0070085	glycosylation	0.024202	1
GO:0015748	organophosphate ester transport	0.02496	1
GO:0005996	monosaccharide metabolic process	0.030167	1
GO:0098732	macromolecule deacylation	0.030217	1
GO:0006090	pyruvate metabolic process	0.031721	1
GO:0009259	ribonucleotide metabolic process	0.032227	1
GO:0009132	nucleoside diphosphate metabolic process	0.03274	1
GO:0007059	chromosome segregation	0.039646	1
GO:0001655	urogenital system development	0.04422	1
GO:0072524	pyridine-containing compound metabolic process	0.053148	1
GO:0002526	acute inflammatory response	0.053767	1

Table S2: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 47 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 4 stages (0, I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S2B. Enriched Pathways

Classification Case 1.1 - 47 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.0032924	0.98659
hsa03460	Fanconi anemia pathway	0.0072892	0.98659
hsa00010	Glycolysis / Gluconeogenesis	0.011375	0.98659
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.011995	0.98659
hsa00740	Riboflavin metabolism	0.019127	1
hsa04066	HIF-1 signaling pathway	0.023619	1
hsa04916	Melanogenesis	0.024062	1
hsa01100	Metabolic pathways	0.026151	1
hsa01200	Carbon metabolism	0.031113	1
hsa00603	Glycosphingolipid biosynthesis	0.035579	1
hsa00770	Pantothenate and CoA biosynthesis	0.044862	1
hsa00340	Histidine metabolism	0.054062	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.014056	1
P02744	Fructose galactose metabolism	0.019289	1
P00012	Cadherin signaling pathway	0.024829	1
P00024	Glycolysis	0.029692	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP4018	Pathways in clear cell renal cell carcinoma	0.037065	1

Table S2: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 47 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 4 stages (0, I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S2C. Enriched miRNA targets

Classification Case 1.1 - 47 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0036484	0.31145
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.0037866	0.31145
GTGACTT,MIR-224	0.0042279	0.31145
ACCGAGC,MIR-423	0.017097	0.71805
ATTACAT,MIR-380-3P	0.019738	0.71805
GGGCATT,MIR-365	0.021975	0.71805
ACTGCAG,MIR-17-3P	0.022744	0.71805
GTGCAAT,MIR-25,MIR-32,MIR-92,MIR-363,MIR-367	0.026861	0.74203
CGCTGCT,MIR-503	0.050472	1

Table S2: c) miRNA target annotation (MSigDB) of 47 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 4 stages (0, I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S3. Gene Annotation - Classification Case 1.2

Classification Case 1.2 - 34 Genes		
Entrez Gene ID	**Gene Symbol**	**Gene Name**
284185	LINC00482	long intergenic non-protein coding RNA 482
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma
54825	CDHR2	cadherin-related family member 2
1261	CNGA3	cyclic nucleotide gated channel alpha 3
196410	METTL7B	methyltransferase like 7B
8347	HIST1H2BC	histone cluster 1, H2bc
10615	SPAG5	sperm associated antigen 5
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
4585	MUC4	mucin 4, cell surface associated
63035	BCORL1	BCL6 corepressor-like 1
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
23534	TNPO3	transportin 3
699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)
23090	ZNF423	zinc finger protein 423
25894	PLEKHG4	pleckstrin homology domain containing, family G (with RhoGef domain) member 4
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3
7125	TNNC2	troponin C type 2 (fast)
79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
142	PARP1	poly (ADP-ribose) polymerase 1
7104	TM4SF4	transmembrane 4 L six family member 4
9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
6690	SPINK1	serine peptidase inhibitor, Kazal type 1
7477	WNT7B	member 7B
23414	ZFPM2	zinc finger protein, multitype 2
54993	ZSCAN2	zinc finger and SCAN domain containing 2
220134	SKA1	spindle and kinetochore associated complex
3026	HABP2	hyaluronan binding protein 2
3853	KRT6A	keratin 6A
5980	REV3L	catalytic subunit
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
26112	CCDC69	coiled-coil domain containing 69
78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
6878	TAF6	protein (TBP)-associated factor, 80kDa

