BIOMEDICAL ENGINEERING MSc PROGRAM

UNIVERSITY OF CRETE
TECHNICAL UNIVERSITY OF CRETE
FOUNDATION FOR RESEARCH AND TECHNOLOGY-HELLAS

Application Grade Thesis

Title: Extraction and validation of radiotranscriptomic signatures for Non-Small Cell Lung Cancer Diagnosis

Τίτλος: Εξαγωγή και επικύρωση ραδιο-μεταγραφικών υπογραφών για τη διάγνωση του Μη Μικροκυτταρικού Καρκίνου του πνεύμονα

Student's Name: Aikaterini Dovrou

Supervisor's Name: Michalis Zervakis

Date of completion: 13/5/2022

This dissertation is submitted as a partial fulfilment of the requirements for the Master's degree of Biomedical Engineering M.Sc. program

Acknowledgements

With the completion of the present thesis, I would like to thank the people that supported me. Firstly, I would like to thank my advisor, Professor Michalis Zervakis, for his guidance and trust in me. His valuable advice helped me evolve and better understand specific aspects of my field.

I would also like to thank Dr. Ekaterini Bei and Dr. Stelios Sfakianakis for their continuous support and help during this dissertation. Their availability and suggestions helped me to complete my thesis.

I owe a thank you to Professor Kostas Marias and Mr. Eleftherios Trivizakis for kindly providing me useful insights for the extraction of the radiomic features.

Furthermore, I would like to thank my fellow students in this program and Stelios for their support and encouragement.

Last but not least, I owe the biggest thank you to my family for their tremendous support, help and encouragement. I would like to thank my parents, Sofia and Nikos, for always being there for me and supporting me to achieve my goals! I would also like to thank my sister, Eleni, for her valuable advice and guidance, pushing me to the right direction to reach my goals!

Abstract

Radiotranscriptomics is an emerging field that aims to combine the radiomic features extracted from the tumor region and the gene expression profiles in order to contribute in the diagnosis, treatment planning and prognosis of the cancer. The integration of the non-invasive radiomic features with the expression profiles of the genomic substrate can lead to the identification of robust biomarkers. In this study, radiomic and transcriptomic signatures were derived based on their predictive relationships and were assessed for their ability to discriminate the malignant and the adjacent normal tissues, as well as the lung cancer staging. Three transcriptomics datasets of DNA microarray and RNAseq data were used in order to validate the differentiation ability of a 73-gene signature by implementing multiple Support Vector Machines (SVM) linear classifiers, boxplots, t-tests and volcano plots. Furthermore, linear regression models of the transcriptomic features based on the non-invasive radiomic markers were developed. A non-small cell lung cancer (NSCLC) radiotranscriptomic dataset that contains 112 patients with Computed Tomography (CT) scans and RNAseq data was used in order to validate the predictive models of the radiomic and the transcriptomic features. The derived radiomics, transcriptomics and joint signatures were used for the implementation of SVM linear and Random Forest classifiers in order to investigate their potential to predict the lung cancer staging. The transcriptomic signature was validated for its potential to discriminate between malignant and adjacent normal tissue by all the machine learning and statistical algorithms, achieving accuracy greater than 89% and statistical significance with p-value less than e^{-12}. The exploration analysis resulted in the identification of 11 radiomic and 9 transcriptomic features that can be predicted through square regression from the transcriptomic and the radiomic features, respectively, in the RNAseq data. All the Random Forest classifiers demonstrated slightly better performance than the SVM classifiers, achieving accuracy ∼70-75%, sensitivity ∼70-75% and specificity ∼75-80%. Thus, the derived radiomic and transcriptomic signatures have the ability to predict the lung cancer staging and aid in the decision making of the treatment planning.

Περίληψη

Η ραδιομεταγραφική (Radiotranscriptomics) είναι ένα αναπτυσσόμενο πεδίο που στοχεύει να συνδυάσει τα ραδιομικά χαρακτηριστικά που εξάγονται από την περιοχή του όγκου και τα προφίλ γονιδιακής έκφρασης, προκειμένου να συμβάλει στη διάγνωση, στο σχεδιασμό της θεραπείας και στην πρόγνωση του καρκίνου. Η συσχέτιση των μη επεμβατικών ραδιομικών χαρακτηριστικών με τα προφίλ έκφρασης του γονιδιωματικού υποστρώματος μπορεί να οδηγήσει στην ταυτοποίηση ισχυρών βιοδεικτών. Σε αυτή τη μελέτη, ραδιομικές και μεταγραφικές υπογραφές εξήχθησαν με βάση τις προγνωστικές τους σχέσεις και αξιολογήθηκαν για την ικανότητά τους να διακρίνουν τους κακοήθεις από τους γειτονικούς φυσιολογικούς ιστούς, καθώς και το στάδιο του καρκίνου του πνεύμονα. Τρία σύνολα μεταγραφικών δεδομένων από μικροσυστοιχίες γονιδίων και RNAseq τεχνολογία, χρησιμοποιήθηκαν για να επικυρώσουν την ικανότητα διαφοροποίησης μιας υπογραφής 73 γονιδίων υλοποιώντας γραμμικούς ταξινομητές Support Vector Machines (SVM), boxplots, t-tests και volcano plots. Επιπλέον, αναπτύχθηκαν μοντέλα γραμμικής παλινδρόμησης των μεταγραφικών χαρακτηριστικών που βασίζονται στα μη επεμβατικά ραδιομικά χαρακτηριστικά. Ένα σύνολο ραδιομεταγραφικών δεδομένων για το μη μικροκυτταρικό καρκίνο του πνεύμονα (ΜΜΚΠ), που περιέχει 112 ασθενείς με εικόνες αξονικής τομογραφίας και δεδομένα RNAseq, χρησιμοποιήθηκε για την αξιολόγηση των προγνωστικών μοντέλων των ραδιομικών και μεταγραφικών χαρακτηριστικών. Οι ραδιομικές, μεταγραφικές και ραδιομεταγραφικές υπογραφές που εξήχθησαν, χρησιμοποιήθηκαν για την υλοποίηση γραμμικών ταξινομητών SVM και Random Forest, προκειμένου να διερευνηθεί η ικανότητά τους να προβλέπουν το στάδιο του καρκίνου του πνεύμονα. Η μεταγραφική υπογραφή επικυρώθηκε για τη δυνατότητά της να διακρίνει τον κακοήθη από τον γειτονικό φυσιολογικό ιστό αξιολογώντας τους αλγόριθμους μηχανικής μάθησης και στατιστικής, επιτυγχάνοντας ακρίβεια μεγαλύτερη από 89% και στατιστική σημαντικότητα με τιμή p-value μικρότερη από e^{-12}. Επίσης, η ανάλυση οδήγησε στην εύρεση 11 ραδιομικών και 9 μεταγραφικών χαρακτηριστικών που μπορούν να προβλεφθούν μέσω μη γραμμικής παλινδρόμησης δευτέρου βαθμού από τα μεταγραφικά και ραδιομικά χαρακτηριστικά, αντίστοιχα, χρησιμοποιώντας το σύνολο δεδομένων από RNAseq τεχνολογία. Όλοι οι ταξινομητές Random Forest είχαν ελαφρώς καλύτερη απόδοση από τους ταξινομητές SVM, επιτυγχάνοντας ακρίβεια ~70-75%, ευαισθησία ~70-75% και ειδικότητα ~75-80%. Συνεπώς, οι ραδιομικές και μεταγραφικές υπογραφές έχουν την ικανότητα να προβλέπουν το στάδιο του καρκίνου του πνεύμονα και να βοηθούν στη λήψη αποφάσεων του σχεδιασμού της θεραπείας.

Table of contents

List of figures

List of tables

Chapter 1: Introduction

Lung cancer is an aggressive type of cancer and constitutes the leading cause of cancer-related deaths worldwide. The majority of the population diagnosed with lung cancer is of age 70 or over, while a small proportion of affected subjects (1% or lower) is of age younger than 45 [1]. Small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) are the two main groups of lung cancer. Non-small cell lung cancer is the most common form of lung cancer, accounting for more than 85% of the cases [2]. The adenocarcinoma and the squamous cell carcinoma are the two major histological types of NSCLC. Surgery, radiation, chemotherapy and targeted drug therapies are traditional strategies that are used for curating the disease. The treatment planning based on specific oncogenic driver alterations has increasingly been used towards the direction of the precision medicine [3]. However, the diagnosis of the cancer at its earliest stages remains the most significant factor for increased probabilities of survival of the patient.

Screening tests, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), are performed in order to detect the lung tumor. The low dose CT scan is the most commonly used screening test for the diagnosis and treatment monitoring of the NSCLC. Radiomics is a high-throughput image analysis technique that derives a huge amount of quantitative imaging features. These features reflect the heterogeneity and the size of tumor, characterizing the tumor phenotype. Thus, these features are extracted from the Region of Interest (ROI) of the scan. The radiomic features are categorized to the following classes: i) first-order statistics; ii) second-order statistics; iii) higher order statistics; and iv) shape-based. The first-order statistics features describe the distribution of the voxel intensities within the image region defined by the mask through commonly used and basic metrics, such as mean and standard deviation. The second-order statistics features are based on the joint probability distribution of pairs of voxels, describing the spatial arrangement of patterns. These features are extracted from the Gray Level Co-occurrence Matrix (GLCM). The higher-order statistics features are based on the relation between a pixel and the neighboring pixels. Hence, these features are extracted from the Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM) and Gray Level Dependence Matrix (GLDM). Finally, the shape-based features describe the 2D or 3D size and shape of the tumor and are independent from the gray level intensity distribution in the region of interest [4].

Tumors are characterized by somatic mutations. Hence, the unveiling of the way that the genetic alternations affect the cell proliferation and the tumor texture and shape is critical for a deeper understanding of the disease. Thus, a lung biopsy is also performed in most cases, by removing a sample of the lung tissue, in order to determine the presence and the spread of the cancer and subsequently provide insight into the biological and molecular functions of the neoplasms. The genotype of the tumor could be examined using either DNA microarray technology or next-generation sequencing methods. The levels of the gene expression data are measured quantitatively using these methods in order to give insights into the complexity of cancer, since they are used to derive profiles or signatures that define concrete cancer phenotypes[5]. These technologies are high-throughput techniques that measure the expression of many thousands of genes, facilitating the implementation of –omics studies. The RNAseq technology [6] is a next-generation sequencing technology and has prevailed over the DNA microarray technology, which depends on the hybridization. Despite the fact that the gene expression levels measured by the RNAseq and the DNA microarray technology are well-correlated [7], the RNAseq presents some significant advantages over the microarrays [8][9]. More specifically, the RNAseq technology has the ability to predict novel and low-abundance transcripts, as it does not require transcript-specific probes for hybridization. Furthermore, the RNAseq has wider dynamic range and higher specificity and sensitivity than the microarrays, resulting in higher resolution.

The combination of the information captured from the phenotype and the genotype of the tumor lies in the field of Radiotranscriptomics/Radiogenomics [10][11]. The radiotranscriptomic studies aim to investigate the associations between the radiomic and the transcriptomic features in order to reveal their underlying biological connection. Furthermore, the imaging and the genomic data are used in order to investigate whether their combination can enhance the predictive power of the models in challenging tasks, such as the prediction of oncogenic mutations, staging, survival and treatment response.

In this study, the predictive relationships between the radiomic and the transcriptomic features were investigated in order to derive regression models for each modality. The ability of the non-invasive radiomic features to simulate the transcriptomic features through linear regression was examined using a dataset with DNA microarray data. Furthermore, the diagnostic potential of the transcriptomic signatures was thoroughly investigated in order to validate their power to discriminate malignant from adjacent normal tissue. The radiotranscriptomic relationships was further investigated in a dataset with the prominent RNAseq technology. The predictive models of radiomic and transcriptomic features in terms of transcriptomics and radiomics, respectively, were validated in order to investigate whether these relationships are preserved in the RNAseq data. All the derived radiomic, transcriptomic and radiotranscriptomic signatures were assessed for their ability to perform the classification task of lung cancer staging. The identification of the cancer staging is very significant for the determination of the treatment planning of the disease.

This thesis is divided in 6 chapters. Chapter 1 is a brief introduction of the basic principles of the Radiotranscriptomics field and the goals of the study. Chapter 2 includes a comprehensive overview of the state-of-the-art studies that have been conducted in the field of Radiogenomics/Radiotranscriptomics. Chapter 3 and Chapter 4 present the proposed methodological framework and the final results of the study, respectively. Chapter 5 includes the interpretation of the findings and a discussion about the limitations of the study. Chapter 6 summarizes the work alongside with proposed future work.

Chapter 2: State-of-the-art

The lung cancer is the leading cause of cancer-related deaths and thus many radiomics and radiogenomics studies have been conducted in order to provide insight into the diagnosis and the prognosis of the disease [12]. The radiomic features can predict the histological subtypes of NSCLC using patients from many clinical centres [13]. Furthermore, the radiomic features extracted from CT scans can improve the prediction of overall survival in NSCLC patients [14][15]. Additionally, solely genomic information can predict the survival of patients with NSCLC. Zhang et al. [16] showed that a glycolysis-related gene signature, consisting of 9 genes, can predict the overall survival in patients with lung adenocarcinoma.

The majority of the radiogenomic/radiotranscriptomic studies are focused on the prediction of oncogenic mutations, survival and histologic subtypes, solely from the non-invasive radiomic features. More precisely, the radiomic signatures have the potential to predict the status of the epidermal growth factor receptor (EGFR) and the Kirsten rat sarcoma viral oncogene homolog (KRAS) mutations, which are major oncogenic driver mutations [17][18][19][20]. However, Pinheiro et al. [21] showed that the radiomic features extracted from CT scans are correlated with the EGFR mutation status but not with the KRAS mutations status. The same conclusion was derived by the Gevaert et al. [22], identifying a statistically significant model for the EGFR mutation status but not for the KRAS mutation status.
The associations between imaging features extracted from PET images and oncogenic signalling pathways were investigated by Kim et al. [23]. More specifically, clusters of PET imaging features were associated with the activation of three oncogenic signaling pathways, the cell cycle, the WNT and the TGFβ. Moreover, Ubaldi et al. [24] showed that solely radiomic features can predict the tumor histology and stage using small datasets of patients with NSCLC. The machine learning classifiers, Random Forest and linear Support Vector machines (SVM), achieved Area Under the Curve (AUC) greater than 70%.

Zhu et al. [25] extracted radiomic features from CT scans using the opensource software pyradiomics in patients with advanced lung adenocarcinoma. Several traditional machine learning classifiers were used in order to predict co-mutations of TP53 and EGFR (Figure 1). Each classifier was implemented using four different types of feature vectors, which are i) clinical features; ii) semantic features (qualitative features reported by radiologists to characterize lung lesions[26]); iii) radiomic features; and iv) an integrated model of clinical and semantic and radiomic features. The results showed that the integrated model achieved the best performance for discriminating the co-mutations of TP53 and EGFR, indicating that the clinical and semantic features can enhance the predictive ability of the CT derived radiomic features.

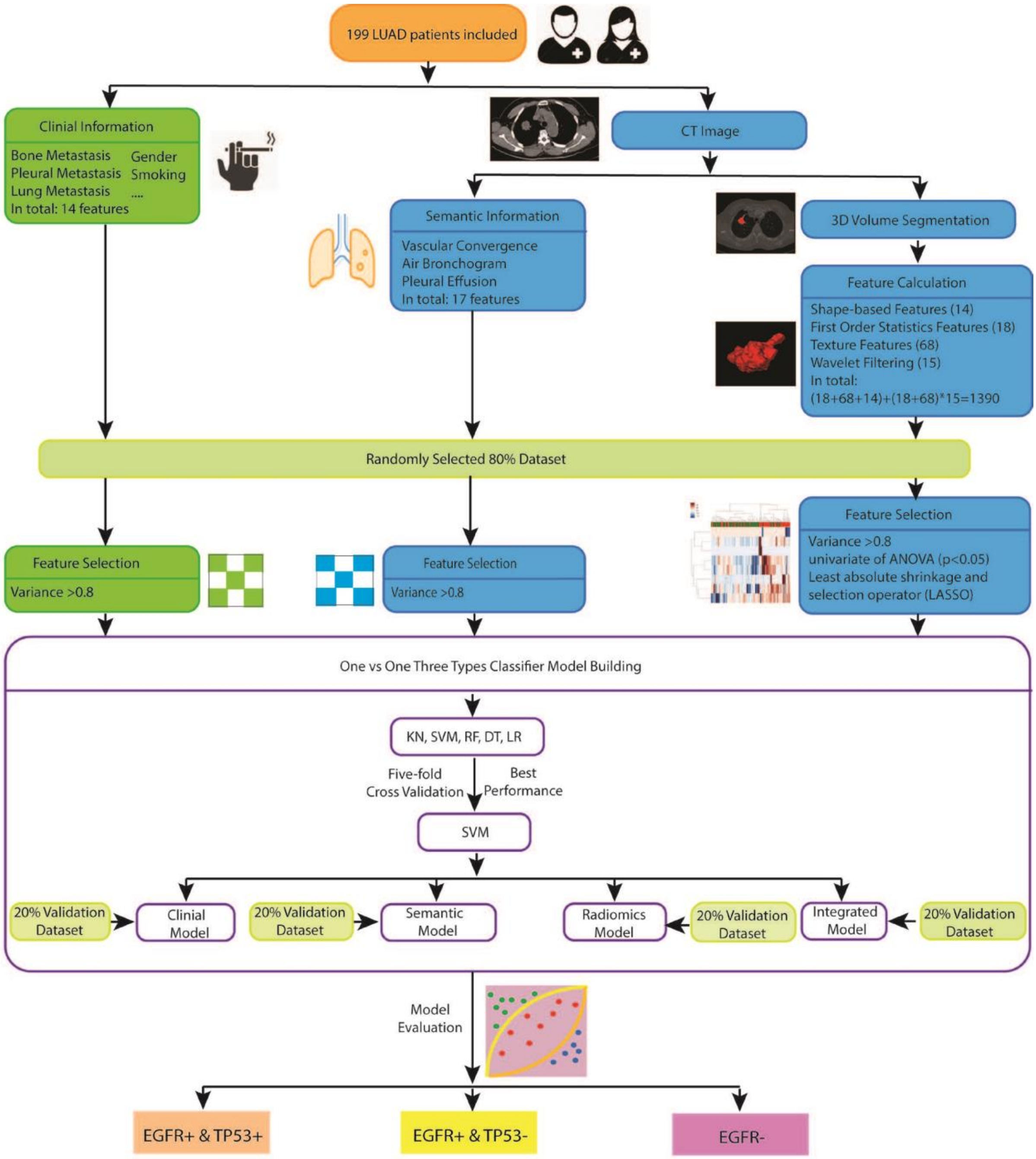

Figure 1. Flowchart of the methodology proposed by Zhu et al. (2021) [25]

However, the use of the combination of selected radiomic and transcriptomic features has been investigated, to a lesser extent, in order to assess whether the joint signatures enhance the predictive strength of the models. More specifically, Fan et al. [27] developed radiotranscriptomic signatures, which consisted of CT radiomic features and miRNA levels, along with basic clinical features for the prediction of the objective response rate (ORR), overall survival (OR) and progression-free survival (PFS) in patients with NSCLC treated with radiotherapy. Furthermore, they developed a radiotranscriptomic signature-based nomogram for predicting the ORR in NSCLC patients. The proposed methodology is presented in Figure 2.

