2020•2021

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The potential of radiomics with PET/CT: study of correlations with the metabolic profile and its discriminative power

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Prof. dr. Liesbet MESOTTEN

BEGELEIDER :

Mevr. Elien DERVEAUX

Laura Deckers, Rani Truyens

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleaire en medisch

Gezamenlijke opleiding UHasselt en KU Leuven

2020•2021

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The potential of radiomics with PET/CT: study of correlations with the metabolic profile and its discriminative power

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Prof. dr. Liesbet MESOTTEN

BEGELEIDER :

Mevr. Elien DERVEAUX

Laura Deckers, Rani Truyens

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleaire en medisch

UHASSELT KU LEUVEN

Preface

We got the chance to write our Master's thesis as a part of the ProLUNG study in collaboration with the hospital Ziekenhuis Oost-Limburg (ZOL) and the University of Hasselt. During this year, we tried, as one of the first research groups, to discover a link between radiomics and metabolomics for patients diagnosed with early-stage and locally advanced non-small cell lung cancer (NSCLC). This was done with the help of blood plasma and PET/CT images of the ProLUNG study. Furthermore, we got to assist in the starting phase of learning to work with and understand radiomics and generate a discriminative model to optimize the treatment plan and make it more patient-specific.

Firstly, we want to thank Ph.D. student Derveaux Elien and Prof. dr. Mesotten Liesbet for their support, advice, and positive energy to improve our work day by day and give us this chance to be an essential part of the ProLUNG study. It has been a pleasure to work and learn together.

Secondly, we want to thank Prof. dr. Reniers Brigitte for the support and revise of our work and Prof. dr. Thomeer Michiel, for the support, revises of our work and new insights into lung cancer.

Thirdly, a special thanks to Prof. dr. Boellaard Ronald, who gave us access to the used tools and helped us every step to reach our study goals.

Furthermore, we want to thank Dr. Ivanova Anna with her help during the statistic process and student Freson Pauline for the obtained metabolomic data.

At last, we want to thank our family and friends for the support and enthusiasm they showed during the whole process.

Table of Contents

List of tables

List of figures

List of supplementary figures and code

List of abbreviations

^{18}F-FDG	2-$^{[18]}$F fluorodeoxyglucose
^{1}H-NMR	Proton nuclear magnetic resonance
2-HG	2-hydroxyglutarate
AABB	Axis-aligned bounding box
Acetyl CoA	Acetyl-coenzyme A
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BMI	Body mass index
CT	Computed Tomography
EANM	European Association of Nuclear Medicine
FA	Factor Analysis
$FADH_2$	Flavin adenine dinucleotide
FBP	Filtered back-projection
FBP	Filtered back-projection
G-6P	Glucose-6-phosphate
GAPDH	Glyceraldehyde-3-phosphate
GBM	Glioblastoma multiforme
GLCM	Grey level co-occurrence matrix
GLUTs	Glucose transporters
GTP	Guanosine triphosphate
HK	Hexokinase
HU	Hounsfield units
IA	Injection administration
LDH	Lactate dehydrogenase
LOR(s)	Line of response(s)
LSO	Lutetium oxyorthosilicate
MR	Magnetic resonance
MRI	Magnetic resonance imaging

MTV	Metabolic tumor volume
NAD+/NADH	Nicotinamide-adenine-dinucleotide
NSCLC	Non-small cell lung cancer
PACS	Picture Archiving and Communications Systems
PCA	Principal Component Analysis
PCs	Principal components
PEP	Phosphoenolpyruvate
PET	Positron emission tomography
PET/CT	Positron emission tomography-computed tomography
PK	Pyruvate kinase
PMT	Photomultiplier tube
ROI(s)	Region of interest(s)
SCLC	Small cell lung cancer
SPECT	Single-photon emission tomography
SUV	Standard Uptake Value
TCA	Tricarboxylic acid cycle
TLG	Total lesion glycolysis
ToF	Time of flight
VOI(s)	Volumes of interest(s)
ZOL	Ziekenhuis Oost-Limburg
χ^2	Chi-square value

Abstract

Treatment of lung cancer is still a challenge, partly due to late-stage diagnosis of patients. Focusing on non-small cell lung cancer (NSCLC), metabolic biomarkers from blood plasma (metabolomics) and features from medical images (radiomics) are examined. This study combines metabolomics and radiomics datasets from NSCLC patients, unravels the underlying correlations and generate discriminative models based on radiomics features.

Two patient cohorts of (i) 39 patients and (ii) 85 patients were used. All patients were diagnosed with early-stage, locally advanced NSCLC and underwent a chirurgical resection of the lung tumor. PET/CT images of all patients were collected and segmented. From each volume of interest, 483 parameters are extracted. From 39 patients, 238 metabolic parameters representing 62 plasma metabolites are determined using proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. A correlation test is used on the total omics-dataset of 39 patients. Logistic regression is used to generate the discriminative models based on radiomics parameters from 85 patients.

The correlation matrices show that glucose and glycerol are strongly correlated with specific radiomics features. These results suggest new insights in PET/CT interpretation and that more plasma metabolites might be correlated with features out of PET/CT images. The discriminative models are built using two radiomics features and can distinguish between malignant/non-malignant PET-positive lung nodules, and between adenocarcinoma/ squamous cell carcinoma.

Abstract (in Dutch)

De behandeling van longkanker is en blijft een uitdaging, deels doordat de diagnose laattijdig wordt gesteld. Gefocust op niet-kleincellige longkanker worden metabole biomarkers vanuit het bloedplasma (metabolomics) en parameters van medische afbeeldingen (radiomics) onderzocht. Dit onderzoek combineert metabolomics en radiomics datasets om correlaties te ontdekken en genereert discriminerende modellen op basis van radiomics parameters.

Twee patiëntengroepen van (i) 39 patiënten en (ii) 85 patiënten zijn gediagnosticeerd met niet-kleincellige longkanker en ondergingen een chirurgische verwijdering van de longtumor. PET/CT beelden worden verzameld en gesegmenteerd. Uit elke volume of interest worden 483 parameters geëxtraheerd. Bij 39 patiënten worden 238 metabole parameters, representatief voor 62 plasma metabolieten, bepaald door proton nucleaire magnetische resonantie (^{1}H-NMR) spectroscopie. Een correlatietest is gebruikt op de twee omics-datasets van 39 patiënten. Logistische regressie is toegepast om de modellen te genereren op basis van de radiomics dataset van 85 patiënten.

De correlatiematrices tonen dat glucose en glycerol sterk gecorreleerd zijn met specifieke radiomics parameters. Deze resultaten suggereren nieuwe inzichten in interpretatie van PET/CT beelden en dat meerdere plasma metabolieten correleren met parameters uit PET/CT beelden. De discriminerende modellen worden gebouwd op basis van twee radiomics parameters, en kunnen maligne/benigne longletsels en adenocarcinoom/spinocellulair carcinoom onderscheiden.

1 Introduction

One of the most common causes of cancer death for men and women worldwide is lung cancer, with almost 25% of all cancer deaths (1). This is partly due to the late-stage diagnosis, which makes the treatment of lung cancer challenging (2). This Master's thesis will focus on the most common type of lung cancer, named non-small cell lung cancer (NSCLC) and, more specifically, stage I-IIIA NSCLC.

This research is part of the ProLUNG study, which is a study at the hospital Ziekenhuis Oost-Limburg (ZOL) in cooperation with UHasselt and funded by 'Kom op tegen kanker'. The ProLUNG study focuses on patients with NSCLC who undergo surgery to remove the primary lung tumor, specifically a lobectomy, as part of their standard-of-care treatment plan.

This study examines the discriminative potential of combining specific metabolic biomarkers from blood plasma (metabolomics) with features out of medical images (radiomics). This way, metabolomics and radiomics might be at the base of developing a more personalized treatment plan for lung cancer patients. This research aims to combine metabolomics and radiomics datasets from NSCLC patients, unravel the underlying correlations between these techniques, and use the radiomics features to generate models to discriminate between malignant and non-malignant lung lesions, and between the pathology of an adenocarcinoma and a squamous cell carcinoma.

To find the correlations between metabolomics and radiomics, a patient cohort was formed with 39 patients, all diagnosed with early-stage and locally advanced NSCLC. All these patients underwent a lobectomy. PET/CT images were collected from these patients using ^{18}F-FDG and the Biograph Horizon camera from Siemens Healthineers. The PET/CT images of all patients were collected and saved in the Picture Archiving and Communications Systems (PACS).

These PET/CT images were then all segmented using a semi-automatic tool (ACCURATE), creating specific volumes of interest (VOIs) of the lung lesions for each patient. By loading the VOIs into the second tool (RADIOMICS), 483 radiomics parameters were extracted from each VOI. The research team of Prof. dr. Boellaard (Amsterdam, VUmc) developed both tools. The segmentation was done in three different ways. Simultaneously, 238 metabolic parameters representing 62 plasma metabolites were determined from the same patients using proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. A correlation coefficient test was used on the total omics dataset to find correlations between these two sets of parameters.

For the second goal, generating the models to discriminate between a malignant lung lesion and a non-malignant lung lesion, and between the pathology of an adenocarcinoma and a squamous cell carcinoma based on radiomics features, a second patient cohort was used. This cohort consisted of 85 NSCLC patients who also underwent lobectomy and had a PET/CT scan using ^{18}F-FDG with the same camera. To generate the model based on adenocarcinoma and squamous cell carcinoma, the patients with a different diagnosis were excluded from this dataset. The resulting dataset contained 66 patients.

The same method for segmentation of the lung lesions as in the first patient cohort was used: 483 parameters per VOI per patient were extracted from the PET/CT images. A Spearman correlation test was used on the radiomics data from the 85 NSCLC patients to determine which parameters can be excluded from the dataset. The threshold for this exclusion is 0.9. This parameter reduction is necessary before building the model.

After the exclusion, 56 radiomics features remained in the dataset. The dataset is split into a 75% training dataset and a 25% test dataset. This is done for all 85 patients and for the 66 patients diagnosed with adenocarcinoma and squamous cell carcinoma. First, forward selection regression is used on the training dataset with a threshold of 0.2 to build the model one variable at a time. After this, backwards stepwise selection regression is used on this model with a threshold of 0.05 to refine it. With these logistic regression methods, the new model differentiates between malignant and non-malignant lung lesions, and between the pathology of an adenocarcinoma and a squamous cell carcinoma. After building this model with the training dataset, the test dataset is used to test the accuracy of the models.

To understand this research, there must be an understanding of what cancer is, how the metabolism in a normal cell differs from that in a tumor cell, and which role this plays in diagnosing cancer. This is explained in chapters two to four.

Chapter five will focus on non-small cell lung cancer, the diagnosis and treatment of NSCLC, and the difficulties that arise with NSCLC.

Chapters six and seven describe the most used imaging techniques for diagnosing NSCLC, PET and CT. The physical process, the tracer, detector, scintillator, and important parameters are explained in the chapter about PET. Then, the link between a PET image and NSCLC is described. The chapter about CT handles the x-ray production, x-ray tube, the data acquisition with some important CT parameters and reconstruction of the images. A short explanation is given as to why combining these two techniques is useful for diagnosis.

In chapter eight, radiomics is discussed thoroughly. After a short introduction to this topic, the paper gives an overview of the results obtained with radiomics in oncological and non-oncological applications. After this, the whole process of radiomics is explained.

Chapter nine and ten both describe statistic methods to reduce parameters and how they work. These methods are respectively Factor Analysis and Principal Component Analysis. The next chapter handles the Pearson test and the used correlation test, the Spearman correlation test. The statistics and use of this method are clarified in this chapter. Chapter twelve will focus on the Chi-square test, used to give an indication of the deviation between two sets. To close the literature study of this research, chapter thirteen describes the logistic regression method to obtain the discriminative model for NSCLC patients.

After this, the transition is made to the actual research, starting with the ProLUNG study. Chapter fourteen explains the study as a whole and focuses on the study's research goals, followed by the section describing the materials and method. The patient cohort, inclusion and exclusion criteria, the specifications of the Biograph Horizon PET/CT camera from Siemens Healthineers and the used ^{18}F-FDG PET/CT protocol are explained.

Next, the radiomics parameters and used tools are discussed, together with the correlation test, used methods for data reduction and the building of the discriminative models. The datasets, which are divided into the datasets for the correlations and the datasets for the discriminative models, are discussed last.

Chapter sixteen shows the results of the research, beginning with the data from all patients. The correlation results are discussed for data from the segmentation of the lung lesions on PET/CT images with a poorly lined lesion on the CT. These results are then compared with the data from the segmentation of the lung lesions on PET/CT images with a properly lined lesion on the CT and the segmentation of the lung lesions based on only the PET images of the patients. There will be looked at the similarities and differences between the three correlation matrices.

Then, the results of the discriminative models are shown. The accuracy and precision of the discrimination between malignant and non-malignant lesions, and between the pathology of an adenocarcinoma and a squamous cell carcinoma are discussed.

Chapter seventeen discusses the possibility to extend this research, the problems, and potential solutions for these problems. At last, chapter eighteen contains the conclusion of this research.

2 The origin of Cancer

Cancer is a disease mainly characterized by uncontrolled cell division and leads to abnormal tissue growth (3).
The two large categories when looking at tumor types are benign tumors and malignant tumors. Between these two, the most apparent difference is that a benign tumor does not spread to or invades other parts of the tissue. This means that this type of tumor is not life-threatening in most cases.

The other type, a malignant tumor, does invade and destroy surrounding tissue, another typical characteristic of cancer. It even spreads to different tissues in the body, a process called metastasis, and can become life-threatening (4).

Cancer cells are derived from normal cells in the body. There are multiple ways that cancer can be established, but cells also have multiple systems to prevent cancer from forming.

The evolution of cancer, carcinogenesis, is influenced by several disruptions such as interferences in growth factors or resistance to apoptosis (5). Carcinogenesis is a multi-step process that can be summarized by several Hallmarks, as shown by figure 1. We will take a deeper look into one of these hallmarks for this research: the deregulation of cellular energetics or the metabolism (6). One of the primary roles of the metabolism that appears in a cell is to convert nutrients into energy.

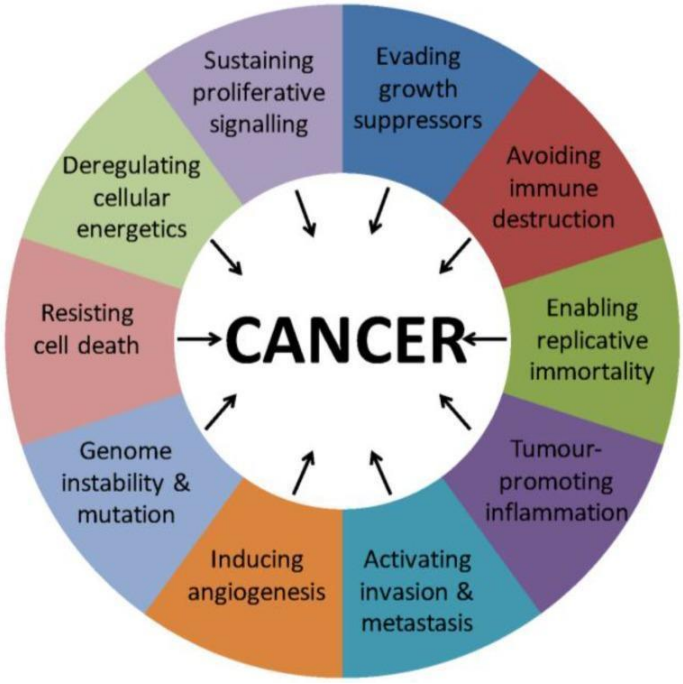

Figure 1: The Hallmarks of Cancer. This figure shows the hallmarks obtained by most tumors (7).

3 Metabolism in a cell

Cells obtain energy through their metabolism. Food that we eat is converted into cellular energy or adenosine triphosphate (ATP). Figure 2 shows one of the most essential and well-known pathways that leads to ATP production in cells (5).

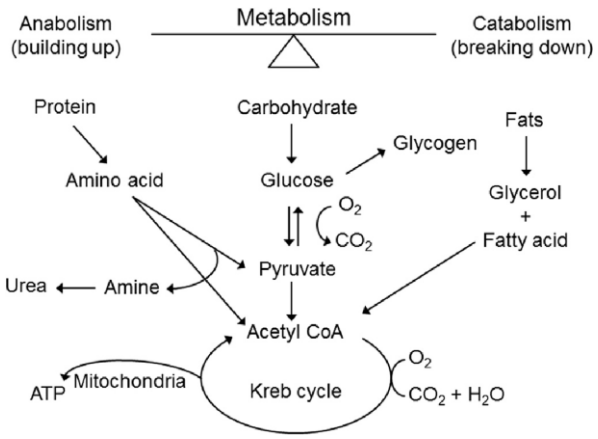

Figure 2: The metabolism in a cell. This figure displays how cellular energy is formed from metabolism (5).

3.1 Glycolysis

First, a deeper look into the glycolysis that takes place in the human body is needed. At first, carbohydrates are broken down to glucose. This is converted to pyruvate and then oxidizes to acetyl-coenzyme A (acetyl CoA). The pathway that comprises the complete conversion of glucose to pyruvate is called glycolysis. Glycolysis is mainly an energy provider for other metabolic pathways, such as nucleic acid synthesis or protein synthesis (5, 8). The oxidation of pyruvate to acetyl CoA is the link between glycolysis and the Krebs cycle. A small amount of ATP is formed during glycolysis. Figure 3 shows the production of pyruvate and lactate via glycolysis in a tumor cell (8). As shown in this figure, glycolysis can be explained in ten steps. The first five steps displayed all use ATP, and the last five all produce ATP (8, 9).
In the first step, the enzyme hexokinase (HK) uses ATP as a source for phosphates and forms, together with glucose, glucose-6-phosphate (G-6P). Secondly, this molecule is converted into one of its isomers, fructose-6-phosphate. This happens with the help of an isomerase, an enzyme that enables the conversion of a molecule in an isomer of that specific molecule (5, 9).

Using another enzyme, phosphofructokinase, and another ATP molecule, fructose-1,6-biphosphate is produced. In the fourth step, the 1,6-biphosphate molecule is split into two three-carbon isomers: dihydroxyacetone-phosphate and glyceraldehyde-3-phosphate. The last step of this first part of glycolysis is converting the dihydroxyacetone-phosphate molecule into another glyceraldehyde-3-phosphate (GAPDH) molecule (5, 9).

In this first part, two ATP molecules are used. The following five steps start with two glyceraldehyde-3-phosphate molecules. These two molecules are oxidized and then phosphorylated. The product of this step is 1,3-bisphosphoglycerate. High-energy electrons are extracted from the sugar (glyceraldehyde-3-phosphate) and picked up by nicotinamide-adenine-dinucleotide (NAD+) during the oxidation. This reduces the molecule and thus forms NADH.

1,3-bisphosphoglycerate then gives a high-energy phosphate to adenosine diphosphate (ADP) molecule, forming one ATP molecule. This reaction takes place with the help of another specific enzyme, phosphoglycerate. In the same step, a carbonyl group on the 1,3-bisphosphoglycerate molecule is oxidized to a carboxyl group. The product of this oxidation is 3-phosphoglycerate.

The eighth step shows the formation of 2-phosphoglycerate by moving the phosphate group in 3-phosphoglycerate to the second carbon, using a mutase, which is also an isomerase.

Next, 2-phosphoglycerate dehydrates to produce phosphoenolpyruvate (PEP) as a product of this step. Enolase is the enzyme used in this step (5, 8, 9).

Lastly, a second ATP molecule is produced by phosphorylation of PEP with the enzyme pyruvate kinase (PK), and a pyruvate molecule is formed. Pyruvate is a critical metabolite of glycolysis when looking at the metabolism of a normal cell versus a cancer cell. In cancer cells, pyruvate preferably forms lactic acid or lactate catalyzed by lactate dehydrogenase (LDH) in cells (5, 8, 9).

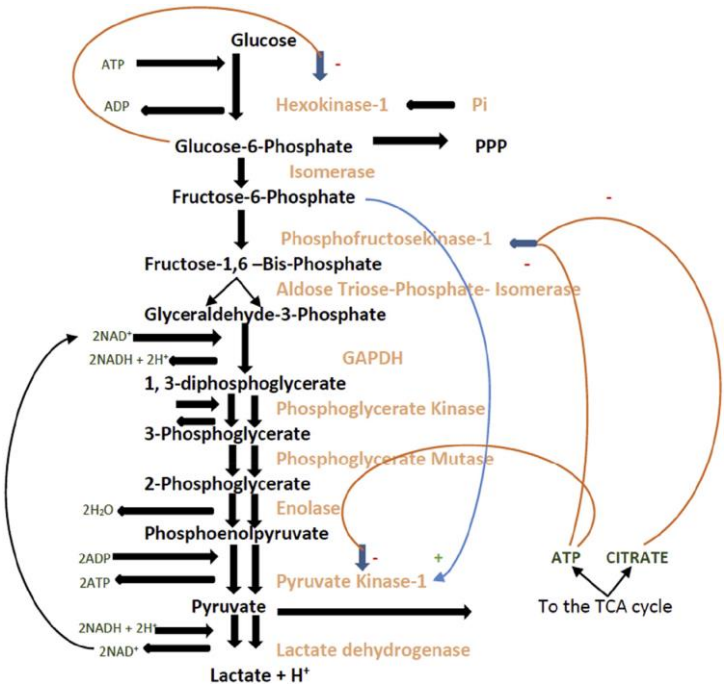

Figure 3: Glycolysis in a normal cell (8). The terms shown in orange are enzymes that catalyze the different steps in glycolysis. The pathway starting from glucose to lactate is shown in this scheme. The last step is more prominently used in a tumor cell.

3.2 Krebs cycle

The glycolysis is only the start of the energy metabolism. The following mechanism is the Krebs cycle. The formed acetyl CoA is then metabolized to the Krebs cycle, also known as the citric acid cycle or the tricarboxylic acid cycle (TCA) (10). The remaining energy from the original glucose molecule will be extracted in this cycle. Acetyl CoA enters the TCA cycle and is converted into citric acid during the first step by combining oxaloacetic acid.
Citric acid is oxidized during several steps, and oxaloacetic acid is again generated. During the oxidation steps, CO_2 and H_2O are formed. Other products of the TCA cycle are NADH, the hydroquinone form of flavin adenine dinucleotide ($FADH_2$), and ATP or guanosine triphosphate (GTP) (depending on the cell type) (10, 11). Figure 4 shows the schematic representation of the citric acid cycle.

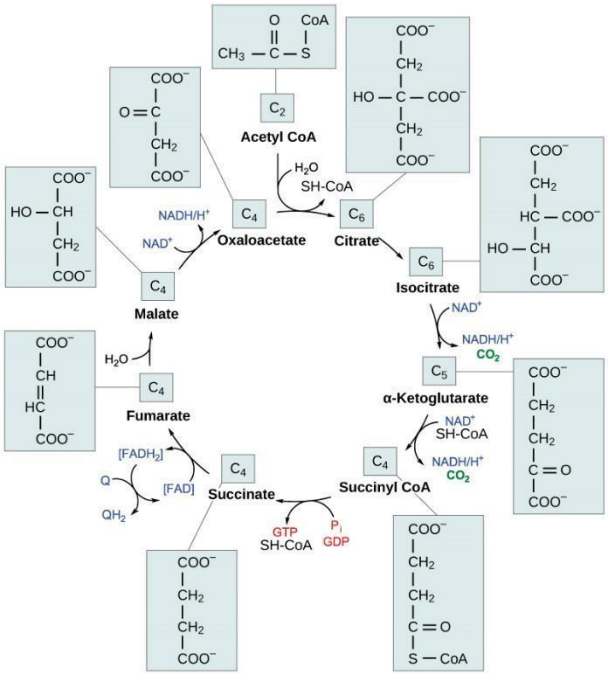

Figure 4: The tricarboxylic acid cycle. The acetyl group from acetyl CoA is attached to a four-carbon oxaloacetate molecule to form a six-carbon citrate molecule. This citrate is then oxidized through multiple steps (14).

While some ATP is produced directly during this cycle, it produces ATP indirectly through NADH and $FADH_2$. These two molecules are electron carriers and pass their electrons into the electron transport chain. In this chain, electrons are transferred through a series of electron acceptors. During each transfer, energy is released (11, 12).

This energy is used to create an electrochemical gradient across the membrane in the next step, chemiosmosis. Due to this gradient, ions can flow across the membrane. In ATP formation, hydrogen ions flow across this membrane throughout cellular respiration (13).

Then, ATP is generated through the stored energy in the gradient (12). This process, the electron transport chain and chemiosmosis, is called oxidative phosphorylation.

4 Metabolism in a cancer cell

In a tumor cell, the same metabolic pathways generate ATP, but the pathways are reprogrammed to meet the requirements for tumor cell proliferation and survival. A tumor cell needs larger amounts of ATP, NADPH, and NADH to survive (14).

One of the main alterations of the pathways occurs during glycolysis. Tumor cells enhance glycolysis to generate more ATP, and they produce lactate despite the abundant presence of oxygen to oxidize the glucose completely. During this process, the formed pyruvate is converted to lactate. A tumor cell uses the same process as displayed in figure 3. In figure 3, the last step shows the conversion from pyruvate to lactate. A tumor cell produces more lactate than a normal cell. This particular phenomenon is called the Warburg effect or aerobic glycolysis (8).

Figure 5 compares the metabolism of a normal cell (left) and the metabolism of a tumor cell (right). It is visible that the glucose in a tumor cell does not fully oxidize and therefore produces lactate and pyruvate (14).

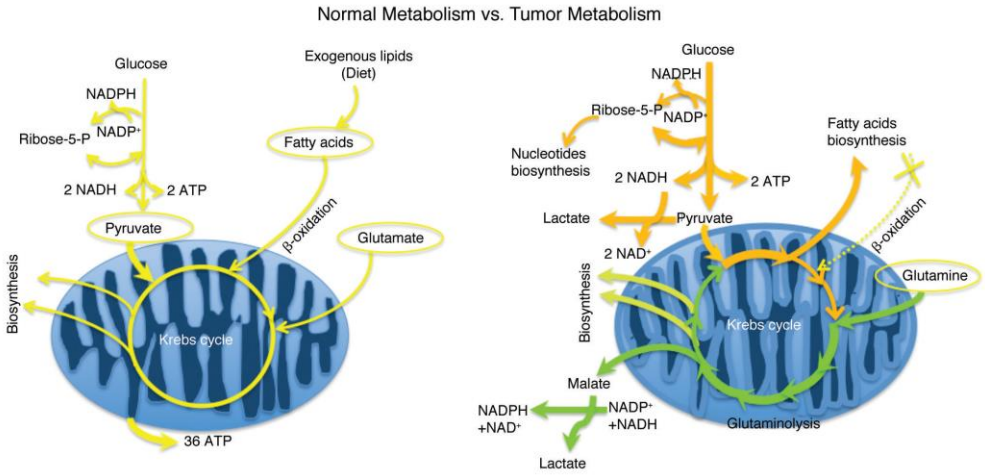

Figure 5: Comparison between the metabolism of a normal cell and a tumor cell. In a tumor cell, more lactate is formed compared to in a normal cell (14).

The ATP yield is less during glycolysis than during the Krebs cycle, but tumor cells still choose this process and make up for the inefficiency by going through the glycolysis much faster than normal cells. One of the consequences of this is that tumor cells need more glucose import (8).

To meet these requirements for a tumor cell, specific transporters are overexpressed. Examples of these transporters are glucose transporters (GLUTs), like GLUT1 and GLUT 3. Both GLUT1 and GLUT3 have a high preference for glucose, 2-deoxyglucose, and 2-$^{[18]}$F fluorodeoxyglucose (^{18}F-FDG) (15). Further, in this literature study, it will become clear that one reason for using positron emission tomography (PET) to stage cancers is that PET measures the FDG uptake in vivo (16). Pyruvate transport in a tumor cell is less than in a normal cell because a tumor cell focuses more on aerobic glycolysis instead of the Krebs cycle.

In conclusion, a tumor cell uses the same metabolic pathways as a normal cell but focuses more on glycolysis. A couple of metabolites, succinate, itaconate, fumarate, 2-hydroxyglutarate (2-HG), and acetyl-CoA, which occur in the Krebs cycle, are further investigated by Ryan et al. (17). This research shows that these play a vital role in the disrupted metabolism of a cancer cell. Vanhove et al. researched the important metabolite glutamate (16). Their research proves that this metabolite seems indispensable for lung tumors and is also overexpressed to meet their metabolic requirements. Thus, the main consequence of the disturbed metabolism in a cancer cell is that the concentration of specific metabolites in the pathways and the concentration of certain end products will differ in a cancer cell compared to a normal cell. In the next section, this knowledge will be applied on the topic of non-small cell lung cancer.

5 What is lung cancer?

Lung cancer is still one of the most common causes of cancer death for men and women worldwide. In Belgium alone, 14.4% of the newly reported cancer cases for men and 8.3% for women were due to lung cancer in 2018 (18). Several risk factors for lung cancer are identified, with smoking being the most important (19).

Primary lung cancer starts in the lungs and can metastasize further in the body to other organs. Lung cancer can be divided into two large groups, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The last one is the most common type of lung cancer, which is the type of focus in this research (20).

NSCLC can be subdivided into different types: squamous cell carcinoma, large cell carcinoma, and adenocarcinoma being the three most common ones (19).

NSCLC can exist in different stages (stage 0 to stage IV). The stage is determined by TNM staging, where T stands for 'size and extent of the primary tumor,' N stands for 'involvement of lymph nodes in the region of the lungs,' and M stands for 'metastatic involvement or spread to distant organs' (21). The M category can be subdivided into M0 and M1. In subcategory M1, cancer has metastasized to distant organs or tissues. Subcategory M0 means that cancer has not spread out to different organs or tissues (22). This research focuses solely on subcategory M0 tumors.

The diagnosis of lung cancer can be determined in different ways, depending on the situation. For example, the patients can undergo a biopsy, where fluid is retracted from a dubious area or the area surrounding the lung, using a needle or surgery. Another way is to examine the secretions of the lung. A PET/CT examination is of crucial importance in the process of lung cancer diagnosis (23-25).