Table S3: Based on the 73 gene list (S1), 34 genes were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 3 stages (I, II, III). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S3A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 1.2 - 34 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.0015491	1
GO:0001570	vasculogenesis	0.007603	1
GO:0034504	protein localization to nucleus	0.01109	1
GO:0090305	nucleic acid phosphodiester bond hydrolysis	0.013595	1
GO:0007059	chromosome segregation	0.016826	1
GO:2000241	regulation of reproductive process	0.025196	1
GO:0048545	response to steroid hormone	0.0297	1
GO:0023019	signal transduction involved in regulation of gene expression	0.034556	1
GO:0030323	respiratory tube development	0.036321	1
GO:0003012	muscle system process	0.036989	1
GO:0007051	spindle organization	0.037483	1
GO:0019932	second-messenger-mediated signaling	0.039227	1
GO:0060541	respiratory system development	0.04475	1
GO:0051051	negative regulation of transport	0.045124	1
GO:0070528	protein kinase C signaling	0.048054	1
GO:0048871	multicellular organismal homeostasis	0.048105	1
GO:0032886	regulation of microtubule-based process	0.049429	1
GO:0000002	mitochondrial genome maintenance	0.049729	1
GO:0032528	microvillus organization	0.051401	1
GO:0017038	protein import	0.052497	1
GO:0010737	protein kinase A signaling	0.053069	1
GO:0002251	organ or tissue specific immune response	0.054735	1

Table S3: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 34 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 3 stages (I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S3B. Enriched Pathways

Classification Case 1.2 - 34 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa04916	Melanogenesis	0.012816	1
hsa00740	Riboflavin metabolism	0.013846	1
hsa00770	Pantothenate and CoA biosynthesis	0.032596	1
hsa04020	Calcium signaling pathway	0.039003	1
hsa04024	cAMP signaling pathway	0.045425	1
hsa00760	Nicotinate and nicotinamide metabolism	0.051016	1
hsa00512	Mucin type O-glycan biosynthesis	0.052674	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P00012	Cadherin signaling pathway	0.039562	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP4240	Regulation of sister chromatid separation at the metaphase-anaphase transition	0.04121	1

Table S3: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 34 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 3 stages (I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S3C. Enriched miRNA targets

Classification Case 1.2 - 34 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
GTGACTT,MIR-224	0.0013462	0.29751
ATTACAT,MIR-380-3P	0.0094586	0.65095
ACTGCAG,MIR-17-3P	0.010937	0.65095
ACCGAGC,MIR-423	0.011782	0.65095
TCATCTC,MIR-143	0.019493	0.86158
CGCTGCT,MIR-503	0.034968	0.9504
GTATTAT,MIR-369-3P	0.035969	0.9504
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.036974	0.9504
GGCAGCT,MIR-22	0.044303	0.9504
AACTGGA,MIR-145	0.044999	0.9504
ATGTTAA,MIR-302C	0.048184	0.9504

Table S3: c) miRNA target annotation (MSigDB) of 34 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 3 stages (I, II, III). The enrichment analysis was performed by WebGestalt 2019.

S4. Gene Annotation - Classification Case 1.3

Classification Case 1.3 - 61 Genes					
Entrez Gene ID	**Gene Symbol**	**Gene Name**	**Entrez Gene ID**	**Gene Symbol**	**Gene Name**
170487	ACTL10	actin-like 10	25894	PLEKHG4	family G (with RhoGef domain) member
10882	C1QL1	complement component 1, q subcomponent-like 1	5169	ENPP3	pyrophosphatase/phosphodiesterase 3
443	ASPA	aspartoacylase	3161	HMMR	(RHAMM)
284185	LINC00482	long intergenic non-protein coding RNA 482	79674	VEPH1	homolog 1 (zebrafish)
638	BIK	BCL2-interacting killer (apoptosis-inducing)	7125	TNNC2	troponin C type 2 (fast)
26232	FBXO2	F-box protein 2	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
2177	FANCD2	Fanconi anemia, complementation group D2	1290	COL5A2	collagen, type V, alpha 2
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	7104	TM4SF4	transmembrane 4 L six family member 4
54825	CDHR2	cadherin-related family member 2	347853	TBX10	T-box 10
83933	HDAC10	histone deacetylase 10	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
1261	CNGA3	cyclic nucleotide gated channel alpha 3	80201	HKDC1	hexokinase domain containing 1
196410	METTL7B	methyltransferase like 7B	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
8347	HIST1H2BC	histone cluster 1, H2bc	7477	WNT7B	wingless-type MMTV integration site family, member 7B
10615	SPAG5	sperm associated antigen 5	65260	SELRC1	Sel1 repeat containing 1
2859	GPR35	G protein-coupled receptor 35	54993	ZSCAN2	zinc finger and SCAN domain containing 2
114907	FBXO32	F-box protein 32	23414	ZFPM2	zinc finger protein, multitype 2
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)	9244	CRLF1	cytokine receptor-like factor 1
2705	GJB1	gap junction protein, beta 1, 32kDa	220134	SKA1	spindle and kinetochore associated complex subunit 1
2524	FUT2	fucosyltransferase 2 (secretor status included)	127845	GOLT1A	golgi transport 1A
4585	MUC4	mucin 4, cell surface associated	3026	HABP2	hyaluronan binding protein 2
63035	BCORL1	BCL6 corepressor-like 1	3853	KRT6A	keratin 6A
55277	FGGY	FGGY carbohydrate kinase domain containing	116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
153768	PRELID2	PRELI domain containing 2	4796	TONSL	tonsoku-like, DNA repair protein
2027	ENO3	enolase 3 (beta, muscle)	25934	NIPSNAP3A	nipsnap homolog 3A (C. elegans)
79035	NABP2	nucleic acid binding protein 2	78990	OTUB2	binding 2
27094	KCNMB3	channel, subfamily M beta member 3	26112	CCDC69	coiled-coil domain containing 69
23534	TNPO3	transportin 3	84300	MNF1	mitochondrial nucleoid factor 1
699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)	6878	TAF6	binding protein (TBP)-associated factor, 80kDa
23090	ZNF423	zinc finger protein 423	23780	APOL2	apolipoprotein L, 2
1741	DLG3	discs, large homolog 3 (Drosophila)			