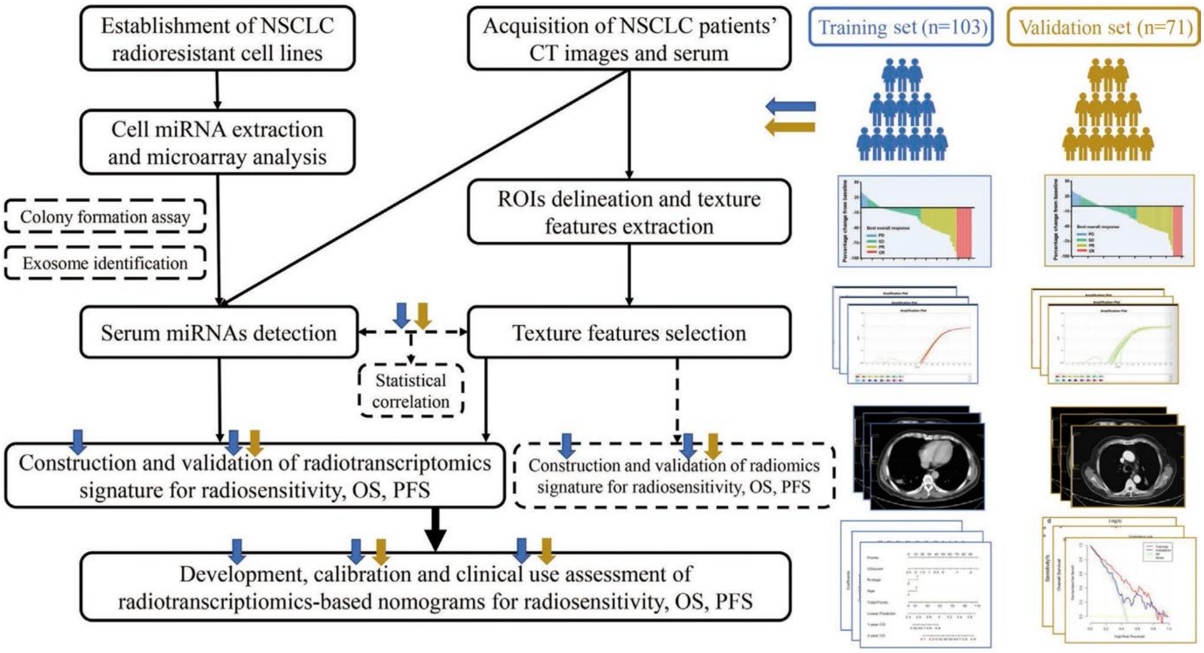
Establishment of NSCLC radioresistant cell lines
Cell miRNA extraction and microarray analysis
Colony formation assay
Exosome identification
Acquisition of NSCLC patients' CT images and serum
ROIs delineation and texture features extraction
Serum miRNAs detection
Texture features selection
Statistical correlation
Construction and validation of radiotranscriptomics signature for radiosensitivity, OS, PFS
Construction and validation of radiomics signature for radiosensitivity, OS, PFS
Development, calibration and clinical use assessment of radiotranscriptiomics-based nomograms for radiosensitivity, OS, PFS
Training set (n=103)
Validation set (n=71)

Figure 2. Flowchart for the extraction of the radiotranscriptomic signatures and the radiotranscriptomics-based nomograms developed by Fan et al. (2020) [27].

Radiotranscriptomics has also been extracted in order to evaluate cardiovascular and brain diseases. More precisely, Oikonomou et al. [28] extracted radiomics and transcriptomics from the adipose tissue. The derived radiotranscriptomics signature resulted in improvement of the cardiac risk prediction in cardiovascular diseases. Furthermore, the investigation of the models that integrate radiomics and multiple –omics data is also of paramount importance in order to assess whether additional –omics information can enhance the predictive power of the models. Chaddad et al. [29] showed that an integrative model consisting of multi-omics features improves the ability to predict the survival of patients with IDH1 wild-type glioblastoma. This model contained radiomic, genomic, transcriptomic and protein expression-immunohistochemical (IHC) features (Figure 3).

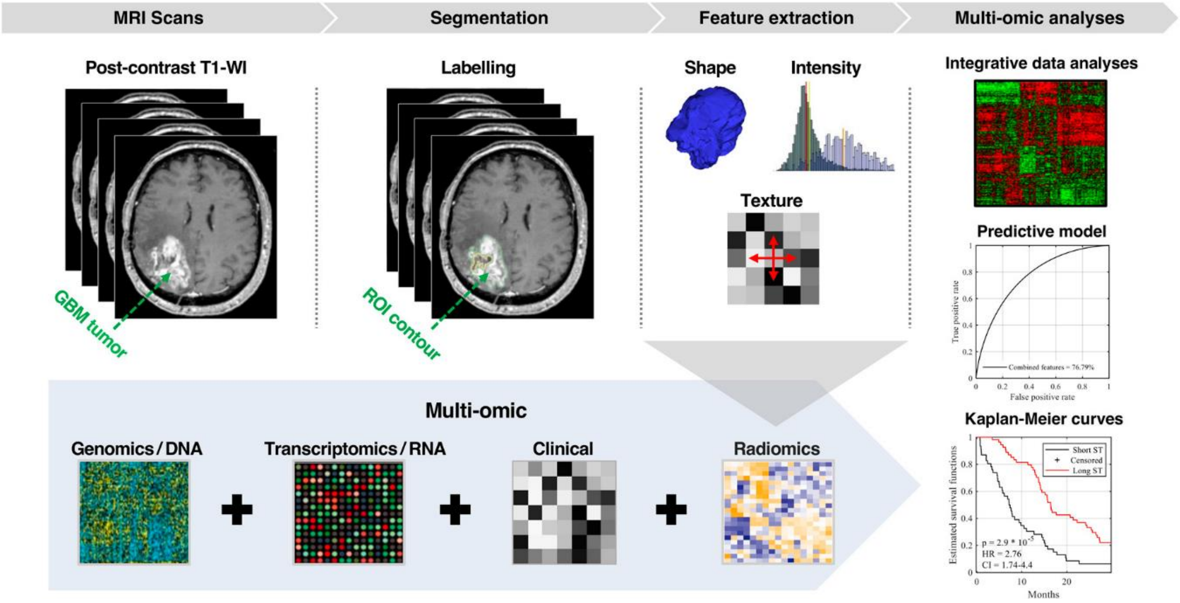

Figure 3. Workflow for the development of the multi-omics model proposed by Chaddad et al. [29]

More recently, with the rapid development of the deep learning networks, many researchers focus on the use of deep neural networks to extract radiomic features from medical images in order to predict and evaluate a disease [30]. Trivizakis et al. [31] proposed a deep radiotranscriptomic model (Figure 4). They combined features extracted from the deep neural networks, which are called deep features, with transcriptomics features into a common space in order to predict the molecular mutations, i.e. EGFR and KRAS mutations, and the histological subtypes, i.e. adenocarcinoma or squamous, for NSCLC patients. The proposed deep radiotranscriptomic model demonstrated high performance achieving an AUC of 83.1% and 92.5% for the prediction of the molecular mutations and the histological subtypes, respectively.

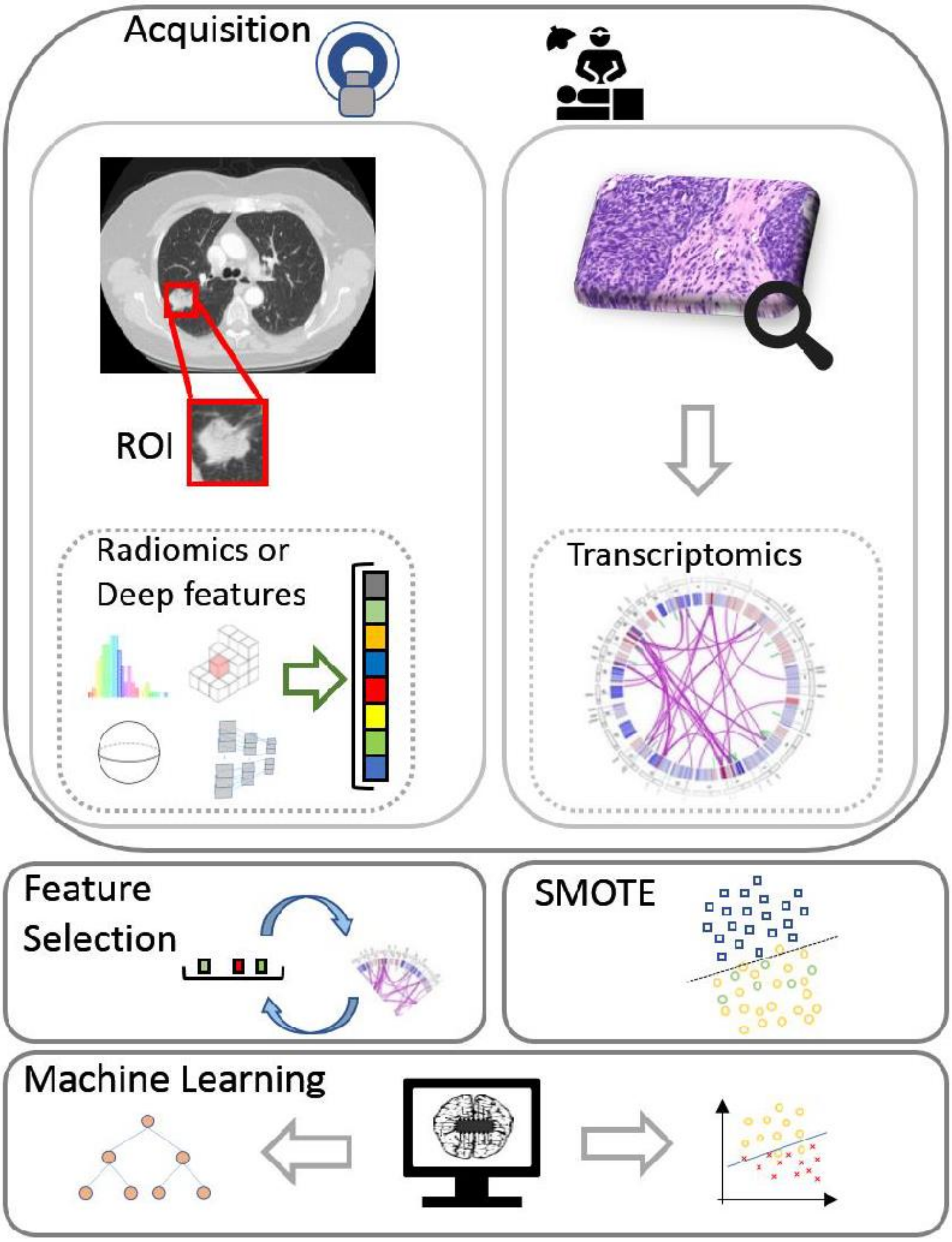

Figure 4. Flowchart of the deep radiotranscriptomic model proposed by Trivizakis et al.[31]

Emphasis has also been given to the investigation of the correlation between the imaging features and the genomic data. These correlations could reveal the underlying biological connection of the imaging features and genes. The first studies in this field were those of Gevaert et al. [32] and Nair et al. [33]. Gevaert et al. [32] extracted the gene expression profiles and imaging features from PET and CT images from patients with NSCLC. The high-throughput gene expression profiles were grouped together to metagenes as clusters of co-expressed genes. The relationships between the imaging features and the metagenes were assessed using the Spearman rank correlation test, the Significance Analysis of Microarrays (SAM) and the false discovery rate (FDR) for multiple comparisons corrections. Furthermore, the generalized linear regression with Lasso Regularization was used in order to produce predictive models of radiomic and genomic features in terms of genomic and radiomic features, respectively. The metagenes were associated with survival and thus the predicted imaging features from genes were also associated with survival. Zhou et al. [34] created a radiogenomics map to link the derived imaging features from CT scans with the RNAseq gene expression profiles for patients with NSCLC (Figure 5). Similarly to the aforementioned studies, Zhou et al. [34] produced metagenes and identified 10 homogeneous metagenes. The results showed that the late cell cycle genes were correlated with nodule attenuation and nodule margins and the genes of the EGFR pathway were associated with nodule margins and ground-glass opacity. Thus, they proposed a method in which specific imaging features could be linked with specific group of genes that describe molecular properties and activate or deactivate specific molecular pathways.

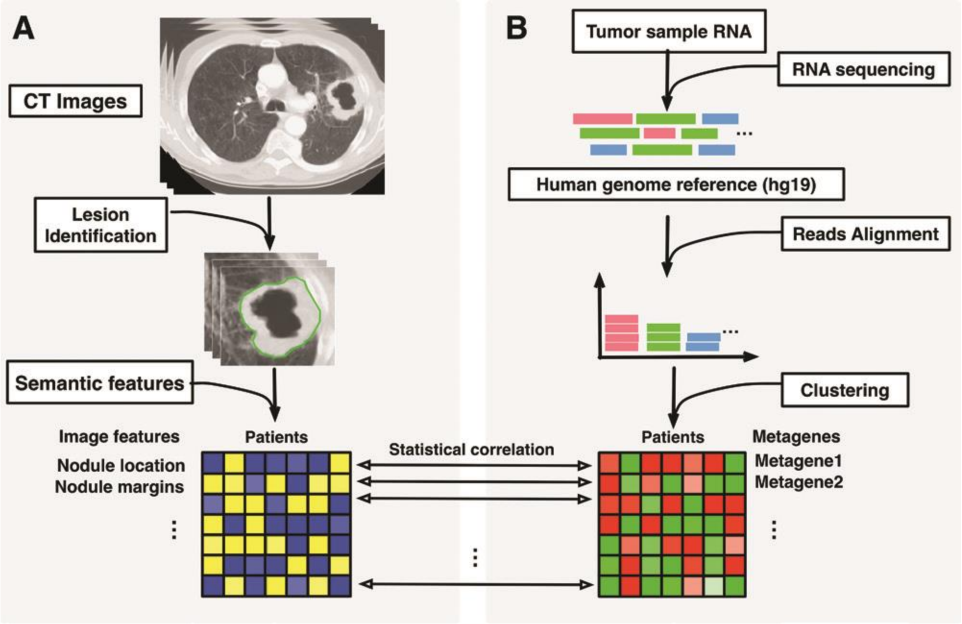

Figure 5. Workflow for the radiogenomic map that was derived by Zhou et al.[34]

In our previous study [35], we investigated the relationships between the radiomic features extracted from CT scans and the gene expression profiles measured by DNA microarray technology. The analysis resulted in a transcriptomic signature of 73 genes that had statistically significant correlation with radiomic features and diagnostic ability to discriminate between malignant and adjacent normal lung tissues. The radiomic features were clustered into homogeneous groups and each cluster was represented by the central radiomic feature. The representative radiomic features were modelled through lasso regression from the 73 genes. Thus, 51 predictive models of radiomics based on subsets of the transcriptomic features were developed. Furthermore, enrichment analysis of the transcriptomic signatures were performed in order to reveal signalling and metabolic pathways related to the oncogenesis.

The results of this study motivated us to further investigate the relationships between the radiomic and the transcriptomic features in order to identify the underlying biological connection of the phenotype and the genotype and build robust models. Hence, we explore these associations in DNA microarray and RNAseq datasets in order to produce regression models for both features. Furthermore, radiomic, transcriptomic and integrated signatures are used to assess their ability to predict the lung cancer stage. To the best of our knowledge, the current work is the first radiotranscriptomic study in NSCLC that integrates radiomic and transcriptomic data for the prediction of lung cancer staging.

Chapter 3: Research methodology

The current study implements statistical analysis and machine learning techniques to validate the relationships between the radiomic and the transcriptomics features and investigate their impact on lung cancer staging prediction. The workflow of the study is presented in Figure 6. More specifically, in our previous work [35] we had extracted a transcriptomics signature and a p-metaomics signature consisting of 73 genes (Appendix A) and 51 radiomic features (Appendix B) modelled by transcriptomics features, respectively. In this study, the reverse modeling of the transcriptomics features from the non-invasive radiomic features were implemented. Furthermore, the transcriptomics and the p-metaomics signatures were utilized to validate their diagnostic potential for differentiating the malignant and the benign lung tumors in DNA microarrays and RNAseq datasets. A radiotranscriptomics dataset with RNAseq data and CT scans was exploited to validate the derived regression models of both modalities, which are the models of radiomics based on transcriptomics and vice versa. Finally, various radiomics, transcriptomics and radiotranscriptomics signatures, which have been derived from the analysis, were used as feature vectors in a SVM classifier and a Random Forest Classifier in order to investigate their ability to perform the challenging task of lung cancer staging prediction.

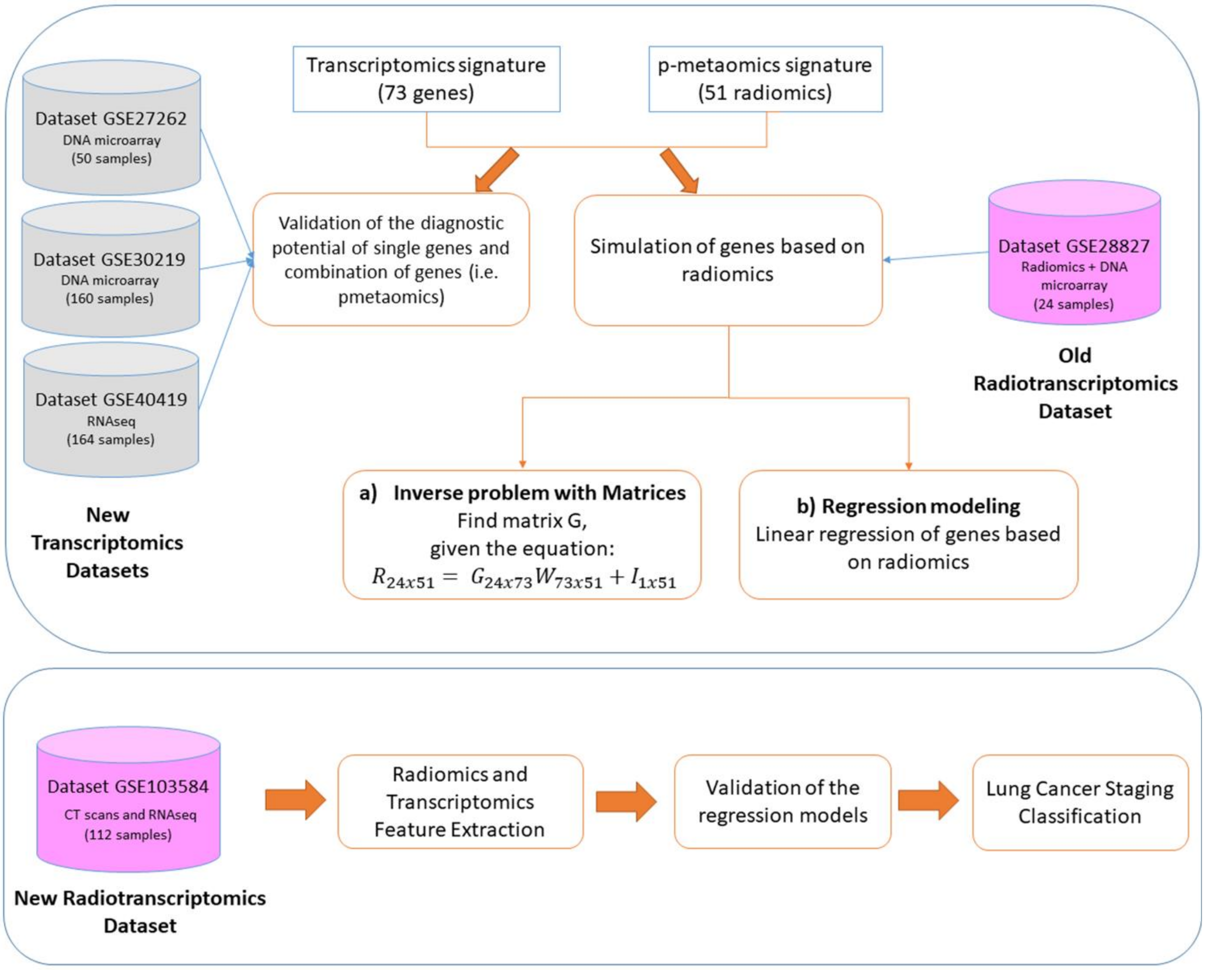
Transcriptomics signature (73 genes)
p-metaomics signature (51 radiomics)
Dataset GSE27262 DNA microarray (50 samples)
Dataset GSE30219 DNA microarray (160 samples)
Dataset GSE40419 RNAseq (164 samples)
New Transcriptomics Datasets
Validation of the diagnostic potential of single genes and combination of genes (i.e. pmetaomics)
Simulation of genes based on radiomics
Dataset GSE28827 Radiomics + DNA microarray (24 samples)
Old Radiotranscriptomics Dataset
a) Inverse problem with Matrices
Find matrix G, given the equation:
$R_{24x51} = G_{24x73}W_{73x51} + I_{1x51}$
b) Regression modeling
Linear regression of genes based on radiomics
Dataset GSE103584 CT scans and RNAseq (112 samples)
New Radiotranscriptomics Dataset
Radiomics and Transcriptomics Feature Extraction
Validation of the regression models
Lung Cancer Staging Classification

Figure 6. Overall workflow of the study

3.1. Datasets Description

Three transcriptomics and two radiotranscriptomics datasets were used for the analysis. A detailed description of the used datasets are presented in Table 1. The transcriptomics data were obtained from the publicly available Gene Expression Omnibus (GEO) database. In each dataset, the probes were mapped into their corresponding Entrez Gene ID according to the Illumina platform. However, one Entrez Gene ID may map to more than one probe; thus, the probe with the higher gene expression value was used to express the corresponding Entrez Gene ID.

