Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The differentiating powers of radiomics: analyses of healthy and tumor tissue in 18F-FDG PET images of NSCLC

Daan Baert
Kobe Verlinden

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

PROMOTOR :
Prof. dr. Brigitte RENIERS

PROMOTOR :
Prof. dr. Liesbet MESOTTEN

COPROMOTOR :
Dr. Elien DERVEAUX

Gezamenlijke opleiding UHasselt en KU Leuven

Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt

2022
2023

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

The differentiating powers of radiomics: analyses of healthy and tumor tissue in 18F-FDG PET images of NSCLC

Daan Baert
Kobe Verlinden

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

PROMOTOR :
Prof. dr. Brigitte RENIERS

PROMOTOR :
Prof. dr. Liesbet MESOTTEN

COPROMOTOR :
Dr. Elien DERVEAUX

UHASSELT KU LEUVEN

Preface

This Master's thesis is the cornerstone of our industrial engineering degree. It resulted from a collaboration between the ProLung study at the hospital Ziekenhuis Oost-Limburg (ZOL) and the University of Hasselt. Over the course of a year, we got fully immersed in cancer research and more importantly became aware of the need for new technologies in this field of study. Cancer is one of the biggest health risks today and we feel proud to have been able to contribute to an already extensive amount of research. The importance of cancer research was also our major driving factor in successfully completing this thesis. Moreover, we experienced first hand the multidisciplinarity of being an engineer. Our background in nuclear technology helped us understand the imaging techniques, knowledge of programming in Python and Matlab helped us perform the statistical analyses and the medical expertise of our promoters helped us evaluate our results.

Firstly, we would like to address our deepest appreciation to our promoters Prof. Dr. Liesbet Mesotten from ZOL and Dr. Elien Derveaux from UHasselt who helped us overcome the many difficulties we encountered. Their optimism, ideas, feedback and knowledge were indispensable and we could always count on their help and encouragement. Next, we would like to extend our gratitude to Prof. Dr. Dirk Valkenborg who guided us through the statistical analyses performed in this research paper and provided us with input and ideas.

Additionally, we want to thank prof Brigitte Reniers who taught us the necessary knowledge in the medical nuclear field. We would also like to thank Prof. Dr. Ronald Boellaard for providing us with both the 'Accurate' and 'Radiomics' tools without which this research would not have been possible.

We would like to thank our families and friends for their unwavering support and encouragement. We would also like to thank our fellow students who were in the same boat, yet willing to provide us with moral support and practical suggestions. Last but not least we would like to offer our gratitude to the PXL campus in Diepenbeek for accommodating us during our research.

Table of contents

List of tables

List of figures

List of supplementary figures, codes and guideline

List of abbreviations

^{18}F-FDG	2-[18]F fluorodeoxyglucose
2Davg	2 dimensional averaged
2DDmrg	2 dimensional; directional merged
2Dmrg	2 dimensional merged
3Davg	3 dimensional averaged
3DWmrg	3 dimensional merged
AC	Adenocarcinoma
AI	Artificial intelligence
AP	Acute pancreatitis
BMI	Body mass index
CAD	Computer-aided diagnosis and detection
COPD	Chronic obstructive pulmonary disease
CT	Computed Tomography
df	Degrees of freedom
EANM	European Association of Nuclear Medicine
ε	An arbitrarily small positive number (≈ 2.2×10-16)
FDG	Fludeoxyglucose
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
H_0	Null hypothesis
IARC	The International Agency for Research on Cancer
KNN	K nearest neighbors
LOR	Lines of response
LSO	Lutetium oxyorthosilicate
MRI	Magnetic resonance imaging
NE	Neuroendocrine
NEC	Necrotizing enterocolitis
N_g	Number of discrete intensity levels in the image
NGTDM	Neighboring Gray Tone Difference Matrix

NMSC	National MS centre
NSCLC	Non-small cell lung carcinoma
μ_x	Mean gray level intensity of p_x
OPLS-DA	Orthogonal partial least squares-discriminant analysis
OPLS-EP	Orthogonal partial least squares-effect projections
p(i)	Normalized first order histogram
p(i,j)	Normalized co-occurrence matrix
PACS	Picture Archiving and Communications Systems
PC	Principal components
PCA	Principal component analysis
PET	Positron Emissie Tomografie
PET/CT	Positron emission tomography-computed tomography
PMT	Photomultiplier tube
RMS	Root mean square
ROI	Region of interest
RUS	Random Undersampling
SCC	Squamous cell carcinoma
SCLC	Small cell lung carcinoma
SUV	Standardized uptake value
SVD	Singular Value Decomposition
SVM	Support vector machine
TB	Tuberculosis
TCA	Tricarboxylic acid
UMC	Universitair medische centra
US	Ultrasound
VOI	Volume of interest
WHO	World Health Organization
X	A set of Np voxels included in the VOI
ZOL	Ziekenhuis Oost-Limburg

Abstract

Lung cancer has the second highest mortality rate among cancer phenotypes and is linked with an increasing incidence of 2.4 million by 2035, raising the need for more efficient diagnosis and accurate prognosis. Radiomics can be a useful tool to support both purposes. This master's thesis aims to uncover the possible distinguishing powers of radiomics data between healthy and lung cancer tissue.