Table S4: Based on the 73 gene list (S1), 61 genes were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 3 stages (0, I, and 'II' (a combination of II and III)). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S4A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 1.3 - 61 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.0094598	1
GO:0005996	monosaccharide metabolic process	0.013606	1
GO:0043954	cellular component maintenance	0.014245	1
GO:0016052	carbohydrate catabolic process	0.014578	1
GO:0009141	nucleoside triphosphate metabolic process	0.01788	1
GO:0007059	chromosome segregation	0.019472	1
GO:0009123	nucleoside monophosphate metabolic process	0.021365	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.02289	1
GO:0001570	vasculogenesis	0.02553	1
GO:0046434	organophosphate catabolic process	0.030226	1
GO:0016073	snRNA metabolic process	0.039404	1
GO:0009100	glycoprotein metabolic process	0.039624	1
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	0.040159	1
GO:0046939	nucleotide phosphorylation	0.040159	1
GO:0015748	organophosphate ester transport	0.044016	1
GO:0098781	ncRNA transcription	0.046394	1
GO:0061458	reproductive system development	0.05259	1
GO:0070085	glycosylation	0.053078	1
GO:0044282	small molecule catabolic process	0.05333	1
GO:0006090	pyruvate metabolic process	0.055503	1

Table S4: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 61 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 3 stages (0, I, and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S4B. Enriched Pathways

Classification Case 1.3 - 61 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.0073505	1
hsa03460	Fanconi anemia pathway	0.016043	1
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.017949	1
hsa00010	Glycolysis / Gluconeogenesis	0.02476	1
hsa00740	Riboflavin metabolism	0.028569	1
hsa04066	HIF-1 signaling pathway	0.050146	1
hsa04916	Melanogenesis	0.051047	1
hsa00603	Glycosphingolipid biosynthesis	0.052922	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.024485	1
P02744	Fructose galactose metabolism	0.033533	1
P00024	Glycolysis	0.051415	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM	0.025722	1

Table S4: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 61 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 3 stages (0, I, and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S4C. Enriched miRNA targets

Classification Case 1.3 - 61 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0036484	0.46718
GTGACTT,MIR-224	0.0042279	0.46718
ACCGAGC,MIR-423	0.017097	0.83773
ATTACAT,MIR-380-3P	0.019738	0.83773
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.022056	0.83773
ACTGCAG,MIR-17-3P	0.022744	0.83773
GTGCAAA,MIR-507	0.031469	0.99352
CGCTGCT,MIR-503	0.050472	1

Table S4: c) miRNA target annotation (MSigDB) of 61 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 24 samples), and 3 stages (0, I, and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S5. Gene Annotation - Classification Case 1.4