The standard-of-care treatment for patients with early-stage and locally advanced lung cancer (TNM I - TNM IIIA) diagnosis is surgery or lobectomy. During this operation, a whole lobe of the lung that contains the tumor is removed, like shown in figure 6 (19, 26, 27).

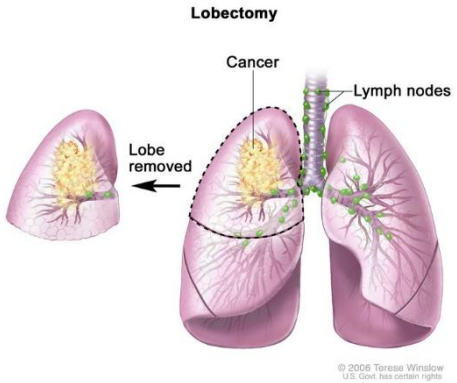

Figure 6: The process of surgery called a lobectomy. A part of the lung lobe that contains the tumor and a small part of healthy tissue surrounding the tumor is removed.

A common problem that arises with the diagnosis and treatment of lung cancer patients is the heterogeneity of tumors. A big reason for the different responses of every patient to treatments is the molecular heterogeneity between lung cancer patients diagnosed with the same histology (28).

The survival rate over five years for patients diagnosed with NSCLC overall is 14% (29). When looking at the different stages of NSCLC, the survival rate can differ drastically. For a patient diagnosed with stage I NSCLC and where the tumor is surgically removed, the survival rate over five years goes up to 70%. In contrast to this, when a patient is diagnosed with an inoperable NSCLC, the survival duration drops to nine months.

It is essential to get an early diagnosis of NSCLC before the tumor metastasizes. Only in this stage, an operation is helpful for the patient and will increase the patients' survival rate (27).

6 PET

One of the many Positron Emission Tomography (PET) applications, a functional imaging technology, is to obtain a non-invasive method for lung cancer imaging. Before a PET examination, compounds labeled with short-living positron-emitting radioisotopes are injected intravenously. Depending on the carrier molecule (isotopic labeling), the injected isotope will distribute to different tissues. External detectors orientated in various directions will detect the radiation emitted from these injected radiopharmaceuticals (30).

6.1 Physical process

More specifically, the injected radiopharmaceutical will undergo positron decay, also known as beta plus decay. A proton will be converted into a neutron during this process and release a positron, also known as a beta particle, and an electron neutrino. Formula 1 describes the process of beta plus decay.

$$p_1^1 \rightarrow n_0^1 + e_{+1}^0 + \nu + \text{energy} \qquad (1)$$

This physical conversion occurs particularly for proton-rich isotopes, such as ^{102}Pd, to correct the imbalance between protons and neutrons (31).

The created positron will interact with a surrounding electron. This complete annihilation will create energy in the form of two photons that speed off in opposite directions (32). The two photons both carry an energy of 511 keV and will be measured by a ring of detectors around the patient during the PET examination. The origin of the photons can be determined using the direction of the annihilation photons; hence, the radioactive decay process that created them can be localized. This process is visualized in figure 7 (33).

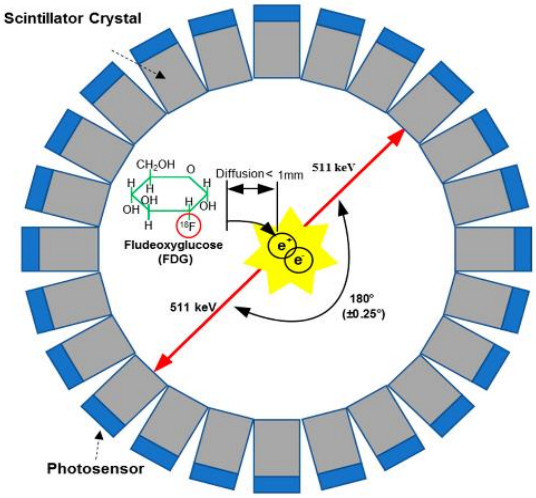

Figure 7: Positron emission tomography process. This figure visualizes the physical process that occurs during a PET-scan. The positron annihilation is shown in yellow, with two photons of each 511 keV that speed in opposite directions. Furthermore, the atomic structure of fluorodeoxyglucose is described.

The annihilation of the positron depends mainly on the range of the positron, which is affected by the electron density of the medium. The typical range of positrons emitted from radionuclides used in PET is about one to two millimeters. Here, the medium is considered water because a significant part of biological tissue consists of water (34). As discussed, ^{18}F is a proton-rich isotope, which is one reason it is generally used as a tracer in PET examinations. Sequentially, it will be explained how this isotope can be used as a tracer in the next section. This will first be described in general and then specific for NSCLC.

6.2 Tracer (^{18}F-FDG)

^{18}F-labeled FDG (^{18}F-FDG) is a commonly used radiopharmaceutical in nuclear medicine and is thus also used during a PET-scan in the nuclear medicine department of the hospital of Genk. ^{18}F-FDG will undergo positron decay for 97% (35). The half-life of ^{18}F-FDG is 110 minutes (30). It is noteworthy that the half-life is relatively short to limit radioactivity in the patient but is long enough to obtain a medically qualified pet scan.
The atomic structure of ^{18}F-FDG is approximately the same as the atomic structure of glucose, and both are represented in figure 8. The difference is that a fluor-18 atom replaces one hydroxyl group of glucose. Because the difference in the atomic structure is small, the characteristics of the chemical components are very similar (36, 37).

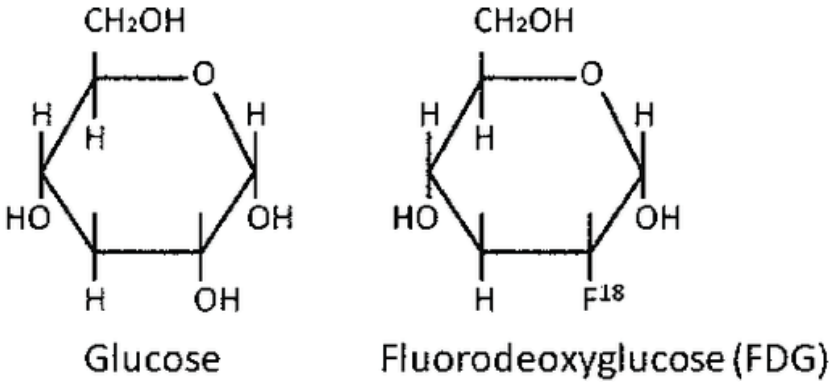

Figure 8: Atomic structure glucose and FDG. This figure gives the atomic structure of both glucose and fluorodeoxyglucose.

Normal cells have a lower glucose consumption than malignant cells. Thus, a typical characterization of malignant cells is a higher uptake of glucose with the consequence that those cells also have a higher uptake of FDG (38). ^{18}F-FDG PET has the significant advantage that a change in cellular metabolism is visible more quickly than a change in tumor size (39).
Small compact medical cyclotrons are created to produce most positron-emitting isotopes, such as ^{18}F-FDG, by starting from [^{18}O]water (40). The energy required to create these isotopes is relatively moderate, and that way less than 20 MeV for protons (35). In a cyclotron and other accelerators, both a magnetic field and an electric field are used. The electric field will accelerate the ions, and the magnetic field is applied to make the ions move in a preferable direction.

The amount of ^{18}F-FDG that is injected is a vital parameter to obtain a medically qualified PET-scan. Another necessary characteristic of ^{18}F-FDG is that it is nontoxic, or it does not harm the patient. Besides, ^{18}F-FDG is chemically incorporated into the specific physiological process, but it will not influence or modify this metabolic process (41).

In the European Association of Nuclear Medicine (EANM), clinical guidelines are published based on standards for the doses of specific radiotracers and their specific applications (36). The dosage of ^{18}F-FDG depends on the used system and the patient's weight (42). The amount of ^{18}F-FDG that is injected into the patient's body, injection administration (IA), is calculated by Boellaard et al. by the following formula 2 (42):

$$IA = 7.2\, patient\ weight\ (kg) \,/\, scan\ time\ (min) \pm 10 \quad (2)$$

In this research, the injected ^{18}F-FDG dose in the hospital ZOL of Genk is described by the following formula 3. Here, the dose is first expressed in millicurie and multiplied by 37 to obtain the dose in megabecquerel. The weight of the patient is, in both formulas, an indispensable parameter.

$$Dose = (patient\ weight\ (kg)\ /10) - 1 \cdot 37 \quad (3)$$

To optimize the image quality, the patient will undergo a complete fasting for a minimum of 6 hours before the scan, whereby only drinking plain water should be allowed (43). This fastening aims to minimize the competition between glucose out of food and the ^{18}F-FDG uptake (43). Furthermore, high-intensity activities should be avoided for a minimum of 24 hours before the injection (43).

6.3 FDG and a cancer cell

The localization of biological mechanisms visible with the help of ^{18}F-FDG and other synthesized radiopharmaceuticals is known as metabolic trapping. This principle is based on the metabolic activity of the tissues in the human body (44). Together with the uniform distribution of ^{18}F-FDG, specific organs have a significantly higher uptake level than others. For example, the brain and heart both have a higher uptake after two hours past injection compared to the lungs (45). In general, this can be summarized by the phosphorylation of glucose, as explained in section '*Glycolysis*' (45).
As explained in the section '*Metabolism in a cancer cell*', the same metabolic pathways as in a normal cell are used to generate ATP. The difference is that the pathways are reprogrammed for a tumor cell (14).

Furthermore, glycolysis appears to happen much faster in a tumor cell than in a normal cell. As explained, this results in a higher need for glucose to optimize the working of a tumor cell (8). Because of the similar anatomic structure and the corresponding similar characteristics of glucose and ^{18}F-FDG, a higher glucose uptake will also lead to a higher ^{18}F-FDG uptake (36). Another reason why ^{18}F-FDG is used for the staging and imaging of cancer is that specific transporters have a high preference for ^{18}F-FDG, par example, GLUT1, and GLUT3 (15, 16). Kaira et al. concluded that the uptake and accumulation of ^{18}F-FDG within lung cancer cells is determined by the metabolism of glucose, hypoxia, and angiogenesis (46). This leads to a further specification of FDG and non-small cell lung cancer.

6.4 FDG and NSCLC

As explained in the section '*Cancer*' and the section '*Lung cancer*,' the three main subtypes of NSCLC are squamous cell carcinoma, adenocarcinoma, and large cell carcinoma (20). Smoking history and squamous cell carcinoma are predominantly associated with each other (47). The subtypes differ in their cell of origin, location in the lung, and growth pattern (48). An essential difference between those three major subtypes of NSCLC is reflected in the ^{18}F-FDG uptake. This uptake of the radiopharmaceutical tracer is related to the SUVmean. De Geus-Oei et al. have shown that the ^{18}F-FDG uptake is the highest in squamous cell carcinomas, followed by adenocarcinomas with significantly higher uptake in ^{18}F-FDG than large cell carcinoma (49).
Furthermore, the plasma glutamate concentration seems to have a complementary role. With the help of the plasma glutamate concentration, a differentiation can be made between inflammation and cancer in the lung (50). ^{1}H-NMR spectroscopy can be used as a tool to determine this difference in concentration. Vanhove et al. showed that a relative glutamate level, in PET-positive patients, less than or equal to 0.31 indicates lung cancer diagnosis, and a level above 0.31 is correlated to inflammation (50).

In the following sections, the device of a PET scanner, the processing of an image, and data acquisition to obtain a medically qualified PET scan will be explained more in detail.

6.5 Device and image processing/ data acquisition

6.5.1 Detector

Around the patients, a ring of detectors made of lutetium oxyorthosilicate crystals will detect the photons emitted back-to-back from the positron annihilation.
The detectors are all electronically coupled with each other. That way, opposite detectors simultaneously identify a pair of emitted photons using coincidence detection circuits (30). Alignment is obtained due to the electronic coincidence detection, which results in no need to use collimators. This is why PET has a better detection efficiency than single-photon emission tomography (SPECT) (34). The first step of four for data acquisition of PET is to generate a list mode of the two detectors that measure a photon. A pair of detectors registering a coincidence will be included in an event list or list mode (51).

A line can connect two detectors that measure a pair of photons simultaneously, and this line is called a line of response (LOR). This is the second step for data acquisition. On this line, the annihilation must have occurred.

The annihilation will be localized along the LOR in conventional PET, but there is no information about where the annihilation has occurred. This indicates an even distribution of events along the LOR and adds noise to the image (52). That is why modern PET scanners rely on the effect of time-of-flight (ToF), where information about the time is directly incorporated to reconstruct the image. ToF was for the first time identified around 1980 (53).

The correlation of two physical events is an essential part of the reconstruction of an image. This will be done with the help of a coincidence window. First, a detected photon will be linked to a specific detector and detection time (53). This is done for all the detected photons. The difference between two detected photons will be compared to a set coincidence window of 540 picoseconds (54). If this difference in time is more significant than the coincidence window, the two detected photons are considered physically uncorrelated. If the difference in time is smaller, the two events are correlated (53).

Characterization of the LOR is the smallest angle between the LOR and the center of the gantry (55). This characterization can be visualized in a graph and is known as the third important step in the process of data acquisition, where the x-axis is known as the shortest distance between the line of response and the center of the gantry, while the angle is plotted on the y-axis (51, 55). Figure 9A is a visual representation of four LORs.

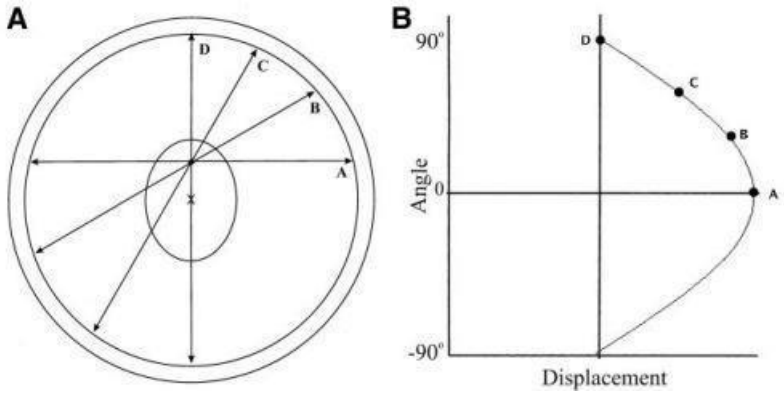

Figure 9: Sinogram formation. Each LOR of coincidence events is plotted as a function of its angular orientations and its displacement from the gantry center. In part A of the figure, the gantry center is marked by a cross (X). The 4 LORS are plotted where the angular orientation is plotted on the y-axis, and the displacement from the center of the gantry is plotted on the x-axis. Half of a sine wave is displayed if all possible LORs pass through the point where the 4 LORs are crossing.

When this is plotted for many LORs from the same point or pixel, half of a sine wave can be obtained. This resulting figure is known as a 'sinogram', where PET data are directly acquired and is visualized in figure 9B (55).

A sinogram consisting of a significant number of sine waves overlapping is the result of a more complex object and is visualized in figure 10. The sum of parallel LORs at a specific angle is a projection and is represented by a specific horizontal row in the sinogram. The sinogram of a PET will take into account all the projection angles and represent the data acquired per slice (55). This indicates that raw PET can be represented as a series of sinograms or as a series of projections, this for respectively a sinogram per slice and a separate view per projection angle. Step four of the data acquisition indicates that the transverse images are reconstructed with the help of the PET system's computer from the sinogram of projection data (30).

Figure 10: Complicated sinogram. This figure visualizes a series of sinograms of more complicated objects. This results in many overlapping sine waves. This is an example of a sinogram from a brain scan.

6.5.2 Scintillator

PET detectors are mostly inorganic scintillators. Around the scintillation crystal, an optical reflector is located to optimize the amount of interaction within the crystal. Rayleigh, Compton, and photoelectric effects will cause interaction of the annihilation photons within the scintillator. A PET scan will focus on events based on Compton and photoelectric effects due to the more considerable energy released in these effects (34). The photoelectric effect appears when a photon collides with a bound orbital electron and ejects the electron from the atom, and the Compton effect appears when a photon interacts with a loosely bound electron (56). A part of the energy from the initial photon is imparted to the electron, and the initial photon is scattered (56). If the annihilation energy is deposited in one location, it is known as the photoelectric effect. In contrast to Compton interactions, where the energy is deposited in several points of the crystal block (34). Because both effects are related to the atomic number, a crystal block with a high atomic number is preferred.

The scintillation process is based on the absorbed energy of the annihilation that results in a higher state of the crystal lattice. This characteristic state will emit lower-energy photons, also known as the scintillation photons, due to the decay that has taken place after a characteristic time (34). A photomultiplier tube (PMT) photocathode detects the scintillation photons, whose amplitude is a linear relationship with the electronic signal. At the front-end, electronics, such as pre-amplifiers, are located to process the signal further. Other properties will influence the detector's efficiency, such as the detector's energy resolution (34, 51).

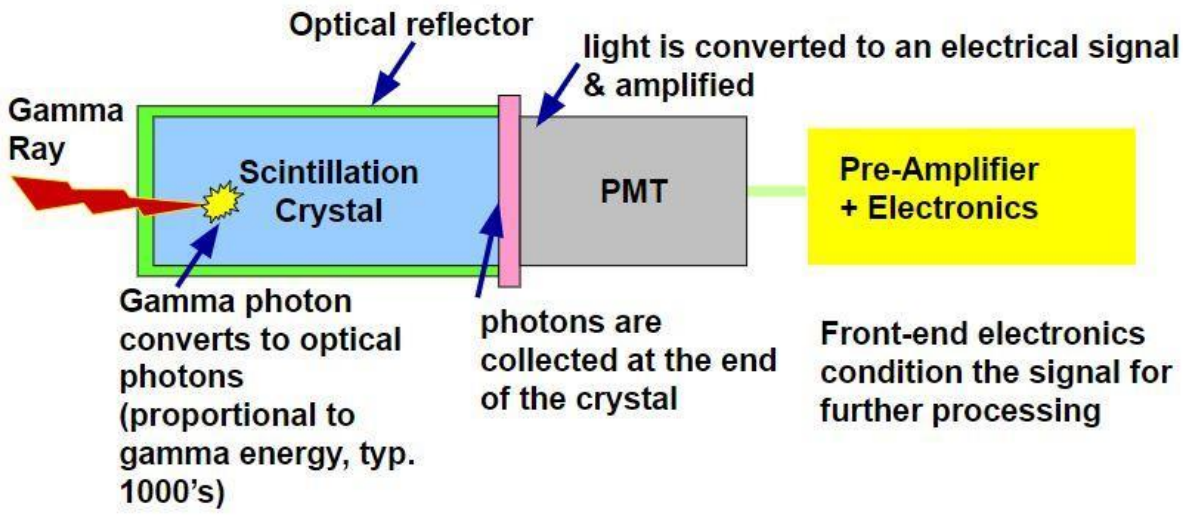

Figure 11: Electronics of a PET detector. A visualization of the electronica in a PET detector. As main parts the scintillation crystal, photomultiplier, pre-amplifier, and other electronics.

During a PET scan, essential parameters can be extracted and analyzed for further research.

6.5.3 PET parameters

An essential parameter during a PET is the relative measure of ^{18}F-FDG uptake, referred to as the Standard Uptake Value (SUV) (56). This is known as the ratio of the activity concentration measured by the PET scan relative to the initial injected activity of ^{18}F-FDG that is decay corrected. This initial activity is divided by the mass of the patient expressed in milliliter (42, 51). Formula 4 describes the ratio that SUV defines. This parameter is dimensionless.

$$SUV = \frac{activity\ concentration\ at\ time\ of\ the\ PET\ [Bq/ml]}{initial\ activity\ injected\ [Bq]\ /\ patient\ weight\ [ml]} \quad (4)$$

The SUVmean is defined as the average of the SUV in a region of interest. Because of image noise and a limited resolution, defining the boundaries of the region over which the average is computed will be challenging (56). Due to this challenge, simply taking the average of the SUV for a region of interest (ROI) is not a reasonable estimate for the SUVmean (56). Accurate measurements of the SUVmean are only good when the central region of a relatively large lesion is used. In addition, a uniform SUV is needed over the region of interest (56).

This leads to a need for a more accurate value, the SUVmax. In contrast to the SUVmean, the SUVmax is invariant to a slight shift of the region of interest. That way, the SUVmax is more stable than the SUVmean. However, a single value of a pixel will bias the SUVmax and lead to more noise than the SUVmean (56).

^{18}F-FDG represents the whole body total metabolic tumor volume (MTV) and thus represents the volume of tumor tissue. More specifically, it is a quantitative measurement of tumor cells. Higher glycolytic activity is an evident characteristic of tumor cells (57). It is a commonly used parameter to give information about the tumor.

The product of the SUVmean and the MTV is known as the total lesion glycolysis (TLG) (58). This is an exciting parameter because it combines the volumetric as the metabolomic information of the patient obtained by a PET/CT-scan (57). Several studies have shown the importance and usefulness of TLG, especially for the treatment response, not only for lung cancers (59-61). After the explanation of PET and the corresponding parameters, the next chapter will focus on CT.

7 Computed tomography

7.1 X-ray production

X-rays result from converting the kinetic energy of accelerated charged particles into electromagnetic radiation by multiple collisions with a target (62).

Bremsstrahlung falls under the category of electromagnetic radiation and is released when a charged particle (e.g., an electron) loses energy due to collisions with atomic particles (63). This phenomenon is shown below in figure 12, where multiple electrons are deflected by a nucleus and lose kinetic energy together with the emission of energy (bremsstrahlung photon) (63). The output that is generated is a continuous spectrum of different x-ray energies (62).

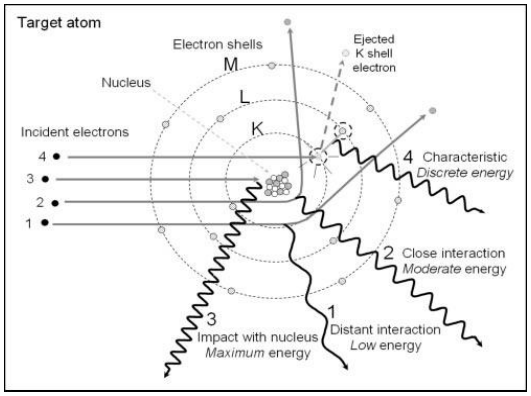

Figure 12: X-ray production. Numbers 1,2, and 3 show incident electrons that interact nearby the nucleus. This results in bremsstrahlung production. Number 4 shows characteristic radiation emission (62).

Another interaction of the accelerated electrons with the atomic shell that can occur, is the ejection of an inner shell electron. This is also shown in figure 12. A K-shell electron has the highest binding energy and lowest number of electrons on the shell. When an incident electron removes a K-shell electron, a vacancy is created. This vacancy is then quickly filled with an electron from a lesser bound shell. During this process, characteristic energy is released. Figure 13 shows the resulting output spectrum of both bremsstrahlung and characteristic radiation. The continuous spectrum of bremsstrahlung is visible, and the characteristic radiation is represented by the monoenergetic spikes on the continuous bremsstrahlung spectrum (62).

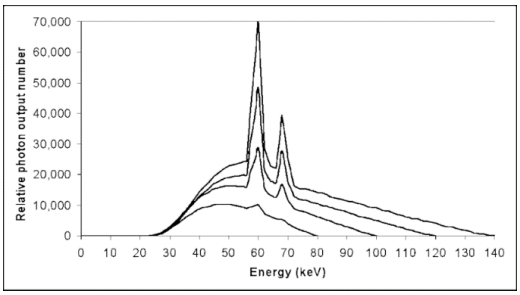

Figure 13: Radiation spectrum with bremsstrahlung and characteristic radiation for a tungsten anode (64).

7.2 X-ray tube

Computed tomography (CT) is a computerized x-ray imaging procedure (65). An x-ray tube is used as the source for x-rays. The x-ray tube is shown in figure 14.

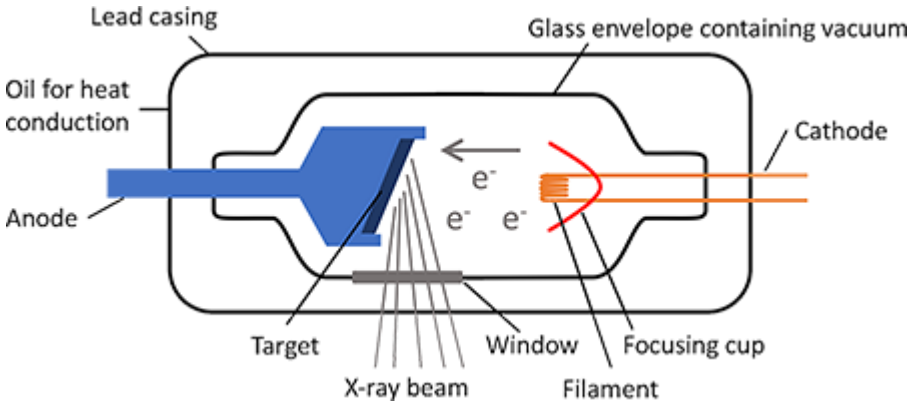

Figure 14: X-ray tube equipment (66).

The two most important parts of the tube are the filament and the target. The filament (cathode) is electrically heated and generates electrons, due to this heat. These electrons are accelerated towards the target (anode) through a high voltage tube. The target is a high atomic number material (e.g., tungsten), so that much bremsstrahlung is produced when the electrons from the heated filament hit the target (67-69).

The collision of electrons from the cathode with the anode is called the focal spot (70). The intensity or electron flow in the x-ray tube is expressed as a current in units of milliamperes (mA).

A vacuum chamber is placed around the filament and target. This makes sure that the electrons can travel from the filament to the target without disappearing by reacting with the air molecules.

The window is used to focus the x-ray beam that leaves the tube. A lead casing surrounds this whole structure to ensure minimal unwanted exposure from the x-ray beam to the environment and for extra filtration (68, 69).

The following section discusses the acquisition of data, followed by the reconstruction of an image. Last, the combination of PET and CT will be shortly discussed.

7.3 Data acquisition

When using x-ray beams for CT imaging of a patient, a couple of other elements are necessary between the x-ray tube and the patient, like a collimator to shape the beam and a detector for dose validation.

An x-ray tube's image is created due to a difference in attenuation between different tissues inside the patient (67).

The attenuation of the x-ray beam is associated with the density of that specific tissue. When the density of tissue is high, the possibility of detecting a photon by the detectors is lower, then when the density of tissue is low. This is because the probability of interaction with the atoms in the tissue rises linearly with the density of the tissue (71).

The attenuation in a specific tissue is expressed by its attenuation coefficient μ and is also directly related to the density of the tissue. This is, however, not the only factor where μ depends on. Two other factors are the thickness of the respective tissue and the energy of the x-ray beam (71).

With computed tomography, the x-ray tube is installed in the gantry, where multiple detectors are installed. This is illustrated in figure 15.

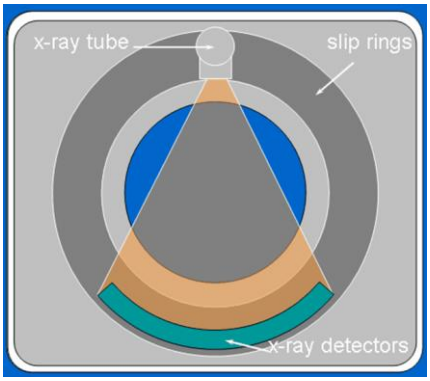

Figure 15: Schematic figure of a CT scanner (72).

During a CT scan, μ is measured in each detector at a certain angle of the gantry to get a projection at this specific angle. Then, the gantry is rotated over a small angle, and the measurements are done again to get another projection at this angle. The combination of all these projections forms a sinogram, as explained in the section about PET.

7.4 Reconstruction

Image reconstruction aims to map the attenuation coefficient distribution of a specific tissue through a volume. The first step of reconstruction is to calculate the CT-number, expressed in Hounsfield units (HU). It can be calculated with the following formula:

$$CT\ number\ [HU] = \frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \times 1000 \quad (5)$$

As displayed in this formula, the calculated CT-number is the relative attenuation compared to the attenuation in water, set to 0 HU (70).

A crucial reconstruction technique within this field is the filtered back-projection (FBP). This is an analytical technique. By back-projecting each available attenuation measurement for each point in space, an image is obtained. It is also called the convolution method since it applies a convolution filter to remove the blurring present in the image with the conventional back-projection method (70, 73).

Since the computational power increased over the last few years, iterative reconstruction techniques were implemented to reconstruct images. The basic idea of this technique is that various reconstructions are necessary to obtain a more accurate image (64, 70). A general flowchart of how these algorithms may look, is shown in figure 16. The first step of this technique is to start with an estimate of the image. Then, the projections are calculated and compared with the actual projections. Last, the result of the calculations is used to modify the current estimate (67). This technique is currently not efficient enough to replace the FBP technique.

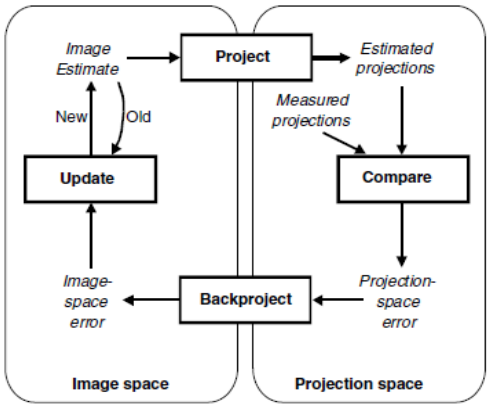

Figure 16: Visualization of the iterative reconstruction algorithm (64).

There can be many artifacts when an image is reconstructed from a CT scan. There are two big categories: physics artifacts and patient-based artifacts. Most of them are fixed with filtration, calibration correction, and correction software use (67).

A combination of positron emission tomography and computed tomography has the significant advantage of collecting, in one scanning, both anatomic analyses and metabolomic or functional images in vivo (74). CT, PET, and their parameters lead to the central part of this research, radiomics.

8 Radiomics

In 2012 the concept of radiomics was presented by Lambin et al. (75). The extraction and analysis of quantitative medical imaging features with high throughput is known as radiomics. A medical image can carry much information. This information can reflect underlying pathophysiology with the help of quantitative image analysis (72). Digital clinical images are obtained for almost every cancer patient, a significant strength of radiomics because all these images are potential radiomics databases (72). These digital images might be the foundation of a specific discriminative clinical model, in this case, for NSCLC.