A cohort of 49 patients was diagnosed with NSCLC by ^{18}F-FDG PET imaging for more accurate staging and underwent a lobectomy. The VOI of the lung lesion was first segmented in the ACCURATE tool (developed by Prof. Dr. Boellaard). Then, the previously segmented VOIs are exactly translated in the ACCURATE tool to the opposite lung including only healthy lung tissue. Next, radiomics features were extracted from both the tumor and the healthy VOI, providing 504 features. A paired t-test unveiled 69 irrelevant features in differentiating both tissue types. Next, a principal component analysis (PCA) was performed to establish clustering, remove noise and uncover relevant features. Lastly, discriminative models were created to distinguish healthy and tumor tissue.

The PCA indicates separate clustering of both tissue types, with the noise-reduced dataset of 269 features showing the best results. A set of 30 features still performed adequately. The Fine and Weighted K-Nearest Neighbors, and Bagged Trees and Subspace Discriminant ensemble learning models show an accuracy of 98.3% in predicting tissue type.

Abstract (in Dutch)

Longkanker heeft het tweede hoogste sterftecijfer onder kankers en een toenemende incidentie van 2,4 miljoen tegen 2035. Dit vergroot de behoefte aan efficiënte diagnose en nauwkeurige prognose. Radiomics kan een hulpmiddel zijn om beide doelen te ondersteunen. Deze masterproef heeft als doel om gezond en longkankerweefsel te kunnen onderscheiden aan de hand van radiomics-data.

Een cohort van 49 patiënten werd gediagnosticeerd met NSCLC door ^{18}F-FDG PET-beeldvorming voor een nauwkeurigere stadiëring, en onderging een lobectomie. De VOI van de longlaesie werd eerst gesegmenteerd in de ACCURATE tool. Vervolgens worden de eerder gesegmenteerde VOI's in deze tool exact vertaald naar de andere long, met enkel gezond longweefsel. Daarna werden radiomics-features geëxtraheerd uit beide VOIs, wat 504 features opleverde. Een gepaarde t-test vond 69 features die irrelevant zijn in het onderscheiden van de twee weefseltypes. Daarna werd een PCA uitgevoerd om clustering vast te stellen, ruis te verwijderen en relevante kenmerken te onthullen. Ten slotte zijn er discriminerende modellen gecreëerd om gezond en tumorweefsel te onderscheiden.

De PCA geeft gescheiden clustering van beide weefseltypen aan, waarbij de dataset met ruisonderdrukking van 269 features de beste resultaten laat zien. Een set van 30 features presteerde nog adequaat. De Fine- en Weighted KNN, en Bagged Trees en Subspace Discriminant ensemble modellen tonen een nauwkeurigheid van 98,3% bij het voorspellen van het weefseltype.

1 Introduction

Every year, over 2 million people worldwide receive a lung cancer diagnosis [1]. This number is estimated to continue to rise in the coming years [2]. The survival rate for lung cancer has been shown to correlate with the disease's stage [3]. Due to a general lack of symptoms in the early stages of lung cancer, a proper diagnosis often takes place when the disease has already reached later stages [4]. Only 17% of lung cancer cases are detected early and mostly by accident [5]. Because of this, lung cancer was responsible for an estimated 1.8 million deaths in 2020 alone [6]. Chapter two will introduce the principles behind cancer and its specific metabolism and will take a closer look at lung cancer and non-small cell lung carcinoma (NSCLC) in particular.

A topic gaining popularity in cancer research is radiomics. It encompasses the extraction of features (higher dimensional data) from images. The image can be the volume of interest (VOI) of a tumor itself. These features contain quantitative information based on the intensity, shape, size or volume, and texture of tumor phenotype and its microenvironment. It is an extension of computer-aided diagnosis and detection (CAD) systems [7-8]. Radiomics and its applications in cancer research are elaborated on in Chapter four.

The ProLung study is a research project at Ziekenhuis Oost-Limburg (ZOL) located in Genk and is funded by ‘Kom op tegen Kanker’. This study focuses on patients diagnosed with NSCLC who underwent a lobectomy as part of their standard-of-care treatment plan. The tumor staging in these patients ranges from stage I-IIIA.

This Master’s thesis focuses on patients with NSCLC, which comprises 80% of lung cancer cases, and the resulting radiomic data retrieved from the tumor volume of interest (VOI). Radiomics in lung cancer has shown promising results concerning diagnosis and prognosis [9]. Following the extensive research done in this area, this study aims to distinguish healthy lung tissue from NSCLC tissue solely using radiomics features extracted from the segmented VOI. Furthermore, this Master’s thesis aims to find a statistical discriminative model for the aforementioned tissues. The purpose of both objectives is to establish whether radiomics can form a basis for assisting in and/or automation of diagnosis and prognosis.

This study’s methodology starts at Chapter six. A patient cohort of 49 patients is used for the purposes of this study. All patients comply with the aforestated conditions, namely stage I-IIIA NSCLC and a lobectomy as part of their standard-of-care treatment plan. ^{18}F-FDG PET/CT images were gathered using the Biograph Horizon camera (Siemens Healthineers). Tumor segmentation was performed semi-automatically on the PET images using the Accurate tool (developed by the research team of Prof. Dr. Boellaard, Amsterdam UMC). The CT images captured concurrently are used to correct the segmented VOIs for breathing artifacts. A total of 504 features were then extracted using the Radiomics tool (developed by the research team of Prof. Dr. Boellaard, Amsterdam UMC). These can be subdivided into 498 radiomics features and 6 PET Uptake Metrics per patient. Afterward, the same procedure

was used to create a second dataset for healthy tissue VOIs resulting in the same number of extracted features.