Classification Case 1.4 - 71 Genes					
Entrez Gene ID	**Gene Symbol**	**Gene Name**	**Entrez Gene ID**	**Gene Symbol**	**Gene Name**
10882	C1QL1	complement component 1, q subcomponent-like 1	196410	METTL7B	methyltransferase like 7B
443	ASPA	aspartoacylase	8347	HIST1H2BC	histone cluster 1, H2bc
284185	LINC00482	long intergenic non-protein coding RNA 482	10615	SPAG5	sperm associated antigen 5
26232	FBXO2	F-box protein 2	3603	IL16	interleukin 16
2177	FANCD2	Fanconi anemia, complementation group D2	5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	2524	FUT2	fucosyltransferase 2 (secretor status included)
1261	CNGA3	cyclic nucleotide gated channel alpha 3	10129	FRY	furry homolog (Drosophila)
2859	GPR35	G protein-coupled receptor 35	153768	PRELID2	PRELI domain containing 2
3866	KRT15	keratin 15	2027	ENO3	enolase 3 (beta, muscle)
114907	FBXO32	F-box protein 32	27094	KCNMB3	potassium large conductance calcium-activated channel, subfamily M beta
2705	GJB1	gap junction protein, beta 1, 32kDa	79035	NABP2	nucleic acid binding protein 2
4585	MUC4	mucin 4, cell surface associated	23534	TNPO3	transportin 3
63035	BCORL1	BCL6 corepressor-like 1	699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)
55277	FGGY	FGGY carbohydrate kinase domain containing	23090	ZNF423	zinc finger protein 423
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	25894	PLEKHG4	family G (with RhoGef domain) member 4
1741	DLG3	discs, large homolog 3 (Drosophila)	3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
2118	ETV4	ets variant 4	79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
55311	ZNF444	zinc finger protein 444	7104	TM4SF4	transmembrane 4 L six family member 4
7125	TNNC2	troponin C type 2 (fast)	347853	TBX10	T-box 10
1290	COL5A2	collagen, type V, alpha 2	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
1747	DLX3	distal-less homeobox 3	7477	WNT7B	wingless-type MMTV integration site family, member 7B
8612	PPAP2C	phosphatidic acid phosphatase type 2C	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
80201	HKDC1	hexokinase domain containing 1	23414	ZFPM2	zinc finger protein, multitype 2
65260	SELRC1	Sel1 repeat containing 1	220134	SKA1	spindle and kinetochore associated complex subunit 1
54993	ZSCAN2	zinc finger and SCAN domain containing 2	3026	HABP2	hyaluronan binding protein 2
9244	CRLF1	cytokine receptor-like factor 1	127845	GOLT1A	golgi transport 1A
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1	3853	KRT6A	keratin 6A
4796	TONSL	tonsoku-like, DNA repair protein	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
84300	MNF1	mitochondrial nucleoid factor 1	25934	NIPSNAP3A	nipsnap homolog 3A (C. elegans)
170487	ACTL10	actin-like 10	26112	CCDC69	coiled-coil domain containing 69
10045	SH2D3A	SH2 domain containing 3A	78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
202374	STK32A	serine/threonine kinase 32A	629	CFB	complement factor B
638	BIK	BCL2-interacting killer (apoptosis-inducing)	6878	TAF6	binding protein (TBP)-associated factor, 80kDa
54825	CDHR2	cadherin-related family member 2	23780	APOL2	apolipoprotein L, 2
83933	HDAC10	histone deacetylase 10			

Table S5: Based on the 73 gene list (S1), 71 genes were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 2 stages (I and 'II' (a combination of II and III)). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S5A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 1.4 - 71 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.014271	1
GO:0043954	cellular component maintenance	0.018971	1
GO:0002526	acute inflammatory response	0.021432	1
GO:0016052	carbohydrate catabolic process	0.021796	1
GO:0005996	monosaccharide metabolic process	0.022577	1
GO:0061458	reproductive system development	0.024112	1
GO:0009141	nucleoside triphosphate metabolic process	0.02938	1
GO:0007059	chromosome segregation	0.031891	1
GO:0001570	vasculogenesis	0.033764	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.033845	1
GO:0009123	nucleoside monophosphate metabolic process	0.034864	1
GO:0046434	organophosphate catabolic process	0.044334	1
GO:0050727	regulation of inflammatory response	0.049985	1
GO:0016073	snRNA metabolic process	0.051755	1
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	0.052728	1
GO:0046939	nucleotide phosphorylation	0.052728	1

Table S5: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 71 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 2 stages (I and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S5B. Enriched Pathways

Classification Case 1.4 - 71 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.010231	1
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.021245	1
hsa03460	Fanconi anemia pathway	0.022156	1
hsa00740	Riboflavin metabolism	0.033781	1
hsa00010	Glycolysis / Gluconeogenesis	0.033988	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.024485	1
P02744	Fructose galactose metabolism	0.033533	1
P00024	Glycolysis	0.051415	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM	0.025722	1

Table S5: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 71 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 2 stages (I and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S5C. Enriched miRNA targets

Classification Case 1.4 - 71 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0080534	0.9714
GTGACTT,MIR-224	0.0092955	0.9714
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.013186	0.9714
ACCGAGC,MIR-423	0.022387	1
ATTACAT,MIR-380-3P	0.033045	1
GGGCATT,MIR-365	0.036694	1
ACTGCAG,MIR-17-3P	0.037945	1
GTGCAAA,MIR-507	0.052025	1
GTGCAAT,MIR-25,MIR-32,MIR-92,MIR-363,MIR-367	0.054785	1

Table S5: c) miRNA target annotation (MSigDB) of 71 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 20 samples), and 2 stages (I and 'II' (a combination of II and III)). The enrichment analysis was performed by WebGestalt 2019.