Regarding the transcriptomics datasets, two datasets with gene expression microarray data, the GSE27262 and the GSE30219, and one dataset with gene expression measured with RNAseq, the GSE40419, were used. These datasets contain gene expression data from lung tumor samples and adjacent normal samples.

Additionally, the radiotranscriptomics dataset GEO28827 [32][36], which was used in our previous study [35], was also utilized in the current study. This dataset was updated in 2018. The new updated dataset [37], which consists of 130 patients with RNAseq and CT scans accompanied by the segmentation masks of the tumor, was used in the present study. The RNAseq data of the dataset can be found under the accession number GSE103584. The CT scans and the corresponding segmentation masks for each patient of the dataset were obtained from the publicly available Cancer Imaging Archive (TCIA) database [38] and can be downloaded from https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics#a99a795ff4454409862a398ffc076b98 (accessed on September 2021).

However, the dataset did not contain segmentation masks for 13 patients; thus, these patients were excluded from the analysis. Moreover, 2 more patients were excluded due to the fact that their 3D segmentation mask was smaller than 10 pixels. Furthermore, the RNAseq data contains NAN values for many patients. Thus, 112 out of 130 patients were used for the analysis of the present study. Detailed description of the number of samples and the extraction of the radiomics and the transcriptomics features is presented in the section *Radiomics and Transcriptomics Features Extraction*.

Table 1. Overview of datasets

GEO series	Platform	Authors	Patients	LUAD	LUSC	Adjacent normal	Radiomics
GEO28827	Illumina HumanHT-12 V3.0 expression beadchip	Gevaert et al. 2012 [32]	24	19	5	-	Yes
GSE27262	Affymetrix Human Genome U133 Plus 2.0 Array	Wei et al. 2012 [39]	25	25	-	25	No
GSE30219	Affymetrix Human Genome U133 Plus 2.0 Array	Rousseaux et al. 2013 [40]	307	85	61	14	No
GSE40419 (RNAseq)	Illumina HiSeq 2000 (Homo sapiens)	Seo et al. 2012 [41]	164	87	-	77	No
GSE103584 (RNAseq)	Illumina HiSeq 2500 (Homo sapiens)	Bakr et al. 2018 [37]	130*	96	31	-	Yes

LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinoma
*3 samples are identified as NSCLC NOS (not otherwise specified)

3.2. Simulation of genes based on radiomics

The radiomic features extracted from the tumor region of the CT scans, had been modeled by their genomic substrate, resulting in the p-metaomics features [35]. During the initial analysis, the 73 genes of the transcriptomics signature were used to model 51 radiomic features. This modeling aims to biologically justify the radiomic features, which describe the phenotype, based on the gene expression microarray data, which link the genotype to phenotype. The inverse modeling is also of paramount importance for the improvement of the diagnosis, treatment planning and prognosis in lung cancer. To this end, the inverse problem, which is the simulation of 73 genes based on the 51 radiomic features, was implemented using the old Radiotranscriptomics Dataset GEO28827, which contains DNA microarray data and CT radiomic features. This modeling was achieved using two different approaches: i) solving the direct inverse problem based on the modeling of the radiomics from the transcriptomics and ii) solving the indirect problem of modeling the transcriptomics based on specific radiomics. The aim of these 2 approaches is the same, which is the modelling of the genes based on the non-invasive radiomic features with respect to the previous modeling of radiomics based on transcriptomics. The genes that can be predicted from the radiomics with both approaches were used for subsequent analysis.

First approach: direct inverse problem using the matrices

More specifically, the linear regression modeling of the radiomics in terms of transcriptomics, which had been implemented in the previous analysis with the GEO28827 dataset, is formulated via the following equation:

$$R_{24x51} = G_{24x73}W_{73x51} + I_{1x51} \tag{1}$$

where **R** is the matrix that contains the values of the 51 radiomics features, **G** is the matrix that contains the values of the 73 genes, **W** is the matrix that contains the weights (i.e. regression coefficients) of the genes for each radiomics regression model and **I** is a vector that contains the intercepts for each radiomics regression model. The number 24 refers to the number of samples of the radiomics and the transcriptomics features of dataset GEO28827.

In order to solve the inverse problem, we should calculate the matrix **G** from the equation (1), given that the matrices **R**, **W** and the vector **I** are known. The values of the **W** and **I** matrices are known from the modeling of the radiomics from the transcriptomics. Thus, the values of the 73 genes can be simulated from the radiomic features based on the already derived regression models. The inverse matrix of **W**, denoted by W^{-1}, should be calculated in order to solve for **G** in the equation (1) based on the property $W * W^{-1} = I$. However, the matrix **W** is not a square matrix and thus the inverse matrix $\mathbf{W}^{-1}$ does not exist. Hence, the Moore-Penrose pseudoinverse of the matrix **W** was calculated in order to overcome this limitation. The Moore-Penrose pseudoinverse was introduced by E. H. Moore [42] and reinvented by R. Penrose [43][44]. Therefore, the target matrix **G** was calculated via the following equation:

$$G_{24x73} = (R_{24x51} - I_{1x51})Wpseudoinverse_{51x73} \tag{2}$$

Hence, the matrix **G** was calculated and the values of the 73 genes were simulated based on the pseudoinverse matrix of the regression weights. In order to evaluate the simulation of the genes, several metrics were calculated. More precisely, the normalized Root Mean Squared Error (RMSE) was computed between the actual and the simulated gene values. The normalized RMSE was calculated using the following equation:

$$NRMSE = \frac{\sqrt{\frac{\sum_{n;observations}(Y_{actual} - Y_{predicted})^2}{N}}}{Y_{max} - Y_{min}} \tag{3}$$

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

Y_{min} is the minimum value of the Y_{actual}

The normalized RMSE measures the goodness of fit of the models. Values closer to 0 correspond to better prediction of the actual values of the gene. Thus, the predictive models that satisfy the criterion of normalized RMSE <= 1, were considered for further investigation.

Second approach: Indirect inverse problem with linear regression

The values of the 73 genes were also modelled with linear regression using the values of specific radiomic features. More specifically, each radiomic feature had been modeled using a particular subset of genes. In the current modelling, for each gene we identified the radiomic models in which this gene was member of their regression models. The subset of the radiomic features, in which the gene participated in their regression models, are the independent variables, i.e. predictors, of the new regression models for each gene.

The regression model of each gene was assessed using the R-squared, which is known as the coefficient of determination. The R-squared represents the proportion of the variance for the gene-dependent variable that is explained by the radiomics-predictors in the regression model. The maximum value of R-squared is equal to 1 and is achieved when the predicted values are identical to the actual values. The R-squared is calculated via the following equation:

$$R^2 = 1 - \frac{\sum_{k;observations}(Y_{actual} - Y_{predicted})^2}{\sum_{k;observations}(Y_{actual} - Y_{mean})^2} \tag{4}$$

The regression models that satisfy the criterion of R-squared >= 0.70, were used for subsequent analysis. Furthermore, the RMSE and the Pearson correlation coefficient with the corresponding p-value were calculated between the actual and the predicted values for each regression model in order to further evaluate the performance.

3.3. Validation of transcriptomics signatures

The transcriptomics signature consisting of the 73 significant genes (Appendix A) had been investigated for the ability to discriminate between malignant and non-malignant lung tissues using a small external dataset. In the current work, 3 independent transcriptomics datasets were exploited in order to further validate the diagnostic potential of this transcriptomics signature. The ability to discriminate between malignant and normal tissue was also investigated for the genes combinations that simulate the radiomic features (i.e. the p-metaomics features). The two DNA microarray datasets, the GSE27262 and the GSE30219, and the RNAseq dataset, GSE40419, were used in various combinations in order to validate the diagnostic potential of the transcriptomics signatures (i.e. single genes and combination of genes).

The 2 DNA microarray datasets had been extracted using the same Affymetrix platform. All the 3 transcriptomics datasets had been preprocessed and normalized in order to obtain the expression of each gene. The values of the 73 genes were identified in each dataset.

The ability of the transcriptomics signature to predict whether a tissue sample is malignant or normal was examined using four different methods. More specifically, several SVM linear classifiers, boxplots, t-tests and volcano plots were implemented in order to assess the discrimination potential of the transcriptomics signatures.

The SVM linear classifier is a very widely used classification algorithm for machine learning applications due to its simple implementation and high performance. The linear kernel was selected due to its superior performance than the rbf, sigmoid and polynomial kernel. Four SVM linear classifiers using different feature vectors were implemented in order to evaluate their diagnostic ability. More precisely, the following different classification schemas were implemented:

- the first classification schema refers to the classifier that had been implemented in our previous analysis and had been trained with the dataset GSE75037, which had been used during the extraction process of the signatures. The feature vector of this classifier was the values of the 73 genes that constitute the transcriptomics signature. The training set was 166 samples of the GSE75037 dataset, which consists of 83 tumor-adenocarcinoma samples and 83 normal samples. The test set was the combination of the 3 new transcriptomics dataset, i.e. GSE27262, GSE30219 and GSE40419. Hence, the test set consisted of 258 tumor, adenocarcinoma and squamous, samples and 116 control samples. Standardization with mean value 0 and standard deviation 1 was applied as pre-processing step.
- The second classification schema utilized only the new external datasets. The 3 new datasets, i.e. GSE27262, GSE30219 and GSE40419, were combined into one group and the 73 genes were used as feature vector. The new combined dataset included 374 samples. The stratified random sampling method was applied to produce the training and testing sets. Stratified sampling is a sampling technique where the samples are divided into groups, called ‘strata’, in the same proportion as they appear in the initial

population. More specifically, the samples were divided randomly into 70% training set (262 samples) and 30% test set (112 samples). The 70% of the tumor samples were used for the training set (181 samples) and the rest 30% (77 samples) in the test set. Similarly, the 70% of the control samples (81 samples) were used for the training set and the rest 30% (35 samples) in the test set. The training set was normalized and subsequently the test set was normalized applying the normalizing parameters used for the training set. This procedure was repeated 100 times to validate the results, and the average values of the metrics were calculated. However, these 3 datasets had been extracted in different centers under different laboratory conditions and staff members, introducing batch effects. To this end, the ComBat harmonization technique [45] was used to remove the batch effects that may exist across different datasets. Thus, the ComBat (*sva* package in R) was applied to combine efficiently the 3 datasets to increase the statistical power. Hence, the second classification schema of combining the 3 transcriptomics datasets and splitting the samples into train and test subset, was repeated using the ComBat corrected gene expression profiles and without further normalization. The first two classification schemas were implemented in order to validate the power of the 73 genes of the transcriptomics signature in discriminating the malignant from the normal samples, independently of the type of the dataset, i.e. microarray or RNAseq.

- The third classification schema exploited the 2 DNA microarray datasets, the GSE27262 and GSE30219, as training set and the RNAseq dataset, the GSE40419, as test set. The 73 genes of the transcriptomics signature were used again as feature vector. The aim of this classification schema was to investigate whether the 73 genes, which had been derived using their microarray expression profiles, can predict the type of lung tissue using their expression profiles extracted by the RNAseq technology. The size of the training set was 210 samples, while the size of the test set was 164 samples. Similarly to the aforementioned classifiers, the training set was normalized and subsequently the test set was normalized applying the normalizing parameters used for the training set. This classification schema was also implemented using the ComBat corrected gene expression profiles without the normalization.
- Finally, the forth classification schema utilized the 51 p-metaomics features (Appendix B) that had been derived in our previous analysis. Thus, this classifier was based on the linear combination of genes that had been derived for microarray datasets, aiming to validate and investigate whether these models can enhance the predictive and diagnostic ability of the transcriptomics data. To this end, the regression coefficients, which had been derived from the modeling of radiomics from transcriptomics in our previous analysis, were used to produce the p-metaomics features for microarray datasets. These 51 features, which had been derived from the combinations of genes, were produced to the 2 new microarray datasets, GSE27262 and GSE30219. The 51 p-metaomics features were used as feature vector of the classifier. The 2 microarray datasets were combined into one, consisting of 210 samples. The stratified random sampling was used and the samples were divided randomly into 70% training set (147

samples) and 30% test set (63 samples). The same normalization procedure was followed in order to train and test the classifier. The procedure was repeated 100 times and the average values of the metrics were calculated.

The performance of the classifiers was assessed in terms of accuracy, sensitivity, specificity, balanced accuracy and geometric mean (g-mean). The accuracy describes the proportion of correct predictions over the total number of samples. The sensitivity refers to the ability of the classifier to predict the tumor samples correctly. Similarly, the specificity refers to the ability of the classifier to predict the normal samples correctly. The last two metrics, balanced accuracy and geometric mean, are calculated when imbalanced datasets are used. An imbalanced dataset is a dataset with unequal class distribution. The balanced accuracy is the arithmetic mean of sensitivity and specificity. The evaluation metrics were calculated according to the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{8}$$

$$Geometric\ Mean = \sqrt{Sensitivity * Specificity} \tag{9}$$

The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were calculated from the confusion matrix of each classifier.

Furthermore, boxplots were implemented in order to visualize and validate further the diagnostic ability of the 73 genes of the transcriptomics signature. The boxplot is a graph for visual inspection of the data distribution. Boxplots were implemented for each new dataset to identify the distribution of the positive (i.e. higher values in cancerous than in normal tissues) and the negative (i.e. higher values in normal than cancerous tissues) genes in control and tumor population. Additionally, the 3 transcriptomics datasets, GSE27262, GSE30219 and GSE40419, were combined into one dataset and the corresponding boxplots were implemented. From the previous analysis, 62 genes out of the 73 were positive significant and 11 genes were negative significant. Thus, the positive and the negative significant genes should be distinguishable also in the new datasets. For instance, the positive significant genes should have higher values than the negative significant in the tumor population of each dataset. Accordingly, the negative significant genes should present higher values than the positive significant in the control population of each dataset. The values of the genes were log transformed in order to better visualize the distributions. Moreover, a two-sided t-test was performed between these 2 groups, i.e. positive and negative, in each population, i.e. tumor and control, to identify whether their values were statistically significantly different.

The last technique used for the validation of the transcriptomics signature was the volcano plots. The volcano plots provide visual identification of genes with large fold change that are also statistically significant. The p-value cutoff was set to 0.01 and the fold change (FC) cutoff to 2 (package *EnhancedVolcano* in R). Hence, the genes that demonstrated 2-fold change between the cancerous and the normal samples and simultaneously the difference in their values between these 2 groups were statistically significant, were identified as significant in the volcano plots. A Volcano plot was implemented for each one of the new dataset.

3.4. Radiomics and Transcriptomics Features Extraction

A publicly available radiotranscriptomics dataset was used in order to validate the associations between the radiomics and the transcriptomics data of patients with NSCLC. The radiomics, the transcriptomics and the radiotranscriptomics markers were subsequently investigated for their ability to predict the lung cancer staging.

The radiotranscriptomics dataset contains CT scans and the corresponding segmentation masks of the tumor, which defines the region of interest (ROI). However, 13 patients did not have the corresponding segmentation mask and thus they were excluded from the analysis, resulting in 117 patients. The 3D segmentation masks were smaller than 10 pixels for 2 patients and they were also excluded from the subsequent analysis. Hence, the total number of patients was reduced to 115. An example of a CT medical image of the dataset is depicted in Figure 7. The 3D slicer software was used in order to load and view the medical images. The CT scans can be viewed from all the planes (axial, coronal and longitudinal) with the 3D slicer. In Figure 8, the CT scan along with the tumor’s mask are presented in order to visualize the imaging data, which are required for the radiomic features extraction. More precisely, the CT scan and the corresponding tumor’s mask are required to calculate the radiomic features, as these features are extracted from the segmented region of the tumor in order to characterize the shape and the texture of the tumor. The radiomic features were calculated using the opensource python software pyradiomics [46], which produces a huge amount of quantitative features from radiological images. The CT scan and the segmentation mask of each patient are saved in DICOM format in the repository. However, the pyradiomics module requires as input the medical images and the segmentation masks in Nifti or Nrrd format. These image formats represent the image as a 3D object. Thus, the dicom series of the CT images and the masks were converted to Nifti format (.nii) using the python library dicom2nifti. For each patient, the nifti files of the CT image and the corresponding mask were given as input to the pyradiomics module in order to extract the radiomic features. The radiomic features were extracted from the original image as well as from filtered images. More specifically, 7 filters were applied to the original image, which are the Laplacian of Gaussian, Wavelet, Square, Square Root, Logarithm, Exponential and Gradient. Hence, 2996 radiomic features were extracted from the original and the filtered images. The radiomic features were classified in 7 classes according to the definitions introduced by the Imaging Biomarker Standardization Initiative (IBSI) [47]. The feature classes are the first-order statistics, shape descriptors and texture classes, such as the Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM).

However, the current study focuses on the 51 radiomic features (Appendix B) that had been modelled from the transcriptomics data in our previous analysis. These are first-order, shape and texture features, representing all the examined feature classes. More specifically, there are 11 first-order, 4 shape and 36 texture features. The majority belongs to the texture class, which is a highly important class, as it describes the heterogeneity and the morphological complexity of the lung tumor. The number of radiomic features that belong to each feature class is presented in Table 2.