Firstly, a paired t-test was performed on both datasets to eliminate features that showed no significance in discriminating between healthy and tumor tissue VOIs. This resulted in 435 remaining features.

Secondly, the new dataset of 435 features is visualized using a dendrogram, which is a tree-based representation of the hierarchical clustering of data. This method of visualization shows relationships between objects of a dataset and can be useful in revealing clustering of those objects. In this case, the objects are the 435 features.

Thirdly, the healthy and tumor VOI datasets with the reduced number of features underwent a principal component analysis (PCA). This analysis is used to visualize high-dimensional datasets by reducing them to principal components (PCs) to form a two-dimensional plot. This could indicate clustering of the data and reveal relevant radiomics features.

Finally, the Matlab classification learner tool is used to find models able to predict certain classifications of the patients, such as tumor or healthy tissue, diabetes, packyears,... Therefore, the patient cohort is split into two groups of 30 and 19 patients each. These groups are used for training the model and using that model to make predictions respectively.

The results of these analyses are laid out in Chapter eight, as well as demographic data collected for the patient cohort. The results consist of the eliminated features based on the paired t-test, a dendrogram visualizing the features, a PCA to see clustering, and uncover the most relevant features and models used to train and test based on distinguishing classifiers.

Chapter nine elaborates on the features showing the most significance in differentiating healthy tissue from NSCLC tissue. This elaboration consists of a general description combined with a mathematical definition for each feature.

The discussion of the used methodology and the resulting findings can be found in Chapter ten. Possible pitfalls and future extensions of this study are also part of this chapter.

Chapter eleven comprises the conclusion of this research concerning the two objectives set forth in this chapter. These objectives are distinguishing healthy lung tissue from NSCLC tissue solely using radiomics features extracted from the segmented VOI and finding a statistical discriminative model for the aforementioned tissues.

2 Cancer

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths. The most common types are breast, lung, colon and rectum, and prostate cancers [6]. Whenever a normal cell's basic traits and behaviors become distorted and a loss of normal biological controls occurs, the cell becomes cancerous. Generally, these cancer cells have three fundamental phases. The first phase of these cells is uncontrolled division. The mechanisms which control cell division fail and the population of malfunctioning cells expands rapidly. Secondly, due to this rapid increase in the population of cancerous cells, the surrounding tissue can be invaded and destroyed. The third and final phase is the colonization of distant body sites, also called metastasis [10]. After the third phase, the process can start over in the current sites. A distinction can be made between two types of cancer, namely benign and malignant cancer. A benign tumor only adopts the first of the aforementioned traits while a malignant tumor invades other nearby and/or distant tissues [11]. A malignant tumor is therefore more life-threatening and poses another problem. What is the primary tumor site? A nodule found in the lungs does not automatically mean the patient started with lung cancer. This can make treatment more difficult and underlines the importance of proper screening and diagnosis [12].

2.1 Biology of cancer

“The Hallmarks of cancer”, as originally defined by Hanahan and Weinberg, are characteristics that describe the transformation of normal cells to cancerous cells. In time, more hallmarks have been discovered and now ten hallmarks are widely accepted. These are visualized in Figure 1 [13].

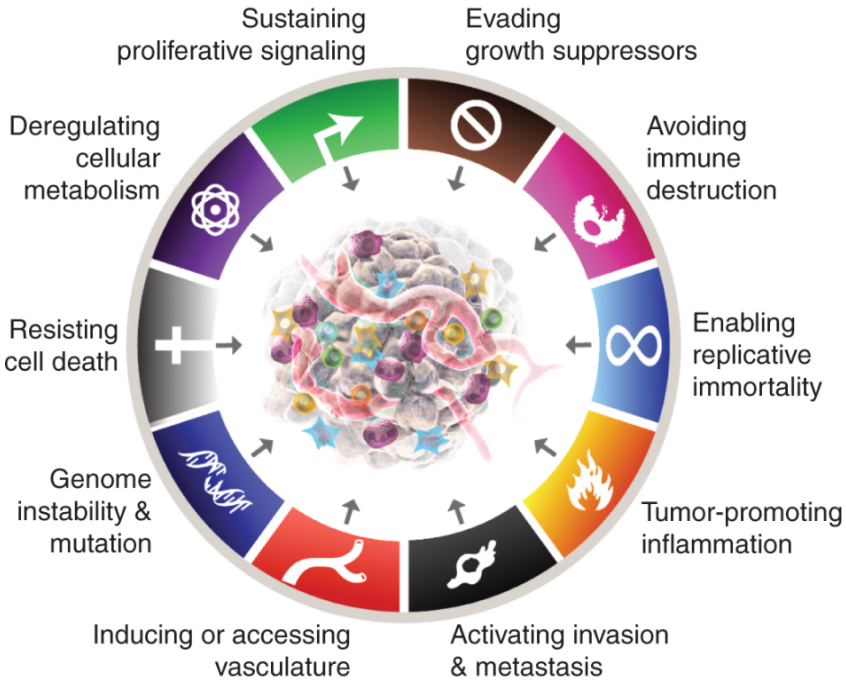


Figure 1: The Ten Hallmarks of Cancer [13].