S6. Gene Annotation - Classification Case 1.5

Classification Case 1.5 - 58 Genes					
Entrez Gene ID	**Gene Symbol**	**Gene Name**	**Entrez Gene ID**	**Gene Symbol**	**Gene Name**
10882	C1QL1	complement component 1, q subcomponent-like 1	5169	ENPP3	pyrophosphatase/phosphodiesterase 3
5883	RAD9A	RAD9 homolog A (S. pombe)	55311	ZNF444	zinc finger protein 444
443	ASPA	aspartoacylase	3161	HMMR	(RHAMM)
284185	LINC00482	long intergenic non-protein coding RNA 482	79674	VEPH1	homolog 1 (zebrafish)
638	BIK	BCL2-interacting killer (apoptosis-inducing)	7125	TNNC2	troponin C type 2 (fast)
26232	FBXO2	F-box protein 2	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	142	PARP1	poly (ADP-ribose) polymerase 1
54825	CDHR2	cadherin-related family member 2	1290	COL5A2	collagen, type V, alpha 2
83933	HDAC10	histone deacetylase 10	1747	DLX3	distal-less homeobox 3
1261	CNGA3	cyclic nucleotide gated channel alpha 3	7104	TM4SF4	transmembrane 4 L six family member 4
196410	METTL7B	methyltransferase like 7B	347853	TBX10	T-box 10
8347	HIST1H2BC	histone cluster 1, H2bc	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
10615	SPAG5	sperm associated antigen 5	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
114907	FBXO32	F-box protein 32	7477	WNT7B	family, member 7B
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)	65260	SELRC1	Sel1 repeat containing 1
2705	GJB1	gap junction protein, beta 1, 32kDa	54993	ZSCAN2	zinc finger and SCAN domain containing 2
10129	FRY	furry homolog (Drosophila)	23414	ZFPM2	zinc finger protein, multitype 2
2524	FUT2	fucosyltransferase 2 (secretor status included)	9244	CRLF1	cytokine receptor-like factor 1
4585	MUC4	mucin 4, cell surface associated	220134	SKA1	spindle and kinetochore associated complex subunit 1
63035	BCORL1	BCL6 corepressor-like 1	3026	HABP2	hyaluronan binding protein 2
55277	FGGY	FGGY carbohydrate kinase domain containing	3853	KRT6A	keratin 6A
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
153768	PRELID2	PRELI domain containing 2	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
79035	NABP2	nucleic acid binding protein 2	4796	TONSL	tonsoku-like, DNA repair protein
699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)	78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
23090	ZNF423	zinc finger protein 423	26112	CCDC69	coiled-coil domain containing 69
1741	DLG3	discs, large homolog 3 (Drosophila)	84300	MNF1	mitochondrial nucleoid factor 1
2118	ETV4	ets variant 4	6878	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
25894	PLEKHG4	(with RhoGef domain) member 4	23780	APOL2	apolipoprotein L, 2

Table S6: Based on the 73 gene list (S1), 58 genes were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 17 samples), and 2 stages (I and III). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S6A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 1.5 - 58 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.0079462	1
GO:0061458	reproductive system development	0.010198	1
GO:0090305	nucleic acid phosphodiester bond hydrolysis	0.012045	1
GO:0043954	cellular component maintenance	0.012626	1
GO:0001570	vasculogenesis	0.022688	1
GO:0000075	cell cycle checkpoint	0.029342	1
GO:0006302	double-strand break repair	0.030041	1
GO:0048545	response to steroid hormone	0.031907	1
GO:0009100	glycoprotein metabolic process	0.032428	1
GO:0016073	snRNA metabolic process	0.035109	1
GO:0060249	anatomical structure homeostasis	0.037343	1
GO:0015748	organophosphate ester transport	0.039249	1
GO:0098781	ncRNA transcription	0.041385	1
GO:0045930	negative regulation of mitotic cell cycle	0.044961	1
GO:0070085	glycosylation	0.045396	1
GO:0098732	macromolecule deacylation	0.047291	1
GO:0009314	response to radiation	0.048077	1
GO:0006310	DNA recombination	0.052635	1

Table S6: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 58 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S6B. Enriched Pathways

Classification Case 1.5 - 58 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00740	Riboflavin metabolism	0.026478	1
hsa04916	Melanogenesis	0.044392	1
hsa00603	Glycosphingolipid biosynthesis	0.049093	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
none	none	none	none
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM	0.001677	0.12913
WP3875	ATR Signaling	0.041682	1
WP4016	DNA IR-damage and cellular response via ATR	0.052646	1

Table S6: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 58 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S6C. Enriched miRNA targets

Classification Case 1.5 - 58 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0070045	0.78429
GTGACTT,MIR-224	0.0080912	0.78429
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.010646	0.78429
ACCGAGC,MIR-423	0.021331	1
ATTACAT,MIR-380-3P	0.030163	1
GGGCATT,MIR-365	0.033512	1
ACTGCAG,MIR-17-3P	0.03466	1
GTGCAAA,MIR-507	0.047607	1
GTGCAAT,MIR-25,MIR-32,MIR-92,MIR-363,MIR-367	0.048403	1