The main goal of radiomics is to create mineable databases to develop descriptive and predictive models for valuable diagnostic, prognostic, or predictive information (76). This means that these systems provide support to make the best clinical decisions to optimize the individual treatment. Combining radiomics data and other patient characteristics may lead to the development of evidence-based clinical decision support tools and models (72).

Radiomics is a reasonably fast-growing, recent, and multidisciplinary technique that will play an increasingly important role in the medical world by combining data, most of all obtained out of images, in combination with genetics and bioinformatics (75). In the next section, an overview of results in radiomics will be explained for both oncological and non-oncological applications.

8.1 Overview results radiomics in oncological and non-oncological applications

8.1.1 Non-oncological applications

Radiomics can have some non-oncological applications. For example, when characterizing damage in the lung due to radiation pre-radiotherapy and post-radiotherapy, CT scans can be used. More specifically, the variation in texture features and corresponding values between those CT scans are used (77).

Imaging features are also often used in neurological applications. This way, Alzheimer's and multiple sclerosis can be diagnosed, staged, or prognosed with the help of imaging features (78, 79). Computed tomographic angiography has been used to analyze textures, more specifically after endovascular prostheses, to predict endovascular leak (80).

8.2 Oncological applications

Several studies have hypothesized a link between tumor characteristics at the cellular level, genetic level, and the phenotypic pattern (81-83). This link can be captured by medical imaging, such as PET and CT (75).

If we take a better look at studies done with ^{18}F-FDG and a PET scan, it is proposed that a non-uniform distribution of ^{18}F-FDG is linked to tumor heterogeneity (81, 84, 85). This gives better insights into the volume and treatment of a tumor.

8.2.1 Breast

Genomic features as microRNA expressions were associated with radiomics phenotypes by a radiogenic study (86). More specifically, the radiomics features showed associations with microRNA expressions and transcriptional activities of pathways (72). These microRNAs expressions were linked to the size of the tumor (87).

The combination of genomics and radiomics data of magnetic resonance imaging (MRI) resulted in a significant improvement in prediction performance compared to analyses done by genomics or radiomics alone (88).

8.2.2 Glioblastoma (GBM) and prostate

Immunohistochemically identified protein patterns of glioblastoma multiforme (GBM), which were predicted with the help of radiomic features obtained from magnetic resonance MR (82). Radiomics features were correlated with a compelling prognostic factor for prostate cancer, the Gleason score. Furthermore, these radiomics features were also related to biochemical recurrence following prostate radiotherapy (89).

8.2.3 Vulvar cancer

In 2019, Collarino et al. examined vulvar cancer based on radiomics with the help of ^{18}F-FDG PET/CT images (90). These radiomics features were identified by principal component analysis (PCA) for 40 women with a primary tumor of at least 2.6 cm in diameter. Furthermore, each woman received an ^{18}F-FDG PET/CT scan followed by surgery. The study concluded that PCA could be used to perform dimensionality reduction for radiomics. This statistical step is essential when a large radiomics database is extracted from PET or CT images. Only the parameters that significantly contribute to the predictive model should be considered and, therefore, be identified using PCA. PCA goes further than radiomics features used in the present clinical workspace (e.g., SUVmean or Grey levels) (90). In this study, the PCA analysis showed that both known (e.g., SUVmean) and unknown parameters (e.g., Moran's I) are essential for the creation of this predictive model (90).

8.2.4 Lung

NSCLC is broadly studied and characterized with the help of radiomics (87). The focus to obtain the images needed for radiomics until now is CT, which is also helpful in treating a patient. Studies show that these radiomics features are related to the tumor stage and histopathology (79, 91).

Conventional, radiomics and combined features are based on multivariate predictive models to create, at the time of surgery, predicting pathological response to neoadjuvant chemoradiation assessed (88).

A radiomics signature as a combination of four features was built by Aerts et al. (92). This combination of features consists of 'statistics energy,' which gives more information about the overall density of the tumor. A second feature is the ‘shape compactness,' which indicates how compact the tumor is. The third feature is a measure of wavelet and heterogeneity and is described in 'gray level non-uniformity.' The last feature describes the intra-tumor heterogeneity and is known as 'gray level non-uniformity HLH.' Importantly, this last feature can only be extracted after decomposing the images in mid-frequencies (87). The most stable features were selected by using the RIDER dataset on a set of 422 lung cancer patients (87). Aerts et al. concluded that the radiomics signature was predictive for survival due to a confidence interval of 0.65.

Furthermore, the signature was successfully tested on different cancers, such as lung, head, and neck cancer (92). Out of multi-center data of 201 patients, Ohri et al. published a predictive radiomics model (93). With the help of the Lasso procedure, one textural feature was identified. This feature was calculated from GLCM and SumMean to predict overall survival, complementary to MTV with an optimal cut point of 9.3 cm (93).

The general process of radiomics, explained in the next section, will provide insight into how these radiomics features related to specific organs can be obtained.

8.3 Process

This process can be defined in several steps, starting with the acquisition of images. This step is followed by segmenting volumes, such as a tumor located in the lungs, as will be performed in this study. Out of these essential features, more specific radiomics-image properties are extracted. The last step includes developing, analyzing, and validating models obtained by this process (75). The basic idea is to collect as much data and information as possible at the front end and use databases to identify radiomics features of which the prognostic value is the highest (72). This philosophy leans on process engineering. A more detailed overview of this process can be found in figure 17 (76).

In general, the first step is to obtain clinical images of study patients. This study is done with the help of PET/CT. Information about tumor phenotype and microenvironment can be extended from image features such as intensity, shape, size, volume, and texture (72). The detailed information about the PET/CT images can be found in the sections describing PET and CT. The images are obtained by using a standard protocol.

Furthermore, the obtained images must be de-identified; this is done by anonymizing the images linked to one patient by a specific code.

The second step, including segmentation of images into volumes of interest (VOIs), is crucial. This is a very challenging step because the borders of tumors are often indistinct. Segmentation of the VOIs should be accurate and reproducible to be time-efficient and reduce the interaction of operators, such as medical experts or doctors (76).

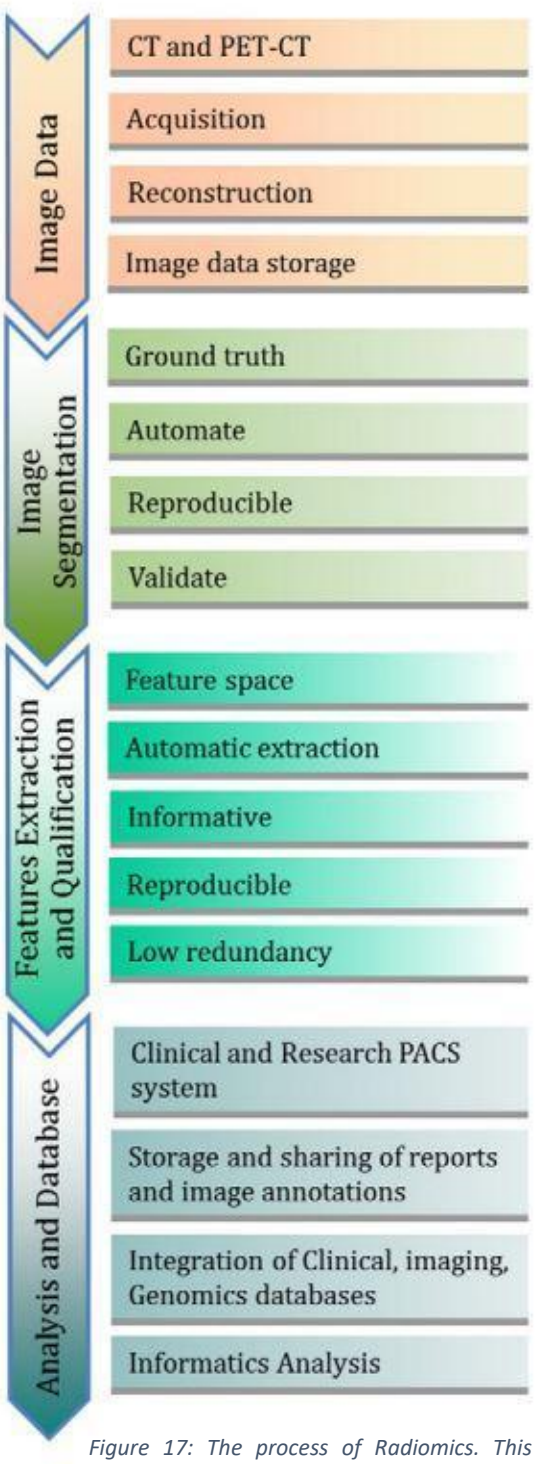

Figure 17: The process of Radiomics. This scheme visualizes the four essential steps starting from the image data, followed by image segmentation, feature extraction, and qualification, and ending with analysis and database; Picture Archiving and Communications Systems (PACS).

The third step is feature extraction and qualification, where semantic and agnostic features are extracted. On the first hand, the features that are commonly used to describe regions of interest are semantic. On the other hand, the features which describe lesion heterogeneity are known as agnostic features. The potential of radiomics to measure intra- and intertumoral quantitative heterogeneity is a central driver for radiomics research (76).

The most systematic approach is first to identify redundant features (76). The extracted features are often visualized in a covariance matrix, where highly correlated features are indicated in the same color. Figure 18 shows an example of 219 features extracted from CT scans in 143 patients diagnosed with NSCLC (64, 76). Highly correlated features can form clusters and can be collapsed into one representative feature. Aerts et al. further ranked features based on different agnostic and semantic classes of features (92). Models can be built out of the two or three features of each class with the highest priority.

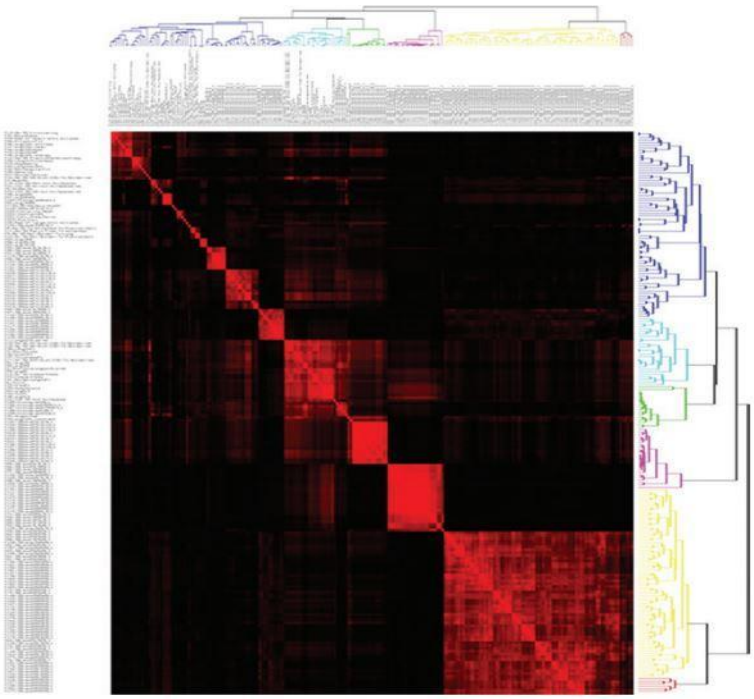

Figure 18: Radiomics features in a covariance matrix. This matrix is an example of 219 features that were extracted from NSCLC tumors in 235 patients. With the help of regression analysis, the features were compared, and correlation coefficients (R^2) were generated. Features with high correlation were clustered and plotted along both axes and shown in red.

The last step of the process is to construct these models and analyze the outcoming results. Important in this step is to minimalize the risk of overfitting (75). When this is the case, this will lead to models that are only specific for the given set of data, and those models will not perform well on a new, independent set of data. Data mining is known as discovering patterns in large data sets and is necessary to develop a model (75). This can be done via artificial intelligence, machine learning, or with the help of a statistical approach (72).

Notably, the model must be validated before it can be used in a clinical setting. Thus, it is essential to test the model using an independent validation dataset and document the statistics methods carefully (75). An overview of the different steps of radiomics is given in figure 19 (64).

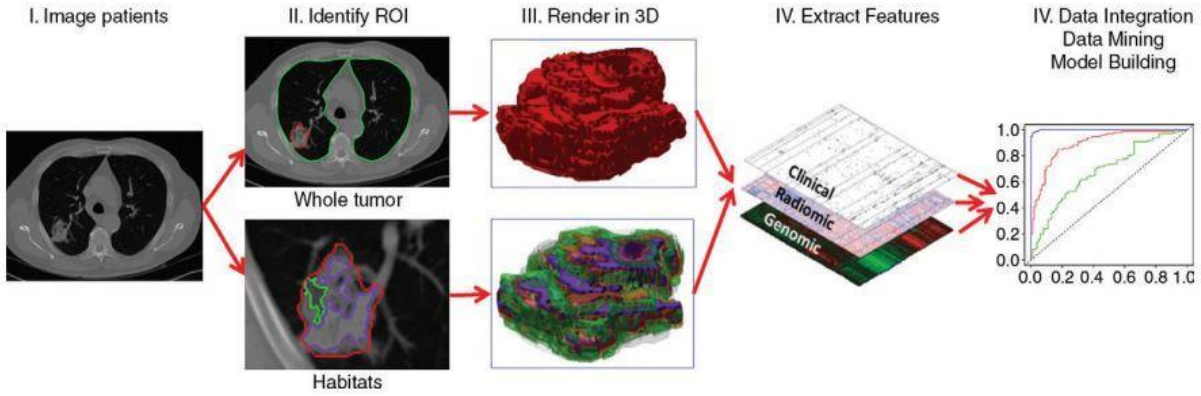

Figure 19: Flowchart of a Radiomics process. Medical images obtain information about the patients. Secondly, the region of interest is identified. Afterward, these are segmented with operator edits and eventually rendered in three dimensions. From these rendered volumes, quantitative features can be extracted. Finally, these data are ready for developing, predicting, or prognostic models; Region of Interest (ROI).

9 Factor analysis

To reduce and correlate the parameters obtained from radiomics techniques, Factor Analysis (FA) is often used. It is a statistical tool to observe a correlation between the initial parameters and combine and reduce these by representing a common factor (94). With the help of factor analysis, the dimensions and structures within the data are also identified.

The technique is based on finding a common variable (e.g., factor I). This factor is responsible for creating two or more variables (e.g., S1, S2, and S3). This indicates that factor I is the reason behind the correlations and association between S1, S2, and S3 (94). Figure 20 gives a visualization of the common factor, variables, and the relationship between these two. Here, factor I and factor II are the common factors, introducing three variables (respectively S1, S2, S3 and S4, S5 and S6). The three variables are reciprocally correlated, indicating a line between the variables (94).

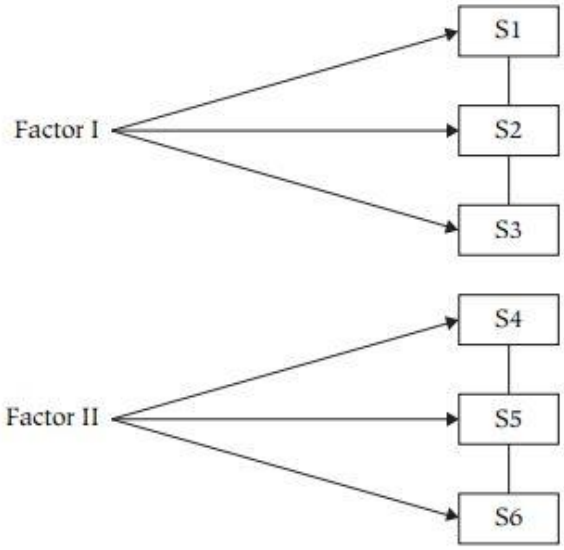

Figure 20: Common factors, variables, and underlying correlation. The common factors are named factor I and factor II. Furthermore, the variables are named from S1 up to and including S6. The correlation between the variables is indicated with a line between the concerned variables. An arrow between the factor and variable indicates that this factor introduces these variables.

Identifying a common factor starts from the correlation matrix, which is filled by variables. The common factor that will be determined first is the one that will explain the most correlation in the matrix with variable. This way, the common factor is generated by a theoretical procedure (94). The correlation matrix can be expressed in the form of a heat map. This is represented in figure 21 (95).

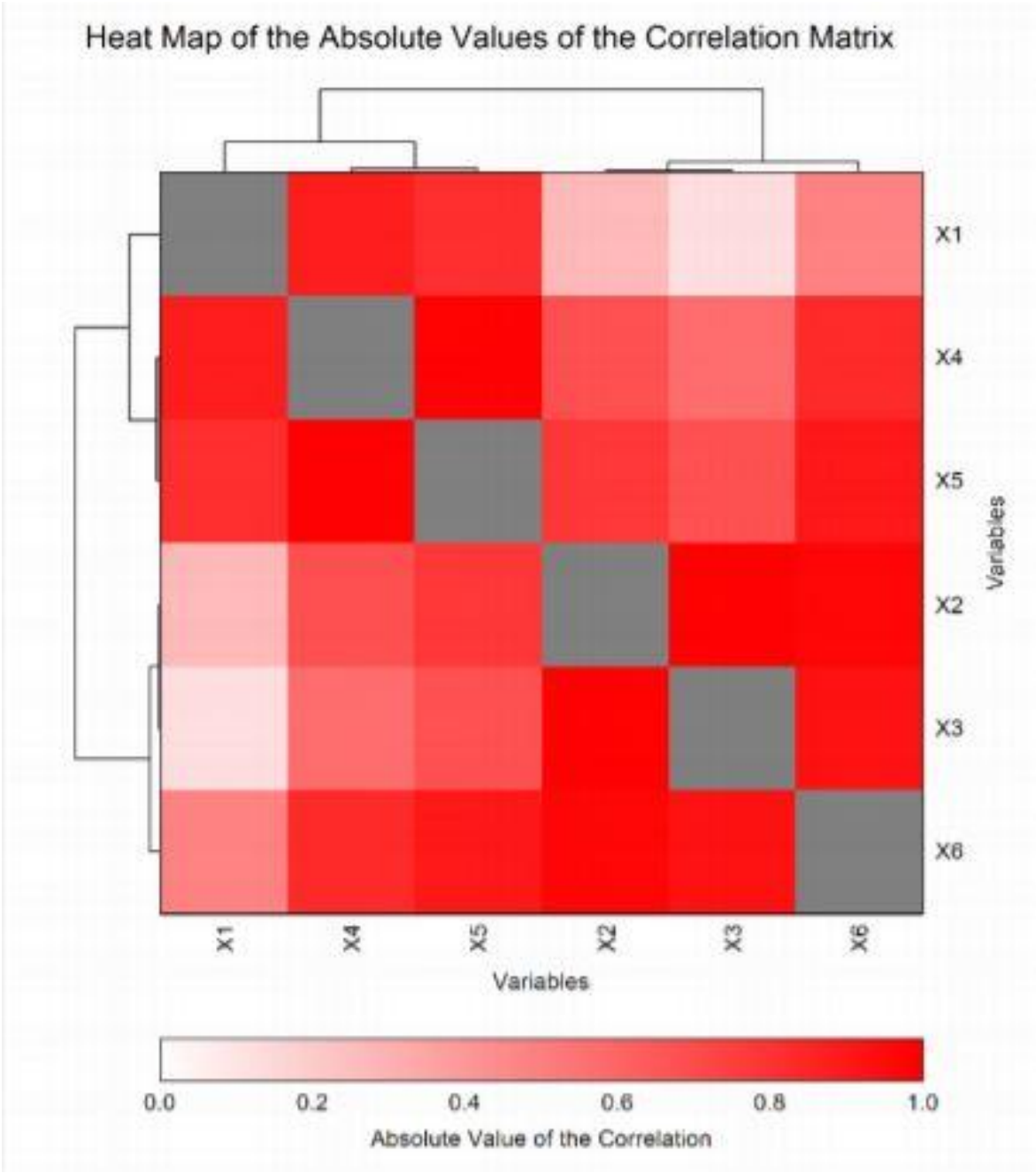

Figure 21: Heat map of the absolute values of the correlation matrix. If the absolute value of correlation increases, the spot will be a deeper red. If the correlation between the variables is less, the spot will turn into a soft spot of red.

10 Principal Component Analysis

Because Principal Component Analysis (PCA) is also often used in literature to reduce parameters, an explanation is given in this section.

PCA, a mathematical algorithm known as Karhunen-Loeve expansion, is based on orthogonal transformations where a conversion of a set of possibly correlated variables into a set of values that are linear correlated occurs (96, 97). The reduction of parameters can be accomplished by identifying directions, which are named principal components (PCs) (97). These principal components are directed to where the variance in data is maximum. That way, PCA requires knowledge of linear algebra and statistics (96).

PCA is based on Eigen vectors and Eigen values. An Eigen vector is a type of vector that does not change after a transformation is applied (98). PCA calculates the Eigen vectors of the covariance matrix. These project the original data onto a feature space with lower dimensions. Eigen vectors define this space with large Eigen values (99). An Eigen value is a scalar that transforms an Eigen vector, and this can be visually seen by stretching the Eigen vector with an own direction and magnitude. In other words, the new Eigen vector is a scaled version of the original vector (98). Figure 22 illustrates an Eigen vector (x), Eigen value (λ), and matrix A (100).

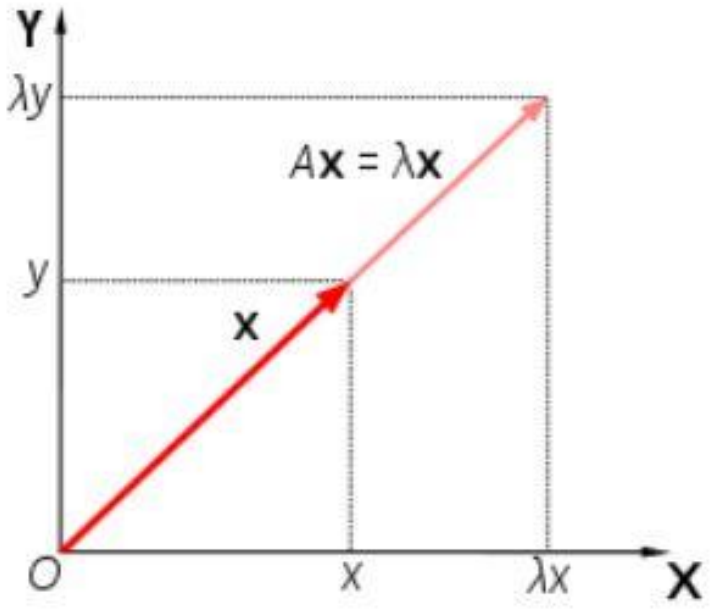

Figure 22: Visualization of an Eigen vector and Eigen value. A is a matrix that stretches the Eigen vector X. There is no change in direction, only in the scale of the Eigen vector.

Relatively few numbers can represent a set of relatively large values by using a few components. After that, the remaining parameters will be plotted to visually asses similarities and differences and determine whether parameters can be grouped together or not (97).

The principal component that is chosen first is the one that minimizes the total distance between the original data and their projection onto the principal component. A minimalization in this distance results in a maximization of the variance of the projected points (101). The second and subsequent components are selected the same way. Each time, there is one additional requirement for these principal components, which indicates that they are not correlated with all the previous PCs (101). Figure 23 provides a visual representation of the data set before and after the principal component analysis (102).

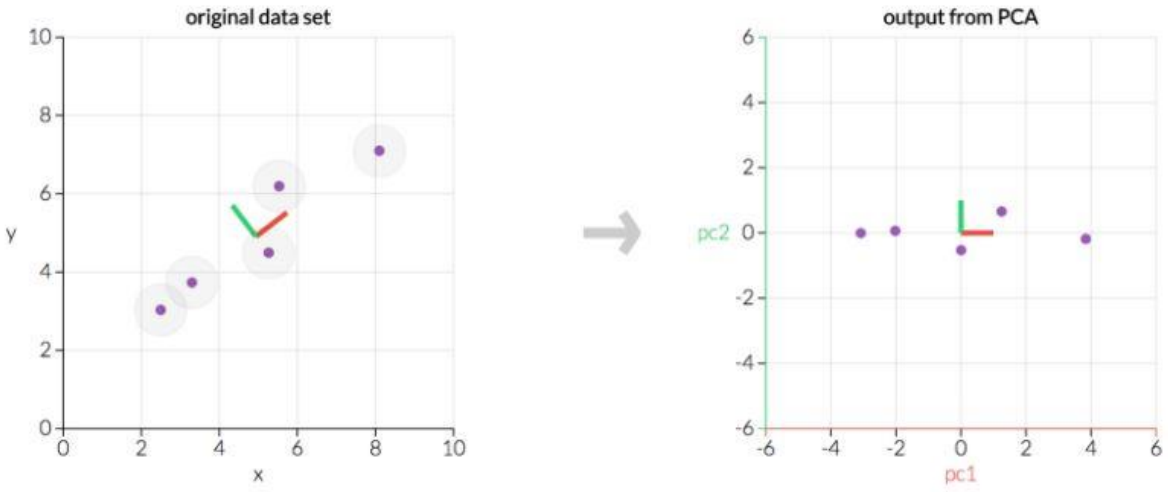

Figure 23: Example of Principal Component Analysis. The left graph represents the original date, and on the right, the transformed dataset is graphically represented. In this example, the PCA is performed on five points. PCA; Principal component analysis.

Like any other statistical technique, PCA has its advantages and drawbacks. A couple advantages are low noise sensitivity and decreased used memory and increased efficiency, both are logical consequences of the smaller dimensions of the data after PCA (99, 103). Furthermore, a lack of redundancy of the data given the orthogonal components is a key advantage. Another essential advantage of PCA is the switch from a difficult evaluation of the covariance matrix to a visual representation accurately. A disadvantage is that the training data explicitly needs every information of the simplest invariance. Otherwise, this invariance could not be captured (99, 103).

When using radiomics to build predictive or discriminative models, PCA can be used right before the actual data analysis. This step is called 'Post Processing.' With the help of PCA, the original features are combined through the transformations explained above, to form new features. These new features are then processed into a model during the actual data analysis (104).

A difference with FA is that in the PCA technique, the common variance, represented by factor I in figure 20, becomes maximized and is not unique for each variable (94). In contrast to FA, where there is assumed that there are a substantial number of unique variances.

11 Correlation tests

11.1 Pearson correlation test

A frequently used method to examine a correlation between two normal continuous distributed datasets is the Pearson correlation coefficient. This statistical method is known as a linear correlation technique (105).
Karl Pearson defined formula 6 as Pearson's correlation coefficient, whereby X and Y are two normal continuous distributed parameters of two different datasets. Random samples of size n for variables X and Y are denoted, respectively X_1,..., X_n, Y_1,..., Y_n and $\bar{X}$ and $\bar{Y}$ (106).

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{[\sum_{i=1}^{n}[(X_i-\bar{X})^2][\sum_{i=1}^{n}(Y_i-\bar{Y})^2]}} \quad (6)$$

The solution of the Pearson correlation analysis is a numerical value that indicates how well all variables of the different datasets are correlated two by two (107).

The Pearson correlation coefficient can take a value starting from -1 to +1. Where a perfect positive correlation is known as +1, and a negative correlation is represented by -1. A correlation value of 0 equals to no correlation between the two variables (108).

11.2 Spearman correlation test

If the two datasets are not normally distributed, another method is needed to obtain a correlation coefficient based on the datasets. An often-used method is the Spearman correlation test. This method is based on the ranked values for each variable and, that way, not based on the raw, original data (109).

No assumptions of the distribution of the data are known and carried by the Spearman correlation test. In addition, it needs an ordinal scale of variables to run the correlation analysis (110).

The Spearman rank correlation is calculated with the help of formula 7. In this formula, ρ stands for the Spearman rank correlation, d_i for the difference between the ranks of corresponding variables and the number of observations is denoted as n.

$$\rho = 1 - \frac{6 \sum d_i^2}{n\,(n^2-1)} \quad (7)$$

Like the Pearson correlation test, the Spearman correlation test returns a value from -1 to +1. A perfect positive correlation is denoted by +1, a perfect negative correlation is linked to a correlation value of -1, and when there is no correlation, the correlation value is similar to 0 (111).

12 The Chi-square test

A statistical test that was used during this study is the Chi-square (χ^2) test. This method is a nonparametric statistical analysis, that is often used to decide if the classifications of two samples is dependent or not. That way, the result of a Chi-square test is an indication of the deviation or distance between two sets (112).

To use a chi-square test, the following requirements need to be fulfilled (113):

1. The measurements levels of all variables need to be nominal or ordinal.
2. The sizes of the sample of the study groups are not equal and the groups of χ^2 may be equally or unequally sized.
3. The original data are measured at an interval or ratio level, but violate one assumption of a parametric test, which are listed below:
    a. The researcher uses a distribution free statistical analysis instead of a parametric analysis, due to a seriously peaked or skewed distribution.
    b. The condition of an equal variance or homoscedasticity could not be met.
    c. The data are no longer a ratio or interval, because the continuous data are collapsed into a smaller number of categories.

The method to look at the association between datasets is based on the difference between the observed and expected frequencies, respectively *O* and *E* (114). The actual count of cases is equal to the observed frequencies. The expected value is determined with the help of the row marginal for that cell M_R, the column marginal for that cell M_C and the total sample size n. Formula 8 represent the expected value and formula 9 describes the Chi-square test (113, 114).

$$E = \frac{M_R \times M_C}{n} \tag{8}$$

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{9}$$

13 Logistic Regression

To go from the dataset with parameters to a discriminative model based on these parameters, a logistic regression model is often used. This is a statistical method used on parameters to classify a malignant lung tumor and a non-malignant lung tumor, or to classify two histology's, based on these parameters. This method is chosen because of its success in prediction and diagnosis in medicine (115).

The logistic function used in this logistic regression model is defined by formula 10.

$$logistic(\eta) = \frac{1}{1 + \exp(-\eta)} \tag{10}$$

The output of this logistic regression model is probabilities between 0 and 1. The equation to interpret these probabilities is given by formula 11.
The term in the log() function describes the probability of an event divided by the probability of no event (116). This way, every feature that is used to base the model on gets a certain weight.