When normal cells undergo some or most of these hallmarks and become cancerous, their population increases dramatically. This explosion in population means an increased energy demand to sustain the continued cell growth and division. New metabolic pathways are created to accommodate this increased need for energy. Warburg demonstrated that tumor cells exhibit a high rate of glucose metabolism compared to normal cells. Two more properties, lactate secretion, and oxygen availability combined with the glucose uptake make up the Warburg effect [14-15]. The mechanism has been visualized in Figure 2. This Warburg effect can clinically be exploited by ^{18}F-FDG PET imaging as will be discussed in a later paragraph.

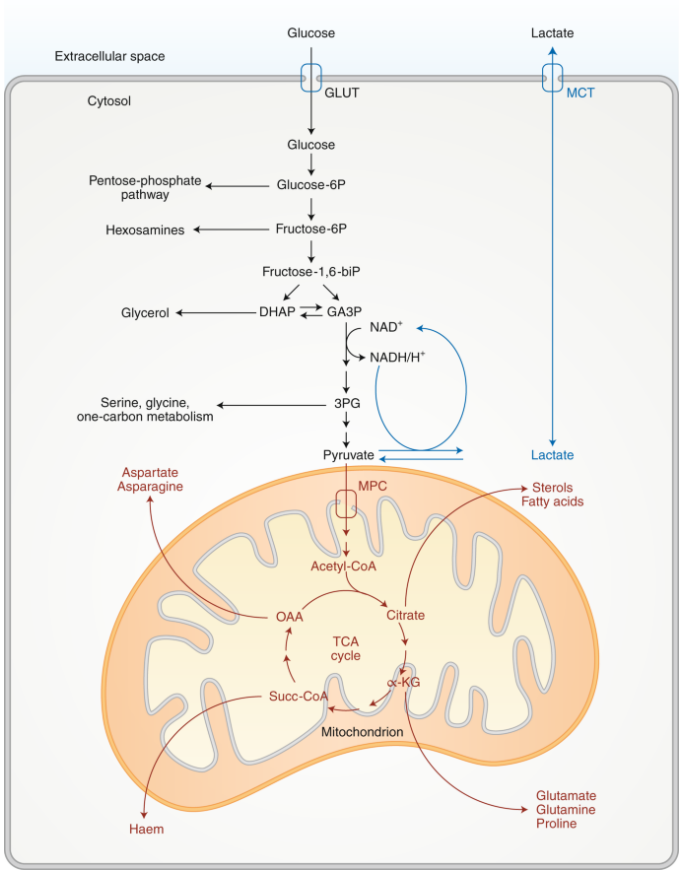


Figure 2: Warburg effect, glycolysis and TCA-cycle metabolism pathways [15].

2.2 Lung Cancer

Lung cancer has the second highest incidence of all types of cancer. The International Agency for Research on Cancer (IARC), a subdivision of the World Health Organization (WHO), estimates an incidence of 2.206.771 new cases of lung cancer in 2020 alone, making it the second most prevalent type of cancer globally behind breast cancer, as can be seen in Figure 3 [1].

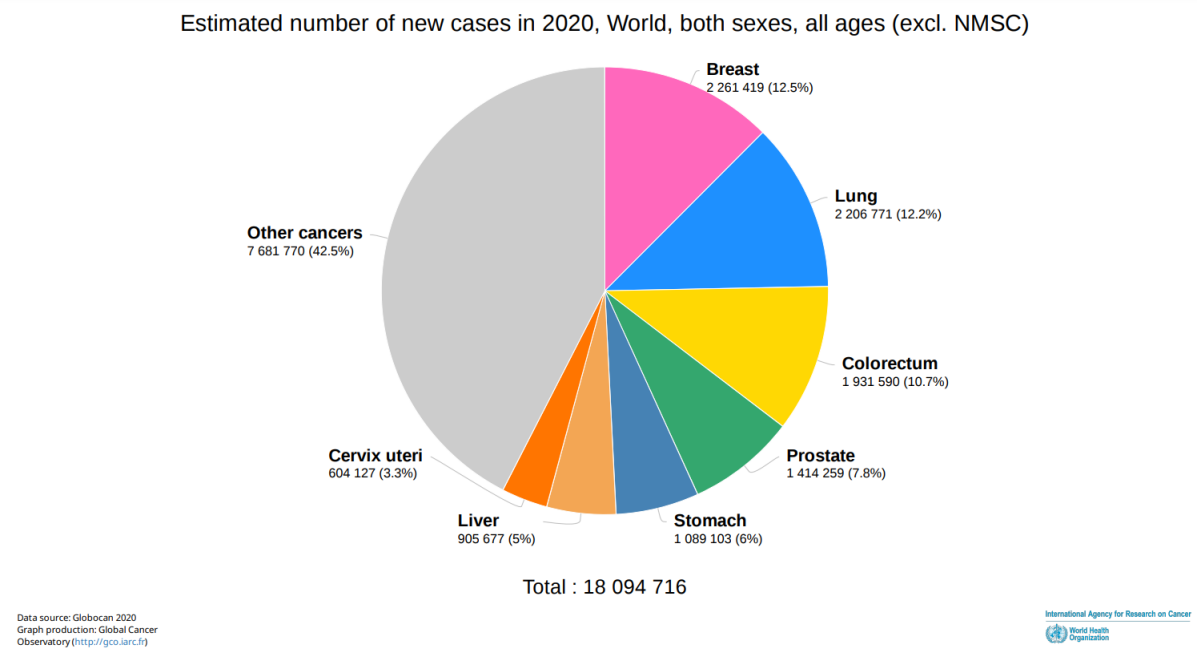


Figure 3: Pie chart of estimated new cancer cases in 2020, World, both sexes, all ages (excl. NMSC) [1].