Table 2. The number of radiomic features that belong to each feature class. (total number of radiomic features = 51)

CLASS	**FIRST-ORDER**	**SHAPE**	**TEXTURE**				
			GLCM	**GLDM**	**GLSZM**	**GLRLM**	**NGTDM**
NO.	11	4	15	10	6	3	2

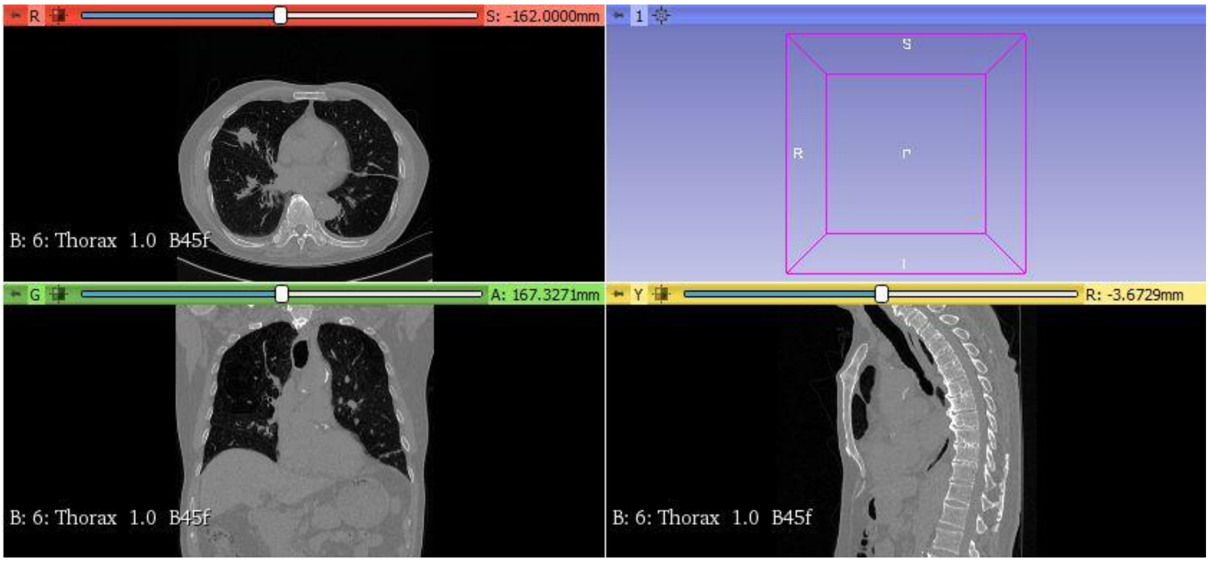

Figure 7. CT scan of a patient with NSCLC obtained with the 3D slicer. The axial view of the image is displayed in the top left figure, while the coronal and the longitudinal plane are depicted in the bottom left and right figure, respectively.

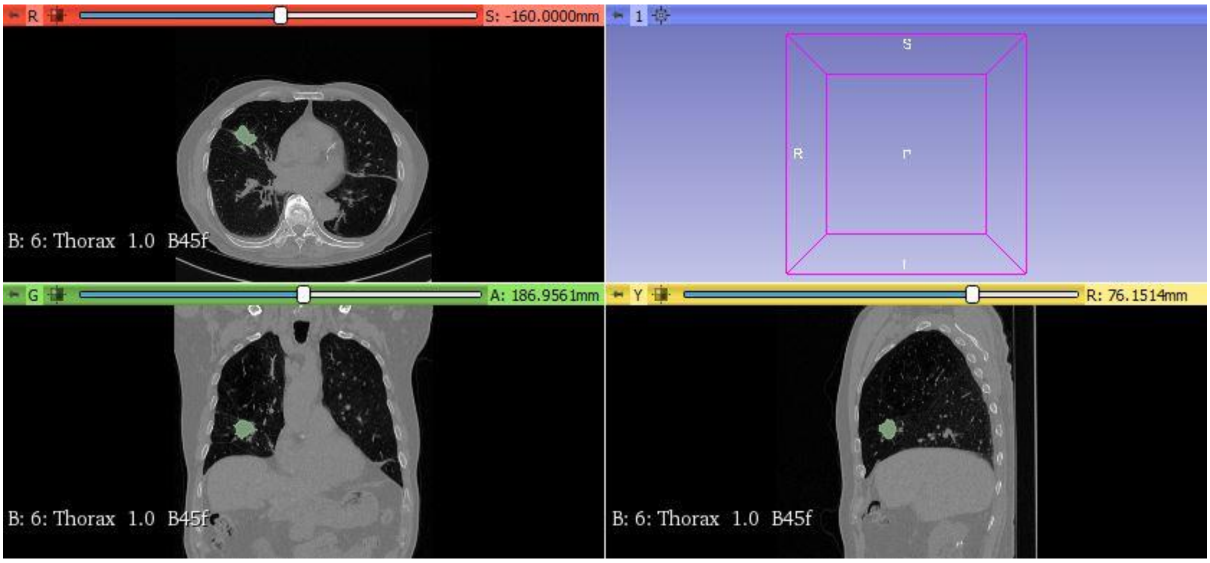

Figure 8. CT scan and the corresponding segmentation mask of the tumor of a patient with NSCLC obtained with the 3D slicer. The segmentation mask corresponds to the region with the green color and is superimposed on the CT scan. The axial view of the image is displayed in the top left figure, while the coronal and the longitudinal plane are depicted in the bottom left and right figure, respectively.

The transcriptomics data of this radiotranscriptomics dataset is the gene expression profiles that were extracted using the RNAseq technology, which is a high-throughput sequencing technology. This dataset consists of 22126 genes. To code the genes, each name was converted and mapped to the corresponding Entrez Gene ID. All the genes were identified by the unique Entrez Gene ID, except 2148 genes that do not map to an Entrez ID. Regarding the significant genes, 4 out of the 73 significant genes did not exist in this RNAseq dataset. The rest of the genes (69 genes) did not have values for all the 115 patients of the new radiotranscriptomics dataset. Thus, the genes that had values for at least the 70% of the number of patients were identified in order to be used for subsequent analysis. To this end, 41 genes satisfied this criterion, having values for at least 0.7*115 = 80 patients. A similar criterion was applied to the patients, ensuring that the patients used in the current analysis have values for at least the 70% of the number of genes. Hence, 112 out of 115 patients satisfy the criterion of having values for at least 0.7*41 = 28 genes.

The strategy for the genes' and patients' selection aimed to identify the genes and the patients that have values for a sufficient proportion of samples and genes, respectively. Thus, the dataset still contains missing values (i.e. Nan) for some genes. An imputation method was applied in order to complete the dataset and use it in machine learning algorithms. The R package missMDA [48] was applied in order to impute the missing values of the 41 genes. The missMDA package performs imputations using iteratively the Principal Component Analysis (PCA). More precisely, the first step of the algorithm is to assign initial values, such as the mean value of the variable across the observed values, to the missing entries, i.e. impute the missing entries. Then, the PCA algorithm is performed on the imputed dataset. A new value for the missing entry is estimated by identifying the value fitted by the PCA. The initial observed values remain the same; only the missing entry is updated by the value fitted by the PCA algorithm. The same procedure, which consists of the implementation of the PCA algorithm and the imputation of the missing entries by the values fitted by the PCA, is repeated until convergence. Thus, the missing entries are imputed and a complete dataset is derived.

Furthermore, the MIPCA (multiple imputation with PCA) method was implemented in order to perform multiple imputations generating several imputed datasets and assess the uncertainty of the predictions (i.e. imputed values). The observed values remain the same across the multiple imputations, but the imputed values of the missing entries differ. Several graphs were produced in order to evaluate the variability of the imputed values.

3.5. Validation of the regression models

The radiomic and the transcriptomics features had been simulated using the values of the dataset GEO28827. This dataset contains genes expression profiles derived from the DNA microarray technology. More specifically, 51 radiomic features and 23 genes had been modeled using linear combinations of the genes expression profiles and radiomic features, respectively, with dataset GEO28827. These regression models, which describe associations between the two modalities, had been extracted in a dataset with DNA microarray data. However, the RNAseq technology presents several advantages and has prevailed over the DNA microarray technology as next-generation sequencing technology. To this end, we aim to validate quantitatively the modeling of the radiomic and the transcriptomics features in an RNAseq dataset and derive the new quantitative relationships between these two modalities in the RNAseq dataset. The RNAseq technology is a sequence-based technique while the DNA microarray technology depends on the hybridization. Thus, the level of the expression profiles may be different between these two technologies. The pipeline for the validation of the regression models of the radiomic and the transcriptomic features is presented in Figure 9.

The regression models of the 51 radiomic features, which had been derived from the DNA microarray GEO28827 dataset, were based on the values of the 73 genes of the transcriptomics signature. Each radiomic feature was modeled based on a specific subset of genes, which was different for each feature, during the previous analysis. However, only the 41 out of the 73 genes were identified in the new RNAseq dataset GSE103584 due to the existence of many Nan values in the other genes. Hence, for each of the 51 radiomic features, we identified which of the genes-predictors exist also in the new dataset. The percentage of the predictors that exist also in the new dataset was calculated for each model. The models that had the 60% of their predictors in the new dataset, were used for subsequent analysis to validate quantitatively the relationship between the 2 modalities. For instance, if a model needed 10 genes to predict the radiomic feature, then this model will be examined in the new dataset, only in the event that at least 6 out of the 10 genes exist in the new dataset. Hence, 25 out of the initial 51 models satisfy this criterion in the new dataset. The modeling of the 25 radiomic features from their genes-predictors was investigated in the new RNAseq dataset.

Regarding the inverse modeling of the transcriptomics features in terms of radiomic features, 23 genes were identified that can be predicted from the radiomic features in the GEO28827 dataset. However, 10 out of the 23 genes did not exist in the new RNAseq dataset. Hence, the models of the 13 genes were examined in the new dataset. All the radiomic features that had been used as predictors for the models exist in the new dataset. Hence, the modeling of the 13 genes from their radiomics-predictors was investigated in the new RNAseq dataset.

The models of the 25 radiomic and the 13 transcriptomics features were validated using linear regression in the RNAseq dataset. The R-squared metric was used to assess the accuracy of the predictive models. However, the models obtained from the linear regression resulted in poor performance. Hence, the polynomial regression was used to evaluate whether the radiomics and the transcriptomics features can be simulated from their same transcriptomics and radiomics predictors, respectively, also in the RNAseq dataset. The degree of the polynomial regression was set equal to 2 in order to prevent overfitting. The root mean squared error (RMSE) was computed in order to evaluate the predicted models of the radiomics and the transcriptomics in the new dataset. The models that achieved RMSE <= 0.60 were considered that they simulated efficiently the feature. These models were used for subsequent analysis. Additionally, the Pearson correlation coefficient and the corresponding p-value between the actual and the predicted values of each feature were calculated as evaluation metrics.

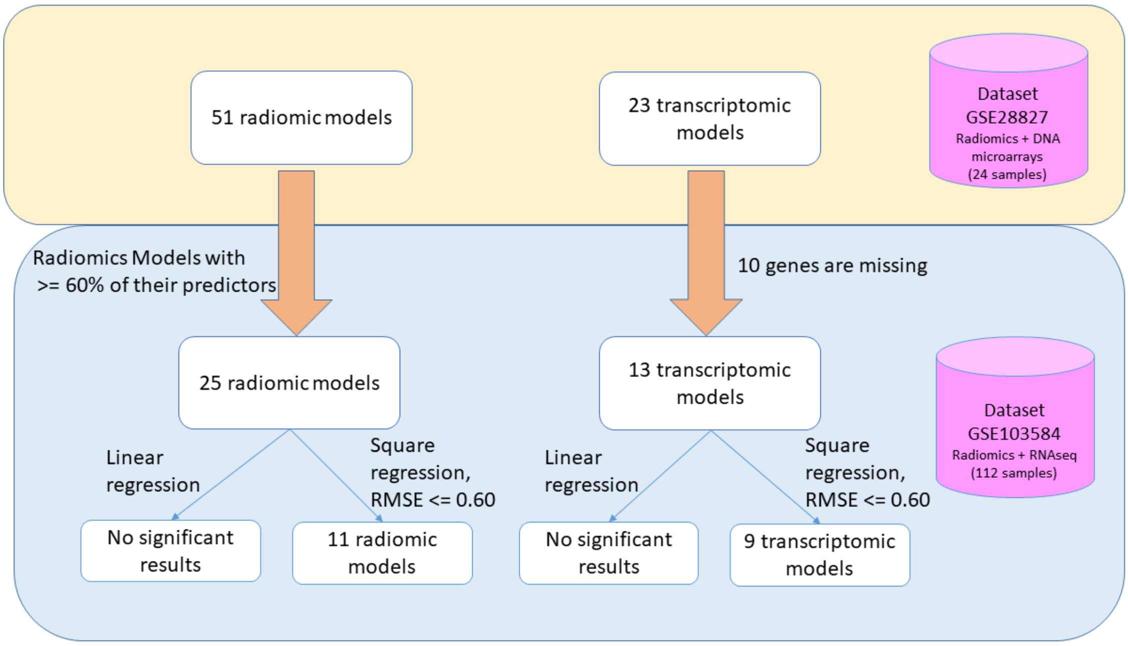

Figure 9. Pipeline for the validation of the regression models of the radiomic and the transcriptomic features. The regression models had been derived using the DNA Microarray dataset GEO28827 (area with orange color in the figure) and the validation is performed on the RNAseq dataset GSE103584 (area with blue color in the figure).

3.6. Lung cancer staging classification

After extracting the radiomics and the transcriptomics features that can be simulated using the RNAseq dataset, we proceeded with the evaluation of the ability of several radiomics, transcriptomics and radiotranscriptomics markers to predict the lung cancer staging. The diagnosis of the lung cancer staging plays crucial role for the treatment planning and consequently the patient's outcome.

The lung cancer staging was expressed in the system of T, N, M staging [49]. The T (tumor) stage describes the size and location of the tumor, 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. The combination of the status of these three descriptors determines the final stage of the tumor, which is expressed with Roman numerals from 0 to IV according to the American Joint Committee on Cancer (AJCC) TNM system. The dataset GSE103584 was used for the implementation of the current task due to the fact that it contains radiomic and transcriptomic features. The dataset contains patients from all stages, but it is imbalanced as there is no equal proportion of patients from each cancer stage (Table 3). Furthermore, the number of patients is restricted to 112. Thus, the stages were divided in two broad categories, which are the early stage and the late stage. The patients who have lung cancer of stages 0 and I are categorized as patients in early stage, while the patients with stages II, III and IV are categorized as patients with late stage. Hence, there are 75 patients with early stage cancer and 37 patients with late stage cancer.

Two state-of-the-art classifiers were implemented in order to evaluate the potential of the single and the joint markers of the 2 modalities in lung cancer staging prediction. The SVM linear classifier and the Random Forest Classifier with 100 trees were used for this task. However, the dataset remains imbalanced, since it contains 75 samples from the one class and 37 samples for the other class. Thus, there are few samples of the minority class, which is the late stage class, to effectively learn the decision boundary. For this reason, the Synthetic Minority Oversampling Technique (SMOTE) [50] was implemented in order to create a balance dataset. SMOTE is an oversampling technique where new samples are generated for the minority class, achieving balance class distribution. More specifically, a random sample of the minority class is selected and the 5 nearest neighbors of the sample are found. Then, a neighbor is randomly selected and a line between these samples is drawn in the feature space. Finally, a synthetic sample is created by randomly selecting a sample at a point along this line. Furthermore, the values of the features that comprise the feature vector were standardized in order to have a mean value equal to 0 and standard deviation equal to 1.

A 3-fold stratified cross validation was used in order to train and test the classifiers due to the small sample size and the lack of an external dataset. This procedure was repeated for 30 times in order to gain a more comprehensive overview of the performance of the classifiers. In order to evaluate the performance of the classifiers, the average values and the standard deviation of the values of the accuracy, the sensitivity and the specificity were calculated.

The two classifiers were implemented using various feature vectors in order to assess the diagnostic ability of the derived radiomic and transcriptomic features as well as of joint signatures.

Several feature vectors were used for the classification of the lung cancer staging in order to investigate their impact on the prediction. Thus, the actual values as well as the predicted values of the radiomic and the transcriptomics features were used for the classification. Furthermore, joint signatures, purely radiomic and purely transcriptomic signatures were used separately in order to assess whether they enhance the accuracy of the prediction or not.

More precisely, the feature vectors, in which the actual values of the features were used, are:

- 41 genes: 41 initial genes that were extracted from the old dataset GEO28827
- 51 radiomics: 51 initial radiomics that were extracted from the old dataset GEO28827
- Combination initial: the combination of the 41 genes + the 51 radiomics
- 9 Selected genes: the actual values of the 9 genes that can also be modeled from radiomics in the new Dataset GSE103584
- 11 Selected Radiomics : the actual values of the 11 radiomics that can also be modeled from genes in the new Dataset GSE103584
- Combination Selected: the actual values of the 9 selected genes + the 11 selected radiomics

The feature vectors, in which the predicted values or combinations of predicted and actual values of the features were used, are:

- 9 genes: the predicted values of the 9 genes that can also be modeled from radiomics in the new Dataset GSE103584
- 11 Radiomics: the predicted values of the 11 radiomics that can also be modeled from genes in the new Dataset GSE103584
- Combination: the predicted values of the 9 genes + 11 radiomics
- Gene signature 1: the actual values of the 41 initial genes + the predicted values of 11 radiomics, which are derived from combination of genes
- Gene signature 2: the actual values of the 9 selected genes + the predicted values of 11 radiomics, which are derived from combination of genes
- Radiomics signature 1: the actual values of the 51 initial radiomics + the predicted values of 9 genes, which are derived from combination of radiomics
- Radiomics signature 2: the actual values of the 11 selected radiomics + the predicted values of 9 genes, which are derived from combination of radiomics

Table 3. Number of patients for each lung cancer stage in dataset GSE103584. The stages, which are marked with green color, are grouped as early stage and the stages marked with orange color are grouped as late stage.

Staging	No. of patients
0	5
I	70
II	19
III	14
IV	4

Chapter 4: Research findings / results

4.1. Genes modelling based on radiomic features

The 73 genes were simulated using the values of the radiomic features from the old dataset GEO28827, which contains gene expression microarray data and CT radiomic features. The modelling of the genes based on the radiomic features was performed using two different approaches. The first approach, which is the direct inverse problem with matrices, resulted in 36 models that achieved normalized RMSE <= 1. Thus, 36 genes can be adequately predicted using the Moore-Penrose pseudoinverse of the weights matrix. The second approach, which is the indirect inverse problem with linear regression, resulted in 30 models that achieved R-squared >= 0.70. These 30 models correspond to 30 genes that can be modelled by specific radiomic features through linear regression. Several validity metrics were computed to assess the selected 30 models. All the models resulted in RMSE <= 0.3414, Pearson correlation coefficient >= 0.8417 and corresponding p-value <= 2.5437e-07, showing statistical significance (p-value <= 0.05). The validity metrics indicated that the derived models had values close to the actual values, strong relationship with the actual values and statistical significant results. The min and the max value of each of the validity metrics for the 30 genes-models are depicted in Table 4.

Table 4. Range of the values of the validity metrics for each of the 30 genes-models predicted by specific radiomic features using the linear regression approach.

VALIDITY METRIC	MIN VALUE	MAX VALUE
RMSE	3.2375e-08	0.3414
PEARSON CORRELATION COEFFICIENT	0.8417	1
P-VALUE OF PEARSON COEFFICIENT	1.0880e-173	2.5437e-07

The common genes between the 36 genes which had been identified by the modelling of the genes using the pseudoinverse matrix and the 30 genes which had been extracted by performing the linear regression algorithm, were considered that can be adequately predicted from radiomic features. Thus, 23 common genes were used for subsequent analysis. The Entrez gene IDs of these 23 genes are depicted in Table 5.