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_{1x1} + \cdots \beta_{pxp} \tag{11}$$

In the case of discrimination between a non-malignant lung lesion and a malignant lung lesion, the logistic regression model works with two classes or possible scenarios, the two types of lesions. One is labeled 0, and the other is labeled 1, or respectively 'no event' and 'event' (117). The same method can be used when building a discriminative model for two histology types, where one of the types is labeled '0', and the other one is labeled '1'.

Another frequently used predictive algorithm is linear regression, but a logistic regression model is chosen in this study. This because a linear regression model can only represent linear correlations. The interpretation of the outcomes of a linear regression model is often more complex since one feature depends on all other features. For the goal of this study, a logistic regression model works best to make the classification between a non-malignant lung lesion and a malignant lung lesion, and between the pathology of an adenocarcinoma and a squamous cell carcinoma (116, 118).

14 The ProLUNG Study

This research on non-small cell lung cancer (NSCLC) patients using radiomics is part of a study named the ProLUNG study, which has two main branches: metabolomics and radiomics. In this ongoing research, metabolomics is used to predict the therapy response of patients with phase I-IIIA NSCLC. More specifically, this means that with the help of a blood sample and proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy, specific metabolites are analyzed. The main goal of the ProLUNG study that focuses on metabolomics is to develop a biomarker based on the metabolism, a so-called metabolic profile, which may predict a possible relapse after surgery for a specific patient. This can optimize the treatment of NSCLC and make it more specific for the patient as an individual. This part of the study is visualized in the central part of figure 24.

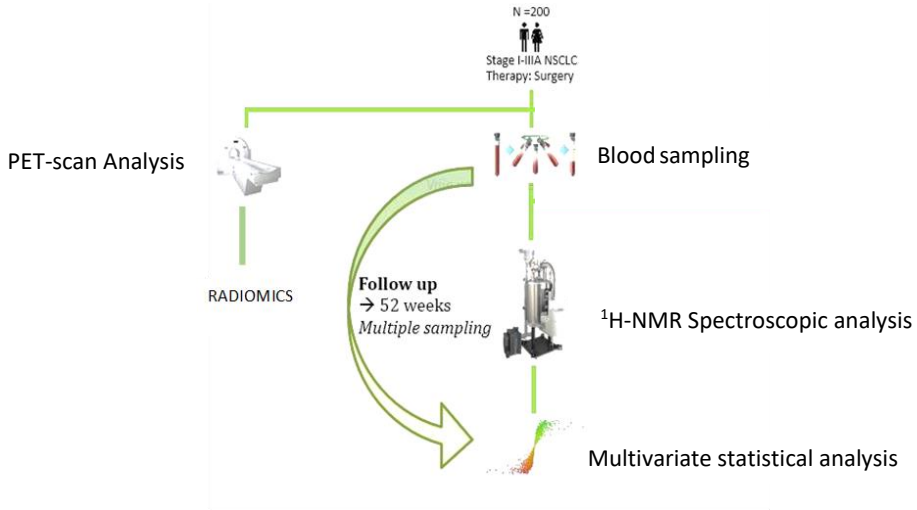

Figure 24: Schematic representation of the ProLUNG study. This scheme visualizes an overview of the different aspects that are being studied in the ProLUNG study. The focus is on radiomics (left) and metabolomics (right).

The patient cohort in the metabolomics study exists of 142 patients. These patients have been diagnosed with stage I-IIIA lung cancer or early-stage and locally advanced NSCLC. The tumor has not metastasized in these stages and all patients received a lobectomy as part of their standard-of-care treatment plan.

This research focuses on the radiomics part of the ProLUNG study, as displayed in the left part of figure 24. The first goal of this research is to find underlying correlations between the metabolomics data of the patients and the radiomics data of these same patients. Another aim of this study is to see if radiomics can provide specific markers to generate models to discriminate between malignant and non-malignant lung lesions, and between the pathology of an adenocarcinoma and a squamous cell carcinoma and optimize the patient treatment plan (e.g., is an operation necessary, does the patient need chemotherapy afterward). This part aims to examine if a combination of metabolomics and radiomics can obtain a better discriminative model. The main goal of metabolomics and radiomics in the ProLUNG study is the same. The main difference between these two research fields is that radiomics uses parameters extracted from PET/CT images of the patients instead of extracted parameters out of the plasma metabolic profile of the patients.

15 Materials and method

Two patient cohorts are used in this study. The first patient cohort consists of 39 patients, and the second cohort consists of 85 patients. All patients are diagnosed with NSCLC stage I-IIIA. The patients' tumor has not metastasized, so the TNM-staging is M0 for all the patients included in this research. Furthermore, all the patients underwent a lobectomy as part of their standard-of-care treatment plan. The included patients underwent treatment in Ziekenhuis Oost-Limburg (ZOL), located in Genk, and gave written informed consent. A patient can drop out of the study at any time.

To limit external influences on the study, exclusion criteria are also introduced. The inclusion and exclusion criteria for the ProLUNG study are shown in table 1.

Table 1: Inclusion and exclusion criteria of the ProLUNG study

Inclusion criteria	**Exclusion criteria**
Stage I-IIIA NSCLC tumor	No fasting starting 6 hours before PET/CT
Lobectomy	Medication intake in the morning of PET/CT
Signed written informed consent	Fasting blood glucose concentration ≥ 200 mg/dl in the morning of PET/CT
	History of cancer during the past five years
	Treatment for cancer during the past five years

All included patients had a PET/CT scan, according to the European Association of Nuclear Medicine (EANM). All the patients were imaged using a Biograph Horizon PET/CT scanner from Siemens with a 16 slice CT in this research. The CT scans are used to correct the PET images for attenuation.

This specific PET/CT scanner uses lutetium oxyorthosilicate (Lu2(SiO4)O, LSO) crystals. These crystals have very high stopping power, high light yield, fast decay time, and good radiation hardness against gamma rays (54, 119). There are other crystals available, but LSO crystals provide better image quality (54).

The Biograph Horizon detectors use 4 mm LSO crystals, placed with no gap between two detector blocks to guarantee high spatial resolution and lesion visualization (54).

Due to the high-quality LSO crystals, ToF is supported, which improves the signal-to-noise ratio and the speed of the scans. Not only this, the injected dose is lowered, which is an excellent quality when looking at the ALARA principle (54).

In the software of the CT technology of Siemens, SAFIRE3 is included to lower the dose by 60%, and iMAR3 reduces metal artifacts of the CT images (54).

Technical specifications are listed in table 2. All the images obtained from the scan are saved in the Picture Archiving and Communications Systems (PACS).

Table 2: Specifications PET/CT Biograph Horizon from Siemens Healthineers

Gantry	
Bore diameter	70 cm
Tunnel length	130 cm
Table capacity	227 kg
CT	
Generator power	55 kW
Rotation times	0.48, 0.6, 1.0 and 1.5 s
Tube voltages	80, 110, and 130 kV
Iterative reconstruction	SAFIRE
Metal artifact reduction	iMAR
Slices	16, 32
PET	
Axial field of view	16.4, 22.1 cm
Crystal size	4 x 4 x 20 mm
Time of flight performance	540

15.1 ^{18}F-FDG PET/CT protocol

The radiopharmaceutical is delivered by UZ Leuven, where the production of ^{18}F-FDG is performed. With the help of an autoinjector, more specifically an Iris automated multidose injection system from Comecer, ^{18}F-FDG is injected in the patient to limit the radiation dose to the present medical staff. The medical images of all the included patients were obtained one hour after the administration of ^{18}F-FDG. After presetting and determining the imaging field, a CT (25 mA, 130 kV) was performed. This CT ranged from the midthighs to the base of the skull. A 512-512 matrix was reconstructed with the obtained CT images. Next, a PET-scan, which covers the same axial field as the CT-scan, was performed for 15 to 20 minutes. The emission time per bed position ranged from one to two minutes depending on the concerned patient's body mass index (BMI) (120). More specifically, a patient with a weight less than 50 kg, a weight between 50-80 kg and a patient with a weight above 80 kg are scanned respectively one minute, a minute and half and two minutes.

15.2 Radiomics and metabolomic parameters

First, the obtained PET/CT images of the included patients are retrieved from the PACS and made anonymous. Then, the images are loaded into the *Accurate* tool, developed by the research team of Prof. dr. R. Boellaard (Amsterdam, VUmc). After this step, the segmentation is done in the *Accurate* tool with the semi-automatic help of the tool itself and afterward controlled and corrected by doctor Mesotten of the nuclear medicine department of ZOL Genk. Examples of this tool and segmentation process are given in figure 25. In the last step of this part, the volumes of interest (VOIs) are saved. These are the VOIs based on the first segmentation method. This method of segmentation is only used for the first patient cohort, consisting of 39 patients.

These steps are then repeated to create new VOIs, based on the same PET/CT images. The difference with the previous VOIs is that the lesion was lined differently by doctor Mesotten on the CT image to fit the lung lesion's outline better. These VOIs are saved again. These are the VOIs based on the second segmentation method. This segmentation method is used for the two patient cohorts.

All the steps are done a third time, now only focusing on the PET-images of the patients. The segmentation to create the VOIs is done only on the PET images, and the VOIs are saved. These are the VOIs based on the third segmentation method. This method of segmentation is only used for the first patient cohort, consisting of 39 patients.

These saved VOIs are then loaded into a second tool, named *Radiomics*, which was also developed by the research team of Prof. dr. R. Boellaard (Amsterdam, VUmc). After the analysis of the VOI, new radiomics parameters are extracted from each VOI. These 483 parameters per VOI are saved in an Excel file per patient. After the 483 parameters are extracted for all the patients in the cohort, they are put together in one Excel file per method for segmentation.

Simultaneously, 238 metabolic parameters representing 62 plasma metabolites were determined from the same patients using proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. These parameters were also saved in an Excel file.

Figure 25: Examples of the 'Accurate' tool and segmentation of a lung lesion.

15.3 Datasets

15.3.1 Correlations

Per method of segmentation, the datasets are created the same way. First, the metabolomics dataset and the radiomics dataset of the first patient cohort were used to look at any underlying correlation between the metabolomics data and the radiomics data. The dataset contained 721 parameters (483 radiomics and 238 metabolomics). The Spearman correlation coefficient test was used on the total omics-dataset to measure the underlying statistical correlations between the radiomics parameters and the metabolomics dataset. This was done for the datasets created out of the three methods of segmentation.

The correlation between the radiomics and metabolomics data were examined and the difference in correlation between the data for three methods of segmentation was analyzed.

After this, the dataset with only the 483 radiomics parameters of 85 patients and the dataset with the 483 radiomics parameters of the 66 patients diagnosed with adenocarcinoma and squamous cell carcinoma, based on the second type of segmentation, were used to see any correlation between these parameters. The Spearman correlation coefficient test was used on these datasets to find the correlations. If the correlation between two parameters is more than 90%, one of the two correlated parameters could be removed without loss of information. When all the correlated parameters were removed, the residual dataset was saved as a new Excel file.

15.3.2 Discriminative models

The dataset with the 483 radiomics parameters of 85 patients and the dataset with 483 radiomics features of 66 patients are used to create discriminative models. The next step to achieve this is to reduce the parameters to a more usable number. This was done in two ways, the first one being with factor analysis and executed in RStudio, with a script created with the help of the research team of Prof. dr. Boellaard (Amsterdam, VUmc). The factor analysis for this dataset is performed for three, four, five, and ten outcome factors. These resulting factors were also saved as an Excel file. The FA reduction method was only used on the dataset with 85 patients.

The second way to reduce the dataset with the 483 radiomics parameters of 85 patients used the Spearman correlation coefficient test on this data set. Then, a threshold of 0.9 was chosen in the resulting correlation matrix, and for every two features that showed a correlation above this threshold, one was removed from the correlation matrix. These features were then removed from the dataset with the 483 radiomics parameters, resulting in a dataset of 85 patients with 56 features. These resulting features were also saved as an Excel file.

Every patient in both resulting datasets of the 85 patients with a malignant lung lesion was labeled as ‘1’ in a column ‘Event,’ and every patient in the dataset with a non-malignant lung lesion was labeled as ‘0’ in the ‘Event’ column.

Every patient in the resulting dataset of the 66 patients with an adenocarcinoma was labeled as ‘1’ in a column ‘Event,’ and every patient in the dataset with a squamous cell carcinoma was labeled as ‘0’ in the ‘Event’ column. All three datasets underwent a forward selection and backward stepwise selection regression to generate the discriminative models. This logistic regression model was built in RStudio. The manually reduced datasets were split into a 75% training dataset, to build the model, and a 25% test dataset, to test the accuracy of the model.

16 Results

16.1 Correlation between metabolomics and radiomics

16.1.1 Included patients

For the first goal of this research, the correlation of metabolomics and radiomics features, 39 patients were included. The most significant part of this group, 70%, consisted out of men. More than 50% of the patients confirm being a current smoker, which is a strong risk factor for lung cancer development (121).

In table 3, the complete patient cohort is shown in more detail. A specification of the types of lung tumors is represented in table 4, more specifically for adenocarcinoma, squamous cell carcinoma, and neuroendocrine lung tumors. Table 5 provides a comparison between the malignant and the non-malignant lesions. Figures 26-33 show an overview of this patient cohort.

Table 3: Specifications of the patient cohort (N=39) used to correlate metabolomics and radiomics.

Variable		
Total patients		39
Sex (N, (%))	*Men*	27 (69.2)
	Women	12 (30.8)
Diabetes (N, (%))	*Yes*	4 (10.2)
	No	35 (89.8)
Smoking status (N, (%))	*Current smoker*	20 (51.3)
	Ex-smoker	18 (46.2)
	Non-smoker	1 (2.5)
Packyears (y)	*Median*	40
	Average	44 ± 31
	Range	0 - 139
	Unknown	3
Age (y)	*Median*	71
	Average	71 ± 8
	Range	49 - 84
BMI (kg/m²)	*Median*	26.04
	Average	26.64 ± 5.65
	Range	14.69 - 51.00
Diameter (mm)	*Median*	23
	Average	30.38 ± 22.45
	Range	11 - 120
Plasma glucose (mg%)	*Median*	97.5
	Average	98.23 ± 16.48
	Range	76 - 164
FEV1 absolute value (l)	*Median*	2.27
	Average	2.28 ± 0.71
	Range	1.09 - 4.25

Table 4: Specifications of the different types of malignant lung lesions (N=35) used to correlate metabolomics and radiomics.

		Adenocarcinoma	Squamous cell carcinoma	Neuroendocrine lung tumor
Total patients		24	8	3
Sex (N, (%))	*Men*	15 (62.5)	5 (62.5)	3 (100.0)
	Women	9 (37.5)	3 (37.5)	0 (0.0)
Diabetes (N, (%))	*Yes*	4 (16.7)	0 (0.0)	0 (0.0)
	No	20 (83.3)	8 (100.0)	3 (100.0)
Smoking status (N, (%))	*Current smoker*	11 (45.8)	4 (50.0)	2 (66.6)
	Ex-smoker	12 (50.0)	4 (50.0)	1 (33.3)
	Non-smoker	1 (4.2)	0	0 (0.0)
Packyears (y)	*Median*	35	40	38
	Average	40 ± 30	40 ± 13	45 ± 19
	Range	0 - 75	15 - 61	26 - 70
	Unknown	1	1	0
Age (y)	*Median*	71	73	75
	Average	71 ± 7	72 ± 9	75 ± 3
	Range	49 - 78	55 - 83	71 - 79
BMI (kg/m²)	*Median*	25.43	27.35	27.29
	Average	26.91 ± 6.33	28.10 ± 2.91	25.81 ± 3.67
	Range	19.00 - 51.00	25.00 - 34.89	20.76 - 29.38
Lobe	*Right upper*	6	3	1
	Right middle	6	3	0
	Right lower	7	2	1
	Left upper	2	0	1
	Left under	3	0	0
Diameter (mm)	*Median*	19	32	32
	Average	32.75 ± 26.54	29.29 ± 7.46	35.67 ± 14.52
	Range	11 - 55	13 - 35	20 - 55
Resection margin	*Positive*	2	1	0
	Negative	22	7	3
Visceral pleural invasion	*Yes*	5	3	0
	No	19	5	3
Lymph vascular invasion	*Yes*	7	4	2
	No	17	4	1
Positive nodes	*Yes*	7	0	0
	No	17	8	3
Plasma glucose (mg%)	*Median*	97	103	85
	Average	99.16 ± 18.44	101.88 ± 11.53	86.33 ± 6.60
	Range	76 - 164	83 - 118	79 - 95
FEV1 absolute value (l)	*Median*	2.27	2.02	2.35
	Average	2.37 ± 0.78	1.93 ± 0.51	2.28 ± 0.43
	Range	1.09 - 4.25	1.16 - 2.60	1.72 - 2.77

		Adenocarcinoma	Squamous cell carcinoma	Neuroendocrine lung tumor
T	*1b*	9	1	1
	1c	2	1	0
	2a	7	2	1
	2b	0	1	0
	3	4	0	1
	4	2	2	0
N	*0*	19	6	3
	1	5	1	0
	2	0	1	0
M	*0*	23	8	3
	1a	1	0	0
R	*0*	23	7	3
	1	1	1	0

Table 5: Comparison between malignant and non-malignant lung lesions (N=39) used to correlate metabolomics and radiomics.

		Malignant	Non-malignant
Total patients		35	4
Sex (N, (%))	*Men*	23 (65.7)	4 (100.0)
	Women	12 (34.3)	0 (0.0)
Diabetes (N, (%))	*Yes*	4 (11.4)	0 (0.0)
	No	31 (88.6)	4 (100.0)
Smoking status (N, (%))	*Current smoker*	17 (48.6)	3 (75.0)
	Ex-smoker	17 (48.6)	1 (25.0)
	Non-smoker	1 (2.8)	0
Packyears (y)	*Median*	38	80
	Average	40 ± 26	91 ± 35
	Range	0 - 130	55 - 139
	Unknown	2	1
Age (y)	*Median*	72	65
	Average	71 ± 8	62 ± 10
	Range	49 - 84	52 - 80
BMI (kg/m²)	*Median*	26.57	24.44
	Average	27.09 ± 5.57	22.81 ± 4.88
	Range	19.00 - 51.00	14.69 - 27.90
Lobe	*Right upper*	10	2
	Right middle	9	0
	Right lower	10	2
	Left upper	3	0
	Left under	3	0
Diameter (mm)	*Median*	25	
	Average	31.61 ± 23.03	
	Range	11 - 120	
Plasma glucose (mg%)	*Median*	97	90
	Average	98.69 ± 16.82	96.5 ± 12.42
	Range	76 - 164	76 - 108
FEV1 absolute value (l)	*Median*	2.25	2.62
	Average	2.26 ± 0.73	2.55 ± 0.29
	Range	1.09 - 4.25	2.17 - 2.86

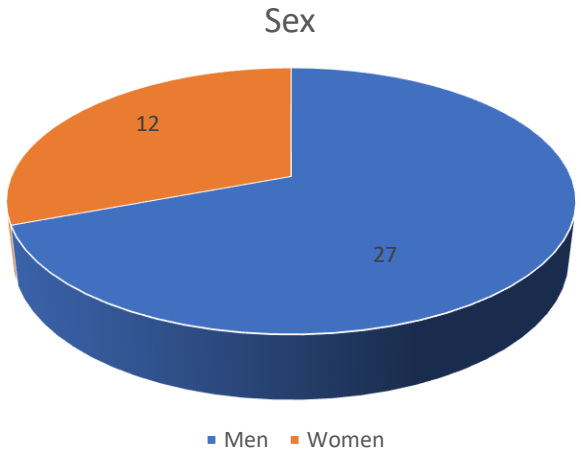

Figure 26: Pie chart of sex of the patients' radiomics features correlated with metabolomics.

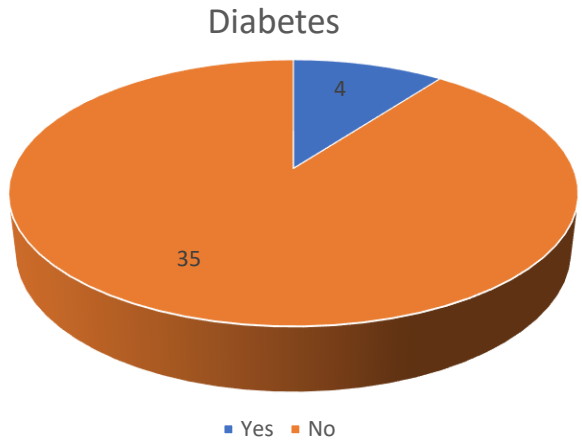

Figure 27: Pie chart of diabetes of the patients' radiomics features correlated with metabolomics.

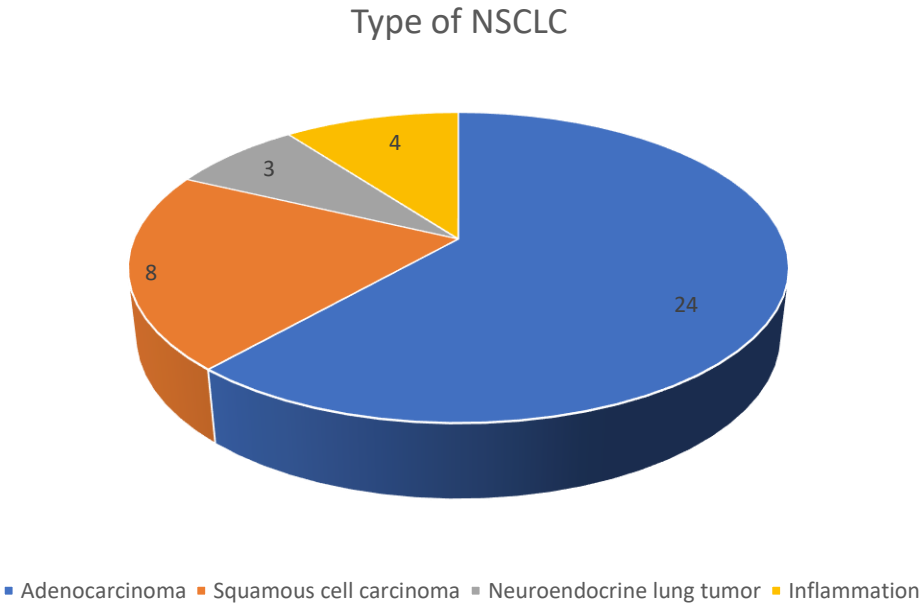

Figure 28: Pie chart of different types of NSCLC of the patients' radiomics features correlated with metabolomics.

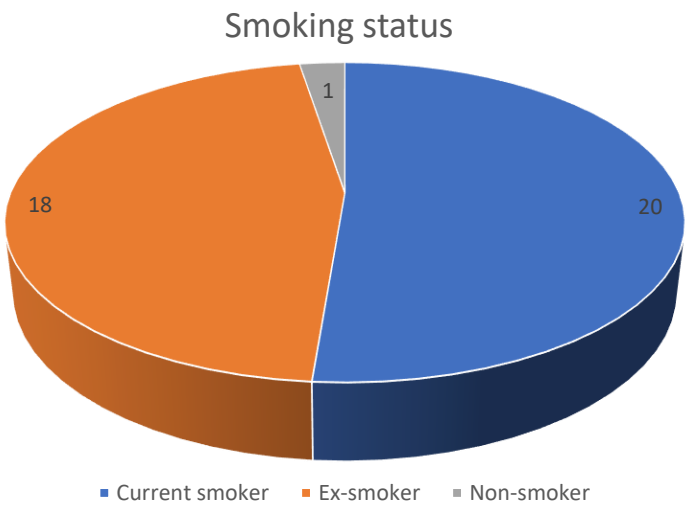

Figure 29: Pie chart of the smoking status of the patients' radiomics features correlated with metabolomics.

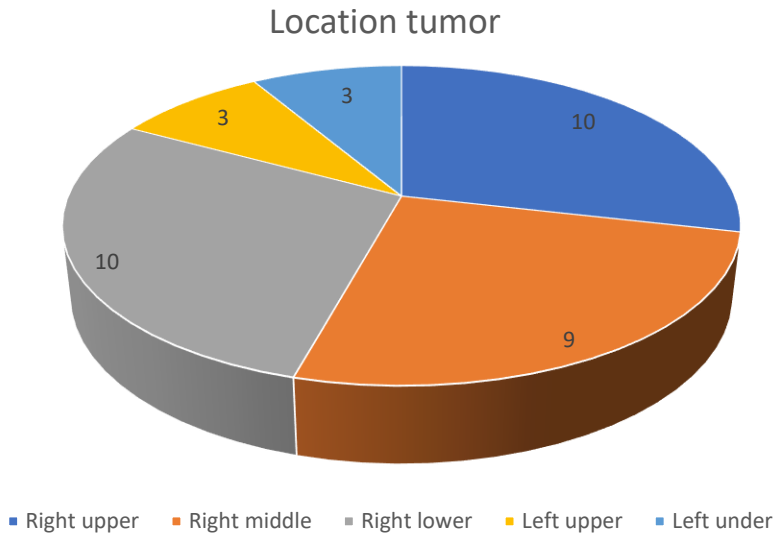

Figure 30: Pie chart of the location of the tumor of the patients' radiomics features correlated with metabolomics.

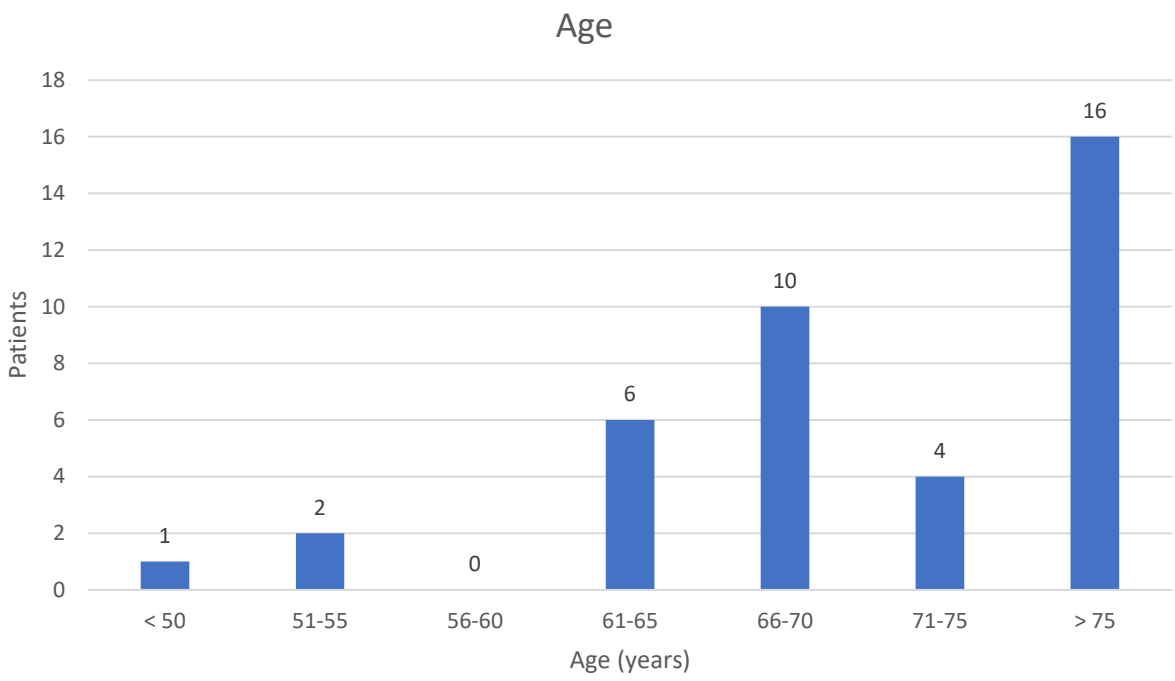

Figure 31: Bar chart of the age of the patients' radiomics features correlated with metabolomics.

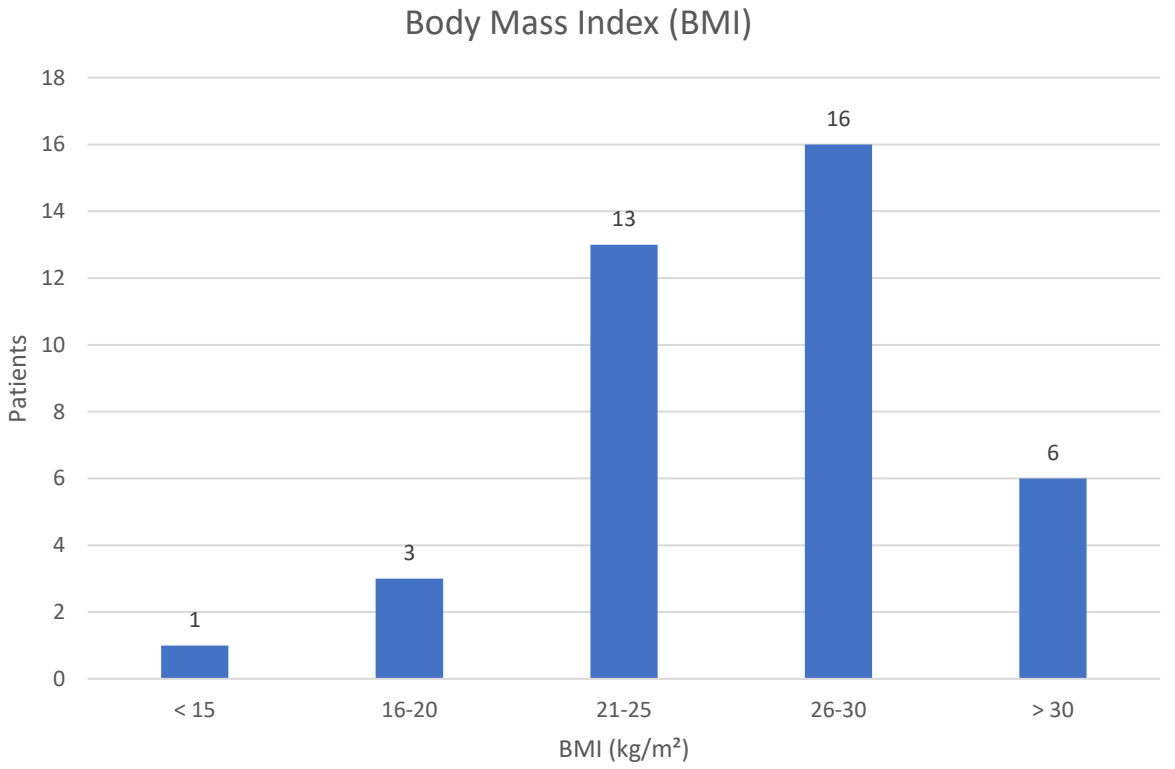

Figure 32: Bar chart of the BMI of the patients' radiomics features correlated with metabolomics.