Table S6: c) miRNA target annotation (MSigDB) of 58 genes that were obtained after the classification approach which involved training and testing at the same dataset (dataset 1-GSE28827, 17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S7. Gene Annotation - Classification Case 2.1

Classification Case 2.1 - 68 Genes					
Entrez Gene ID	**Gene Symbol**	**Gene Name**	**Entrez Gene ID**	**Gene Symbol**	**Gene Name**
10882	C1QL1	complement component 1, q subcomponent-like 1	196410	METTL7B	methyltransferase like 7B
5883	RAD9A	RAD9 homolog A (S. pombe)	8347	HIST1H2BC	histone cluster 1, H2bc
443	ASPA	aspartoacylase	10615	SPAG5	sperm associated antigen 5
284185	LINC00482	long intergenic non-protein coding RNA 482	5733	PTGER3	EP3)
26232	FBXO2	F-box protein 2	2524	FUT2	fucosyltransferase 2 (secretor status included)
2177	FANCD2	Fanconi anemia, complementation group D2	10129	FRY	furry homolog (Drosophila)
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	2027	ENO3	enolase 3 (beta, muscle)
1261	CNGA3	cyclic nucleotide gated channel alpha 3	153768	PRELID2	PRELI domain containing 2
2859	GPR35	G protein-coupled receptor 35	27094	KCNMB3	potassium large conductance calcium-activated channel, subfamily M beta member 3
3866	KRT15	keratin 15	79035	NABP2	nucleic acid binding protein 2
114907	FBXO32	F-box protein 32	23534	TNPO3	transportin 3
2705	GJB1	gap junction protein, beta 1, 32kDa	699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)
4585	MUC4	mucin 4, cell surface associated	23090	ZNF423	zinc finger protein 423
63035	BCORL1	BCL6 corepressor-like 1	25894	PLEKHG4	containing, family G (with RhoGef
55277	FGGY	FGGY carbohydrate kinase domain containing	3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
1741	DLG3	discs, large homolog 3 (Drosophila)	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
2118	ETV4	ets variant 4	7104	TM4SF4	transmembrane 4 L six family member 4
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3	347853	TBX10	T-box 10
55311	ZNF444	zinc finger protein 444	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
7125	TNNC2	troponin C type 2 (fast)	7477	WNT7B	wingless-type MMTV integration site family, member 7B
142	PARP1	poly (ADP-ribose) polymerase 1	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
1290	COL5A2	collagen, type V, alpha 2	23414	ZFPM2	zinc finger protein, multitype 2
80201	HKDC1	hexokinase domain containing 1	220134	SKA1	spindle and kinetochore associated complex subunit 1
65260	SELRC1	Sel1 repeat containing 1	3026	HABP2	hyaluronan binding protein 2
54993	ZSCAN2	zinc finger and SCAN domain containing 2	127845	GOLT1A	golgi transport 1A
9244	CRLF1	cytokine receptor-like factor 1	3853	KRT6A	keratin 6A
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
4796	TONSL	tonsoku-like, DNA repair protein	25934	NIPSNAP3A	nipsnap homolog 3A (C. elegans)
84300	MNF1	mitochondrial nucleoid factor 1	26112	CCDC69	coiled-coil domain containing 69
202374	STK32A	serine/threonine kinase 32A	78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
638	BIK	BCL2-interacting killer (apoptosis-inducing)	629	CFB	complement factor B
54825	CDHR2	cadherin-related family member 2	6878	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
83933	HDAC10	histone deacetylase 10	23780	APOL2	apolipoprotein L, 2

Table S7: Based on the 73 gene list (S1), 68 genes were obtained after the classification approach which involved training at dataset 1-GSE28827 (17 samples), testing at dataset 2-GSE75037 (61 samples), and 2 stages (I and III). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S7A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 2.1 - 68 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0009141	nucleoside triphosphate metabolic process	0.0053053	1
GO:0009123	nucleoside monophosphate metabolic process	0.0066533	1
GO:0007586	digestion	0.012961	1
GO:0043954	cellular component maintenance	0.017735	1
GO:0002526	acute inflammatory response	0.019506	1
GO:0016052	carbohydrate catabolic process	0.019839	1
GO:0005996	monosaccharide metabolic process	0.020068	1
GO:0090305	nucleic acid phosphodiester bond hydrolysis	0.021994	1
GO:0009314	response to radiation	0.023932	1
GO:0007059	chromosome segregation	0.028439	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.030892	1
GO:0001570	vasculogenesis	0.031619	1
GO:0009259	ribonucleotide metabolic process	0.037536	1
GO:0046434	organophosphate catabolic process	0.040547	1
GO:0050727	regulation of inflammatory response	0.044794	1
GO:0000075	cell cycle checkpoint	0.046147	1
GO:0006302	double-strand break repair	0.047205	1
GO:0016073	snRNA metabolic process	0.04855	1
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	0.049468	1
GO:0046939	nucleotide phosphorylation	0.049468	1