Research has shown a clear link between the survival rate for lung cancer and the stage of the disease at the time of diagnosis [3]. Due to a general lack of symptoms in the early stages of lung cancer, a proper diagnosis often takes place when the disease has already reached later stages [4]. Only 17% of lung cancer cases are detected early and mostly by accident [5]. Because of this, lung cancer is responsible for an estimated 1.8 million deaths in 2020 alone [6].

It has long been recognized that smoking is the leading cause of lung cancer [16-17]. This is also highlighted in a study by Alberg et al. which comprised Table 1 with known causes for lung cancer [18].

Table 1: Key factors associated with risk of lung cancer [18].

Factor	Description
A. Single most important causal determinant of individual population risk, most valuable indicator of clinical risk[1]	Active smoking of cigarettes and other tobacco products: - Individual risk increases with greater number of cigarettes smoked per day and greater number of years of smoking. Population risk increases with the prevalence of current smokers because population prevalence predicts lung cancer occurrence with a latency period of about 20 years.
B. Other risk factors causally associated with lung cancer[2]	- Secondhand smoke exposure - Ionizing radiation, including radon - Occupational exposures, e.g. arsenic, chromium, nickel, asbestos, tar and soot - Indoor and outdoor air pollution
C. Additional clinical risk factors[3]	The risk factors noted above, plus: - Older age - Male sex, particularly among those of African American ancestry - Family history of lung cancer - Acquired lung disease, e.g. COPD, TB, pneumoconiosis, idiopathic pulmonary fibrosis and systemic sclerosis - Occupational exposures, such as to silica dust - HIV infection
D. Examples of association with consistent evidence but causal role not presently established	- Fruit and vegetable intake (decreased risk) - Physical activity (decreased risk) - Marijuana smoking (not associated with risk)

[1]COPD, chronic obstructive pulmonary disease; TB, tuberculosis.

[2]The evidence for factors listed in these categories is extremely strong to meet epidemiologic criteria for causality.

[3]The factors listed under clinical risk indicators are all strongly associated with increased risk of lung cancer but are listed in this category either because they are intrinsic characteristics of the patient (age, sex, ethnic ancestry and family history) or are factors with consistent evidence of increased risk that presently falls short of being rated as causal

A study by luo et al. shows that the incidence of lung cancer by 2035 is projected to increase dramatically in most countries. In general, the findings indicate an increase in new lung cancer patients among females, while the incidence for males declines. A change in smoking and working habits among genders is put forth as a driving force for this change. [19].

2.2.1 Non-small cell lung carcinoma

Lung cancer is split into two categories: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). In 80% of the cases, it concerns NSCLC. There are five stages of cancer regarding NSCLC, starting with stage 0, categorized by the TNM system [20]. In stage 0, abnormal cells are detected only in the lungs. In stage I, the abnormal cells become cancer cells where the size of the tumor is a maximum of 3 cm. If the size of the tumor is between 4 and 5 cm, it is classified as stage II. In the next stage, stage III, tumor cells have reached some lymph nodes. Finally, in stage IV the tumor can be any size and has reached many lymph nodes and even other organs [21]. The link between the stage of lung cancer and the survival rate becomes apparent here. The 5-year survival for NSCLC remains below 25% [22]. A study from 2019 showed that in stage I, there is around 55% survival rate while the percentage in stage 4 is only around 5% [23]. This is also highlighted by data from IARC as seen in Figure 4.