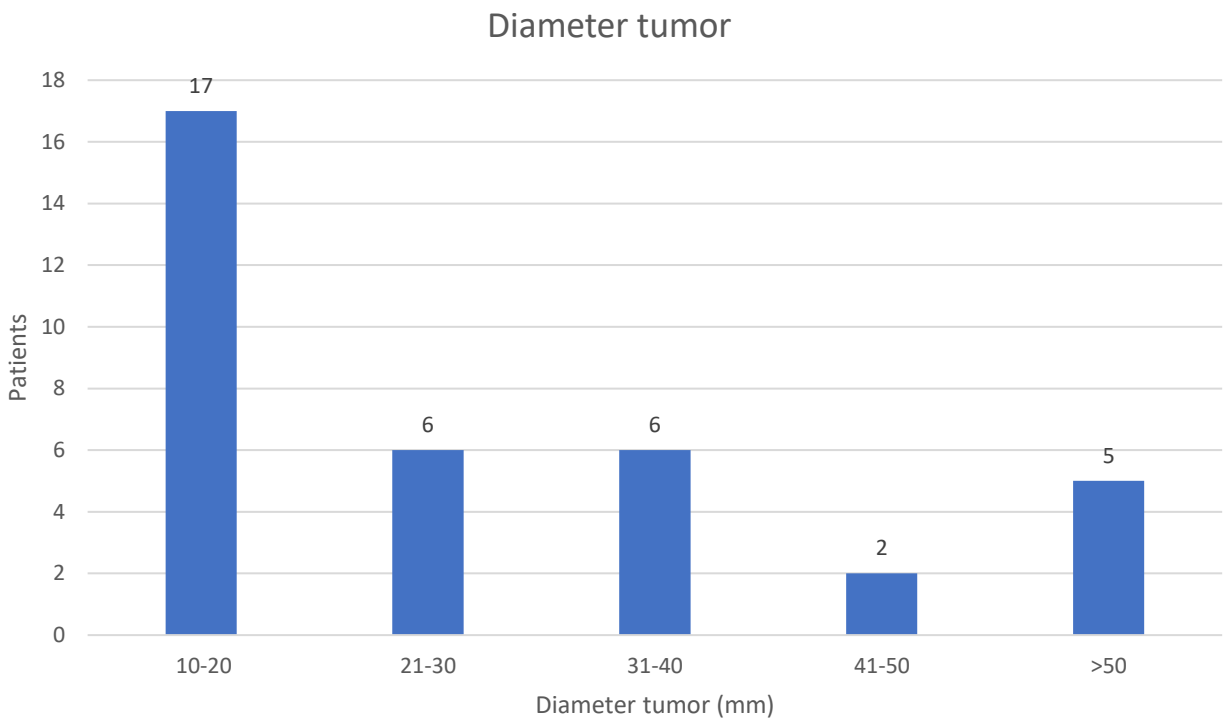

Figure 33: Bar chart of the diameter of the tumor of the patients' radiomics features correlated with metabolomics.

16.1.2 Radiomics features

Out of the radiomics tool, image features were extracted based on the PET/CT images of the included patients. In total, 483 image features per patient were obtained and divided into eleven subgroups.

These subgroups of features are morphological, local intensity, intensity-based statistical, intensity histogram, intensity- volume histogram, gray-level co-occurrence based, grey level run length based, grey level size zone-based, grey level distance zone-based, neighborhood grey tone difference based, and neighborhood grey level dependence based features (122).
In these eleven subgroups, there are multiple features, including volume, mean, and variance. In this section, the radiomics features that showed correlation with the metabolites are explained further. The 13 radiomics features that are important for this study are described in the following pages. Parameters that are highly correlated to each other and divided into the same subgroup are not described. The parameter with the highest correlation to metabolomics was chosen. This was done for the features *quarter coefficient* and *zone distance non-uniformity,* respectively related to the features *small zone grey level* and *dependence count* described in the following pages.

16.1.3 Radiomics features: Moran's I index

A positive correlation is noticed by the *Moran's I* index radiomics feature, classified into the subgroup morphological features. This feature indicates spatial autocorrelation. Formula 12 describes the *Moran's I* index feature.

$$F_{morph.moran.i} = \frac{N_{v,gl}}{\sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2}} \frac{\sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2} (X_{gl,k_1} - \mu)\,(X_{gl,k_2} - \mu)}{\sum_{k=1}^{N_{v,gl}} (X_{gl,k_1} - \mu)^2} \; and \; k_1 \neq k_2 \qquad (12)$$

To understand formula 12, specific parameters need to be known. First, the number of voxels in the intensity mask is represented by $N_{v,gl}$ and the corresponding intensity's is denoted by X_{gl}. Furthermore, the mean of X_{gl} is represented by μ and a weight factor is represented by w_{k_1,k_2}. This weight factor is equal to the inverse Euclidean distance between the corresponding voxels k_1 and k_2 of the region of interest intensity mask (123).

The *Moran's I* index can range between 1.0 and -1.0. A value close to zero means no spatial autocorrelation, a value close to 1.0 indicates a high spatial autocorrelation, and a value close to -1.0 indicates a high spatial anti-autocorrelation (122).

16.1.4 Radiomics features: Grey level non-uniformity

The radiomics feature *grey level non-uniformity* is divided in the subgroup grey level run length based features and measures the distribution of runs over the levels of grey values. A low *grey level non-uniformity* indicates that there is an equal distribution of runs along with grey levels. Formula 13 describes the *grey level non-uniformity* feature. N_s represents the total number of zones in the region of interest and N_g is a fixed number of the discretized intensities into bins. The marginal sum of runs for a certain grey level *i* over run lengths *j* is denoted by r_i (122).

$$F_{rlm.glnu} = \frac{1}{N_s} \sum_{i=1}^{N_g} r_i^2 \tag{13}$$

16.1.5 Radiomics features: Inverse difference

The radiomic feature named *inverse difference* is dived into the subgroup grey level co-occurrence-based features and measures the homogeneity in the region of interest (124).
A significant difference in grey levels co-occurrences results in a lower weight. This leads to a lower total feature value. The *inverse difference* feature is maximal when all the grey levels in the region of interest have all the same value and are described by formula 14.

p_{ij} refers to a probability, more specifically, it corresponds to the probability of common grey level co-occurrence present in the grey level co-occurrence matrix, with grey level *i,* that is discretized, and size *j* (122).

$$F_{cm.inv.diff} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1 + \| i - j \|} \tag{14}$$

16.1.6 Radiomics features: Area density AABB

The *area density axis-aligned bounding box* feature is divided into the subgroup of morphological radiomics features and is described by formula 15. This feature is a ratio of the surface area A of the considered region of interest and the surface area of the axis-aligned bounding box A_{aabb} that encloses the region of interest mesh (122).

The axis-aligned bounding box (AABB) is the smallest box enclosing the vertex set and must be in line with the reference frame axes. A vertex set is a group of vertices, points where multiple curves, edges, or lines come together (125).

$$F_{morph.a.dens.aabb} = \frac{A}{A_{aabb}} \tag{15}$$

16.1.7 Radiomics features: The first measure of information correlation

The radiomics feature named *the first measure of information correlation,* is classified in the subgroup grey level co-occurrence-based features. This feature uses two different measures and is described by formula 16 (126).

To understand this radiomics feature, the following variables need to be known: HXY, HXY_1 and HX. The entropy for the joint probability is represented by HXY, in formula 17. HXY_1 is a specific type of entropy and is represented by formula 18. The entropy of the row marginal probability is known as HX and described by formula 19. As a result of symmetry, the entropy of the row marginal probability is equal to the column's marginal probability entropy (122).

$$F_{cm.info.corr.1} = \frac{HXY - HXY_1}{HX} \tag{16}$$

$$HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij}\log_2(p_{ij}) \tag{17}$$

$$HXY_1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij}\log_2(p_{i.}p_{j.}) \tag{18}$$

$$HX = -\sum_{i=1}^{N_g} p_i\log_2(p_i) \tag{19}$$

16.1.8 Radiomics features: coefficient of variation

The correlation test between the radiomics and metabolomics datasets leads to a high negative correlation value for the radiomics feature named *coefficient of variation*. This feature is divided into the intensity-based statistical features subgroup.

This feature is linked to the dispersion of the corresponding voxel intensities, denoted X_{gl}, where the intensities are again discretized to a fixed number of N_g bins, based on the fixed bin number method. Formula number 20 describes the *coefficient of variance* feature $F_{stat.cov}$.

$$F_{stat.cov} = \frac{\sigma}{\mu} \tag{20}$$

The standard deviation σ is known as the square root out of $F_{stat.var}$ and μ is equal to the mean of the intensity distribution and is represented by $F_{stat.mean}$.

The feature $F_{stat.var}$ is known as the intensity variance of X_{gl} and is calculated with formula 21. The mean intensity of the corresponding voxels is computed with formula 22. N_v denotes the set of intensities of voxels included in the concerned region of interest (122).

$$F_{stat.var} = \frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2 \tag{21}$$

$$F_{stat.mean} = \frac{1}{N_v} \sum_{k=1}^{N_v} X_{gl,k} \tag{22}$$

16.1.9 Radiomics features: Small zone low grey level emphasis

In the subgroup grey level size zone-based features, the radiomics feature *small zone emphasis* is present. The *small zone emphasis* feature is described by formula 23, where the maximum size of a zone of any group linked voxels is represented by N_z. The number of zones with a size *j* and independent of a certain grey level is denoted by $s_{.j}$ (122).

$$F_{szm.sze} = \frac{1}{N_s} \sum_{j=1}^{N_z} \frac{s_{.j}}{j^2} \tag{23}$$

16.1.10 Radiomics features: Minimum histogram gradient

The *minimum histogram gradient* is classified in the subgroup intensity histogram features. To understand this feature, it is needed to know what an intensity histogram is. This histogram is generated based on the discretizing of the original intensity distribution. That way X_{gl} is divided into intensity bins.

To calculate the histogram gradient of the intensity histogram, formula 24 needs to be executed.

$$H_i' = \begin{cases} n_2 - n_1 & i = 1 \\ (n_{i+1} - n_{i-1})/2 & 1 < i < N_g \\ n_{N_g} - n_{N_{g-1}} & i = N_g \end{cases} \quad (24)$$

The *minimum histogram gradient* is afterward described by formula 25 (122).

$$F_{ih.min.grad} = min(\boldsymbol{H}') \quad (25)$$

16.1.11 Radiomics features: Coarseness

The radiomic feature named *coarseness* is divided into the subgroup neighbourhood grey tone difference-based features and calculated by formula 26. Due to large-scale patterns, the different grey levels in coarse textures are, most of the time, relatively small. To indicate the spatial rate of change in intensity, often, a sum of the differences in grey levels is obtained. This is done with the radiomic feature *coarseness (127).*

Note that the sum in the denominator can lead to zero as a result. That way, the maximum value of the *coarseness* feature is set to 10^6, an arbitrary number (122).

$$F_{ngt.coarseness} = \frac{1}{\sum_{i=1}^{N_g} p_i s_i} \quad (26)$$

16.1.12 Radiomics features: Small distance emphasis

The *small distance emphasis* is divided in the subgroup grey level distance zone based on all the radiomics features described by formula 27. The number of zones with grey level *i* that is discretized, independent of a certain distance *j*, is denoted by $d_{.j}$ and N_d stands for the largest distance of any zone (122).

$$F_{dzm.sde} = \frac{1}{N_s} \sum_{j=1}^{N_d} \frac{d_{.j}}{j^2} \tag{27}$$

16.1.13 Radiomics features: Dependence count energy

The *dependence count energy* is divided into the subgroup neighbouring grey level dependency-based features and is named the second moment by Sun and Wee (128).

This feature is described by formula 28, where $p_{ij} = s_{ij}/N_s$. The number of zones with a discretization of grey level i and a size j is represented by s_{ij} and N_s represents the total number of zones in the region of interest (122).

$$F_{ngl.dc.energy} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_n} p_{ij}^2 \tag{28}$$

16.1.14 Radiomics features: Surface to volume ratio

The *surface-to-volume ratio* radiomics feature is divided into the subgroup morphological features. This feature is relatively simple and is the ratio of the surface *A* and the volume *V* of the region of interest. Formula 29 describes the *surface-to-volume feature* (122).

$$F_{morph.av} = \frac{A}{V} \tag{29}$$

16.1.15 Radiomics features: Joint maximum

The radiomics feature *joint maximum* is linked to a probability, more specifically, it corresponds to the most common grey level co-occurrence present in the grey level co-occurrence matrix (GLCM). This matrix shows how discretized intensities, also named grey levels, are combined with neighbouring pixels or voxels within a 3D volume, and in that way, distributed along with one of the directions of the image. Formula 30 represents the *joint maximum* feature divided in the subgroup grey level co-occurrence-based features (122).

$$F_{cm.joint.max} = max(p_{ij}) \tag{30}$$

16.1.16 Radiomics features: Angular second moment

The energy of P_{Δ} is represented by the *angular second moment*. Like the joint maximum feature, this feature is divided into the subgroup grey level co-occurrence-based features. The *angular second momentum* is represented by formula 31. Synonyms of this feature are named energy or uniformity (122).

$$F_{cm.energy} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} p_{ij}^2 \tag{31}$$

16.1.17 Results

The Spearman correlation coefficient test was performed on three datasets. Each dataset consisted of 721 parameters (238 metabolomics parameters and 483 radiomics features), and each was based on a different segmentation method. Segmentation methods one and two were based on the PET/CT images, and segmentation method three was based on only the PET images of the patients.

The 238 metabolomics parameters (obtained by ^{1}H-NMR spectroscopy analysis of blood plasma samples) in each dataset have a named variable and a corresponding number to make it easier to work with during the data reduction (e.g., Var001). One variable of the metabolomics dataset represents a metabolite or a combination of multiple metabolites.

The significant correlation values found between the radiomics and metabolomics parameters in each correlation matrix showed R^2 values between 0.3 and 0.7 (positive correlation) or between -0.3 and -0.7 (negative correlation). The positive correlations found in the metabolomics dataset were mainly related to the concentration of plasma glucose. The radiomics features positively correlated to glucose were identified as Morans I, inverse difference normalized, grey level non-uniformity GLSZM, grey level non-uniformity GLDZM, area density AABB, and first measure of information correlation.

The negative correlations found in the metabolomics dataset were mainly related to glycerol. The radiomics features that were negatively correlated to these metabolites were identified as a coefficient of variation, quartile coefficient, small zone low grey level emphasis, low dependence low grey level emphasis, the surface to volume ratio, minimum histogram gradient, coarseness, zone distance non-uniformity normalized GLDZM, small distance emphasis GLDZM, dependence count energy, joint maximum, angular second moment and grey level non-uniformity normalized.

The correlation output matrices suggested more or less the same correlations between the metabolomics variables and the radiomics features for each segmentation method used. Only the strength of correlation differs between the segmentation methods.

Table 6 shows some of the metabolomics variables that are linked to the concentration of plasma glucose. Table 7 shows some of the metabolomics variables related to glycerol. These metabolomics variables were chosen because there is more certainty that other metabolites do not influence plasma glucose and glycerol in these specific variables. The complete assignment of all 238 metabolomics variables is out of the scope of this thesis.

Table 6: Metabolomic variables (N=20) related to plasma glucose

METABOLOMIC VARIABLES				
Var026	Var029	Var057	Var064	Var065
Var070	Var071	Var074	Var076	Var094
Var095	Var097	Var098	Var099	Var100
Var101	Var104	Var107	Var108	Var122

Table 7: Metabolomic variables (N=12) related to glycerol

METABOLOMIC VARIABLES					
Var068	Var069	Var070	Var071	Var080	Var081
Var082	Var083	Var084	Var091	Var092	Var093

First segmentation method (PET/CT)

By studying the correlation coefficient trend of these metabolomic variables, the correlation matrix of the radiomics and metabolomics features created using the first segmentation method shows a correlation between 10 of the 20 studied metabolomic variables related to plasma glucose and 12 of the 45 radiomics features.

When looking at the metabolomics variables related to plasma glucose concentration, the correlations with four radiomics features are positive, and the R^2-values lie between 0.3 and 0.36.
For glycerol, all 12 metabolomic variables have a negative correlation with eight radiomics features between -0.5 and -0.3.

Table 8 shows the variables related to plasma glucose correlated to the described radiomics features using the first segmentation method in green and those not correlated in grey.

All the metabolomic variables related to glycerol (N=12), displayed in table 7, are correlated to radiomics features.

Table 8: The metabolomic variables related to plasma glucose. The variables colored in green (N=10) are positively correlated to radiomics features after using the first segmentation method.

METABOLOMIC VARIABLES				
Var026	Var029	Var057	Var064	Var065
Var070	Var071	Var074	Var076	Var094
Var095	Var097	Var098	Var099	Var100
Var101	Var104	Var107	Var108	Var122

The correlation output matrix shows that metabolomic variable VAR071 (related to plasma glucose) has the highest correlation with the radiomics features.

For the variables related to glycerol, the correlation output matrix shows that metabolomic variable VAR093 has the highest correlation with the radiomics features.

A heatmap is shown in figure 25. This heatmap shows a summary of the correlations between the 10 metabolomics variables and 12 radiomics features.

term	Var029	Var057	Var065	Var068	Var069	Var070	Var071	Var080	Var081	Var082	Var083	Var084	Var091	Var092	Var093	Var098	Var099	Var100	Var108	Var122
Morans.I	0.323121	0.195536	0.221578	0.00186	0.00186	0.178466	0.095962	-0.06664	-0.11259	-0.04191	-0.115	-0.17628	-0.16752	-0.29883	-0.40716	0.307145	0.200131	0.33647	0.303425	0.301893
inverse.di	0.2137	0.354853	0.284167	0.314805	0.307802	0.303206	0.360105	0.285042	0.269942	0.271474	0.311303	0.28351	0.316555	0.251559	0.359886	0.239085	0.325747	0.224642	0.222891	0.288981
Grey.level	0.165344	0.293484	0.301362	0.282854	0.236897	0.302128	0.368332	0.35157	0.369296	0.335814	0.349382	0.297954	0.368859	0.326403	0.256593	0.182524	0.205942	0.186464	0.193467	0.235378
Grey.level	0.165344	0.293484	0.301362	0.282854	0.236897	0.302128	0.368332	0.35157	0.369296	0.335814	0.349382	0.297954	0.368859	0.326403	0.256593	0.182524	0.205942	0.186464	0.193467	0.235378
Coefficier	-0.14061	-0.2811	-0.21282	-0.42423	-0.42488	-0.22924	-0.24259	-0.35923	-0.36186	-0.32903	-0.38396	-0.42641	-0.37696	-0.44764	-0.53649	-0.1638	-0.24828	-0.13229	-0.14936	-0.06642
Quartile.c	-0.17606	-0.31656	-0.23099	-0.46143	-0.47806	-0.25681	-0.25594	-0.40059	-0.39052	-0.3612	-0.40759	-0.45092	-0.41044	-0.45508	-0.54897	-0.19269	-0.27607	-0.16008	-0.1765	-0.09224
Coefficier	-0.15877	-0.29686	-0.22311	-0.43276	-0.43473	-0.24084	-0.24937	-0.35945	-0.35529	-0.32728	-0.38155	-0.4286	-0.37871	-0.4483	-0.54196	-0.18	-0.26403	-0.14739	-0.16271	-0.07867
Quartile.c	-0.18476	-0.30939	-0.20874	-0.44836	-0.4638	-0.25638	-0.23612	-0.36261	-0.31366	-0.30654	-0.34421	-0.37674	-0.32844	-0.33293	-0.43905	-0.21345	-0.30544	-0.17227	-0.19439	-0.11313
Small.zon	-0.1498	-0.2555	-0.31656	-0.36689	-0.30102	-0.32509	-0.3148	-0.31918	-0.37148	-0.36689	-0.4216	-0.4216	-0.50213	-0.43889	0.106467	-0.18394	-0.16162	-0.19991	-0.21523	-0.17672
Small.zon	-0.0964	-0.20779	-0.27038	-0.32947	-0.25659	-0.27454	-0.2671	-0.324	-0.3833	-0.3612	-0.41985	-0.41306	-0.48966	-0.41635	0.134916	-0.13907	-0.11084	-0.15571	-0.16971	-0.13601
Small.zon	-0.18241	-0.2555	-0.3078	-0.34544	-0.30846	-0.31984	-0.27848	-0.23055	-0.25637	-0.27651	-0.31437	-0.31634	-0.3844	-0.30321	0.216544	-0.22289	-0.18328	-0.24281	-0.259	-0.21042
Low.depe	-0.05723	-0.19444	-0.22245	-0.35398	-0.27016	-0.22595	-0.21764	-0.33669	-0.35945	-0.36032	-0.40606	-0.40015	-0.42948	-0.37761	-0.00208	-0.10669	-0.10121	-0.11128	-0.11588	-0.06773

Figure 34: A heatmap of the correlations between 10 metabolomics variables and 12 radiomics features obtained using the first segmentation method.

term	Var057	Var065	Var068	Var069	Var070	Var071	Var080	Var081	Var082	Var083	Var084	Var091	Var092	Var093	Var094	Var099	Var104	Var107
area.densi	0.323777	0.373454	0.31371	0.276288	0.303206	0.313054	0.258343	0.287668	0.269285	0.325966	0.290294	0.32334	0.311303	0.110844	0.288544	0.210855	0.300799	0.276288
inverse.dif	0.418317	0.326184	0.439326	0.368859	0.342817	0.349382	0.413065	0.439107	0.407375	0.453988	0.425539	0.488565	0.409345	0.246964	0.333844	0.330124	0.284604	0.30058
Grey.level.	0.348288	0.356166	0.355509	0.262501	0.343692	0.393588	0.379363	0.445016	0.399059	0.437575	0.397089	0.495787	0.435387	0.018054	0.30255	0.231863	0.253529	0.255498
Grey.level.	0.348288	0.356166	0.355509	0.262501	0.343692	0.393588	0.379363	0.445016	0.399059	0.437575	0.397089	0.495787	0.435387	0.018054	0.30255	0.231863	0.253529	0.255498
Surface.to	-0.14936	-0.15505	-0.24609	-0.14498	-0.10887	-0.17124	-0.3299	-0.34741	-0.30342	-0.34435	-0.28285	-0.37871	-0.30627	-0.01499	-0.09837	-0.03775	-0.05679	-0.07123
Quartile.cc	-0.33559	-0.28067	-0.33144	-0.37783	-0.2776	-0.31218	-0.28592	-0.29861	-0.25637	-0.30496	-0.28767	-0.30561	-0.3426	-0.40256	-0.25769	-0.2776	-0.25659	-0.25309
Minimum.I	-0.19715	-0.0908	-0.30668	-0.32673	-0.10756	-0.16977	-0.31709	-0.3161	-0.28751	-0.29441	-0.25279	-0.26616	-0.30832	-0.4644	-0.08784	-0.12519	-0.06846	-0.07284
coarsenes	-0.23383	-0.24062	-0.32334	-0.259	-0.23055	-0.29095	-0.35726	-0.39884	-0.37411	-0.38265	-0.33144	-0.39052	-0.39009	-0.17584	-0.18416	-0.11828	-0.13973	-0.14608
coarsenes	-0.23164	-0.22792	-0.29817	-0.23143	-0.21326	-0.26863	-0.34325	-0.39206	-0.35223	-0.37061	-0.32246	-0.39031	-0.3844	-0.17146	-0.17409	-0.10887	-0.1336	-0.14104
small.dista	-0.13557	-0.18897	-0.2555	-0.13229	-0.14345	-0.18547	-0.34172	-0.38046	-0.32925	-0.36448	-0.32815	-0.42641	-0.36558	0.088303	-0.10472	-0.0163	-0.06883	-0.08962
Zone.dista	-0.1266	-0.17759	-0.26403	-0.15855	-0.14433	-0.1916	-0.33428	-0.37324	-0.32268	-0.36295	-0.32028	-0.4065	-0.3612	0.056352	-0.09465	-0.01477	-0.05788	-0.08108
small.dista	-0.16468	-0.19554	-0.27891	-0.14783	-0.15549	-0.18416	-0.37608	-0.41241	-0.37302	-0.41569	-0.3879	-0.475	-0.37761	0.10581	-0.13317	-0.036	-0.09552	-0.10712
Zone.dista	-0.15724	-0.18678	-0.27038	-0.13732	-0.1487	-0.17912	-0.37389	-0.41438	-0.37389	-0.41635	-0.38396	-0.4739	-0.37499	0.115658	-0.12638	-0.0279	-0.08721	-0.10012
small.dista	-0.08776	-0.15123	-0.23789	-0.13066	-0.0929	-0.10702	-0.31187	-0.32609	-0.32576	-0.32259	-0.30355	-0.41035	-0.29753	0.197188	-0.09542	0.022761	-0.05296	-0.06095
Zone.dista	-0.08995	-0.14576	-0.24599	-0.13109	-0.09356	-0.10636	-0.31734	-0.33594	-0.33715	-0.3331	-0.31449	-0.42151	-0.30848	0.193686	-0.08864	0.025387	-0.04662	-0.05198
dependenc	-0.22289	-0.2555	-0.28898	-0.1638	-0.23602	-0.24084	-0.33516	-0.37521	-0.36207	-0.37871	-0.33231	-0.43714	-0.32378	0.220046	-0.18941	-0.10143	-0.15483	-0.16577

Figure 35: A heatmap of the correlations between 8 metabolomics variables and 16 radiomics features obtained using the second segmentation method.

Second segmentation method (PET/CT)

The correlation output matrix created with the second segmentation method describes a correlation between eight of the 20 studied metabolomics variables related to plasma glucose and 16 of the 45 studied radiomics features.

For the eight metabolomics variables related to plasma glucose concentration, the correlations with four radiomics features are positive, and the R^2-values lie between 0.3 and 0.41. Fewer variables are correlated with radiomics features using the second segmentation method than with the first segmentation method, but those correlated show a higher R^2-value using the second segmentation method.

All 12 metabolomic variables related to glycerol are correlated to 12 radiomics features, and all the variables have several negative correlations to the radiomics features. With these negative correlations, the R^2-values lay between -0.57 and -0.3. These negatively correlated variables with radiomics features show an increase in R^2-value when using the second segmentation method compared to using the first segmentation method.

Table 9 shows the variables related to plasma glucose correlated to the described radiomics feature after the second segmentation method in green and those not correlated in grey.

All the metabolomic variables related to glycerol (N=12), displayed in table 7, are correlated to radiomics features.

Table 9: The metabolomic variables related to plasma glucose. The variables colored green (N=8) are positively correlated to radiomics features after using the second segmentation method.

METABOLOMIC VARIABLES				
Var026	Var029	Var057	Var064	Var065
Var070	Var071	Var074	Var076	Var094
Var095	Var097	Var098	Var099	Var100
Var101	Var104	Var107	Var108	Var122

For the variables related to plasma glucose, the correlation output matrix shows that metabolomic variable VAR057 has the highest correlation with the radiomics features.

For the variables related to glycerol, the correlation output matrix shows that metabolomic variable VAR093 has the highest correlation with the radiomics features.

A heatmap is shown in figure 26. This heatmap shows a summary of the correlations between the 8 metabolomics variables and 16 radiomics features.

Third segmentation method (PET)

The correlation output matrix created with the third segmentation method describes a correlation between all 20 of the metabolomics variables related to plasma glucose and 17 of the 45 studied radiomics features.

For the 20 metabolomics variables related to plasma glucose concentration, the correlations with two radiomics features are positive, and the R^2-values lie between 0.3 and 0.5. All the metabolomic variables related to glucose plasma correlated to radiomics features. All the correlated variables show a higher R^2-value using the third segmentation method than the two previous methods.

All 12 metabolomic variables related to glycerol are correlated to 15 radiomics features, and all the variables have several negative correlations to the radiomics features. With these negative correlations, the R^2-values lay between -0.53 and -0.3. The R^2-values obtained using the third segmentation method are very close to the R^2-values obtained using the second segmentation method, but the overall trend shows lower values using the third segmentation method.

Table 10 shows the variables related to plasma glucose correlated to the described radiomics feature after the third segmentation method in green and those not correlated in grey.
All the metabolomic variables related to glycerol (N=12), displayed in table 7, are correlated to radiomics features.

Table 10: The metabolomic variables related to plasma glucose. The variables colored green (N=20) are correlated positively to radiomics features after using the third segmentation method.

METABOLOMIC VARIABLES				
Var026	Var029	Var057	Var064	Var065
Var070	Var071	Var074	Var076	Var094
Var095	Var097	Var098	Var099	Var100
Var101	Var104	Var107	Var108	Var122

For the variables related to plasma glucose, the correlation output matrix shows that metabolomic variable VAR026 has the highest correlation with the radiomics features. For the variables related to glycerol, the correlation output matrix shows that metabolomic variable VAR091 has the highest correlation with the radiomics features.

A heatmap is shown in figures 27 and 28. This heatmap shows a summary of the correlations between the 20 metabolomics variables and 17 radiomics features.