Table S7: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 68 genes that were obtained after the classification approach which involved training at dataset 1-GSE28827 (17 samples), testing at dataset 2-GSE75037 (61 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S7B. Enriched Pathways

Classification Case 2.1 - 68 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.010231	1
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.021245	1
hsa03460	Fanconi anemia pathway	0.022156	1
hsa00740	Riboflavin metabolism	0.033781	1
hsa00010	Glycolysis / Gluconeogenesis	0.033988	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.02794	1
P02744	Fructose galactose metabolism	0.03824	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	**DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM**	**0.00017467**	**0.01345**
WP4016	DNA IR-damage and cellular response via ATR	0.0094637	0.36435
WP2516	ATM Signaling Pathway	0.020857	0.53533
WP3875	ATR Signaling	0.050732	0.88186

Table S7: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 68 genes that were obtained after the classification approach which involved training at dataset 1-GSE28827 (17 samples), testing at dataset 2-GSE75037 (61 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S7C. Enriched miRNA targets

Classification Case 2.1 - 68 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0080534	0.9714
GTGACTT,MIR-224	0.0092955	0.9714
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.013186	0.9714
ACCGAGC,MIR-423	0.022387	1
ATTACAT,MIR-380-3P	0.033045	1
ACTGCAG,MIR-17-3P	0.037945	1
GTGCAAA,MIR-507	0.052025	1
GTGCAAT,MIR-25,MIR-32,MIR-92,MIR-363,MIR-367	0.054785	1

Table S7: c) miRNA target annotation (MSigDB) of 68 genes that were obtained after the classification approach which involved training at dataset 1-GSE28827 (17 samples), testing at dataset 2-GSE75037 (61 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S8. Gene Annotation - Classification Case 2.2

Classification Case 2.2 - 70 Genes					
Entrez Gene ID	**Gene Symbol**	**Gene Name**	**Entrez Gene ID**	**Gene Symbol**	**Gene Name**
10882	C1QL1	complement component 1, q subcomponent-like 1	83933	HDAC10	histone deacetylase 10
5883	RAD9A	RAD9 homolog A (S. pombe)	196410	METTL7B	methyltransferase like 7B
443	ASPA	aspartoacylase	8347	HIST1H2BC	histone cluster 1, H2bc
284185	LINC00482	long intergenic non-protein coding RNA 482	10615	SPAG5	sperm associated antigen 5
26232	FBXO2	F-box protein 2	3603	IL16	interleukin 16
2177	FANCD2	Fanconi anemia, complementation group D2	5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	2524	FUT2	fucosyltransferase 2 (secretor status included)
1261	CNGA3	cyclic nucleotide gated channel alpha 3	10129	FRY	furry homolog (Drosophila)
2859	GPR35	G protein-coupled receptor 35	2027	ENO3	enolase 3 (beta, muscle)
3866	KRT15	keratin 15	153768	PRELID2	PRELI domain containing 2
114907	FBXO32	F-box protein 32	27094	KCNMB3	activated channel, subfamily M beta
2705	GJB1	gap junction protein, beta 1, 32kDa	79035	NABP2	nucleic acid binding protein 2
4585	MUC4	mucin 4, cell surface associated	23534	TNPO3	transportin 3
63035	BCORL1	BCL6 corepressor-like 1	699	BUB1	1 homolog (yeast)
55277	FGGY	FGGY carbohydrate kinase domain containing	23090	ZNF423	zinc finger protein 423
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)	25894	PLEKHG4	family G (with RhoGef domain) member 4
1741	DLG3	discs, large homolog 3 (Drosophila)	3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
2118	ETV4	ets variant 4	79674	VEPH1	ventricular zone expressed PH domain homolog 1 (zebrafish)
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3	22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
55311	ZNF444	zinc finger protein 444	7104	TM4SF4	transmembrane 4 L six family member 4
7125	TNNC2	troponin C type 2 (fast)	347853	TBX10	T-box 10
1290	COL5A2	collagen, type V, alpha 2	9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
1747	DLX3	distal-less homeobox 3	7477	WNT7B	wingless-type MMTV integration site family, member 7B
8612	PPAP2C	phosphatidic acid phosphatase type 2C	6690	SPINK1	serine peptidase inhibitor, Kazal type 1
80201	HKDC1	hexokinase domain containing 1	23414	ZFPM2	zinc finger protein, multitype 2
65260	SELRC1	Sel1 repeat containing 1	220134	SKA1	spindle and kinetochore associated complex subunit 1
54993	ZSCAN2	zinc finger and SCAN domain containing 2	3026	HABP2	hyaluronan binding protein 2
9244	CRLF1	cytokine receptor-like factor 1	127845	GOLT1A	golgi transport 1A
116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1	3853	KRT6A	keratin 6A
4796	TONSL	tonsoku-like, DNA repair protein	5980	REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit
84300	MNF1	mitochondrial nucleoid factor 1	25934	NIPSNAP3A	nipsnap homolog 3A (C. elegans)
170487	ACTL10	actin-like 10	26112	CCDC69	coiled-coil domain containing 69
202374	STK32A	serine/threonine kinase 32A	78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2
638	BIK	BCL2-interacting killer (apoptosis-inducing)	6878	TAF6	binding protein (TBP)-associated factor, 80kDa
54825	CDHR2	cadherin-related family member 2	23780	APOL2	apolipoprotein L, 2