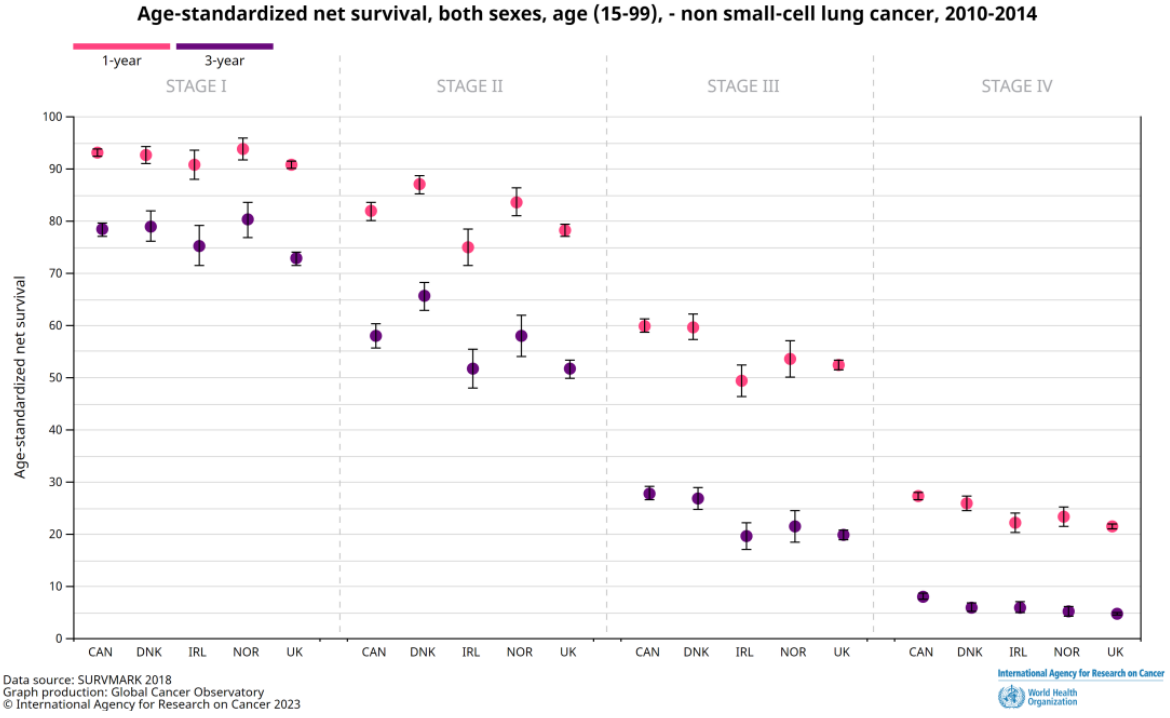


Figure 4: Survival by stage and time since diagnosis of NSCLC for both sexes, age 15-99 [24].

It is evident from these statistics that early diagnosis and better prognostic tools could aid in formulating a treatment plan- and increase a patient's chances of survival.

3 Medical imaging for lung cancer

Due to being less invasive than a biopsy, medical imaging has been the main methodology for screening, diagnosing, and staging lung cancer. Screening is often done by taking a low-dose computed tomography (CT) image [25]. For staging, the merger of ^{18}F-FDG PET and CT is shown to improve its diagnostic accuracy in NSCLC [26]. The use of ^{18}F-FDG PET and CT merged images has shown prognostic capabilities over other modalities. More specifically, SUV is shown to be a predictor of overall survival [27]. This study will examine PET imaging and the potential of data-mining of these images. Before this can be addressed, the principles of positron-emitting tomography will be discussed in this paragraph.

3.1 Computed tomography (CT)

Computed tomography (CT) is a form of medical imaging created by Godfrey Hounsfield in 1972. The principle of this form of imaging is shooting X-rays, photons, through a patient. The X-rays interact with the different tissues it goes through and undergo a process called attenuation, the gradual loss of intensity [28]. This attenuation μ is density-based and thus tissue specific. On the opposite side of the body, the resulting X-rays are collected in detectors. With a known flux of X-rays entering the body and the remaining rays being detected at the other end, a density map can be created of the body. More precisely, the source of the X-rays (X-ray tube) and the detectors are mounted on a ring opposite to each other as seen in Figure 5. This ring rotates creating a slice. Translating the patient through the gantry creates multiple slices which make up the final 3D image [29]. For lung cancer, a CT image is taken of the thorax and since this modality can capture lung, soft tissue, and bone detail simultaneously, it is often the preferred way of imaging [30].

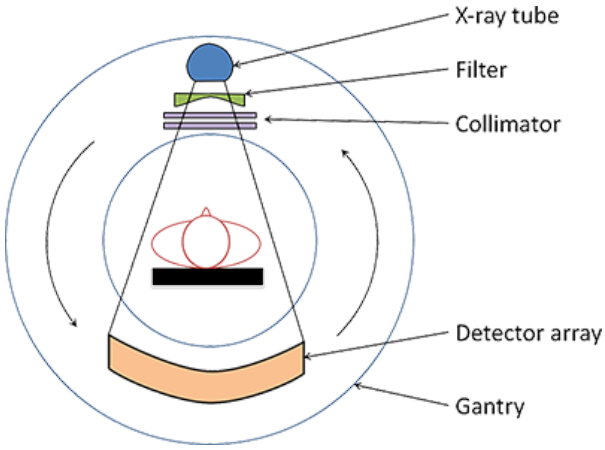


Figure 5: Layout of a CT gantry [31].

3.1.1 The X-ray tube

The source of the X-rays for a CT is an X-ray tube. This device consists of four main components, namely: the tube, the high-voltage generator, the control console, and the cooling system as represented in Figure 6. The tube, containing the anode and cathode, is under vacuum so the electrons are not obstructed while accelerating.