Although out of the scope of this thesis, it is noteworthy that these variables are linked to a single metabolite (VAR026 for glycerol) and only two metabolites (glucose and a small link with another metabolite). This contrasts with the results from the previous segmentation methods, where the variables with the highest correlations represent a combination of multiple metabolites.

term	Var026	Var029	Var057	Var064	Var065	Var068	Var069	Var070	Var071	Var074	Var076	Var080	Var081	Var082	Var083
area.densi	0.307053	0.379056	0.434316	0.388357	0.478744	0.372162	0.331236	0.434863	0.463424	0.394047	0.38398	0.340428	0.415167	0.350933	0.409586
first.meas	0.241711	0.34216	0.441952	0.367327	0.343473	0.301455	0.351789	0.354415	0.432323	0.30999	0.297516	0.260532	0.283948	0.196849	0.245213
Surface.to	-0.06609	-0.03108	-0.15944	-0.09947	-0.16206	-0.27827	-0.15944	-0.12781	-0.18876	-0.09641	-0.10954	-0.38157	-0.42283	-0.3726	-0.4039
joint.maxi	-0.07867	-0.0732	-0.16687	-0.11807	-0.18941	-0.25725	-0.1336	-0.17737	-0.15264	-0.11435	-0.12047	-0.30736	-0.38243	-0.36711	-0.42357
angular.se	-0.13535	-0.12966	-0.22552	-0.16096	-0.23821	-0.28898	-0.18262	-0.24565	-0.22377	-0.16862	-0.1719	-0.30299	-0.38702	-0.36776	-0.43451
joint.maxi	-0.03162	-0.01915	-0.10034	-0.08349	-0.15505	-0.21545	-0.09071	-0.11588	-0.11588	-0.06795	-0.07977	-0.30693	-0.37499	-0.34238	-0.38002
coarsenes	-0.18	-0.18416	-0.29183	-0.22267	-0.29664	-0.31524	-0.23252	-0.28592	-0.34107	-0.21654	-0.22574	-0.35376	-0.41306	-0.36952	-0.39227
coarsenes	-0.20013	-0.20495	-0.3345	-0.24609	-0.31437	-0.39227	-0.324	-0.331	-0.35835	-0.24127	-0.25331	-0.36952	-0.41	-0.39009	-0.41503
small.dista	-0.0743	-0.01061	-0.12244	-0.06095	-0.16008	-0.26403	-0.15264	-0.12813	-0.16183	-0.08721	-0.09531	-0.36667	-0.4135	-0.38658	-0.41657
Zone.dista	-0.11128	-0.04519	-0.14039	-0.08633	-0.18656	-0.24346	-0.14017	-0.15789	-0.18569	-0.11435	-0.1196	-0.32662	-0.38133	-0.34391	-0.39052
small.dista	-0.08086	-0.02725	-0.15658	-0.09356	-0.17365	-0.28023	-0.15593	-0.14258	-0.16577	-0.10975	-0.11894	-0.39249	-0.43954	-0.40409	-0.44261
Zone.dista	-0.07649	-0.02112	-0.15199	-0.0883	-0.16468	-0.27016	-0.1487	-0.13907	-0.16096	-0.1045	-0.11281	-0.38046	-0.43101	-0.39096	-0.43254
small.dista	-0.13141	-0.08064	-0.16555	-0.16096	-0.21808	-0.31612	-0.22946	-0.17387	-0.19619	-0.14411	-0.14279	-0.36383	-0.40475	-0.39446	-0.4019
Zone.dista	-0.1347	-0.08633	-0.17693	-0.1649	-0.22223	-0.3124	-0.22552	-0.18087	-0.20298	-0.15045	-0.14958	-0.35595	-0.39818	-0.38877	-0.39665
Grey.level.	-0.09005	0.004705	-0.08414	-0.04869	-0.13951	-0.21764	-0.12004	-0.11062	-0.15899	-0.07189	-0.0859	-0.29336	-0.35179	-0.31809	-0.34829
dependenc	-0.05022	-0.03162	-0.14345	-0.10297	-0.17168	-0.24281	-0.12507	-0.15483	-0.16315	-0.0999	-0.11084	-0.32137	-0.37761	-0.35135	-0.37696
dependenc	-0.11653	-0.09246	-0.19641	-0.14061	-0.23033	-0.28964	-0.16774	-0.21195	-0.2288	-0.14608	-0.16337	-0.36536	-0.42182	-0.39074	-0.4181

Figure 36: A heatmap of the correlations between 20 metabolomics variables and 15 radiomics features obtained using the third segmentation method (first part).

term	Var084	Var091	Var092	Var093	Var094	Var095	Var097	Var098	Var099	Var100	Var101	Var104	Var107	Var108	Var122
area.densi	0.372052	0.427641	0.45281	0.088745	0.395141	0.376867	0.374679	0.360781	0.316354	0.36669	0.374569	0.402582	0.370739	0.35728	0.354325
first.meas	0.222016	0.296641	0.356385	0.432323	0.343035	0.331874	0.328811	0.289638	0.390086	0.257687	0.29117	0.32334	0.307583	0.268629	0.308677
Surface.to	-0.33813	-0.44033	-0.39262	-0.04673	-0.09991	-0.08032	-0.05121	-0.03491	-0.03425	-0.04322	-0.05592	-0.05789	-0.07113	-0.04607	-0.06347
joint.maxi	-0.3984	-0.5041	-0.39359	0.151329	-0.13141	-0.09465	-0.07364	-0.05745	-0.05526	-0.07539	-0.08436	-0.09531	-0.09159	-0.09706	-0.01543
angular.se	-0.4111	-0.53562	-0.42817	0.110187	-0.18284	-0.14433	-0.12988	-0.1115	-0.11807	-0.12507	-0.13514	-0.15067	-0.14936	-0.14739	-0.07649
joint.maxi	-0.35989	-0.43495	-0.34785	0.222672	-0.09137	-0.06773	-0.03578	-0.0209	0.020681	-0.04147	-0.04497	-0.05285	-0.04191	-0.05198	0.009738
coarsenes	-0.33866	-0.43736	-0.40475	-0.1034	-0.2485	-0.21545	-0.20079	-0.17803	-0.18109	-0.18372	-0.19597	-0.20407	-0.20713	-0.19028	-0.20363
coarsenes	-0.36536	-0.44633	-0.39643	-0.09049	-0.27519	-0.23887	-0.23121	-0.20998	-0.21939	-0.21523	-0.22705	-0.2393	-0.24718	-0.21786	-0.22245
small.dista	-0.37389	-0.4715	-0.38702	0.137761	-0.08743	-0.05876	-0.02834	-0.02221	0.000985	-0.04366	-0.04847	-0.04782	-0.05854	-0.0592	-0.04672
Zone.dista	-0.34785	-0.45355	-0.37148	0.148047	-0.1161	-0.08655	-0.06204	-0.05701	-0.03075	-0.07911	-0.08349	-0.08174	-0.0929	-0.09684	-0.07342
small.dista	-0.403	-0.50213	-0.38593	0.133384	-0.1161	-0.08458	-0.05898	-0.04847	-0.02637	-0.06467	-0.06992	-0.07649	-0.08327	-0.07233	-0.06598
Zone.dista	-0.39227	-0.49622	-0.38002	0.1417	-0.11216	-0.07867	-0.05504	-0.04432	-0.02396	-0.05963	-0.06489	-0.07277	-0.07824	-0.06576	-0.05985
small.dista	-0.37148	-0.48047	-0.3531	0.148922	-0.17781	-0.15352	-0.10931	-0.10121	-0.05176	-0.12069	-0.12988	-0.122	-0.12376	-0.1522	-0.0616
Zone.dista	-0.36842	-0.48112	-0.3531	0.148703	-0.18394	-0.15986	-0.1161	-0.1069	-0.06292	-0.1266	-0.13514	-0.12988	-0.1301	-0.15549	-0.0697
Grey.level.	-0.32443	-0.42554	-0.39928	0.07364	-0.06467	-0.04957	-0.00952	-0.00274	0.008207	-0.02571	-0.03294	-0.02987	-0.04432	-0.05263	-0.02024
dependenc	-0.33384	-0.43692	-0.34128	0.192691	-0.11435	-0.07561	-0.0546	-0.04869	-0.01302	-0.06576	-0.07211	-0.08218	-0.08218	-0.06817	-0.03031
dependenc	-0.36426	-0.46406	-0.37324	0.134041	-0.16271	-0.12551	-0.10625	-0.09706	-0.06817	-0.11435	-0.12266	-0.1231	-0.13163	-0.1185	-0.10406

Figure 37: A heatmap of the correlations between 20 metabolomics variables and 15 radiomics features obtained by using the third segmentation method (second part).

16.2 Discriminative model

For the second goal of this research, building the discriminative models, 85 patients were included.

In table 11, the complete patient cohort is shown in more detail. A specification of the types of lung tumors is represented in table 12, more specifically for adenocarcinoma, squamous cell carcinoma, and neuroendocrine lung tumors. Table 13 provides a comparison between the malignant and the non-malignant lesions. Figures 38-45 show an overview of this patient cohort.

16.2.1 Included patients

Table 11: Specifications total patient cohort (N=85) for the discriminative radiomics model.

Variable		
Total patients		85
Sex (N, (%))	*Men*	53 (62.4)
	Women	32 (37.6)
Diabetes (N, (%))	*Yes*	10 (11.8)
	No	75 (88.2)
Smoking status (N, (%))	*Current smoker*	39 (45.9)
	Ex-smoker	44 (51.8)
	Non-smoker	2 (2.3)
Packyears (y)	*Median*	37
	Average	40 ± 24
	Range	0 - 139
	Unknown	8
Age (y)	*Median*	69
	Average	68 ± 9
	Range	40 - 84
BMI (kg/m²)	*Median*	26.02
	Average	26.30 ± 4.89
	Range	14.69 - 51.00
Diameter (mm)	*Median*	25
	Average	31.03 ± 19.70
	Range	9 - 120
Plasma glucose (mg%)	*Median*	97
	Average	99.75 ± 16.76
	Range	76 - 168
FEV1 absolute value (l)	*Median*	2.27
	Average	2.38 ± 0.75
	Range	1.09 - 4.65

Table 12: Specifications of the malignant lung lesions (N=72) for the discriminative radiomics model.

		Adenocarcinoma	Squamous cell carcinoma	Neuroendocrine lung tumor
Total patients		46	21	5
Sex (N, (%))	*Men*	25 (54.3)	14 (66.7)	5
	Women	21 (45.7)	7 (33.3)	0
Diabetes (N, (%))	*Yes*	7 (15.2)	2 (9.5)	1 (20.0)
	No	39 (84.8)	19 (90.5)	4 (80.0)
Smoking status (N, (%))	*Current smoker*	21 (45.7)	7 (33.3)	3 (60.0)
	Ex-smoker	23 (50.0)	14 (66.7)	2 (40.0)
	Non-smoker	2 (4.3)	0	0
Packyears (y)	*Median*	35	40	38
	Average	36 ± 23	40 ± 21	41 ± 16
	Range	0 - 130	0 - 80	26 - 70
	Unknown	4	1	0
Age (y)	*Median*	70	70	75
	Average	69 ± 8	70 ± 8	75 ± 4
	Range	49 - 84	55 - 83	71 - 80
BMI (kg/m²)	*Median*	25.63	26.85	27.29
	Average	26.31 ± 5.60	26.74 ± 3.31	26.30 ± 3.18
	Range	16 .00 - 51.00	19.72 - 34.89	20.76 - 29.38
Lobe	*Right upper*	19	3	1
	Right middle	3	1	2
	Right lower	5	4	2
	Left upper	12	7	1
	Left under	12	6	0
Diameter (mm)	*Median*	22	30	32
	Average	31.07 ± 22.96	30.70 ± 12.47	37.60 ± 12.97
	Range	9 - 120	12 - 60	20 - 55
Resection margin	*Positive*	4	1	1
	Negative	42	19	4
	Unknown	0	1	0
Visceral pleural invasion	*Yes*	10	8	0
	No	36	13	5
Lymph vascular invasion	*Yes*	10	7	2
	No	36	14	3
Positive nodes	*Yes*	10	2	0
	No	36	19	5
Plasma glucose (mg%)	*Median*	96	100	95
	Average	99.80 ± 18.20	102.81 ± 14.03	101.20 ± 20.75
	Range	76 - 168	83 - 141	79 - 137
FEV1 absolute value (l)	*Median*	2.22	2.32	2.35
	Average	2.35 ± 0.74	2.26 ± 0.71	2.63 ± 0.80
	Range	1.09 - 4.49	1.16 - 3.94	1.72 - 4.08

		Adenocarcinoma	Squamous cell carcinoma	Neuroendocrine lung tumor
T	*1a*	2	1	0
	1b	13	7	2
	1c	9	2	0
	2a	13	5	1
	2b	0	2	1
	3	9	2	1
	4	2	3	0
N	*0*	37	17	5
	1	8	4	0
	2	1	1	0
M	*0*	45	22	5
	1a	1	0	0
R	*0*	44	21	4
	1	2	1	1

Note: due to the COVID-19 crisis, the lobectomy of the last patient is not fulfilled. This means that the type of tumor and other tumor related features are unknown. Demography features of this patients are included.

Table 13: Comparison between malignant and non-malignant lung lesions (N=85) for the discriminative radiomics model.

		Malignant	Non-malignant
Total patients		73	12
Sex (N, (%))	*Men*	44 (60.3)	9 (75.0)
	Women	29 (39.7)	3 (25.0)
Diabetes (N, (%))	*Yes*	10 (13.7)	0 (0.0)
	No	63 (86.3)	12 (100.0)
Smoking status (N, (%))	*Current smoker*	32 (43.8)	7 (58.3)
	Ex-smoker	39 (53.5)	5 (41.7)
	Non-smoker	2 (2.7)	0
Packyears (y)	*Median*	37	48
	Average	38 ± 22	51 ± 38
	Range	0 - 130	8 - 139
	Unknown	6	2
Age (y)	*Median*	70	61
	Average	70 ± 8	61 ± 11
	Range	49 - 84	40 - 80
BMI (kg/m²)	*Median*	26.57	24.44
	Average	26.47 ± 4.87	25.27 ± 4.86
	Range	16.00 - 51.00	14.69 - 33.00
Lobe	*Right upper*	23	6
	Right middle	6	2
	Right lower	10	2
	Left upper	19	2
	Left under	18	1
Diameter (mm)	*Median*	25	
	Average	31.43 ± 19.94	
	Range	9-120	
Plasma glucose (mg%)	*Median*	97	90
	Average	100.66 ± 17.25	94.25 ± 12.04
	Range	76 - 168	76 - 120
FEV1 absolute value (l)	*Median*	2.25	2.31
	Average	2.34 ± 0.74	2.60 ± 0.82
	Range	1.09 - 4.49	1.71 - 4.65

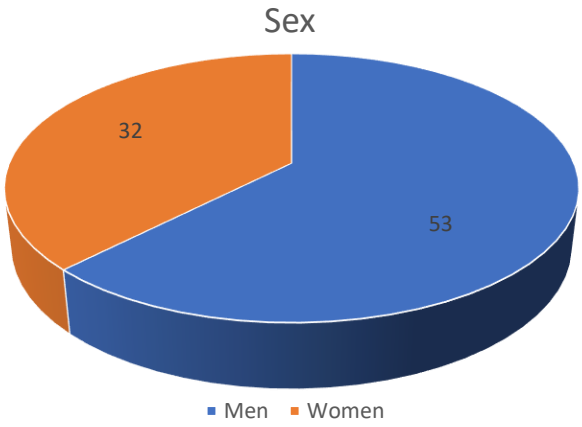

Figure 38: Pie chart of sex of the patients with radiomics features.

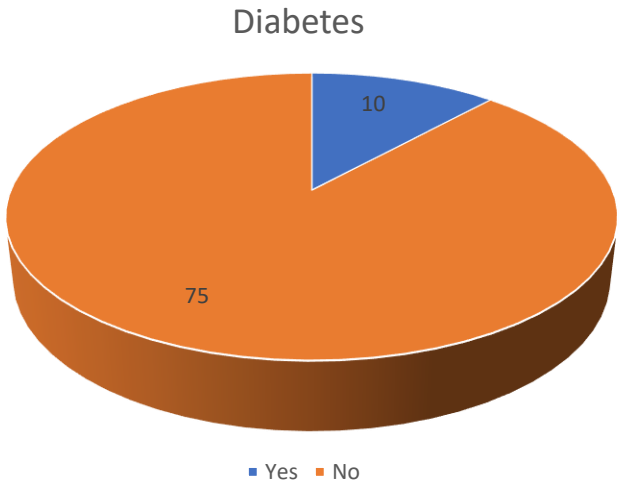

Figure 39: Pie chart of diabetes of the patients with radiomics features.

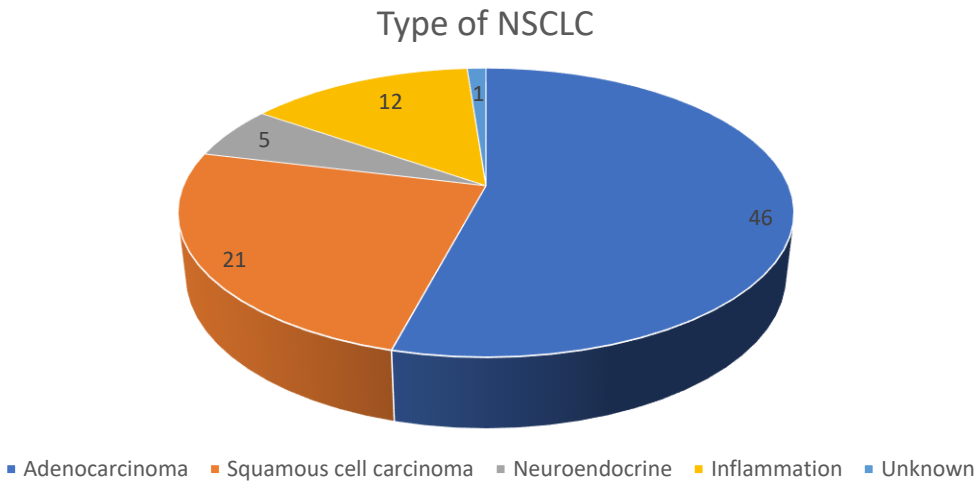

Figure 40: Pie chart of different types of NSCLC of the patients with radiomics features.

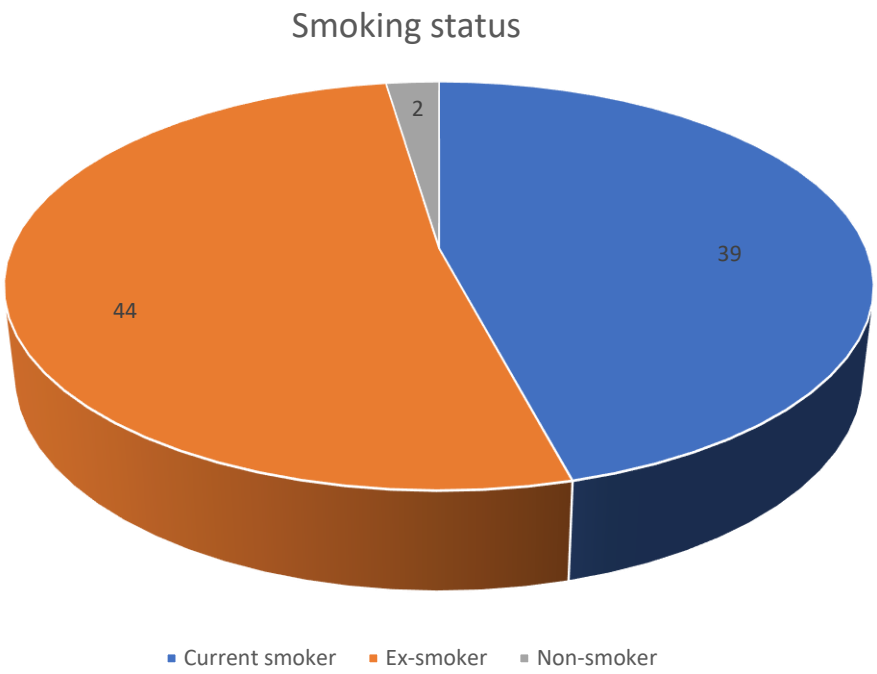

Figure 41: Pie chart of the smoking status of the patients with radiomics features.

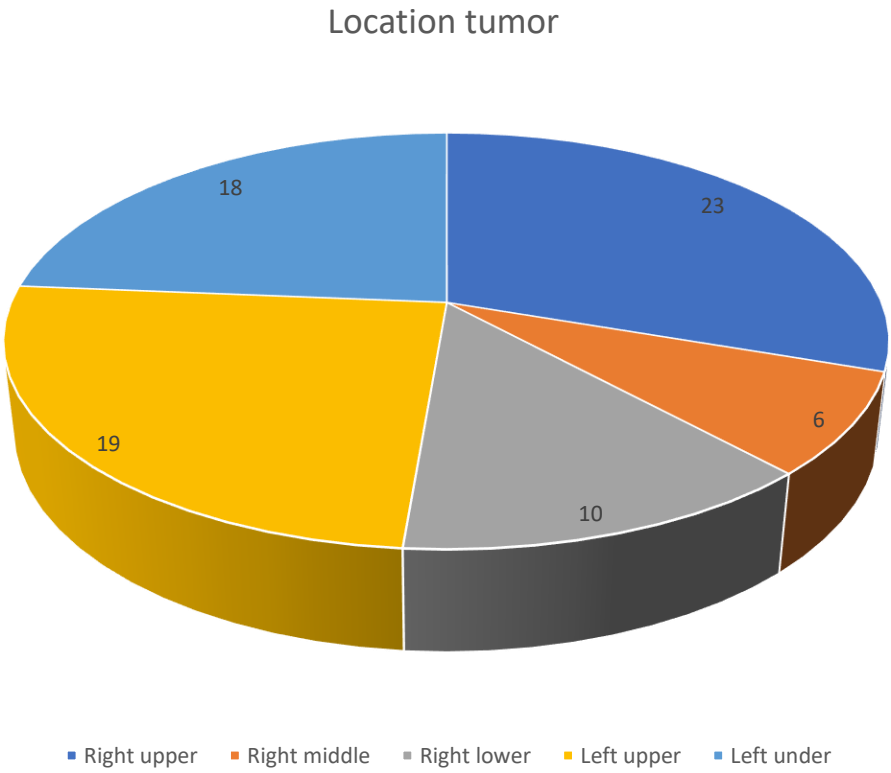

Figure 42: Pie chart of the location of the tumor of the patients with radiomics features.

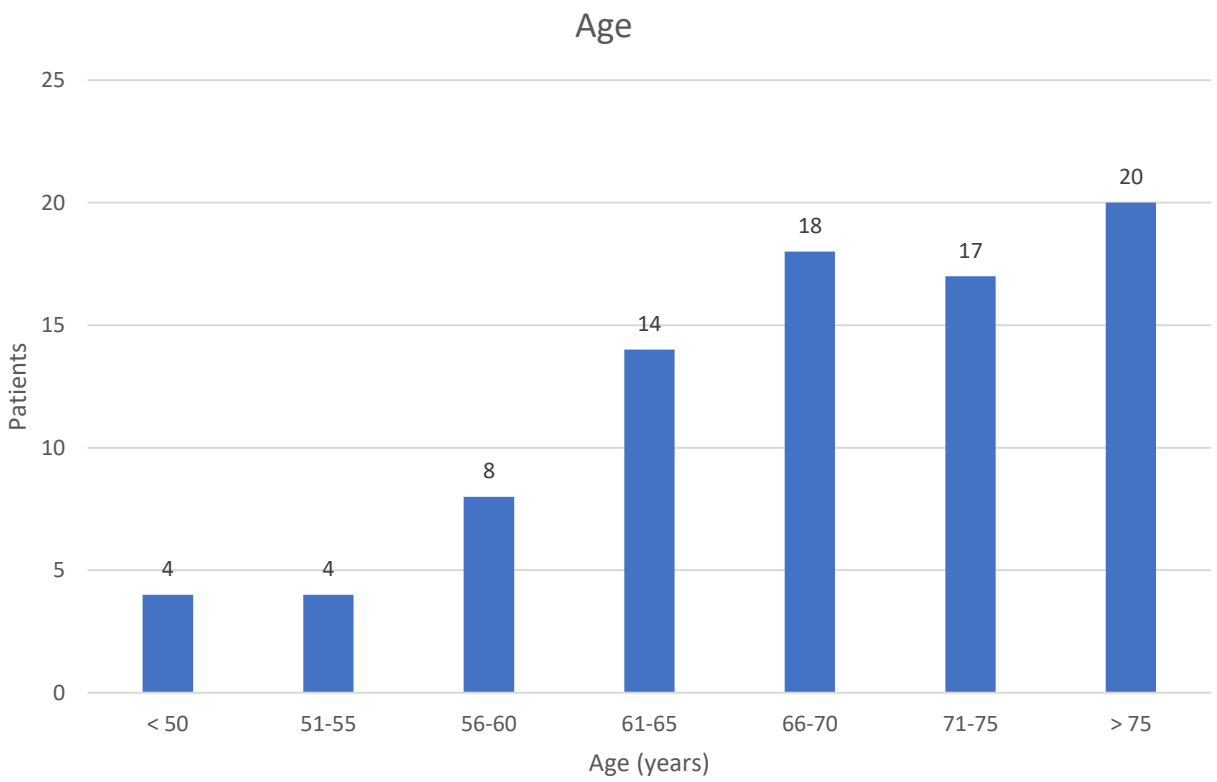

Figure 43: Bar chart of the age of the patients with radiomics features.

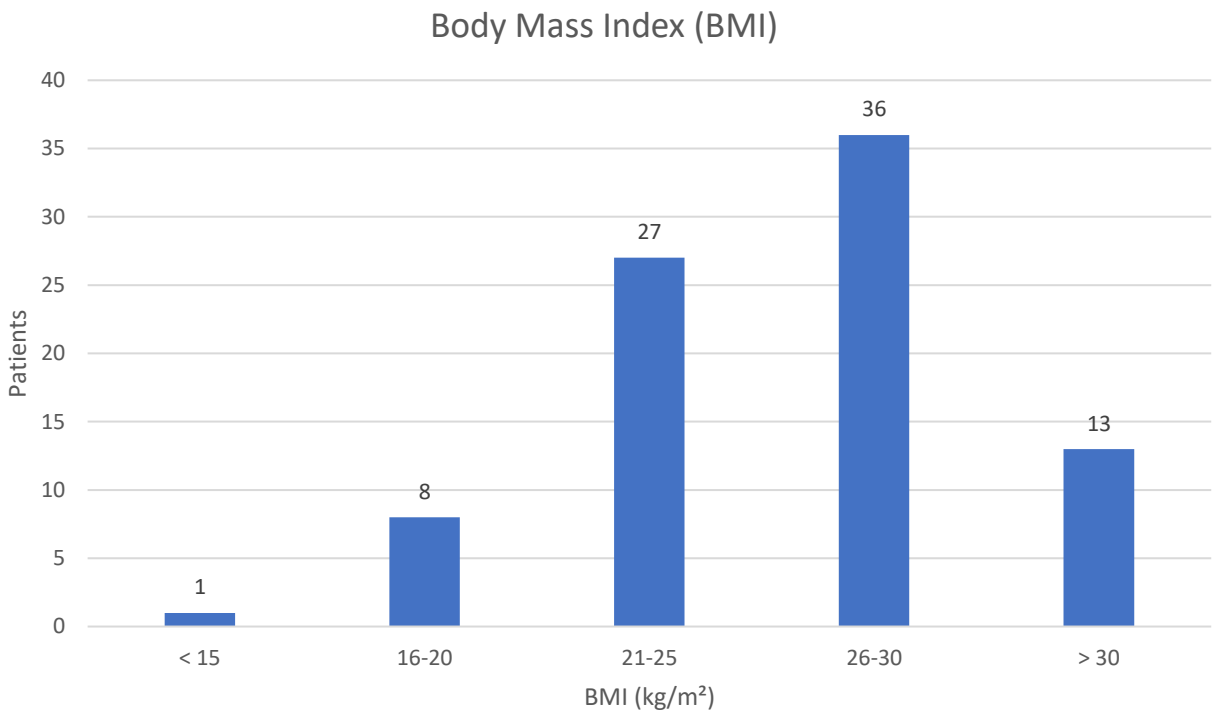

Figure 44: Bar chart of the BMI of the patients with radiomics features.

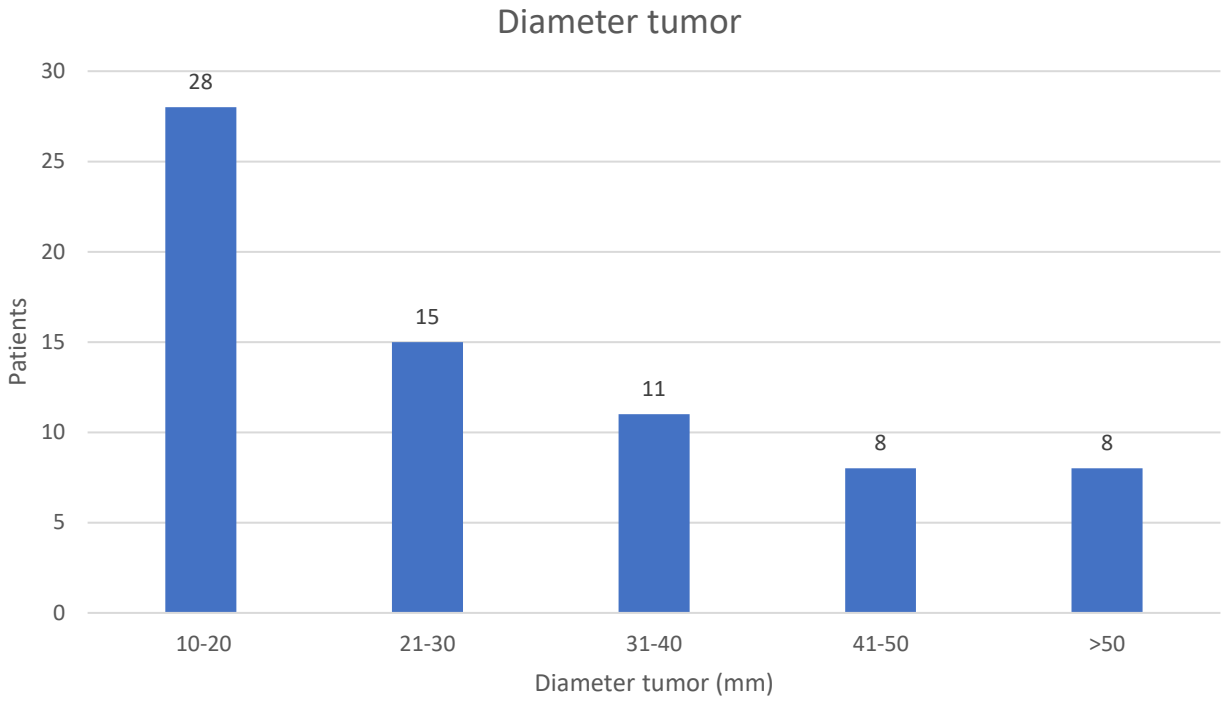

Figure 45: Bar chart of the diameter of the tumor of the patients with radiomics.