Table S8: Based on the 73 gene list (S1), 70 genes were obtained after the classification approach which involved training at dataset 2-GSE75037 (61 samples), testing at dataset 1-GSE28827 (17 samples), and 2 stages (I and III). The genes are described by their gene symbols and gene names using WebGestalt 2013.

S8A. Enriched Gene Ontology (GO) biological processes (BP)

Classification Case 2.2 - 70 Genes			
A. Gene Ontology-Biological Process-noRedundant (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
GO:0007586	digestion	0.013607	1
GO:0043954	cellular component maintenance	0.018348	1
GO:0016052	carbohydrate catabolic process	0.020805	1
GO:0005996	monosaccharide metabolic process	0.021299	1
GO:0061458	reproductive system development	0.022513	1
GO:0009141	nucleoside triphosphate metabolic process	0.027751	1
GO:0007059	chromosome segregation	0.030136	1
GO:0034404	nucleobase-containing small molecule biosynthetic process	0.032351	1
GO:0001570	vasculogenesis	0.032684	1
GO:0009123	nucleoside monophosphate metabolic process	0.03296	1
GO:0046434	organophosphate catabolic process	0.042419	1
GO:0000075	cell cycle checkpoint	0.048253	1
GO:0016073	snRNA metabolic process	0.050143	1
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	0.051089	1
GO:0046939	nucleotide phosphorylation	0.051089	1

Table S8: a) Gene Ontology (GO) annotation in the category of biological process-no redundant of 70 genes that were obtained after the classification approach which involved training at dataset 2-GSE75037 (61 samples), testing at dataset 1-GSE28827 (17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S8B. Enriched Pathways

Classification Case 2.2 - 70 Genes			
B. Pathways			
KEGG (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
hsa00500	Starch and sucrose metabolism	0.010231	1
hsa00524	Neomycin, kanamycin and gentamicin biosynthesis	0.021245	1
hsa03460	Fanconi anemia pathway	0.022156	1
hsa00740	Riboflavin metabolism	0.033781	1
hsa00010	Glycolysis / Gluconeogenesis	0.033988	1
Panther (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
P02762	Pentose phosphate pathway	0.024485	1
P02744	Fructose galactose metabolism	0.033533	1
P00024	Glycolysis	0.051415	1
Wikipathway cancer (p value ≤ 0.05)			
Gene Set	**Description**	**P Value**	**FDR**
WP3959	DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM	0.002347	0.18072
WP2516	ATM Signaling Pathway	0.017292	0.66574
WP3875	ATR Signaling	0.046217	0.88906
WP707	DNA Damage Response	0.04795	0.88906

Table S8: b) Pathway annotation (KEGG, Panther, Wikipathway cancer) of 70 genes that were obtained after the classification approach which involved training at dataset 2-GSE75037 (61 samples), testing at dataset 1-GSE28827 (17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.

S8C. Enriched miRNA targets

Classification Case 2.2 - 70 Genes		
C. miRNA targets		
miRNA targets (p value ≤ 0.05)		
Gene Set	**P Value**	**FDR**
TCATCTC,MIR-143	0.0091899	1
GTGACTT,MIR-224	0.010599	1
TGGTGCT,MIR-29A,MIR-29B,MIR-29C	0.016106	1
ACCGAGC,MIR-423	0.023442	1
ATTACAT,MIR-380-3P	0.03603	1
GGGCATT,MIR-365	0.039988	1
ACTGCAG,MIR-17-3P	0.041343	1

Table S8: c) miRNA target annotation (MSigDB) of 70 genes that were obtained after the classification approach which involved training at dataset 2-GSE75037 (61 samples), testing at dataset 1-GSE28827 (17 samples), and 2 stages (I and III). The enrichment analysis was performed by WebGestalt 2019.