Table 5. Entrez Gene IDs of the 23 common genes that can be adequately predicted from radiomic features in Dataset GEO28827.

Entrez Gene ID

638	4796	9245	26232	79674
2177	5980	11142	27094	83933
2859	6878	23090	54825	84300
3026	7104	23414	63035	
3853	9244	26112	79035	

4.2. Validation of transcriptomics signatures

The transcriptomics signature which consists of 73 genes was validated for its ability to discriminate between malignant and normal lung tissues in three external datasets. The potential of these genes was investigated by performing classification tasks with an SVM linear classifier, boxplots, t-tests and volcano plots.

4.2.1. SVM classification

Several SVM linear classifiers were implemented based on single genes and combination of genes. The performance of each classifier was assessed by calculating specific evaluation metrics. The results are depicted in Table 6.

- Classifier 1 refers to the classifier that was trained on dataset GSE75037 and tested on the combination of the 3 new datasets.
- Classifier 2 refers to the classifier that was trained and tested on samples from the 3 new datasets.
- Classifier 3 refers to the classifier that was trained and tested on samples from the 3 new datasets utilizing the ComBat corrected gene expression profiles.
- Classifier 4 refers to the classifier that was trained on the 2 new microarray datasets and tested on the new RNAseq dataset.
- Classifier 5 refers to the classifier that was trained on the 2 new microarray datasets and tested on the new RNAseq dataset utilizing the ComBat corrected gene expression profiles.
- Classifier 6 refers to the classifier that used the 51 p-metaomics features (i.e. linear combination of genes), which were derived for the 2 new microarray datasets.

Table 6. Evaluation metrics of the performance of the SVM linear classifiers.

METRIC	ACCURACY	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY	GEOMETRIC MEAN
CLASSIFIER 1	86.10%	80.62%	98.28%	89.45%	89.01%
CLASSIFIER 2	97.67% ± 1%	98.18% ± 2%	96.54% ± 3%	97.33% ± 1%	97.31% ± 1%
CLASSIFIER 3	97.42% ± 1%	97.45% ± 2%	97.34% ± 3%	97.39% ± 1%	97.38% ± 1%
CLASSIFIER 4	89.02%	79.31%	**100%**	89.66%	89.06%
CLASSIFIER 5	98.17%	97.70%	98.70%	**98.20%**	**98.20%**
CLASSIFIER 6	**98.19% ± 1%**	**99.02% ± 1%**	94.67% ± 6%	96.84% ± 3%	96.76% ± 3%

4.2.2. Boxplots and t-tests

The Boxplots for the control and the tumor population of each dataset are presented in the following Figures 10-13. The distribution of the values of the positive and the negative genes are demonstrated in the boxplots.

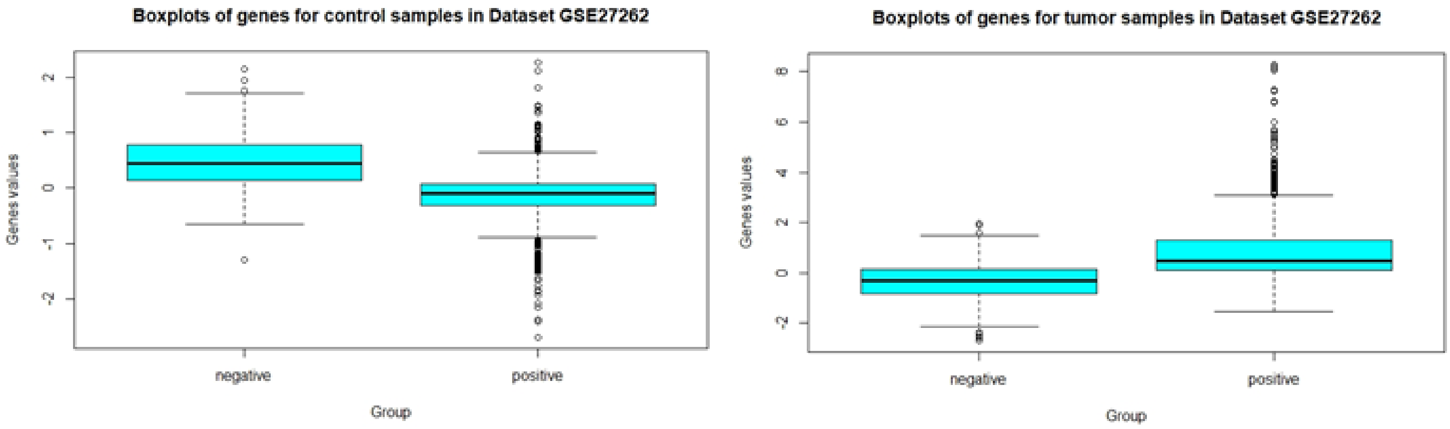

Figure 10. Boxplots of positive and negative genes for control and tumor samples in Dataset GSE27262 in left and right figure, respectively.

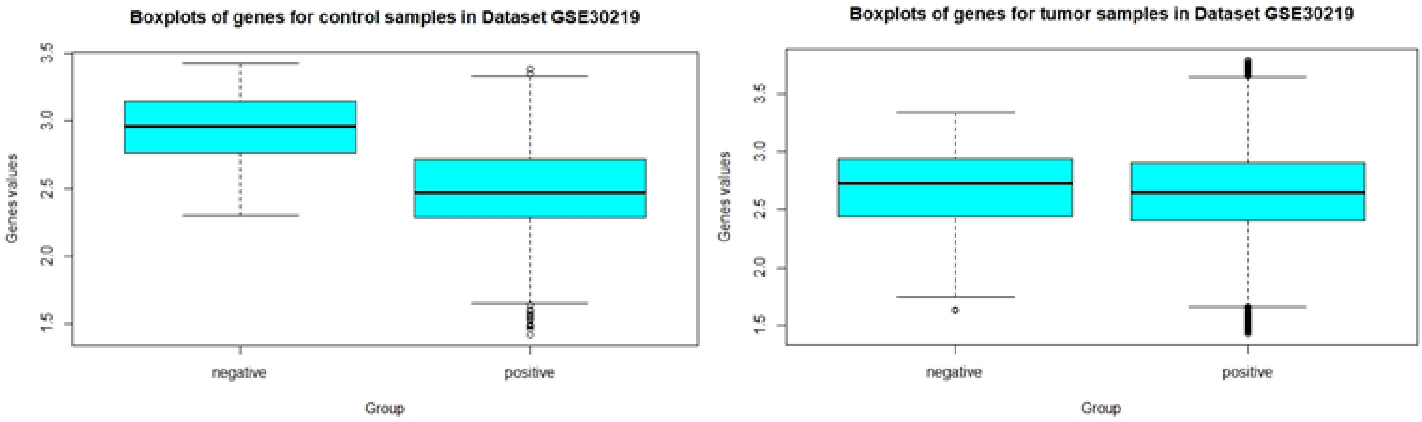

Figure 11. Boxplots of positive and negative genes for control and tumor samples in Dataset GSE30219 in left and right figure, respectively.

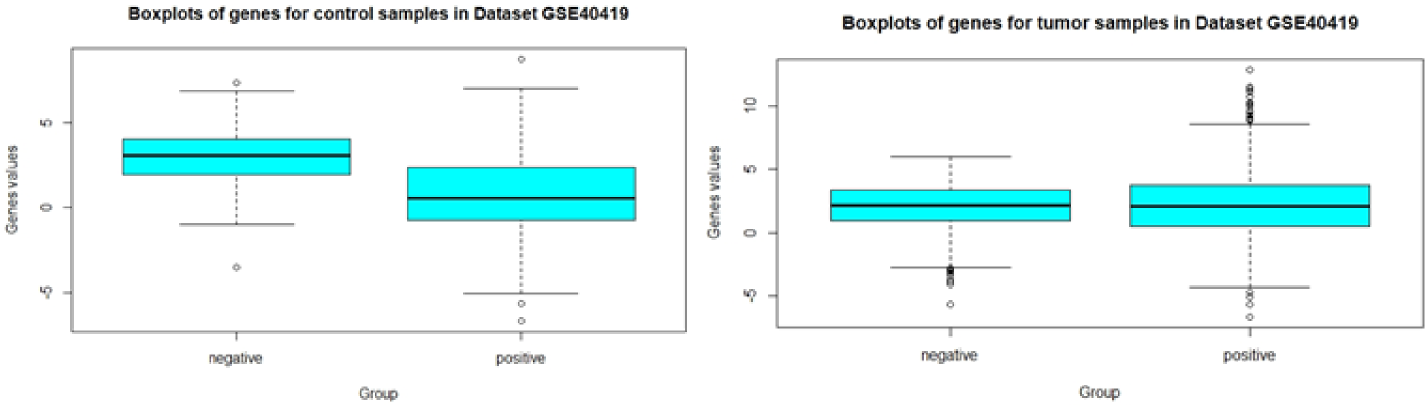

Figure 12. Boxplots of positive and negative genes for control and tumor samples in Dataset GSE40419 in left and right figure, respectively.

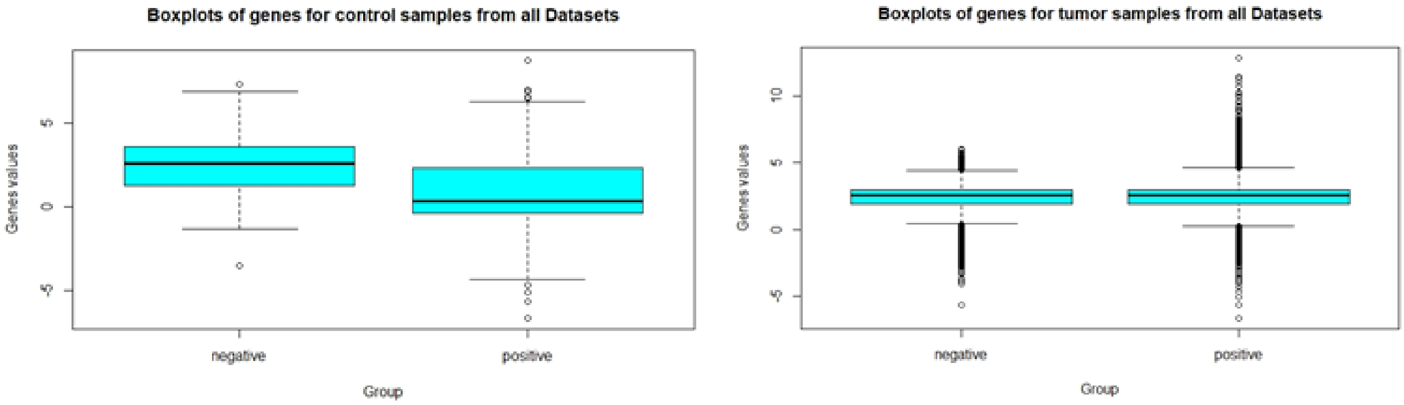

Figure 13. Boxplots of positive and negative genes for control and tumor samples in all datasets (GSE27262, GSE30219 and GSE40419) in left and right figure, respectively.

Two-sided t-tests were performed between the positive and the negative genes in each population, i.e. control or tumor, for each dataset. Hence, two t-tests were performed for each dataset and the results are depicted in Table 7.

Table 7. T-tests results between the positive and the negative genes in each population, control and tumor, of each dataset.

	GSE27262		**GSE30219**		**GSE40419**		**Combined dataset**	
	Control	**Tumor**	**Control**	**Tumor**	**Control**	**Tumor**	**Control**	**Tumor**
Mean of positive	0.95	3.79	5.78	6.50	4.52	16.52	3.9	9.62
Mean of negative	1.49	0.89	7.80	6.52	14.91	7.77	11.16	6.39
p-value	2.2e-16	8.8e-12	2.2e-16	0.6825	2.2e-16	6.647e-07	2.2e-16	6.809e-08
t	-15.76	6.8763	-17.94	-0.40	-16.69	4.97	-16.41	5.39

4.2.3. Volcano Plots

The Volcano plots are depicted in Figures 14-16. The Volcano plot presents the genes that are not significant with gray color, the genes that have p-value < 0.01 with blue color and the genes that correspond to fold change > 2 with green color. The significant genes that satisfy both criteria, i.e. p-value < 0.01 and fold change > 2, are depicted with red color in the graph.

In the microarray datasets GSE27262 and GSE30219, 30 genes and 25 genes out of 73, respectively, were identified as significant from the Volcano plots. In the Volcano plot of the RNAseq dataset GSE40419, 44 genes out of 73 were identified as significant. However, there are 16 common significant genes from the volcano plots of GSE27262, GSE30219 and GSE40419, which are presented in Table 8.

Table 8. Entrez Gene IDs of the common genes that were identified as significant (p-value < 0.01 and FC >2) in all the volcano plots of each dataset.

Entrez Gene ID

699	3161	9244	202374
1290	3866	10615	443
2118	6690	80201	11142
3026	8612	196410	23090

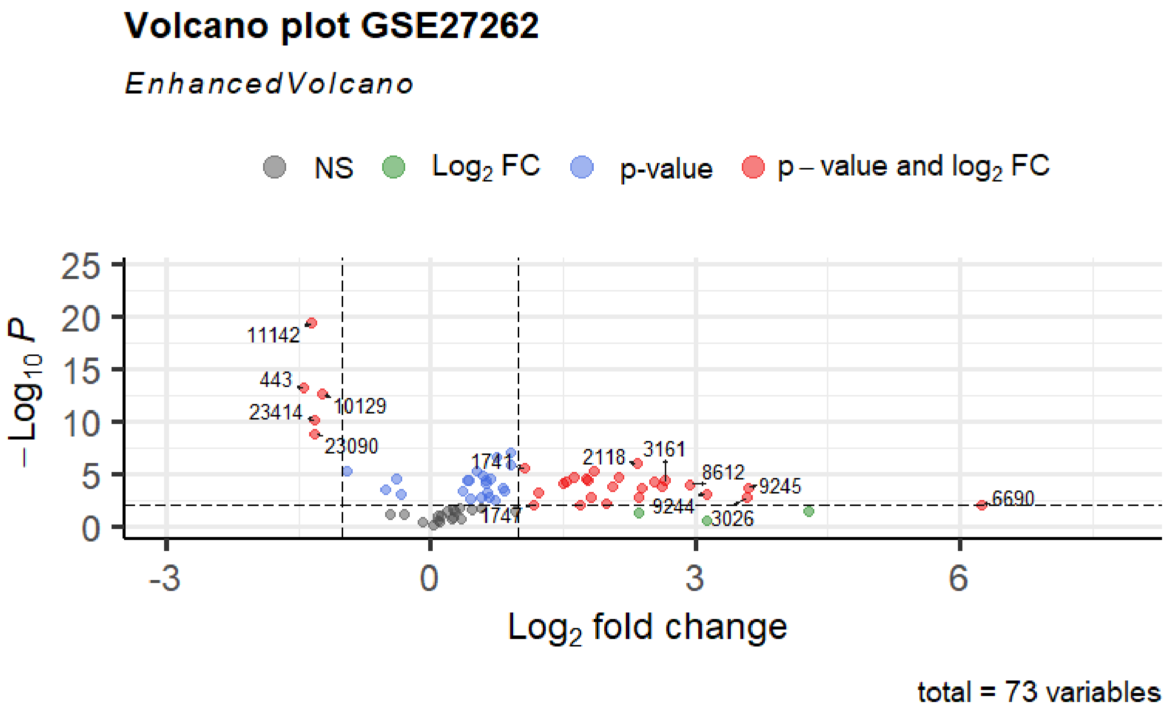

Figure 14. Volcano plot of GSE27262 dataset.

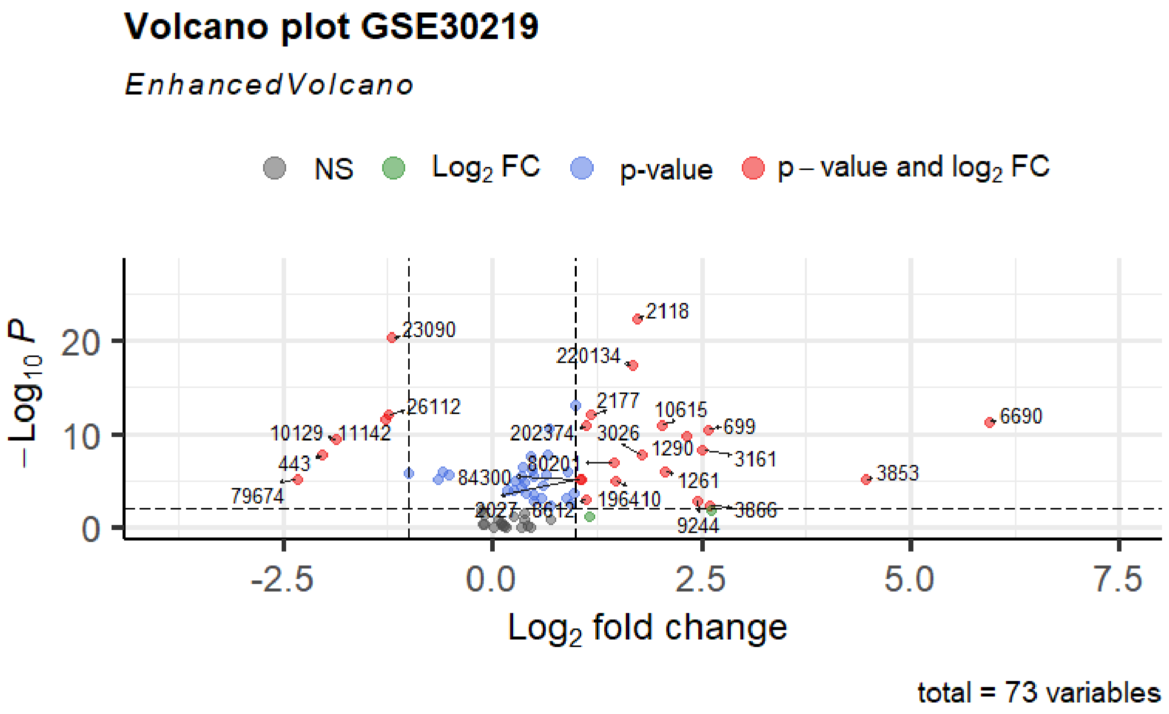

Figure 15. Volcano plot of GSE30219 dataset.