16.2.2 Results

The obtained dataset with radiomics features of the 85 included patients is used in this part of the study, together with a dataset that contains 66 of these 85 patients, diagnosed with adenocarcinoma and squamous cell carcinoma. The datasets were obtained using the second type of segmentation (PET/CT). First, the largest dataset was reduced in two ways: Spearman correlation coefficient test combined with manual data reduction, and FA. The dataset of 66 patients was only reduced using the Spearman correlation coefficient test combined with manual reduction.

All features that showed a correlation of at least 90% were excluded from the dataset. Afterwards, only 56 features remained per patient. FA was executed multiple times in such a way that 3,4,5 and 10 factors per patient remained in the output matrices. The distribution of the radiomics features into the factors is shown by figure 46 for the FA performed with three factors. The distribution for the FA with 4, 5 and 10 factors can be found in the annex (S1-S3).

Every patient in both resulting datasets of the 85 patients with a malignant lung lesion was labeled as '1,' and every patient with a non-malignant lung lesion was labeled as '0' in the 'Event' column. Both datasets underwent a forward selection and backwards stepwise selection regression to generate the discriminative models using the two different data reduction methods. The regression was executed in RStudio. The manually reduced dataset was split into a 75% training dataset and a 25% test dataset. The regression methods were used on the training dataset. The exact same method was used for the manually reduced dataset containing radiomics features of 66 patients. Every patient in this dataset with adenocarcinoma was labeled as '1,' and every patient with squamous cell carcinoma was labeled as '0' in the 'Event' column.

The forward logistic regression was used first, with a threshold of 0.2. For the datasets that were reduced manually, the models started with zero variables. A Chi-squared test was used to determine which parameter with the lowest p-value was added to the model first. After this variable was added to the model, another Chi-squared test was used to determine which parameter with the lowest p-value must be added next. This was done until no other variables were significant enough to add to the model. After these steps, a model based on eight radiomics features was built for the dataset of 85 patients. For the dataset of 66 patients, the model was build based on seven radiomics features.

The next step for these models is to refine to a threshold of 0.05. This is done with a backwards stepwise logistic regression, which started from the model build on respectively eight or seven features. Then a z-test was ran on this model, and, based on the p-value, the feature above the threshold of 0.05 was removed from the model. This was done until it was not significant anymore to remove features from the model. At the end of this step, the refined model was built based on two radiomics features for the dataset of 85 patients; inverse difference normalized and quartile coefficient.

For the other dataset, the model was also built on two radiomics features; inverse difference normalized and zone distance variance GLDZM. The test datasets were used after this to determine the accuracy of the model.

Table 14 and 15 shows the results of these tests, respectively for the model build based on dataset of 85 patients and based on the dataset of 66 patients.

The patients that were labeled with '1' in 'Events' in the 25% test dataset of the 85 patients were placed correctly for 94.74% (sensitivity) using the model generated with the 75% training dataset. Unfortunately, the patients that were labeled with '0' in 'Events' (specificity) were never placed correctly using this model. This phenomenon can be explained by the unbalanced distribution of the patients with malignant and non-malignant PET-positive lung nodules. For the datasets that were created with FA, the logistic regression did not give a significant result for any of the four created datasets. After going through the forward logistic regression with the threshold of 0.2, a model was generated consisting of only one radiomics feature.

Table 14: Summary of the results from the discriminative model for malignant and non-malignant PET-positive lung nodules.

Accuracy	0.8182
95% CI	(0.5972, 0.9481)
P-Value [Acc > NIR]	0.8283
Sensitivity	0.9474
Specificity	0.0000

Table 15: Summary of the results from the discriminative model for adenocarcinoma and squamous cell carcinoma.

Accuracy	0.7647
95% CI	(0.501, 0.9319)
P-Value [Acc > NIR]	0.4093
Sensitivity	0.4000
Specificity	0.9167

The patients that were labeled with '1' in 'Events' in the 25% test dataset of the 66 patients were placed correctly for 40.00% (sensitivity) using the model generated with the 75% training dataset. The patients that were labeled with '0' in 'Events' were placed right using this model in 91.67% of the cases (specificity). Again, the reason for this very low sensitivity will probably be due to the small patient cohort.

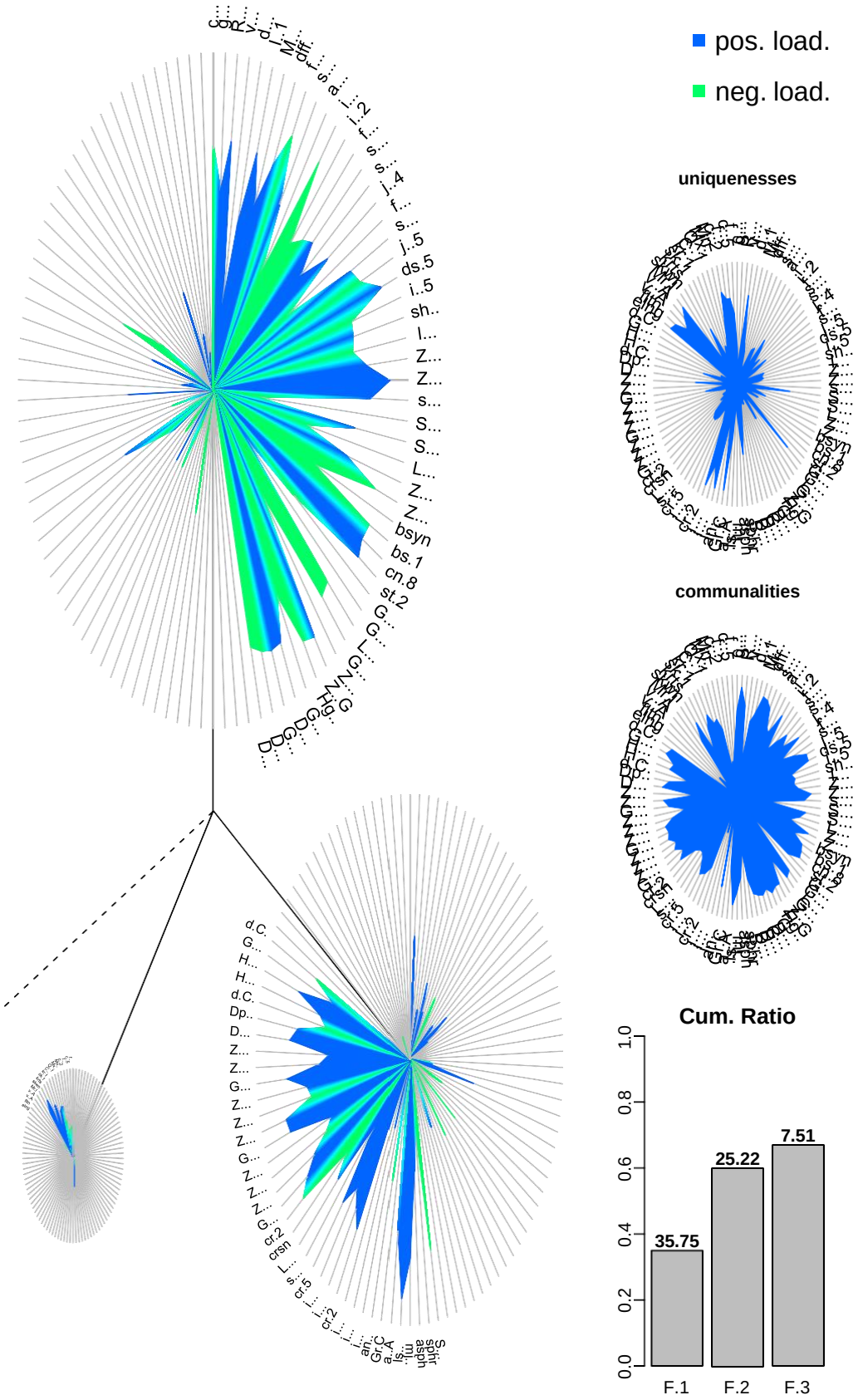

Figure 46: Distribution of the radiomics features after Factor Analysis with an output of three factors.

17 Discussion

This study reports the unique methodology of combining two techniques, metabolomics from plasma samples and radiomics from PET/CT images in NSCLC. We show that metabolomics and radiomics data from patients diagnosed with early-stage and locally advanced NSCLC are correlated and that discriminative models based on radiomics features can be built. The combination of these two techniques leads to our main findings:

1. Correlations are found between radiomics data and metabolomics data of the included NSCLC patients. Increased plasma glucose uptake correlates significantly positively with six radiomics features, whereas 13 radiomics features are negatively correlated with decreased plasma glycerol.
2. A discriminative model based on radiomics features distinguishes between a non-malignant and a malignant lesion, and between the pathology of an adenocarcinoma and a squamous cell carcinoma.

The first findings show positive as well as negative correlations between radiomics features and metabolomics variables. It is crucial to notice that metabolites might be more or less dominant in plasma samples of cancer cells due to the reprogramming of these cells, compared to cells in a healthy person (129). These effects can be represented in the metabolic data. Next to the metabolic data, radiomics features are obtained with the help of PET/CT imaging and are evaluated if they could lead to better insights in the understanding of patients with NSCLC.

The positive correlations found in the total omics-datasets are mainly related to the concentration of plasma glucose. The six radiomics features positively correlated to glucose are identified as Morans I, inverse difference normalized, grey level non-uniformity GLSZM, grey level non-uniformity GLDZM, area density AABB, and first measure of information correlation 1.
The negative correlations found in the total omics-datasets are mainly related to glycerol and phospholipids. The 13 radiomics features that were negatively correlated to the metabolites are identified as coefficient of variation, quartile coefficient, small zone low grey level emphasis, low dependence low grey level emphasis, surface to volume ratio, minimum histogram gradient, coarseness, zone distance non-uniformity normalized GLDZM, small distance emphasis GLDZM, dependence count energy, joint maximum, angular second moment and grey level non-uniformity normalized.

Smolle et al. already showed that the glycolysis and gluconeogenesis processes are highly activated in tumor cells of patients with NSCLC (130). PET/CT imaging is used in this study due to the high glucose uptake revealed by PET imaging (131). It is essential to notice that cancer cells need a higher glucose uptake to support the increased glycolysis cycle.

Vanhove et al. previously suggested that a high glucose level in plasma is needed for cancer cells to survive. Cancer cells support the ongoing glycolysis, and therefore, the compensatory role of gluconeogenesis is crucial (132). This compensatory role indicates that more glucose is made in other cells of the human body by gluconeogenesis. The glucose formed from this process is transported within the blood plasma to sustain the glycolysis process within tumor cells to provide energy. An increase in glucose uptake leads to a decrease in glycerol due to the need for glycerol to form glucose. The hypothesis of Vanhove et al. is confirmed by our independent study based on the correlations found between the radiomics and metabolomics datasets.

Louis et al. already showed that lung cancer patients have an increased plasma glucose and lactate level, which is clearly in line with the result of this study. In addition, it confirms the supporting role of gluconeogenesis in normal non-cancer cells (133). Interestingly, decreased phospholipid levels are also detected in the plasma of lung cancer patients (133). This decrease is accordant to an enhanced membrane synthesis in cancer cells (134-136).

In summary, increased plasma glucose uptake correlates significantly positively with six radiomics features, whereas 13 radiomics features are negatively correlated with decreased plasma glycerol.

The impact of the method of segmentation on these results is not negligible. As shown in the '*Results'* section, the correlation coefficient variation depends on the segmentation method of the lung lesion to obtain the volume of interest. Therefore, a suitable segmentation method is crucial. Lu et al. examined the influence of segmentation methods and the impact of the used tracer. They concluded that slight differences in segmentation methods and the used tracers, ^{18}F-FDG and ^{11}C-choline, are present (137). The segmentation method using PET (third segmentation method as discussed in '*Results*') shows the most and strongest correlations between the metabolomics variables related to the concentration of plasma glucose in blood. This method is the easiest to use when looking at underlying correlations between radiomics features and metabolic parameters. This result is in line with our expectations. Since we are looking at variables related to the glucose plasma uptake in blood, it is logical that higher correlation values are obtained using only PET images (131).

The second part of this research consists of building the discriminative models based on the radiomics features extracted from PET/CT images. Two types of data reduction methods are used on the large radiomics datasets, manual reduction, and FA. The two different methods for data reduction were used to see if this difference was also visible in the resulting models.

However, the models built based on the factors are unsuccessful. The one generated based on the manually reduced dataset of 85 patients shows better results, although still far from optimal. The sensitivity of this model is 0.9474, the specificity is 0.00. Table 14 suggests that the generated model and results of the discriminative model are too fragile for both reduction methods, due to the small patient cohort. Despite these results, the method to generate this discriminative model is correct.

The model that was generated based on the manually reduced dataset of 66 patients also gave a similar result. The patients labeled with '1' in 'Events' in the 25% test dataset are placed correctly for 40% using the model generated with the 75% training dataset. The patients labeled with '0' in 'Events' in the 25% training dataset are placed correctly for 91.67%.
Again, these results also show that a small dataset is far from optimal when building a discriminative model. It was evident that, although the used method is correct, the results would not be optimal because this test dataset only consisted of 17 patients. Independently of the method used for data reduction, the small patient cohort remains a problem.

Following methodological biases can have influenced our results; the small patient cohort, the heterogeneity of lung cancer images on PET/CT, the impact of the PET/CT camera itself, and the focus from mainly plasma glucose and glycerol (and not, for example, glutamine and glutamate (16)). Future research is necessary to verify the concluded results.

The first step to achieve this goal is to enlarge the patient cohort to increase the reliability of this study. Although the included patient cohort is homogeneous, heterogeneity within the patient cohorts is still present. This heterogeneity is made visible in the section named '*Results*' with the help of pie charts of the included patients (figures 75-78, 97-100).

A solution to tackle the problem of cancer heterogeneity with a patient is a quantitative scoring system of the characteristics of the tumor on PET/CT images. This system was already executed as part of this study for four characteristics of the tumor. First, the heterogeneity or homogeneity was scored on an observer-based scale from one to five. A score of five indicated a fully heterogeneous tumor, and one indicated a very homogeneous tumor. Secondly, a score from one to three was given to the shape of the tumor, a score of one was a very nodular shape of the tumor, and a score of three for a fanciful shape. Thirdly, the presence of retro-obstructive pneumonia in the lung lesion was scored. Lastly, the presence of central necrose was evaluated. However, this scaling solution has to be validated by its inter-rater reliability (e.g., Cohens' kappa coefficient (138)) and its usefulness in understanding radiomics before it can be used in further research.

Another possible bias we can not exclude in our results is the impact of the used PET/CT camera. We used a Biograph Horizon from Siemens but could not compare it with another PET/CT camera. The research group has already collected metabolic parameters from another 300 patients. These patients all have PET/CT images made with a different camera, specifically a GEMINI TF Big bore from Philips. The results obtained using the PET/CT camera we used, and the results obtained using this camera could be examined and compared to each other. Our hypothesis is that a different PET/CT camera will not lead to significant results. Minor differences might be visible from the radiomics tool, but there will not be a different outcome.

Looking at the long-term goals of this study, the results are auspicious.
In the nearby future, further research could compare healthy lung tissue and tissue out of a lung tumor. Metabolomics features are parameters out of the blood plasma and radiomics features are parameters out of tissues. Because of the chirurgical resection of the tumor from the included patients, tissue out of a lung tumor is already collected. Combining the results of healthy lung tissue and tissue out of a lung tumor with the metabolomics of the patients could give a good insight into possibly change in tissue (radiomics) that could lead to any changes in the parameters of the blood plasma (metabolomics), or the other way around. This research hypothesizes that the metabolism in the blood will adapt to the metabolism in the tissue if it changes and vice versa.

In the far future, this whole process could be automated. The PET/CT images of a patient could be loaded into a program that would automatically extract relevant radiomics features. Concurrently with the processing of images, blood samples of that patient could be taken and analyzed, focusing on specific metabolites that show an underlying correlation with the radiomics features of the PET/CT image. The automated process could immediately show if specific important metabolic parameters and features of the PET/CT images impact patients' clinical outcome.

To this day, several patients underwent surgery that was not necessary. This type of surgery happens in up to 11% of the cases with lung cancer (139). If this automatization and correlation between radiomics features and metabolic parameters could be linked directly to a diagnosis, unnecessary treatment and operations could be avoided, and the patient could get an earlier diagnosis.

18 Conclusion

For the first goal of this research, the correlation of metabolomics and radiomics features, 39 patients were included. Three different types of segmentation of the lung lesions were used to create VOIs. Out of every VOI, 483 radiomics features were extracted. Simultaneously, 238 metabolic parameters representing 62 plasma metabolites were determined from the same patient cohort using proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. On each dataset, the Spearman correlation coefficient was used to examine the underlying correlations between the radiomics features and the metabolic parameters, using a threshold of 0.3 (or -0.3).

Our research suggests that metabolic parameters of NSCLC patients are correlated to features of the PET/CT images of those patients. In all three segmentation methods combined, high glucose uptake is in line with the six radiomics features significantly positively correlated to increased plasma glucose. In contrast, 13 radiomics features are negatively correlated with decreased plasma glycerol, proving the supporting role of gluconeogenesis. The third segmentation method (PET) shows the strongest correlations with metabolomics.

The second aim of our research was to generate a model that could discriminate between a non-malignant lesion and a malignant lesion, and between an adenocarcinoma and a squamous cell carcinoma for NSCLC patients based on radiomics features. However, the models based on the two reduction methods proved to be not helpful when only a small patient cohort could be obtained.

Based on the datasets reduced with FA, this method did not give a useful model. The model generated based on the manually reduced dataset of 85 patients gave a specificity of 0.00 and a sensitivity of 0.9474. The model generated based on the manually reduced dataset of 66 patients had a specificity of 0.9167 and a sensitivity of 0.40. The models are too fragile to build further conclusions due to the small patient cohort.

References

1. society Ac. Key statistics for lung cancer 2021 [Available from: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html.
2. Burotto M, Thomas A, Subramaniam D, Giaccone G, Rajan A. Biomarkers in early-stage non-small-cell lung cancer: current concepts and future directions. J Thorac Oncol. 2014;9(11):1609-17.
3. Compton C. Cancer: The Enemy From Within: Springer Nature Switzerland AG; 2020 [cited 2021.
4. Dubois L. Tumor Biology. 2020.
5. Kalyanaraman B. Teaching the basics of cancer metabolism: Developing antitumor strategies by exploiting the differences between normal and cancer cell metabolism. Redox Biol. 2017;12:833-42.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
7. El-Tanani M, Dakir el H, Raynor B, Morgan R. Mechanisms of Nuclear Export in Cancer and Resistance to Chemotherapy. Cancers (Basel). 2016;8(3).
8. Fadaka A, Ajiboye B, Ojo O, Adewale O, Olayide I, Emuowhochere R. Biology of glucose metabolization in cancer cells. Journal of Oncological Sciences. 2017;3(2):45-51.
9. College O. Cellular Respiration: Oxidation of pyruvate and the citric acid cycle. 2013. In: Biology [Internet]. [214-7]. Available from: https://cnx.org/contents/GFy_h8cu@9.85:weAHBat1@7/Oxidation-of-Pyruvate-and-the-Citric-Acid-Cycle.
10. Clark MA, Douglas M, Choi J. Oxidation of Pyruvate and the Citric Acid Cycle. Biology 2e. 2 ed. Houston, Texas: OpenStax; 2018. p. 206-9.
11. Khan S. The citric acid cycle 2009 [Available from: https://www.khanacademy.org/science/biology/cellular-respiration-and-fermentation/pyruvate-oxidation-and-the-citric-acid-cycle/a/the-citric-acid-cycle.
12. Rye C, Wise R, Jurukovski V, DeSaix J, Choi J, Avissar Y. Cellular respiration: oxidative phosphorylation. 2017. In: Biology [Internet]. [213-7].
13. Lee O, O'Brien PJ. Modifications of Mitochondrial Function by Toxicants. 2010. In: Comprehensive Toxicology [Internet]. Elsevier. 1. [411-45]. Available from: https://www.sciencedirect.com/science/article/pii/B9780080468846001196.
14. Romero-Garcia S, Lopez-Gonzalez JS, Baez-Viveros JL, Aguilar-Cazares D, Prado-Garcia H. Tumor cell metabolism: an integral view. Cancer Biol Ther. 2011;12(11):939-48.
15. Martinez CAS, Claudio. Heterogeneity of Glucose Transport in Lung Cancer. Biomolecules. 2020;10(868).
16. Vanhove K, Derveaux E, Graulus G-J, Mesotten L, Thomeer M, Noben J-P, et al. Glutamine Addiction and Therapeutic Strategies in Lung Cancer. International Journal of Molecular Sciences. 2019;20(252).
17. Ryan DG, Murphy MP, Frezza C, Prag HA, Chouchani ET, O'Neill LA, et al. Coupling Krebs cycle metabolites to signalling in immunity and cancer. Nature Metabolism. 2019;1:16-33.
18. Cancer IAfRo. Cancer Today 2020 [Available from: https://gco.iarc.fr/today/home.
19. Institute NC. PDQ Non-Small Cell Lung Cancer Treatment 2020 [updated 12/3/2020. Available from: https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq.
20. Prevention CfDCa. What is Lung Cancer? 2020 [Available from: https://www.cdc.gov/cancer/lung/basic_info/what-is-lung-cancer.htm.
21. Stöppler MC. What Is Non-Small-Cell Lung Cancer (NSCLC)?2020. Available from: https://www.emedicinehealth.com/non-small-cell_lung_cancer/article_em.htm#what_are_the_stages_of_non-small-cell_lung_cancer.
22. Society AC. Cancer staging2020. Available from: https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html.
23. Society AC. Lung Cancer Early Detection, Diagnosis, and Staging2019.

24. Research MFfMEa. Lung cancer - Diagnosis. Available from: https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627?p=1.
25. Oncology ASoC. Lung Cancer - Non-Small Cell: Diagnosis2020. Available from: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/diagnosis#:~:text=Often%2C%20the%20radiologist%20uses%20a,for%20diagnosis%20and%20molecular%20testing.
26. Association AL. Lobectomy 2020 [updated 2/19/2020. Available from: https://www.lung.org/lung-health-diseases/lung-procedures-and-tests/lobectomy.
27. Service NH. Lung Cancer 2019 [Available from: https://www.nhs.uk/conditions/lung-cancer/.
28. Marino FZ, Bianco R, Accardo M, Ronchi A, Cozzolino I, Morgillo F, et al. Molecular heterogeneity in lung cancer: from mechanisms of origin to clinical implications2019; 16:[981-9 pp.].
29. Oncology ASoC. Lung Cancer - Non-Small Cell: Statistics2021. Available from: https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics#:~:text=The%205%2Dyear%20survival%20rate%20for%20NSCLC%20is%2024%25%2C,and%20the%20stage%20of%20disease.
30. Omami G, Tamimi D, Branstetter BF. Basic principles and applications of (18)F-FDG-PET/CT in oral and maxillofacial imaging: A pictorial essay. Imaging Sci Dent. 2014;44(4):325-32.
31. S. KK. Introductory Nuclear Physics: Wiley; 1987.
32. Groves AM. Non-[18F]FDG PET in clinical oncology. The lancet oncology. 2007;8:822-30.
33. Jiang W, Chalich Y, Deen MJ. Sensors for Positron Emission Tomography Applications. Sensors (Basel). 2019;19(22).
34. Humm JL, Rosenfeld A, Del Guerra A. From PET detectors to PET scanners. Eur J Nucl Med Mol Imaging. 2003;30(11):1574-97.
35. Le Bars D. Fluorine-18 and medical imaging: Radiopharmaceuticals for positron emission tomography. Journal of Fluorine Chemistry. 2006;127(11):1488-93.
36. Zheng X, Yu CL, Sha W, Radu C, Huang SC, Feng D. Study of an image-derived SUV and a modified SUV using mouse FDG-PET. Nucl Med Biol. 2011;38(3):353-62.
37. Bury T, Dowlati A, Paulus P, Corhay JL, Hustinx R, Ghaye B, et al. Whole-body 18FDG positron emission tomography in the staging of non-small cell lung cancer. Eur Respir J. 1997;10(11):2529-34.
38. Saunders CAB, Dussek JE, O'Doherty MJ, Maisey MN. Evaluation of Fluorine-18-Fluorodeoxyglucose whole body PET imaging in the staging of lung cancer. The Society of Thoracic Surgeons. 1999.
39. Hicks RJ. Role of 18F-FDG PET in assessment of response in non-small cell lung cancer. J Nucl Med. 2009;50 Suppl 1:31S-42S.
40. Zigler SS. Production and Quality Control of 18F-FDG.
41. Alsanea E, Alhalabi W. Prediction of radioactive injection dosage for PET imaging. Soft Computing. 2021.
42. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42(2):328-54.
43. Surasi DS, Bhambhvani P, Baldwin JA, Almodovar SE, O'Malley JP. (1)(8)F-FDG PET and PET/CT patient preparation: a review of the literature. J Nucl Med Technol. 2014;42(1):5-13.
44. Joanna S. Fowler JL, Nora D. Volkow, Gene-Jack Wang, Robert R. MacGregor, , Ding Y-S. Monoamine Oxidase: Radiotracer Development and Human Studies.
45. Brian M. Gallagher JSF, Neal I. Gutterson, Robert R. MacGregor,, Chung-NanWan APW. Metabolic Trapping as a Principle of Radiopharmaceu

tical Design: Some Factors Responsible for the

Biodistribution of [18F]2-Deoxy-

2-Fluoro-D-Glucose. THE JOURNAL OF NUCLEAR MEDICINE.

46. Kaira K, Serizawa M, Koh Y, Takahashi T, Yamaguchi A, Hanaoka H, et al. Biological significance of 18F-FDG uptake on PET in patients with non-small-cell lung cancer. Lung Cancer. 2014;83(2):197-204.
47. Langer CJ, Besse B, Gualberto A, Brambilla E, Soria J-C. The evolving role of histology in the management of advanced non-small-cell lung cancer. Clin Oncol. 2010.
48. Pikor LA, Ramnarine VR, Lam S, Lam WL. Genetic alterations defining NSCLC subtypes and their therapeutic implications. Lung Cancer. 2013;82(2):179-89.
49. de Geus-Oei LF, van Krieken JH, Aliredjo RP, Krabbe PF, Frielink C, Verhagen AF, et al. Biological correlates of FDG uptake in non-small cell lung cancer. Lung Cancer. 2007;55(1):79-87.
50. Vanhove K, Giesen P, Owokotomo OE, Mesotten L, Louis E, Shkedy Z, et al. The plasma glutamate concentration as a complementary tool to differentiate benign PET-positive lung lesions from lung cancer. BMC Cancer. 2018;18(1):868.
51. Reniers B. Les 4 PET-SPECT. 2020.
52. Ullah MN, Pratiwi E, Cheon J, Choi H, Yeom JY. Instrumentation for Time-of-Flight Positron Emission Tomography. Nucl Med Mol Imaging. 2016;50(2):112-22.
53. Conti M. State of the art and challenges of time-of-flight PET. Phys Med. 2009;25(1):1-11.
54. Healthineers S. Biograph Horizon 2020 [Available from: https://www.siemens-healthineers.com/molecular-imaging/pet-ct/biograph-horizon.
55. Fahey FH. Data acquisition in PET imaging. Nucl Med Technol. 2002;30.
56. Kinahan PE, Fletcher JW. Positron Emission Tomography-Computed Tomography Standardized Uptake Values in Clinical Practice and Assessing Response to Therapy. Seminars in Ultrasound, CT and MRI. 2010;31(6):496-505.
57. Chung HH, Kim JW, Han KH, Eo JS, Kang KW, Park NH, et al. Prognostic value of metabolic tumor volume measured by FDG-PET/CT in patients with cervical cancer. Gynecol Oncol. 2011;120(2):270-4.
58. Chen HHW, Chiu N-T, Su M-C, Guo H-R, Lee B-F. Prognostic Value of Whole-Body Total Lesion Glycolysis at Pretreatment FDG PET-CT in Non–Small Cell Lung Cancer. Radiology. 2012.
59. Erdi YE, Macapinlac H, Rosenzweig KE, Humm JL, Larson SM, Erdi AK, et al. Use of PET to monitor the response of lung cancer

to radiation treatment. European Journal of Nuclear Medicine. 2000;27.
60. Benz MR, Allen-Auerbach MS, Eilber FC, Chen HJ, Dry S, Phelps ME, et al. Combined assessment of metabolic and volumetric changes for assessment of tumor response in patients with soft-tissue sarcomas. J Nucl Med. 2008;49(10):1579-84.
61. Flamen P, Vanderlinden B, Delatte P, Ghanem G, Ameye L, Van Den Eynde M, et al. Multimodality imaging can predict the metabolic response of unresectable colorectal liver metastases to radioembolization therapy with Yttrium-90 labeled resin microspheres. Phys Med Biol. 2008;53(22):6591-603.
62. Seibert AJ. X-Ray Imaging Physics for Nuclear Medicine Technologists. Part 1: Basic Principles of X-Ray Production. 2004. In: Nuclear Medical Technology [Internet]. [139-47]. Available from: https://tech.snmjournals.org/content/jnmt/32/3/139.full.pdf.
63. L'Annunziata MF. Nuclear radiation, its interaction with matter and radioisotope decay. 2003. In: Handbook of Radioactivity Analysis [Internet]. [1-121]. Available from: https://click.endnote.com/viewer?doi=10.1016%2Fb978-012436603-9%2F50006&token=WzMyMDAzOTYsIjEwLjEwMTYvYjk3OC0wMTI0MzY2MDMtOS81MDAwNiJd.MNO2wieKfXjT8wv8tUD9fBNVYUk.
64. Lalush DS, Wernick MN. Iterative Image Reconstruction. 2004. In: Emission Tomography [Internet]. [443]. Available from: https://www.sciencedirect.com/science/article/pii/B9780127444826500247.
65. Bioengineering NIoBIa. Computed Tomography (CT) [Available from: https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct#:~:text=How%20does%20CT%20work%3F,-Unlike%20a%20conventional&text=During%20a%20CT%20scan%2C%20the,opposite%20the%20x%2Dray%20source.