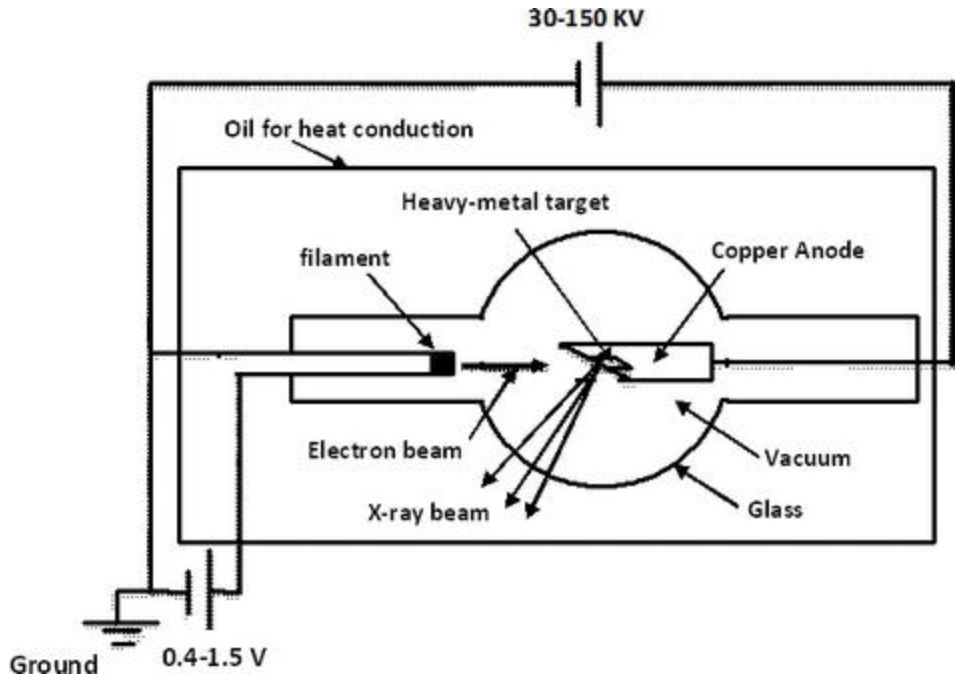


Figure 6: Components of the X-ray tube [29].

The production of X-rays starts by heating up the filament (cathode) and in turn releasing free electrons. When these electrons are released, they feel the voltage in the tube and are accelerated toward the anode (often made from tungsten). The voltage used to accelerate the electrons is between 30 and 150 *kV* [32]. The higher the voltage, the more energy the resulting rays will have. The higher the current (in *mA* range), the higher the flux of X-rays. Finally, the electrons reach the anode and can travel close to the tungsten nuclei, feeling their attractive positive force. They deflect, losing a large portion of their energy in the form of heat and radiation. The generation of heat is the reason an X-ray tube needs proper cooling. The radiation that is produced when charged particles undergo acceleration is called *Brehmsstrahlung* and these are the useful X-rays [29]. Another effect can occur when the electrons approach the tungsten atoms. They can interact with the electron shell, exciting the atom. The return to its ground state, the atom can release an X-ray photon. These are characteristic X-rays and are mono-energetic [33]. The final spectrum of the produced X-rays, both characteristic as from *Bremsstrahlung* is visualized in Figure 7.

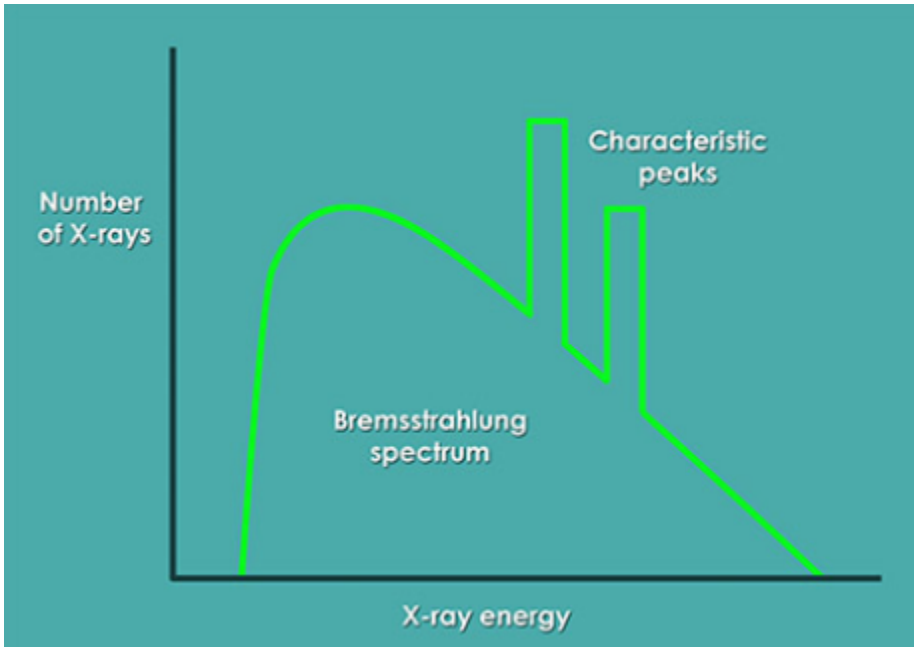


Figure 7: X-ray spectrum produced in an X-ray tube [33].

3.2 Positron-emitting tomography (PET)

PET is a form of nuclear medicine. Whereas CT and magnetic resonance imaging (MRI) give a physician morphological information, a PET scan gives information on metabolic activity. It can be defined as a combination of nuclear medicine and biochemical analysis [34]. The main principle behind this way of imaging is the detection of radiation originating from an intravenously injected radiopharmaceutical. The radiation has to be positron emission also known as β^+-decay. This form of radioactive decay consists of an unstable nucleus converting a proton into a neutron, releasing a positron in the process [35].

$$ {}^{1}_{1}\rho \rightarrow +{}^{0}_{+1}\beta + {}^{1}_{0}n $$

This positron is a positively charged electron and can be seen as the electron's antiparticle. Therefore, when an electron and a positron meet, they undergo annihilation. The particles disappear resulting in two γ-rays flying away at about 180° from each other. Figure 8 visualizes this process.