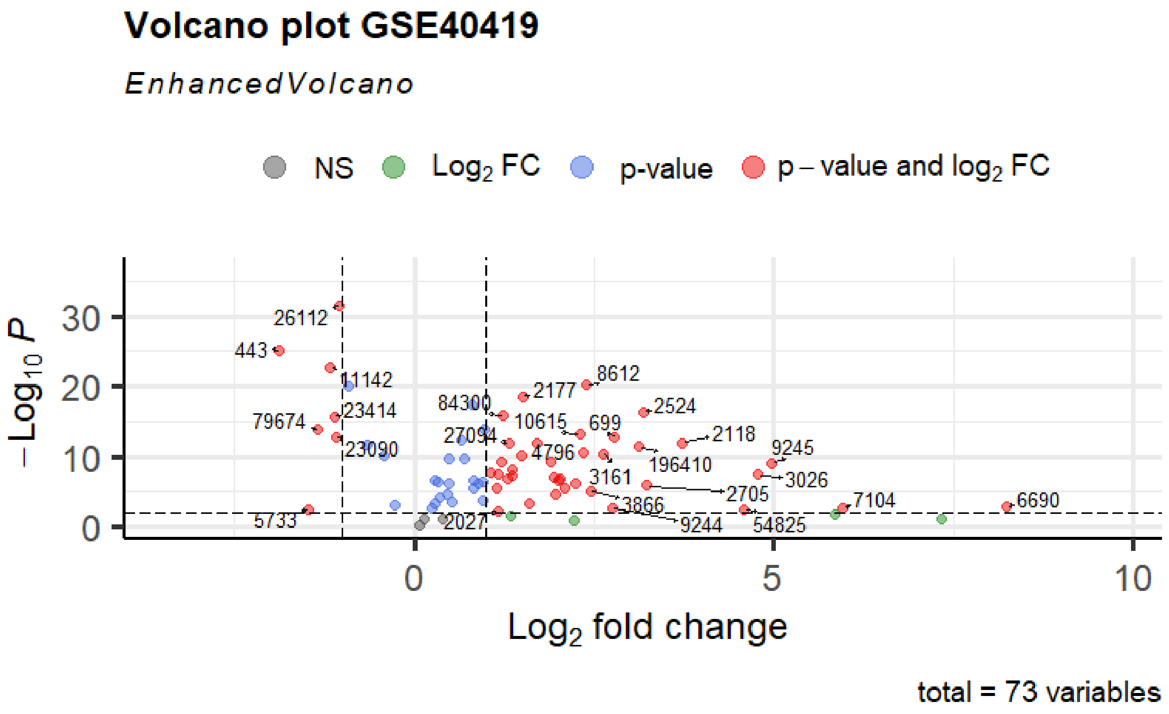

Figure 16. Volcano plot of GSE40419 dataset.

4.3. Genes Missing Values Imputation

In the RNAseq dataset (GSE103584), which contains the transcriptomics data of the new radiotranscriptomics dataset, many genes from the transcriptomics signature have missing entries for some patients. Thus, 41 genes satisfied the criteria and selected for subsequent analysis. The Entrez Gene IDs of the 41 genes are presented in Table 9.

Table 9. Entrez Gene IDs of the 41 genes of the GSE103584 dataset.

Entrez Gene ID					
142	3866	10045	54993	3603	79674
699	4157	10615	55311	5980	
1290	4585	22874	63035	10129	
1741	4796	23534	83933	11142	
2118	6878	23780	114907	23414	
2177	7477	25894	116092	25934	
2524	8347	26232	153768	26112	
3161	9245	27094	202374	55277	

The missMDA package was applied to impute the missing values of the genes. Moreover, the MIPCA was applied to assess the uncertainty of the predictions of the imputed values. The multiple imputed datasets that were produced from MIPCA, were projected on the mean imputed dataset (i.e. reference dataset). The MIPCA derived the graphs that are depicted in Figure 17. The cycles, the ellipses and the clouds represent the confidence area of the predictions. The graphs in figure 17A and 17B are derived after applying PCA to each imputed dataset. The representation of the individuals (i.e. patients) that are obtained after applying Procrustes rotations are presented in Figure 17A. In Figure 17B, the first two principal components of each imputed dataset are projected onto the first two principal components of the reference dataset. The variability of the position of the individuals (patients) and the variables (genes) obtained by each imputed dataset are depicted in Figure 17C and 17D, respectively. The confidence area of the individual 14 is restricted to 1 point in Figure 17C, as this individual does not have missing values. However, the confidence area of the individual 14 is not restricted in 1 point in Figure 17A, since the PCA components are not identical from one PCA of an imputed dataset to another PCA of another imputed dataset due to the different imputed values of the other individuals.

The small variability of the two principal components in Figure 17B indicates that the uncertainty in the predictions is small. This assumption is confirmed also by the small variability in the confidence areas of the position of the individuals, which are represented by cycles and ellipses in Figure 17A and 17C, and the small variability of the predictions of the variables, which are represented by the clouds in Figure 17D. Hence, there is no significant uncertainty in the predictions of the missing entries of the dataset and the use of the imputed dataset for the analysis is encouraged.

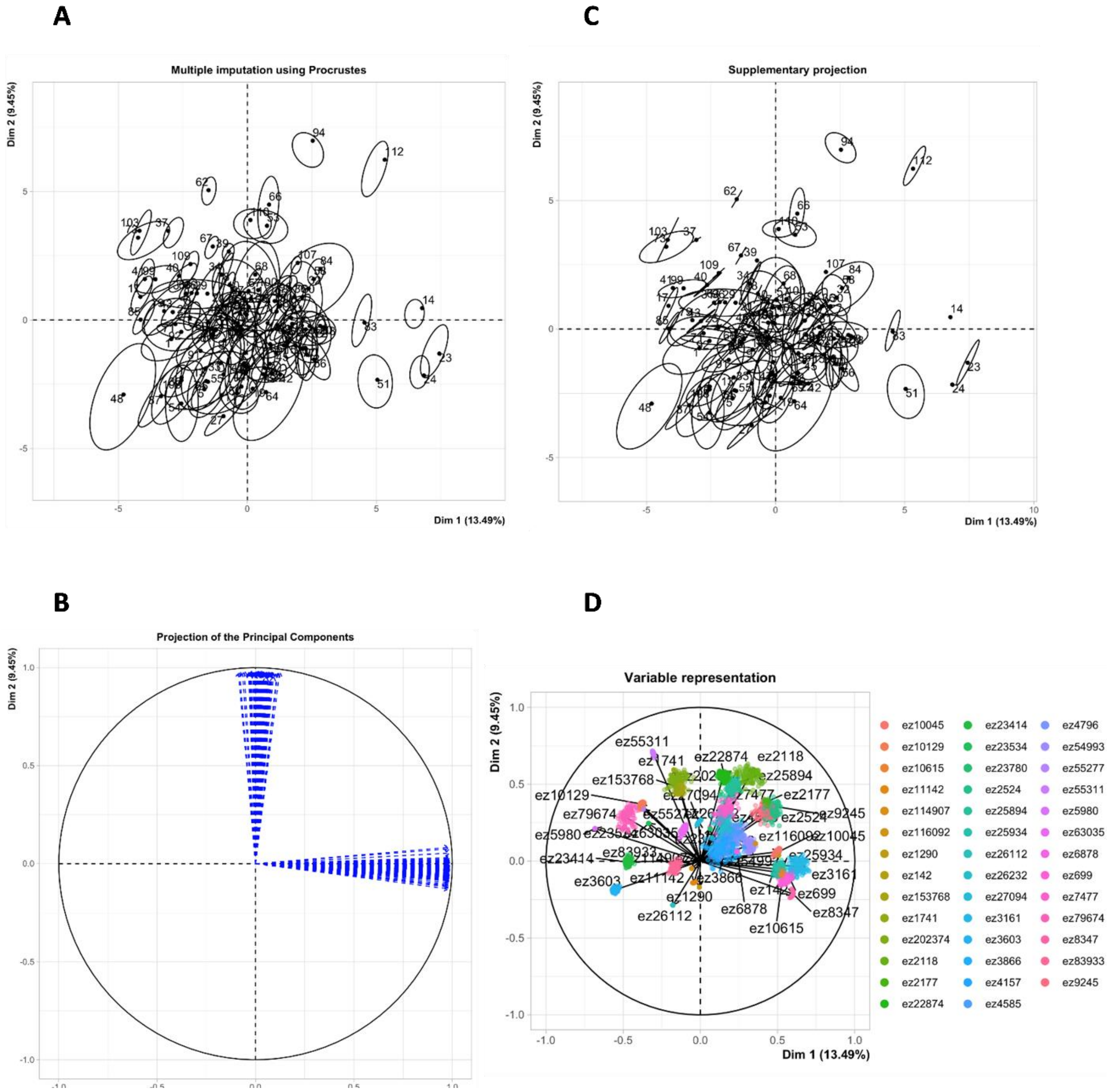

Figure 17. Graphs derived from the MIPCA package after applying multiple imputations to assess the uncertainty of the predictions. ***A.*** *Procrustes rotations to obtain individual graph after applying PCA to each imputed dataset.* ***B.*** *Projection of the 2 principal components of each imputed dataset on the first two principal components of the reference dataset after applying PCA to each imputed dataset.* ***C.*** *Projections of the individuals (patients) of the imputed datasets* ***D.*** *Projections of the variables (genes) of the imputed datasets.*

4.4. Validation of the regression models

The 51 radiomic features had been simulated based on a subset of genes that were identified in the previous analysis using the GEO28827 dataset. However, the 25 out of the 51 radiomic features satisfied the criterion of preserving at least the 60% of their genes-predictors in the RNAseq dataset GSE103584. Hence, for each radiomic feature, the subset of genes that were used as predictors in the previous analysis, were also exploited in the current work. Linear regression was performed using each radiomic feature as dependent variable and the corresponding subset of genes as independent variables-predictors. Thus, we aimed to validate the relationships between the radiomic and the transcriptomics features in an RNAseq data. However, none of the 25 models resulted in good performance (R-squared < 0.32). To this end, square regression was performed to the 25 radiomic features using the same genes-predictors. With this algorithm, 11 out of 25 models demonstrated RMSE <= 0.60. Hence, 11 radiomic features can be predicted through second degree polynomial combinations of the transcriptomics data in the RNAseq dataset. The relationship between only 11 radiomic features and their corresponding genes-predictors was preserved in the RNAseq data. However, more complex combinations of the genes-predictors were required. The 11 selected radiomic features are depicted in Table 10.

All the 11 models achieved RMSE <= 0.60, Pearson correlation coefficient >= 0.81 and p-value <= 2.47e-27, showing statistical significant results (p-value <= 0.05). The range of the values of the validity metrics of the 11 radiomic models are depicted in Table 11.

Table 10. The 11 radiomic features that can be predicted from transcriptomics data in the RNAseq dataset.

Radiomic Feature	
'''log_1_original_glcm_Autocorrelation'''	'''grad_1_original_gldm_LargeDependenceLowGrayLevelEmphasis'''
'''sq_1_original_glcm_DifferenceAverage'''	'''log_1_original_glrlm_RunPercentage'''
'''wavelet_1_original_glszm_SmallAreaLowGrayLevelEmphasis'''	'''original_1_original_firstorder_RootMeanSquared'''
'''sqrt_1_original_firstorder_Minimum'''	'''log_1_original_glcm_ClusterShade'''
'''log_1_original_ngtdm_Busyness'''	'''log_1_original_glcm_JointEnergy'''
'''original_1_original_glcm_Autocorrelation'''	

Table 11. Range of the values of the validity metrics for each of the 11 radiomics-models predicted by specific transcriptomics features using the square regression approach in the RNAseq dataset.

VALIDITY METRIC	MIN VALUE	MAX VALUE
RMSE	2.64e-15	0.58
PEARSON CORRELATION COEFFICIENT	0.81	1
P-VALUE OF PEARSON COEFFICIENT	0	2.47e-27

Regarding the modelling of the transcriptomics data, 23 genes had been simulated from a subset of radiomic features using the GEO28827 dataset. However, 10 out of the 23 genes did not exist in the new dataset. Thus, linear regression between the 13 genes and their corresponding radiomics-predictors were performed in order to validate their relationship in the RNAseq data. Similarly to the aforementioned radiomic models, none of the 13 models could simulate the genes efficiently (R-squared < 0.27). Thus, square regression was performed between each gene and the corresponding radiomics-predictors in order to assess whether the gene can be modelled, even with more complex associations, from the corresponding radiomic features in an RNAseq dataset. The use of the polynomial regression with degree equal to 2 resulted in 9 out of 13 models that achieved RMSE <= 0.60. Thus, there are 9 genes that can be simulated from their corresponding radiomic features in the RNAseq data. The Entrez Gene IDs of these 9 genes are depicted in Table 12.

The validity metrics, which were used for the assessment of the predicted radiomics models, were also used for the evaluation of the predicted transcriptomics models. All the 9 models achieved RMSE <= 0.60, Pearson correlation coefficient >= 0.80 and p-value <= 3.22e-27, showing statistical significant results (p-value <= 0.05). The range of the values of the validity metrics of the 9 transcriptomics models are depicted in Table 13.

Table 12. The 9 transcriptomics features that can be predicted from radiomics data in the RNAseq dataset.

Entrez Gene ID	
2177	26232
4796	27094
6878	63035
23414	79674
26112	

Table 13. Range of the values of the validity metrics for each of the 9 transcriptomics-models predicted by specific radiomics features using the square regression approach in the RNAseq dataset.

VALIDITY METRIC	MIN VALUE	MAX VALUE
RMSE	1.59e-14	0.58
PEARSON CORRELATION COEFFICIENT	0.80	1
P-VALUE OF PEARSON COEFFICIENT	0	3.22e-27

4.5. Evaluation of Lung Cancer Staging Classification

The performance of the SVM linear classifier and the Random Forest Classifier using various feature vectors was assessed quantitatively by computing the accuracy, the sensitivity and the specificity of the classifier. The average value of each metric is depicted in the following barplots (Figures 18-20) along with the error bar, which reflect the standard deviation of the values of the evaluation metric.

The Random Forest Classifier achieved slightly better performance than the SVM linear classifier in all metrics in most examined cases. All the classifiers using different feature vectors demonstrated good performance in predicting the lung cancer staging. More specifically, the Random Forest classifiers had similar performance independently of their feature vector in terms of accuracy (~ 70-75%), sensitivity (~70-75%) and specificity (~75-80%). The SVM linear classifiers demonstrated slightly worse performance, achieving accuracy ~65-70%, sensitivity ~ 60-65% and specificity ~ 70-80%.

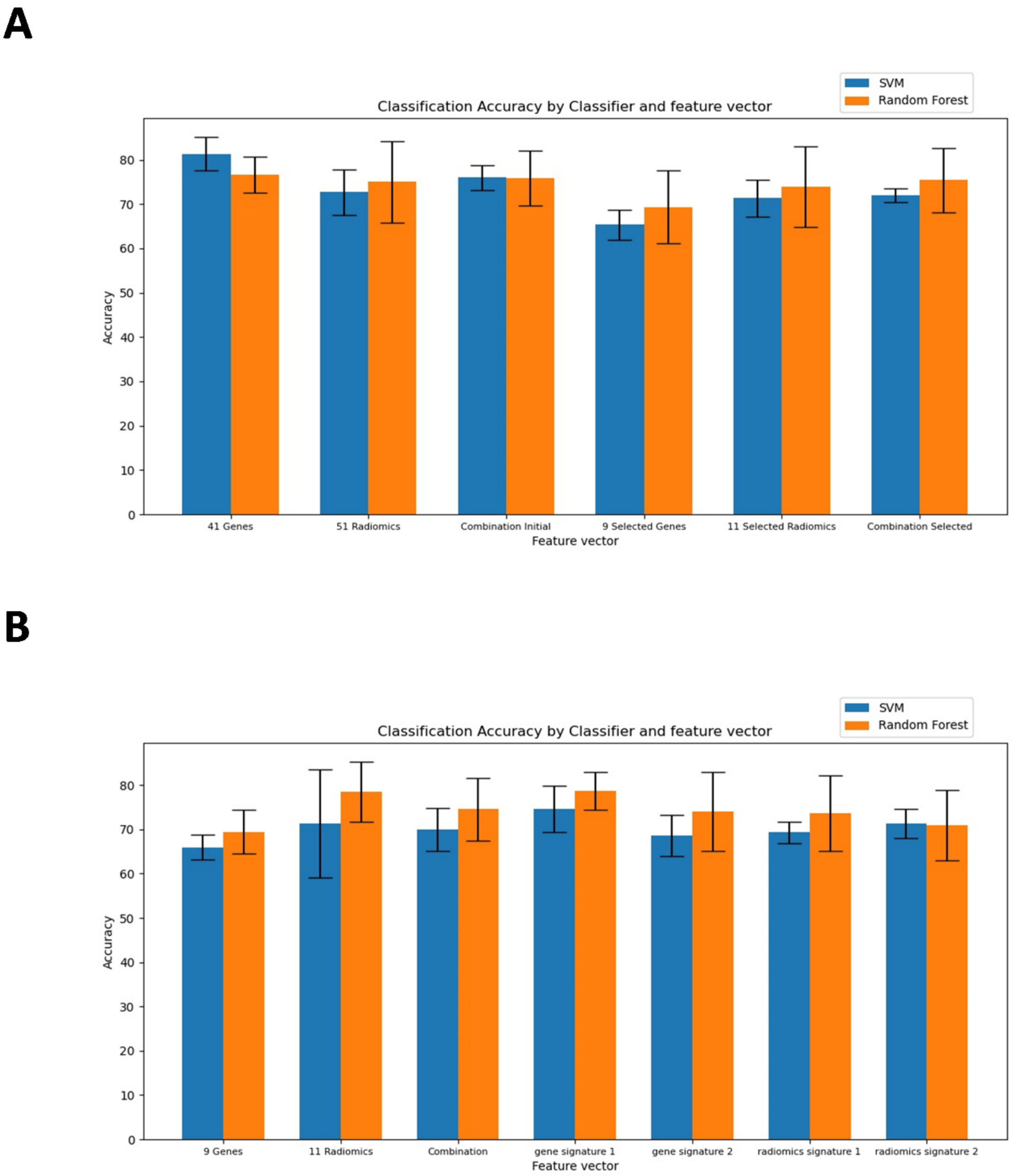

Figure 18. Barplots of the accuracy of the SVM linear and Random Forest classifiers. The black lines represent the error bars. ***A.*** *The accuracy of the classifiers that use actual values of the features;* ***B.*** *The accuracy of the classifiers that use predicted values or combinations of predicted and actual values of the features.*