66. Reid JD. Sarcoidosis in coroner's autopsies: a critical evaluation of diagnosis and prevalence from Cuyahoga County, Ohio. Sarcoidosis, vasculitis, and diffuse lung diseases : official journal of WASOG. 1998;15(1):44-51.
67. Reniers B. X-ray, Computed Tomography and cone beam CT. 2019.
68. Munro L. Basics of Radiation Protection for Everyday Use: World Health Organization; 2004. Available from: https://apps.who.int/iris/bitstream/handle/10665/42973/9241591781_eng.pdf;jsessionid=0DC8DCEECBD9C8F65FEE18AD94CFF118?sequence=1.
69. Bell DJ, Nadrljanski MM. X-ray tube [Available from: https://radiopaedia.org/articles/x-ray-tube-1.
70. Nieman K, Coenen A, Dijkshoorn M. Computed Tomography. 2015. In: Advanced Cardiac Imaging [Internet]. [97-125]. Available from: https://click.endnote.com/viewer?doi=10.1016%2Fb978-1-78242-282-2.00005-6&token=WzMyMDAzOTYsIjEwLjEwMTYvYjk3OC0xLTc4MjQyLTI4Mi0yLjAwMDA1LTYiXQ.YHakqsSMplbMla9D09R6e1u0AYs.
71. Pelberg R. Basic Principles in Computed Tomography (CT). In: Cardiac CT Angiography Manual [Internet]. [19-58].
72. Robert J. Gillies PEK, Hedvig Hricak. Radiomics Images Are More than pictures, they are data.pdf. Radiology. 2016;278.
73. Bernal J, Sanchez J. Use of Filtered Back projection Methods to Improve. 2009.
74. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50 Suppl 1:11S-20S.
75. Leijenaar RTH, Jong EECd, Larue RTHM, Timmeren JEv, Lambin P. Radiomics: de toekomst in medische

beeldvorming. NED TIJDSCHR ONCOL. 2017.
76. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30(9):1234-48.
77. Cunliffe A, Armato SG, 3rd, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. Int J Radiat Oncol Biol Phys. 2015;91(5):1048-56.
78. Zhang J, Yu C, Jiang G, Liu W, Tong L. 3D texture analysis on MRI images of Alzheimer's disease. Brain Imaging Behav. 2012;6(1):61-9.
79. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging. 2010;10:137-43.
80. Garcia G, Maiora J, Tapia A, De Blas M. Evaluation of texture for classification of abdominal aortic aneurysm after endovascular repair. J Digit Imaging. 2012;25(3):369-76.
81. Henriksson E, Kjellen E, Wahlberg P, Ohlsson T, Wennerberg J, Brun E. 2-Deoxy-2-[18F] fluoro-D-glucose uptake and correlation to intratumoral heterogeneity. ANTICANCER RESEARCH. 2007;27:2155-60.
82. Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. PNAS. 2008;105:5213-8.
83. Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. Eur J Nucl Med Mol Imaging. 2011;38(6):987-91.
84. Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, Metges JP, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med. 2011;52(3):369-78.
85. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. Eur J Nucl Med Mol Imaging. 2013;40(1):133-40.

86. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, et al. Deciphering Genomic Underpinnings of Quantitative MRI-based Radiomic Phenotypes of Invasive Breast Carcinoma. Sci Rep. 2015;5:17787.
87. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: The promise of radiomics. Phys Med. 2017;38:122-39.
88. Guo W, Li H, Zhu Y, Lan L, Yang S, Drukker K, et al. Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data. J Med Imaging (Bellingham). 2015;2(4):041007.
89. Gnep K, Fargeas A, Gutierrez-Carvajal RE, Commandeur F, Mathieu R, Ospina JD, et al. Haralick textural features on T2 -weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. J Magn Reson Imaging. 2017;45(1):103-17.
90. Collarino A, Garganese G, Fragomeni SM, Pereira Arias-Bouda LM, Ieria FP, Boellaard R, et al. Radiomics in vulvar cancer: first clinical experience using (18)F-FDG PET/CT images. J Nucl Med. 2018.
91. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non_Small Cell Lung Cancer_ Histopathologic Correlates for Texture Parameters at CT. Radiology. 2013;266.
92. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
93. Ohri N, Duan F, Snyder BS, Wei B, Machtay M, Alavi A, et al. Pretreatment 18F-FDG PET Textural Features in Locally Advanced Non-Small Cell Lung Cancer: Secondary Analysis of ACRIN 6668/RTOG 0235. J Nucl Med. 2016;57(6):842-8.
94. Kim H-J. Common Factor Analysis Versus Principal Component Analysis: Choice for Symptom Cluster Research. Asian Nursing Research. 2008;2.
95. Software NS. Factor Analysis. NCSS.
96. Karamizadeh S, Abdullah SM, Manaf AA, Zamani M, Hooman A. An Overview of Principal Component Analysis. Journal of Signal and Information Processing. 2013;04(03):173-5.
97. Ringnér M. What is principal component analysis. Nature biotechnology. 2008;26.
98. Malik F. What are Eigenvalues and Eigenvectors 2019 [Available from: https://medium.com/fintechexplained/what-are-eigenvalues-and-eigenvectors-a-must-know-concept-for-machine-learning-80d0fd330e47.
99. Asadi S, Rao DCDVS, V.Saikrishna. A Comparative study of Face Recognition with Principal Component Analysis and Cross-Correlation Technique. International Journal of Computer Applications. 2010;10.
100. Kumar H. A visual introduction to eigenvectors and eigenvalues 2018 [Available from: https://kharshit.github.io/blog/2018/05/11/a-visual-introduction-to-eigenvectors-and-eigenvalues.
101. Lever J, Krzywinski M, Altman N. Principal component analysis. Nature Methods. 2017;14(7):641-2.
102. Brems M. A One-Stop Shop for Principal Component Analysis 2017 [Available from: https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c.
103. Phillips PJ, Flynn PJ, Scruggs T, Bowyer KW, Chang J, Hoffman K, et al. Overview of the Face Recognition Grand Challenge. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005.
104. Bianconi F, Palumbo I, Spanu A, Nuvoli S, Fravolini ML, Palumbo B. PET/CT Radiomics in Lung Cancer: An Overview. Applied Science. 2020.
105. Xu H, Deng Y. Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient. IEEE Access. 2018;6:11634-40.
106. Dutilleul P, Stockwell JD, Frigon D, Legendre P. The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies. Journal of Agricultural. 2000.
107. Mansson R, Tsapogas P, Akerlund M, Lagergren A, Gisler R, Sigvardsson M. Pearson correlation analysis of microarray data allows for the identification of genetic targets for early B-cell factor. J Biol Chem. 2004;279(17):17905-13.
108. statisctics L. Pearson Product-Moment Correlation 2018 [Available from: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

109. Thirumalai C, Chandhini SA, M V. Analysing the Concrete Compressive Strength using Pearson and Spearman. International Conference on Electronics, Communication and Aerospace Technology. 2017.
110. Solutions S. Correlation (Pearson, Kendall, Spearman) 2021 [Available from: https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/.
111. to Sh. Spearman Rank Correlation (Spearman's Rho): Definition and How to Calculate it 2021 [Available from: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/spearman-rank-correlation-definition-calculate/.
112. Zibran MF. CHI-Squared Test of Independence.
113. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013;23(2):143-9.
114. Pandis N. The chi-square test. Am J Orthod Dentofacial Orthop. 2016;150(5):898-9.
115. Chhatwal J, Alagoz O, Lindstrom JM, Kahn ECJ, Shaffer AK, Burnside SE. A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis2009.
116. Molnar C. Interpretable Machine Learning 2021. Available from: https://christophm.github.io/interpretable-ml-book/index.html.
117. Hoffman IEJ. Logistic Regression. 2019. In: Basic Biostatistics for Medical and Biomedical Practitioners [Internet]. Available from: https://reader.elsevier.com/reader/sd/pii/B9780128170847000334?token=B58C84E5725EE5CC809E37977EE5340353AE16E22C98E30E2776DD40006DD9E634CCAC2A1E5F0A83126FF5668BD62893&originRegion=eu-west-1&originCreation=20210520124255.
118. Holdnack AJ, Millis S, Larrabee JG, Iverson LG. Assessing Performance Validity with the ACS. 2013. In: WAIS-IV, WMS-IV, and ACS: Advanced Clinical Interpretation [Internet]. Available from: https://reader.elsevier.com/reader/sd/pii/B9780123869340000079?token=3E665F1FEAB92528B1538C39423500998617F4DC5A45500E7A62FB7CB902C1C3A56FD838E3101A4EEEBB0DC28C996600&originRegion=eu-west-1&originCreation=20210520125252.
119. Mao R, Zhang L, Zhu R-Y. LSO/LYSO Crystals for Future HEP Experiments. Journal of Physics: Conference Series. 2011;293.
120. Vanhove K, Thomeer M, Derveaux E, Shkedy Z, Owokotomo OE, Adriaensens P, et al. Correlations between the metabolic profile and (18)F-FDG-Positron Emission Tomography-Computed Tomography parameters reveal the complexity of the metabolic reprogramming within lung cancer patients. Sci Rep. 2019;9(1):16212.
121. Loeb LA, Emster VL, Warner KE, Abbotts J, Laszlo J. Smoking and Lung Cancer: An Overview. 1984.
122. Zwanenburg A, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, Ashrafinia S, et al. Image biomarker standardisation initiative. 2016.
123. da Silva EC, Silva AC, de Paiva AC, Nunes RA. Diagnosis of lung nodule using Moran's index and Geary's coefficient in computerized tomography images. Pattern Analysis and Applications. 2007;11(1):89-99.
124. Clausi DA. An analysis of co-occurrence texture statistics as a function of grey level quantization. Canadian Journal of Remote Sensing. 2014;28(1):45-62.
125. Nykamp DQ. Vertex definition: Math Insight; [Available from: https://mathinsight.org/definition/graph_vertex.
126. Haralick RM, Shanmugam K, Dinstein IH. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973;SMC-3(6):610-21.
127. Peerlings J, Woodruff HC, Winfield JM, Ibrahim A, Van Beers BE, Heerschap A, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep. 2019;9(1):4800.
128. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. 1982.
129. Marien N. H-NMR based metabolomics for earlier diagnosis of NSCLC: analysis of pH-dependent chemical shifts: University of Hasselt; 2019.
130. Smolle E, Leko P, Stacher-Priehse E, Brcic L, El-Heliebi A, Hofmann L, et al. Distribution and prognostic significance of gluconeogenesis and glycolysis in lung cancer. Mol Oncol. 2020;14(11):2853-67.

131. Christen T, Sheikine Y, Rocha VZ, Hurwitz S, Goldfine AB, Di Carli M, et al. Increased glucose uptake in visceral versus subcutaneous adipose tissue revealed by PET imaging. JACC Cardiovasc Imaging. 2010;3(8):843-51.
132. Vanhove K, Graulus GJ, Mesotten L, Thomeer M, Derveaux E, Noben JP, et al. The Metabolic Landscape of Lung Cancer: New Insights in a Disturbed Glucose Metabolism. Front Oncol. 2019;9:1215.
133. Louis E, Adriaensens P, Guedens W, Bigirumurame T, Baeten K, Vanhove K, et al. Detection of Lung Cancer through Metabolic Changes Measured in Blood Plasma. J Thorac Oncol. 2016;11(4):516-23.
134. Santos CR, Schulze A. Lipid metabolism in cancer. FEBS J. 2012;279(15):2610-23.
135. Baenke F, Peck B, Miess H, Schulze A. Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. Dis Model Mech. 2013;6(6):1353-63.
136. Currie E, Schulze A, Zechner R, Walther TC, Farese RV, Jr. Cellular fatty acid metabolism and cancer. Cell Metab. 2013;18(2):153-61.
137. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of Radiomic Features in [(11)C]Choline and [(18)F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. Mol Imaging Biol. 2016;18(6):935-45.
138. Jans M, Soffer P, Jouck T. Building a valuable event log for process mining: An experimental exploration of a guided process2019. Available from: https://click.endnote.com/viewer?doi=10.1177%2F0013164460020001O4&token=WzMyMDAzOTYsIjEwLjExNzcvMDAxMzE2NDQ2MDAyMDAwMTA0Il0.jHPwnvy2zKfOnXVdY1Sy5-adoYc.
139. R. F, T. B, R. A, S. A, L. K, B. S, et al. Balancing curability and unnecessary surgery in the context of computed tomography screening for lung cancer. The Journal of thoracic and cardiovascular surgery. 2014:10.

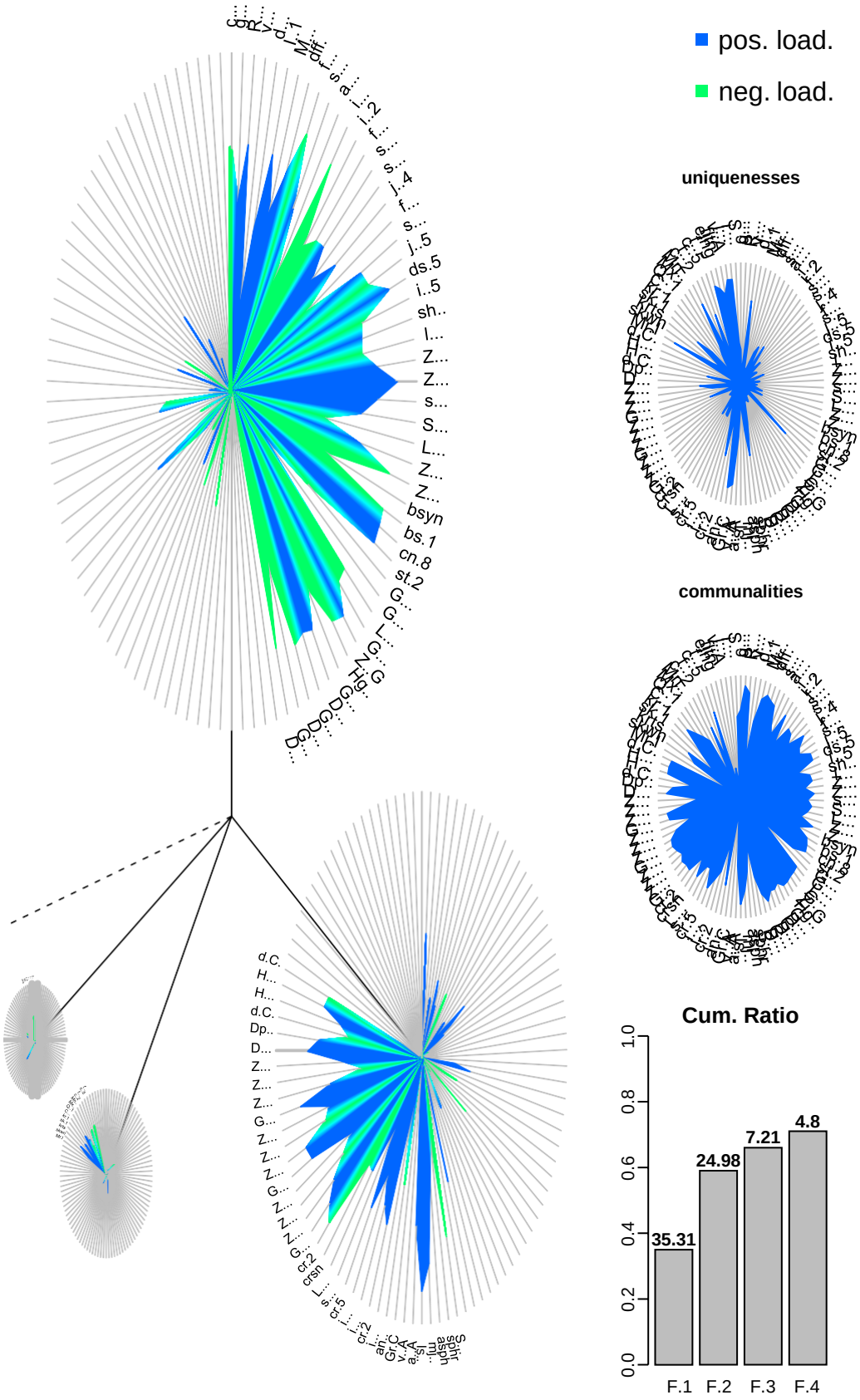

Figure S1: Distribution of the radiomics features after Factor Analysis with an outcome of four factors.

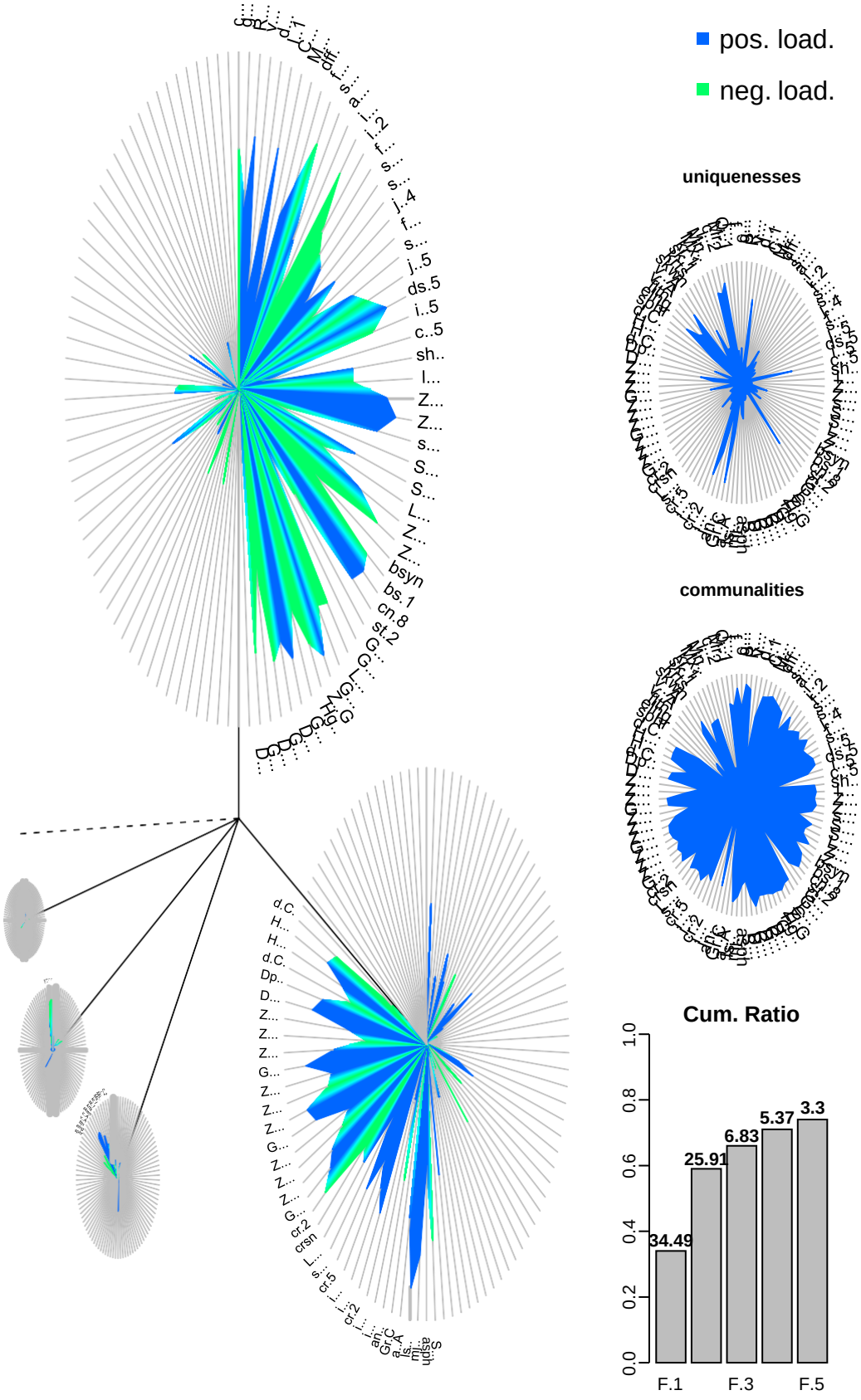

Figure S2: Distribution of the radiomics features after Factor Analysis with an outcome of five factors.

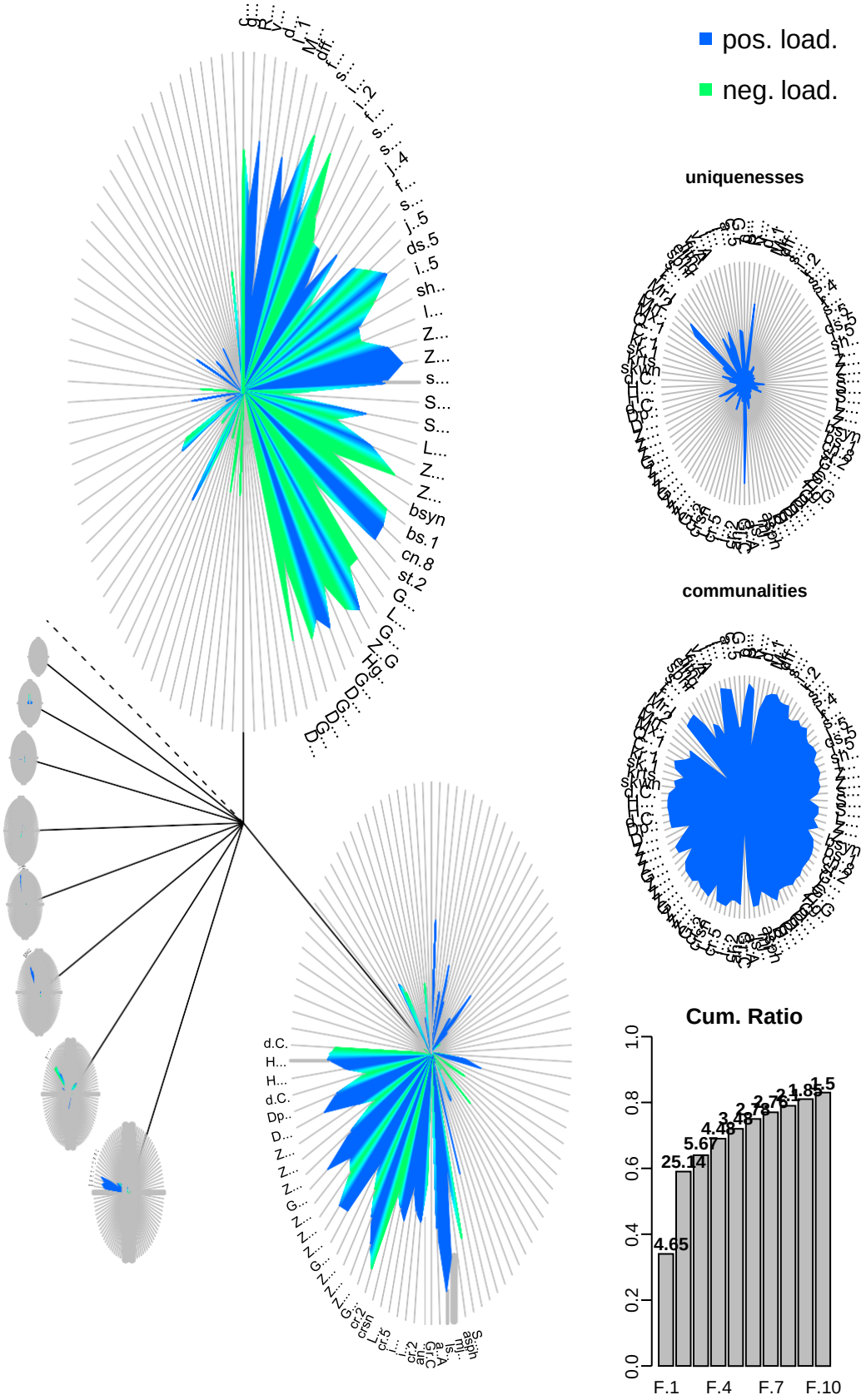

Figure S3: Distribution of the radiomics features after Factor Analysis with an outcome of 10 factors.

Supplementary code:

```
################################################################################
#DATA PARTITIONING
################################################################################
library(dplyr)
library(stringr)
library(caTools)
library(caret)
library (MASS)
radiomics_ds <- read.csv(file=file. choose (), header=TRUE)
#In the dialog window select the file Radiomics_data_totaal_Hasselt_tweedeCT_reductie.csv
colnames(radiomics_ds)

## 75% of the sample size
smp_size <- floor (0.75 * nrow(radiomics_ds))

## set the seed to make the partition reproducible
set. seed(123)
train_ind <- sample(seq_len(nrow(radiomics_ds)), size = smp_size)

trainSet <- radiomics_ds [train_ind, ]
testSet <- radiomics_ds [-train_ind, ]

################################################################################
#LOGISTISCHE REGRESSIE: MANUEEL GEREDUCEERDE DATA
################################################################################

# Fit the model with intercept only
model_null <- glm (Events ~1, data = trainSet, family = binomial)
summary(model_null)
ds_vars <- trainSet [,-c(1, 57)]
add1(model_null, scope=ds_vars, test="Chisq")

model_one <- glm (Events ~ first.measure.of.information.correlation.3, data = trainSet, family =
binomial)
summary(model_one)
add1(model_one, scope=ds_vars, test="Chisq")

model_two <- glm (Events ~ first.measure.of.information.correlation.3+busyness , data = trainSet,
family = binomial)
summary(model_two)
add1(model_two, scope=ds_vars, test="Chisq")

model_three <- glm (Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10 , data = trainSet,
family = binomial)
summary(model_three)
add1(model_three, scope=ds_vars, test="Chisq")

model_four <- glm (Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.l
evel.emphasis.2   , data = trainSet, family = binomial)
summary(model_four)
```

```
add1(model_four, scope=ds_vars, test="Chisq")

model_five <- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 + inverse.difference.normalised.2  , data = trainSet, family = binomial)
summary(model_five)
add1(model_five, scope=ds_vars, test="Chisq")

model_six <- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 + inverse.difference.normalised.2+Quartile.coefficient , data = trainSet, family = binomial)
summary(model_six)
add1(model_six, scope=ds_vars, test="Chisq")

model_seven <- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 +
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2, data = trainSet, family = binomial)
summary(model_seven)
add1(model_seven, scope=ds_vars, test="Chisq")

model_eight <- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 +
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurtosis , data = trainSet, family = binomial)
summary(model_eight)
add1(model_eight, scope=ds_vars, test="Chisq")

#model_nine <- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 +
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurtosis +Grey.level.non.uniformity.GLSZM, data = trainSet, family = binomial)
#summary(model_nine)
#add1(model_nine, scope=ds_vars, test="Chisq")

#model_ten<- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 + inverse.difference.normalised.2+Quartile.coefficient+Quartile.coefficient+kurtosis
+Grey.level.non.uniformity.GLSZM+inverse.difference.normalised.2, data = trainSet, family = binomial)
#summary(model_ten)
#add1(model_ten, scope=ds_vars, test="Chisq")


###Remove variable 1 by 1 starting with less significant
model_full<- glm(Events ~
first.measure.of.information.correlation.3+busyness+volume.at.int.fraction.10+Large.zone.low.grey.level.emphasis.2 +
```

```
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurt
osis, data = trainSet, family = binomial)
summary(model_full)

#min volume.at.int.fraction.10
model_min1 <- glm(Events ~
first.measure.of.information.correlation.3+busyness+Large.zone.low.grey.level.emphasis.2 +
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurt
osis, data = trainSet, family = binomial)
summary(model_min1)

#min busyness
model_min2 <- glm(Events ~
first.measure.of.information.correlation.3+Large.zone.low.grey.level.emphasis.2 +
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurt
osis, data = trainSet, family = binomial)
summary(model_min2)

#min Large.zone.low.grey.level.emphasis.2
model_min3 <- glm(Events ~ first.measure.of.information.correlation.3+
inverse.difference.normalised.2+Quartile.coefficient+first.measure.of.information.correlation.2+kurt
osis, data = trainSet, family = binomial)
summary(model_min3)

#min first.measure.of.information.correlation.2
model_min4 <- glm(Events ~ first.measure.of.information.correlation.3+
inverse.difference.normalised.2+Quartile.coefficient+kurtosis, data = trainSet, family = binomial)
summary(model_min4)

#min first.measure.of.information.correlation.3
model_min5 <- glm(Events ~ inverse.difference.normalised.2+Quartile.coefficient+kurtosis, data =
trainSet, family = binomial)
summary(model_min5)

#min kurtosis
model_min6 <- glm(Events ~ inverse.difference.normalised.2+Quartile.coefficient, data = trainSet,
family = binomial)
summary(model_min6)

#min volume.at.int.fraction.10
#model_min7 <- glm(Events ~
volume.at.int.fraction.90+Grey.level.non.uniformity.GLSZM.1+Quartile.coefficient, data = trainSet,
family = binomial)
#summary(model_min7)



#Stop with model 6
model_fin <- glm(Events ~ inverse.difference.normalised.2+Quartile.coefficient, data = trainSet,
family = binomial)
summary(model_fin)
```

```
##############################################################################
#MODEL TESTEN
##############################################################################

probabs <- predict(model_fin, testSet, type='response')
preds <- ifelse(probabs >= 0.5, 1, 0)

confusionMatrix(factor(preds), factor(testSet$Events))
table(testSet$Events,preds)

##############################################################################
#LOGISTISCHE REGRESSIE: FACTOR ANALYSE
##############################################################################

factors_ds <- read.csv(file=file.choose(), header=TRUE)

model_null <- glm(Events ~ 1, data = factors_ds, family = binomial)
summary(model_null)
ds_vars <- factors_ds[,3:7]
add1(model_null, scope=ds_vars, test="Chisq")

model_one <- glm(Events ~ Factor4, data = factors_ds, family = binomial)
summary(model_one)
add1(model_one, scope=ds_vars, test="Chisq")

model_two <- glm(Events ~ Factor3+ Factor4, data = factors_ds, family = binomial)
summary(model_two)
add1(model_two, scope=ds_vars, test="Chisq")
```