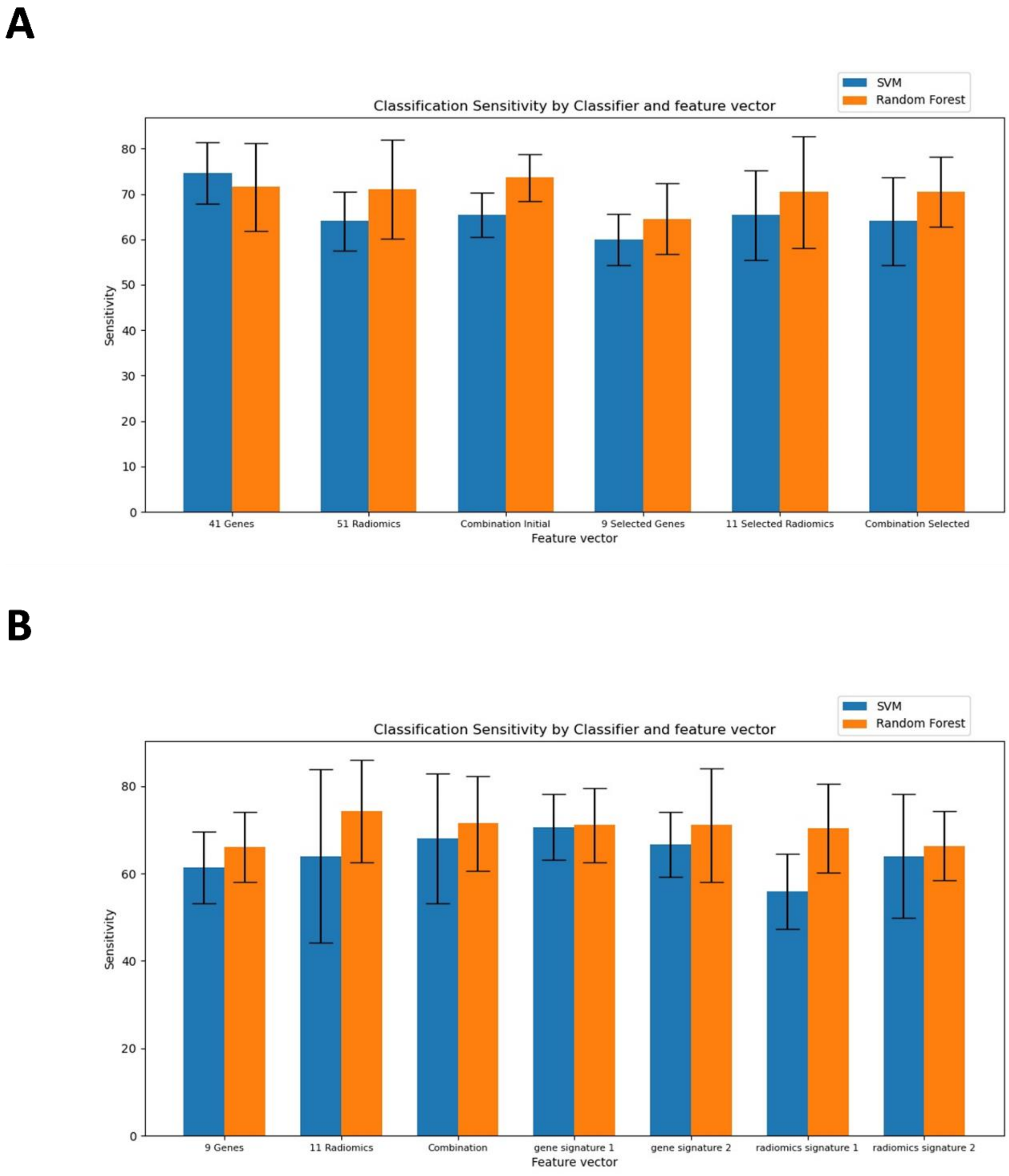

Figure 19. Barplots of the sensitivity of the SVM linear and Random Forest classifiers. The black lines represent the error bars. ***A.*** *The sensitivity of the classifiers that use actual values of the features;* ***B.*** *The sensitivity of the classifiers that use predicted values or combinations of predicted and actual values of the features.*

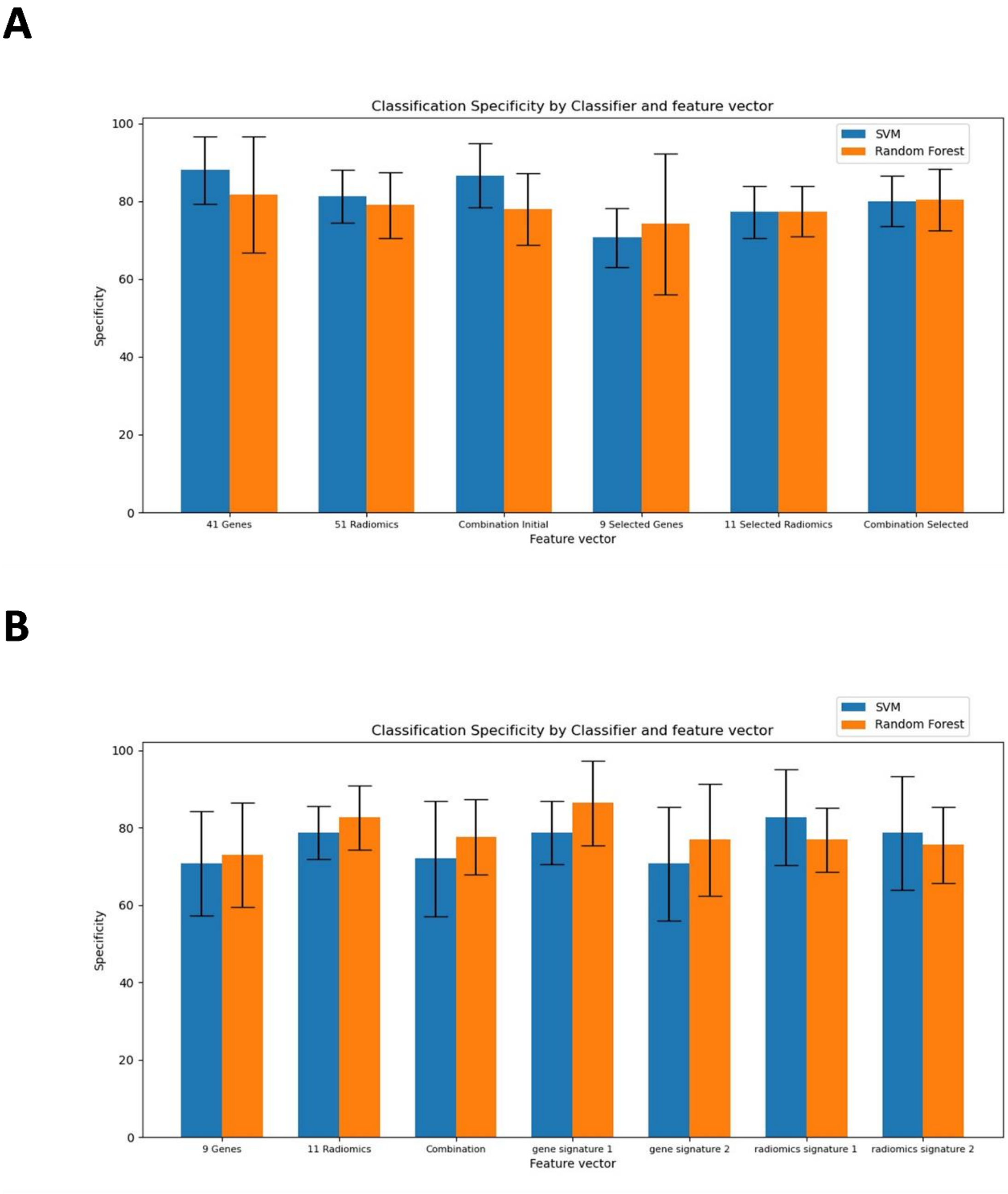

Figure 20. Barplots of the specificity of the SVM linear and Random Forest classifiers. The black lines represent the error bars. ***A.*** *The specificity of the classifiers that use actual values of the features;* ***B.*** *The specificity of the classifiers that use predicted values or combinations of predicted and actual values of the features.*

Chapter 5: Discussion

In the current study, the connection between the radiomic features extracted from the tumor region of CT scans and the transcriptomics features was investigated in order to identify descriptive signatures that have the potential to predict the NSCLC tumor staging. Transcriptome, i.e. the subset of genes that are expressed in a given tissue under specific conditions, links genotype to phenotype. The radiomic features are extracted from the non-invasive CT examination; also, the transcriptome conveys significant information about the tumor appearance and the clinical outcome [51]. Thus, the investigation of the interplay between these two modalities is significant in order to understand the underlying biological mechanism.

The 73 genes that had been extracted from the previous analysis, constitute the transcriptomics signature and is used for further investigation in the current study. From the analysis in the previous study, these genes demonstrated significant statistical and predictive associations with radiomic features and discrimination ability between malignant and adjacent normal lung tissues. The genes simulated 51 radiomic features that characterize the shape and the texture of the tumor (Table 2). The present study initiates with the implementation of the inverse regression direction, which is the modeling of the genes from the radiomic features, using the old dataset GEO28827. The radiomic features are extracted from the CT scan of the patient, which is a non-invasive examination. Thus, the identification of simple linear combinations of radiomic features that can simulate the expression of the genes, would be beneficial for the non-invasive characterization of the tumor. The analysis resulted in 23 genes that can be simulated from radiomic features.

However, the 73 genes had been evaluated for their differential ability between malignant and adjacent normal tissues using a small external dataset. The datasets used for the extraction and the validation of the genes contained DNA microarray data. Thus, these genes should be evaluated further for their diagnostic ability in larger DNA microarray and RNAseq datasets. To this end, two DNA microarray datasets and one RNAseq dataset were used to validate the discrimination potential of the 73 genes in characterizing a sample tissue as malignant or normal. All the examined classification schemas resulted in high performance achieving accuracy > 86%, sensitivity > 79%, specificity > 94%, balanced accuracy > 89% and geometric mean > 89%. More precisely, the overall performance of the classifier that had been trained in the GSE75037 dataset and tested in all the three new datasets (i.e. classifier 1) was good, discriminating the tumor from the non-malignant samples. However, a small portion of tumor samples was misclassified as non-malignant. This may happen due to the fact that the training set did not contain squamous cell carcinoma samples, while the test set contained samples from this class. To this end, the three new datasets were combined and split into training and test set for the implementation of a new SVM linear classifier (i.e. classifier 2). Additionally, the same classifier was implemented with the only difference that the ComBat harmonization technique was applied to the gene expression profiles (i.e. classifier 3). The high performance of the classifier using the ComBat corrected gene expression profiles confirmed the discrimination ability of the 73 genes (i.e. transcriptomics signature) between tumor and non-malignant samples. However, the performance of the classifier was high with either normalization or ComBat harmonization. In the case of the classifiers that had been trained in the DNA microarray datasets and tested in the RNAseq dataset (i.e. classifiers 4 and 5), the classifier 5 had better performance than classifier 4, indicating that the ComBat harmonization improved the diagnostic accuracy of the classifier. Hence, the 73 genes, which was purely investigated in DNA microarray datasets, could discriminate the malignant from the adjacent normal tissues not only in a DNA microarray dataset, but also in an RNAseq dataset. Furthermore, the discrimination ability of the combination of genes were examined in the new DNA microarray datasets. The 73 genes were combined through the appropriate linear equations in order to produce the 51 p-metaomics features in the two DNA microarray datasets. The performance of the classifier 6, which use the 51 p-metaomics features, was high and almost identical to the performance of the classifiers that use single genes. Thus, the combinations of genes that simulate the behavior of the radiomic features had similar impact on the lung cancer detection with the single genes. This may pave the way to investigate the impact of the combination of genes and the single genes in a more challenging task, such as the prediction of the tumor aggressiveness and the cancer prognosis.

Furthermore, the differentiation ability of the 73 genes were further examined implementing the boxplots and the t-tests. The boxplots and the t-tests showed that there was obvious difference in the distribution of the positive and the negative genes in the control population in all cases. As shown in the boxplots of all datasets (Figures 10-13), the median value of the distribution of the values of the negative genes is higher than the median value of the distribution of the values of the positive genes in the control population. This confirms the fact that the negative significant genes have higher values than the positive significant genes in the normal samples. Additionally, the p-value was less than 2.2e-16 in the control population of all the examined datasets, indicating that the mean value of the positive genes differs significantly from the p-value of the negative significant genes. Hence, it is validated in external datasets that the negative genes have significantly higher values than the positive genes in the control population. However, this difference was less obvious in the tumor population in most cases (apart from GSE27262) due to the outliers. In the boxplots, the median value of distribution of the values of the positive significant genes was at the same level with the median value of the negative significant genes. The visual inspection of the distributions with the boxplots may be obstructed due to outliers and batch effects. However, the t-tests resulted in statistically significant difference (p-value < e-12) between the mean values of these 2 groups (positive and negative) in tumor population, except for GSE30219 (p-value=0.6825). Hence, the boxplots and the t-tests in the new datasets in most cases confirm the initial finding that the positive and the negative genes of the extracted transcriptomics signature have statistically significant different expression values across tumor and non-malignant samples. Furthermore, the volcano plots demonstrated that a subset of 16 out of 73 genes satisfy the twofold criterion of p-value <= 0.01 and FC > 2 in all 3 transcriptomics datasets. This finding pave the way to future research efforts to focus on the potential of these 16 genes in NSCLC diagnosis, prognosis and survival.

The validation of the derived radiotranscriptomics models in an RNAseq dataset is very important due to the increased use of the RNAseq dataset in the measurement of the gene expression profiles. After the analysis on the GSE103584 dataset, 11 radiomic features and 9 transcriptomics features could be simulated from the transcriptomics and the radiomics features, respectively, in the RNAseq dataset. Hence, a significantly decreased number of features (11 out of 51 radiomics and 9 out of 23 transcriptomics) could be modeled in the RNAseq dataset. Furthermore, a square regression was required in order to build efficient models, indicating that more complex associations exist between the radiomics and the transcriptomics. This may happen due to the fact that many genes were missing from the RNAseq dataset, removing significant information from the existing models. Additionally, the RNAseq technology is based on sequencing, while the DNA microarray is based on the hybridization. Thus, the two techniques may not result in the same level of expression profile of the same genes, requiring different associations between the two modalities.

The derived radiomic signature consists of 11 radiomic features that can be simulated by transcriptomics markers. These radiomic features describe mainly the texture of the tumor, since 4 of them are higher-order statistics, 5 are second-order statistics features and the rest 2 are first-order statistics features (Table 10). The texture of the tumor reflects the genomic mutations and describes the heterogeneity of the tumor. The textural features are very important due to the fact that they could be used as non-invasive diagnostic and prognostic biomarkers [52].

The analysis had identified 4 discrete radiomic and transcriptomics signatures based on the relationships between these two modalities. More specifically, a transcriptomic signature of 41 genes and a radiomic signature of 51 radiomic features were identified in the RNAseq dataset based on the findings of the previous analysis. Two more signatures, a transcriptomic signature of 9 genes and a radiomic signature of 11 radiomics, were extracted after the validation of the regression models in the RNAseq dataset. Hence, the ability of these 4 signatures, separately or as joint signatures, to predict the lung cancer staging was investigated. The Random Forest classifier achieved slightly better performance than the SVM classifier in all the metrics in most cases. The performance of the classifiers were satisfactory achieving accuracy ∼70-75%, sensitivity ∼70-75% and specificity ∼75-80%, using all the examined feature vectors (Figures 18-20). More specifically, the classifiers that use the selected features (i.e. classifier with the 11 radiomics as feature vector and classifier with the 9 genes as feature vector) achieved similar classification performance using either their actual values or their predicted values. The classifier using the predicted values of the selected 11 radiomic features demonstrated slightly better performance than using the predicted values of only the selected 9 genes or the joint signature of the 9 genes and the 11 radiomics. Furthermore, the classifier using the predicted values of the selected 11 radiomics features showed better performance than the classifier using their actual values, revealing the additive value stemming from the combination of genes. However, the classifiers that use the radiomic features, they had similar performance in all metrics using either the initial 51 radiomic features or the selected 11 radiomic features. The classifiers that use the overall radiomic signatures (either radiomic signature 1 or radiomic signature 2) had good performance but slightly lower than the classifier with the gene signature 1. According to our findings, the signatures that combined the radiomic and the transcriptomic features did not enhance the classification power.

The classifier using the overall gene signature 1 which consists of the 41 initial single genes and the predicted 11 radiomic features, which were derived from the combination of genes, had the best performance among all the classifiers, resulting in high mean accuracy (=78.75%), mean sensitivity (=71.15%) and mean specificity (=86.35%) as well as relative low standard deviation. Furthermore, the classifier using the 41 initial single genes and the classifier using the predicted 11 radiomic features had similarly high performance achieving mean accuracy equal to 76.66% and 78.48%, mean sensitivity equal to 71.55% and 74.31%, and mean specificity equal to 81.77% and 82.6%, respectively. Thus, the transcriptomic data, which had been thoroughly investigated and validated for the diagnostic potential to discriminate malignant from benign lung tumors, had also the potential to perform the more challenging task of classifying the lung cancer staging. However, none of the classifiers showed significantly higher performance than the others. Thus, all the derived signatures and the combination of them can adequately predict the lung cancer staging. Their same behavior in terms of the classification ability in the lung cancer staging can be explained from the pipeline. The derived radiomic and transcriptomic signatures were validated in terms of the same evaluation metrics. They can discriminate the malignant from the adjacent normal lung tissues and they can be predicted from the complementary modality. Hence, all the derived signatures are significant and can be used separately.

A direct comparison of our findings with the radiotranscriptomics studies of Gevaert et al. [32] and Nair et al. [33] that introduced the radiotranscriptomics dataset used in the current study, could not be performed due to the following two reasons: i) these studies utilized different radiomic features, such as semantic features and features extracted with different calculations or from the PET/CT examinations, while our study focused on the CT extracted radiomics utilizing the pyradiomics software and ii) the clinical question was different as these studies focused on the association of the features with the survival, while our study explored the impact of the features in the NSCLC staging prediction. However, the initial 51 radiomic features that had been extracted from the previous analysis, were mainly texture features (Table 2). Similarly, the derived radiomic signature in the RNAseq dataset resulted in 11 radiomic features, which are textural features (Table 10). In both cases, the majority of them are second-order statistics features. The second-order statistics features are features derived from the GLCM matrix, which describes the relationship between two pixels. The GLCM-based textural features are most commonly extracted and have been reported to lead to the most significant results [4][52]. This is in concordance with our results, which reveal that the majority of the significant radiomic features was GLCM-based features which can predict efficiently the lung cancer staging.

The current study has several limitations. The radiotranscriptomics dataset is relatively small consisting of 112 samples. This is a major drawback of many radiotranscriptomics studies, since the data is not easily accessible from the researchers. The lack of publicly available radiotranscriptomics datasets restrict the research efforts towards the use of artificial intelligence (AI) in the medical domain. The machine learning algorithms and the deep neural networks require an abundance of data in order to be robust and trustable. Although there are many purely radiomics or purely transcriptomics datasets, the simultaneous availability of both imaging and transcriptomic data is limited. Furthermore, the lack of data from several clinical centers reduces the variability of the used data and did not reflect the heterogeneity of the patients with NSCLC among different centers. Additionally, there is no a standardization protocol for the extraction of the radiomic features, restricting the robustness and the reproducibility of radiomic features and making it difficult to compare the results with other studies. Another limiting factor is that many genes from the initial extracted transcriptomics signature did not exist in the radiotranscriptomics dataset or had many Nan values. Thus, the accurate investigation and validation of the initial transcriptomics and radiomic signatures are not possible, as many genes do not exist in the dataset. Moreover, the imbalanced dataset and the use of the SMOTE algorithm that produce synthetic samples restrict the robustness of the results. Finally, the lack of the pixel-based annotations of the available PET/CT examinations hampered the extraction of radiomic features from this functional examination, which could provide beneficial information for the lung tumor. These annotations, which include the delineation of the tumors, should be performed by at least two clinicians and a consensus should be achieved in order to use them for the radiomic features extraction.

Chapter 6: Conclusions

This thesis aims to validate the diagnostic potential of the transcriptomics features in NSCLC and investigate the interplay between the radiomic and the transcriptomic markers in order to extract signatures that can predict the lung cancer staging. The transcriptomic features were simulated by linear combinations of the non-invasive radiomic features using the gene expressions that were measured with DNA microarrays. Thus, regression models were derived for both radiomic and transcriptomic features, which were modelled using transcriptomic and radiomic features, respectively. The ability of one modality to simulate the other was further investigated and validated in a dataset with gene expressions measured by the RNAseq technology in order to extract useful and robust biomarkers for NSCLC. The connection between the radiomics and the transcriptomics was validated for a smaller number of features in the RNAseq dataset, requiring more complex associations through square regression. All the derived signatures, which were purely radiomic, purely transcriptomic and joint radiotranscriptomics, were tested for their ability to predict the lung cancer staging. The Random Forest classifiers using all the different derived signatures resulted in similar good performance. The classifiers that used either the single or the combinations of the well-validated transcriptomic features achieved slightly better performance than the others with purely radiomic or radiotranscriptomics markers. However, all the single or joint extracted markers have the potential to predict the NSCLC staging.

Radiotranscriptomics is an emerging field that aims to investigate the complementarities of the non-invasive radiomic features with the expressions of the genomic substrate of the tumor. The investigation of the interplay between these two spaces and the identification of significant biomarkers can contribute to better diagnosis, treatment planning and prognosis of the disease. In order to be used in the clinical practice, more collaboration between the AI-researchers and the clinicians should be developed in order to derive explainable and trustworthy AI models that assist the doctors in their decisions. Thus, larger datasets that combine imaging and genomic/transcriptomic data, should be publicly available in order to develop robust models. To this end, multi-institutional studies are essential in order to obtain a larger portion of samples that reflects the variability that exist between the patients with NSCLC. Furthermore, the collaboration with expert clinicians, who have the ability to perform the delineation of the tumor in the PET/CT examinations and extract SUV metrics, will be achieved in a future radiotranscriptomic study in order to incorporate these crucial examination in the study framework. Additionally, the association of the radiomic features and/or the -omics data with the survival or other clinical data could be examined, when these information is provided. The relationship between the radiotranscriptomic data and other molecular indicators that determine targeted treatments, such as the status of the EGFR, the KRAS and the ALK mutations, will be investigated in a future study. Finally, the deep learning algorithms are very widely used and has gained attention also in the medical applications. These algorithms can extract automatically the radiomic features from the medical images in order to explore their ability to perform several demanding tasks, such as the prediction of the survival, the stage of the cancer and the cancer aggressiveness. In many studies, the deep learning models have outperformed the traditional radiomic studies.

References

1. Bade BC, Dela Cruz CS. Lung Cancer 2020: Epidemiology, Etiology, and Prevention. Clin Chest Med. 2020;41: 1–24. doi:10.1016/j.ccm.2019.10.001

2. Lo Gullo R, Daimiel I, Morris EA, Pinker K. Combining molecular and imaging metrics in cancer: radiogenomics. Insights Imaging. 2020;11: 1–17. doi:10.1186/s13244-019-0795-6

3. Halliday PR, Blakely CM, Bivona TG. Emerging Targeted Therapies for the Treatment of Non-small Cell Lung Cancer. Curr Oncol Rep. 2019;21. doi:10.1007/s11912-019-0770-x

4. Alobaidli S, McQuaid S, South C, Prakash V, Evans P, Nisbet A. The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning. Br J Radiol. 2014;87: 5–14. doi:10.1259/bjr.20140369

5. Nevins JR, Potti A. Mining gene expression profiles: Expression signatures as cancer phenotypes. Nat Rev Genet. 2007;8: 601–609. doi:10.1038/nrg2137

6. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. Brief Funct Genomics. 2015;14: 130–142. doi:10.1093/bfgp/elu035

7. Kogenaru S, Qing Y, Guo Y, Wang N. RNA-seq and microarray complement each other in transcriptome profiling. BMC Genomics. 2012;13. doi:10.1186/1471-2164-13-629

8. Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, et al. Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Front Genet. 2019;10: 1–16. doi:10.3389/fgene.2018.00636

9. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9. doi:10.1371/journal.pone.0078644

10. Bodalal Z, Trebeschi S, Nguyen-Kim TDL, Schats W, Beets-Tan R. Radiogenomics: bridging imaging and genomics. Abdom Radiol. 2019;44: 1960–1984. doi:10.1007/s00261-019-02028-w

11. Katrib A, Hsu W, Bui A, Xing Y. “Radiotranscriptomics”: A synergy of imaging and transcriptomics in clinical assessment. Quant Biol. 2016;4: 1–12. doi:10.1007/s40484-016-0061-6

12. Anagnostopoulos AK, Gaitanis A, Gkiozos I, Athanasiadis EI, Chatziioannou SN, Syrigos KN, et al. Radiomics / Radiogenomics in Lung Cancer : Basic Principles and Initial Clinical Results. Cancers (Basel). 2022;13. doi:https://doi.org/10.3390/cancers14071657

13. Yang F, Chen W, Wei H, Zhang X, Yuan S, Qiao X, et al. Machine Learning for Histologic Subtype Classification of Non-Small Cell Lung Cancer: A Retrospective Multicenter Radiomics Study. Front Oncol. 2021;10: 1–12. doi:10.3389/fonc.2020.608598

14. Le VH, Kha QH, Hung TNK, Le NQK. Risk score generated from CT-based radiomics signatures for overall survival prediction in non-small cell lung cancer. Cancers. 2021;13. doi:10.3390/cancers13143616

15. Luna JM, Barsky AR, Shinohara RT, Roshkovan L, Hershman M, Dreyfuss AD, et al. Radiomic Phenotypes for Improving Early Prediction of Survival in Stage III Non-Small Cell Lung Cancer Adenocarcinoma after Chemoradiation. Cancers. 2022;14. doi:10.3390/cancers14030700

16. Zhang L, Zhang Z, Yu Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. J Transl Med. 2019;17. doi:10.1186/s12967-019-02173-2

17. Li X, Yin G, Zhang Y, Dai D, Liu J, Chen P, et al. Predictive Power of a Radiomic Signature Based on 18F-FDG PET/CT Images for EGFR Mutational Status in NSCLC. Front Oncol. 2019;9: 1–11. doi:10.3389/fonc.2019.01062

18. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine learning-based radiomics signatures for egfr and kras mutations prediction in non-small-cell lung cancer. Int J Mol Sci. 2021;22. doi:10.3390/ijms22179254

19. Shiri I, Maleki H, Hajianfar G, Abdollahi H, Ashrafinia S, Hatt M, et al. Next-Generation Radiogenomics Sequencing for Prediction of EGFR and KRAS Mutation Status in NSCLC Patients Using Multimodal Imaging and Machine Learning Algorithms. Mol Imaging Biol. 2020;22: 1132–1148. doi:10.1007/s11307-020-01487-8

20. Moreno S, Bonfante M, Zurek E, Cherezov D, Goldgof D, Hall L, et al. A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC. Tomogr (Ann Arbor, Mich). 2021;7: 154–168. doi:10.3390/tomography7020014

21. Pinheiro G, Pereira T, Dias C, Freitas C, Hespanhol V, Costa JL, et al. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. Sci Rep. 2020;10: 1–9. doi:10.1038/s41598-020-60202-3

22. Gevaert O, Echegaray S, Khuong A, Hoang CD, Shrager JB, Jensen KC, et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. Sci Rep. 2017;7. doi:10.1038/srep41674

23. Kim G, Kim J, Cha H, Park WY, Ahn JS, Ahn MJ, et al. Metabolic radiogenomics in lung cancer: associations between FDG PET image features and oncogenic signaling pathway alterations. Sci Rep. 2020;10. doi:10.1038/s41598-020-70168-x

24. Ubaldi L, Valenti V, Borgese RF, Collura G, Fantacci ME, Ferrera G, et al. Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. Phys Medica. 2021;90: 13–22. doi:10.1016/j.ejmp.2021.08.015

25. Zhu Y, Guo Y-B, Xu D, Zhang J, Liu Z-G, Wu X, et al. A computed tomography (CT)-derived radiomics approach for predicting primary co-mutations involving TP53 and epidermal growth factor receptor (EGFR) in patients with advanced lung adenocarcinomas (LUAD). Ann Transl Med. 2021;9: 545–545. doi:10.21037/atm-20-6473

26. Wu W, Pierce LA, Zhang Y, Pipavath SNJ, Randolph TW, Lastwika KJ, et al. Comparison of prediction models with radiological semantic features and radiomics in lung cancer diagnosis of the pulmonary nodules: a case-control study. Eur Radiol. 2019;29: 6100–6108. doi:10.1007/s00330-019-06213-9

27. Fan L, Cao Q, Ding X, Gao D, Yang Q, Li B. Radiotranscriptomics signature-based predictive nomograms for radiotherapy response in patients with nonsmall cell lung cancer: Combination and association of CT features and serum miRNAs levels. Cancer Med. 2020;9: 5065–5074. doi:10.1002/cam4.3115

28. Oikonomou EK, Williams MC, Kotanidis CP, Desai MY, Marwan M, Antonopoulos AS, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CTangiography. Eur Heart J. 2019;40: 3529–3543. doi:10.1093/eurheartj/ehz592

29. Chaddad A, Daniel P, Sabri S, Desrosiers C, Abdulkarim B. Integration of Radiomic and Multi-omic Analyses Predicts Survival of Newly Diagnosed IDH1 Wild-Type Glioblastoma. Cancers (Basel). 2019;11. doi:10.3390/cancers11081148

30. Papadimitroulas P, Brocki L, Christopher Chung N, Marchadour W, Vermet F, Gaubert L, et al. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. Phys Medica. 2021;83: 108–121. doi:10.1016/j.ejmp.2021.03.009

31. Trivizakis E, Souglakos I, Karantanas AH, Marias K. Deep radiotranscriptomics of non-small cell lung carcinoma for assessing molecular and histology subtypes with a data-driven analysis. Diagnostics. 2021;11. doi:10.3390/diagnostics11122383

32. Gevaert O, Leung AN, Quon A, Rubin DL, Napel S, Xu J, et al. Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data. Radiology. 2012;264: 387–396. doi:10.1148/radiol.12111607/-/DC1

33. Nair VS, Gevaert O, Davidzon G, Napel S, Graves EE, Hoang CD, 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. 2012;72: 3725–3734. doi:10.1158/0008-5472.CAN-11-3943

34. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, et al. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology. 2018;286: 307–315. doi:10.1148/radiol.2017161845

35. Dovrou A. Combined analysis of phenotype and genotype in lung cancer using Radiogenomics framework. Technical University of Crete. 2020. Available: https://dias.library.tuc.gr/view/86490

36. Napel, Sandy, & Plevritis SK. NSCLC Radiogenomics: Initial Stanford Study of 26 Cases. The Cancer Imaging Archive. 2014. doi:http://doi.org/10.7937/K9/TCIA.2014.X7ONY6B1

37. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, et al. Data descriptor: A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018;5: 1–9. doi:10.1038/sdata.2018.202

38. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging. 2013;26: 1045–1057. doi:10.1007/s10278-013-9622-7

39. Wei TYW, Juan CC, Hisa JY, Su LJ, Lee YCG, Chou HY, et al. Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade. Cancer Sci. 2012;103: 1640–1650. doi:10.1111/j.1349-7006.2012.02367.x

40. Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013;5: 1–24. doi:10.1126/scitranslmed.3005723

41. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22: 2109–2119. doi:10.1101/gr.145144.112

42. Moore EH. On the Reciprocal of the General Algebraic Matrix. Bull Am Math Soc. 1920;26: 394–395.

43. Penrose R. A generalized inverse for matrices. Math Proc Cambridge Philos Soc. 1955;51: 406–413. doi:https://doi.org/10.1017/S0305004100030401

44. Penrose R. On best approximate solutions of linear matrix equations. Math Proc Cambridge Philos Soc. 1956;52: 17–19. doi:https://doi.org/10.1017/S0305004100030929

45. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8: 118–127. doi:10.1093/biostatistics/kxj037

46. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77: e104–e107. doi:10.1158/0008-5472.CAN-17-0339

47. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295: 328–338. doi:10.1148/radiol.2020191145

48. Josse J, Husson F. missMDA: A package for handling missing values in multivariate data analysis. J Stat Softw. 2016;70. doi:10.18637/jss.v070.i01

49. Lim W, Ridge CA, Nicholson AG, Mirsadraee S. The 8th lung cancer TNM classification and clinical staging system: Review of the changes and clinical implications. Quant Imaging Med Surg. 2018;8: 709–718. doi:10.21037/qims.2018.08.02

50. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16: 321–357.

51. Cao S, Wang JR, Shuangxi J, Peng Y, Jingxiao C, Montierth MD, et al. Tumor cell total mRNA expression shapes the molecular and clinical phenotype of cancer. bioRxiv Prepr. 2021. doi:https://doi.org/10.21203/rs.3.rs-600171/v1

52. Chitalia RD, Kontos D. Role of Texture Analysis in Breast MRI as a Cancer Biomarker: A Review. J Magn Reson Imaging. 2019;49: 927–938. doi:10.1002/jmri.26556

Appendices

Appendix A. The Entrez Gene IDs, symbols and names of the 73 genes that had been extracted in previous analysis and are used as data in the current analysis. The 11 under-expressed genes (negative significant) are highlighted with bold.

Appendix B. The names of the 51 radiomic features that had been extracted in previous analysis and are used as data in the current analysis.

Appendix A. The Entrez Gene IDs, symbols and names of the 73 genes that had been extracted in previous analysis and are used as data in the current analysis. The 11 under-expressed genes (negative significant) are highlighted with bold.

Entrez Gene ID	Gene Symbol	Gene Name
10882	C1QL1	complement component 1, q subcomponent-like 1
5883	RAD9A	RAD9 homolog A (S. pombe)
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**	**protein kinase (cAMP-dependent, catalytic) inhibitor gamma**
1261	CNGA3	cyclic nucleotide gated channel alpha 3
2859	GPR35	G protein-coupled receptor 35
3866	KRT15	keratin 15
114907	FBXO32	F-box protein 32
2705	GJB1	gap junction protein, beta 1, 32kDa
4585	MUC4	mucin 4, cell surface associated
63035	BCORL1	BCL6 corepressor-like 1
55277	**FGGY**	**FGGY carbohydrate kinase domain containing**
4157	MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
1741	DLG3	discs, large homolog 3 (Drosophila)
2118	ETV4	ets variant 4
5169	ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3
55311	ZNF444	zinc finger protein 444
7125	TNNC2	troponin C type 2 (fast)
142	PARP1	poly (ADP-ribose) polymerase 1
1290	COL5A2	collagen, type V, alpha 2
1747	DLX3	distal-less homeobox 3
8612	PPAP2C	phosphatidic acid phosphatase type 2C
80201	HKDC1	hexokinase domain containing 1
65260	COA7	cytochrome c oxidase assembly factor 7
54993	ZSCAN2	zinc finger and SCAN domain containing 2
9244	CRLF1	cytokine receptor-like factor 1

116092	DNTTIP1	deoxynucleotidyltransferase, terminal, interacting protein 1
4796	TONSL	tonsoku-like, DNA repair protein
84300	UQCC2	ubiquinol-cytochrome c reductase complex assembly factor 2
170487	ACTL10	actin-like 10
10045	SH2D3A	SH2 domain containing 3A
638	BIK	BCL2-interacting killer (apoptosis-inducing)
202374	STK32A	serine/threonine kinase 32A
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
3603	**IL16**	**interleukin 16**
5733	PTGER3	prostaglandin E receptor 3 (subtype EP3)
10129	**FRY**	**furry homolog (Drosophila)**
2524	FUT2	fucosyltransferase 2 (secretor status included)
2027	ENO3	enolase 3 (beta, muscle)
153768	PRELID2	PRELI domain containing 2
27094	KCNMB3	potassium large conductance calcium-activated channel, subfamily M beta member 3
79035	NABP2	nucleic acid binding protein 2
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
3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
79674	**VEPH1**	**ventricular zone expressed PH domain homolog 1 (zebrafish)**
22874	PLEKHA6	pleckstrin homology domain containing, family A member 6
7104	TM4SF4	transmembrane 4 L six family member 4
347853	TBX10	T-box 10
9245	GCNT3	glucosaminyl (N-acetyl) transferase 3, mucin type
7477	WNT7B	wingless-type MMTV integration site family, member 7B
6690	SPINK1	serine peptidase inhibitor, Kazal type 1
23414	**ZFPM2**	**zinc finger protein, multitype 2**
3026	HABP2	hyaluronan binding protein 2
127845	GOLT1A	golgi transport 1A
220134	SKA1	spindle and kinetochore associated complex subunit 1
3853	KRT6A	keratin 6A
5980	**REV3L**	**REV3-like, polymerase (DNA directed), zeta, catalytic subunit**
25934	**NIPSNAP3A**	**nipsnap homolog 3A (C. elegans)**
26112	**CCDC69**	**coiled-coil domain containing 69**
78990	OTUB2	OTU domain, ubiquitin aldehyde binding 2

629	CFB	complement factor B
6878	TAF6	TAF6 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 80kDa
23780	APOL2	apolipoprotein L, 2

Appendix B. The names of the 51 radiomic features that had been extracted in previous analysis and are used as data in the current analysis.

Radiomic Feature
"sq_1_original_firstorder_Variance'''
'''log_1_original_glszm_SizeZoneNonUniformityNormalized'''
'''log_1_original_glcm_Autocorrelation'''
'''sq_1_original_glcm_Autocorrelation'''
'''sq_1_original_glcm_DifferenceAverage'''
'''wavelet_1_original_glszm_SmallAreaLowGrayLevelEmphasis'''
'''sqrt_1_original_firstorder_Minimum'''
'''sqrt_1_original_glcm_MCC'''
'''log_1_original_gldm_GrayLevelVariance'''
'''original_1_original_ngtdm_Coarseness'''
'''sq_1_original_gldm_LargeDependenceEmphasis'''
'''sq_1_original_glcm_SumEntropy'''
'''grad_1_original_firstorder_90Percentile'''
'''wavelet_1_original_glrlm_GrayLevelVariance'''
'''sq_1_original_gldm_LowGrayLevelEmphasis'''
'''sq_1_original_glszm_GrayLevelNonUniformity'''
'''log_1_original_ngtdm_Busyness'''
'''log_1_original_gldm_SmallDependenceEmphasis'''
'''sqrt_1_original_glcm_Id'''
'''exp_1_original_gldm_LargeDependenceEmphasis'''
'''original_1_original_glcm_Autocorrelation'''
'''original_1_original_glcm_Imc1'''
'''wavelet_1_original_glrlm_RunLengthNonUniformityNormalized'''
'''sq_1_original_firstorder_10Percentile'''
'''grad_1_original_gldm_LargeDependenceLowGrayLevelEmphasis'''
'''original_1_original_gldm_GrayLevelNonUniformity'''
'''exp_1_original_firstorder_MeanAbsoluteDeviation'''
'''log_1_original_glszm_ZoneEntropy'''
'''sqrt_1_original_firstorder_Median'''
'''original_1_original_glszm_ZonePercentage'''
'''original_1_original_shape_Flatness'''
'''original_1_original_shape_MinorAxisLength'''
'''log_1_original_gldm_SmallDependenceLowGrayLevelEmphasis'''
'''sqrt_1_original_firstorder_Kurtosis'''
'''sqrt_1_original_gldm_DependenceVariance'''

'''original_1_original_shape_Maximum3DDiameter'''
'''wavelet_1_original_glcm_Contrast'''
'''sq_1_original_firstorder_Kurtosis'''
'''sqrt_1_original_glcm_JointEnergy'''
'''original_1_original_firstorder_RobustMeanAbsoluteDeviation'''
'''log_1_original_firstorder_RootMeanSquared'''
'''wavelet_1_original_glcm_SumEntropy'''
'''log_1_original_glcm_DifferenceEntropy'''
'''log_1_original_glrlm_RunPercentage'''
'''original_1_original_firstorder_RootMeanSquared'''
'''log_1_original_glcm_ClusterShade'''
'''grad_1_original_glcm_SumAverage'''
'''original_1_original_shape_SurfaceVolumeRatio'''
'''log_1_original_glcm_JointEnergy'''
'''sqrt_1_original_gldm_HighGrayLevelEmphasis'''
'''log_1_original_glszm_GrayLevelVariance'''