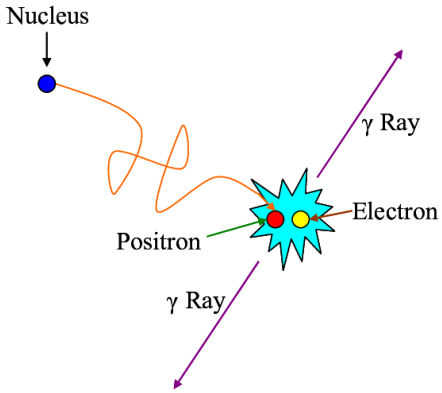


Figure 8: The process of positron-electron annihilation [36].

For PET imaging, this annihilation is the key. The formed photons flying away at 180° form a line of reference. A PET scanner is a ring of detectors made to detect the coincidences of two photons and reconstruct the lines of response (LOR). This way, the origin of the annihilation can be found [37]. The detection is based on a ring of photon counters and scintillators. The scintillator comprises a scintillation crystal, a photomultiplier tube (PMT), and amplifiers as seen in Figure 9.

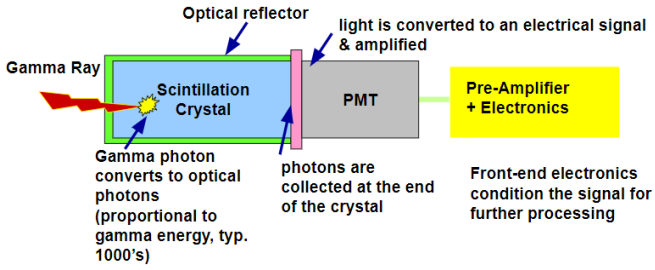


Figure 9: Principles of a scintillation detector for gamma rays [38].

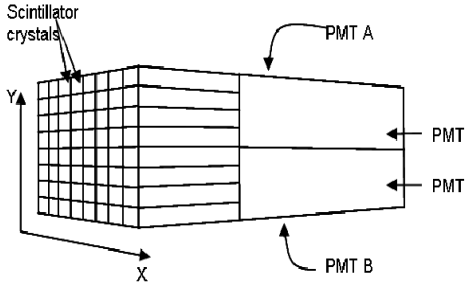


Figure 10: Block detector setup [39]

The main component is the scintillation crystal. Here, the spectrum of the radiation that enters is absorbed and re-emitted in the visible spectrum. This re-emitted light is proportional to the energy deposited by the radiation. The PMT creates a signal by producing photoelectrons for each entering photon. These electrons are then multiplied to create the signal. In modern detectors, multiple pairs of a scintillator and a PMT are put together to form blocks as visualized in Figure 10. Finally, amplifiers are used to further strengthen the signal [38][40]. However, the detection is not as straightforward. Not all coincidences are true coincidences. Figure 11 represents the different events that can occur.

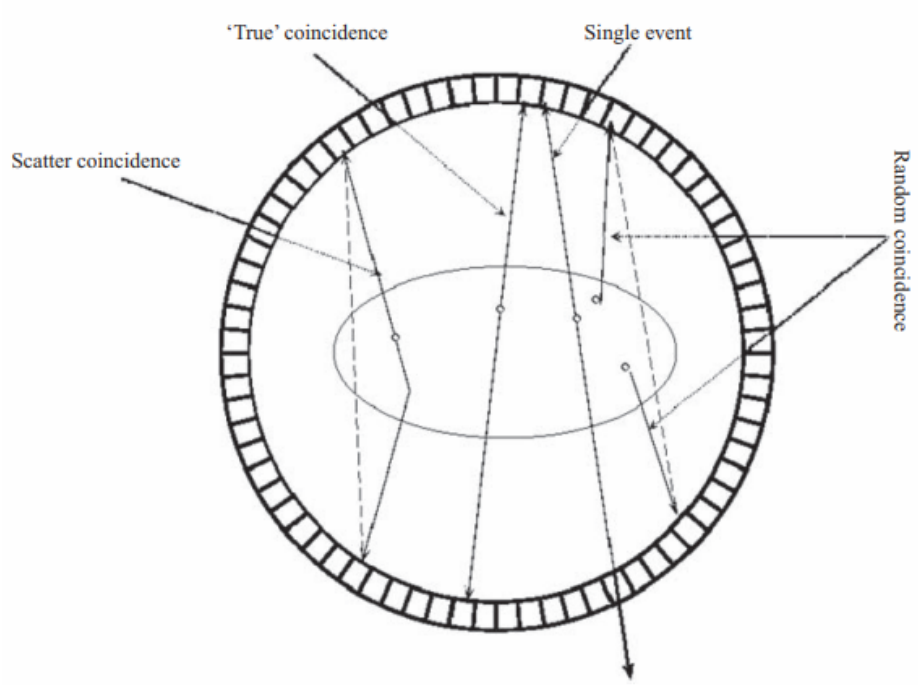


Figure 11: Illustration of the events detected in a PET scanner [41].

As shown in Figure 11, some photons can undergo Compton scattering with electrons resulting in a false LOR. A certain time interval is used to detect a coincidence. This interval can be large enough for two separate events to produce a false measurement, a random coincidence. Finally, only one of the two photons is detected resulting in a single event [42].

The PET image reconstruction process involves several steps, including [43]:

- Attenuation correction: PET gamma rays are attenuated (absorbed or scattered) as they pass through the body, which affects the accuracy of the reconstructed image. Attenuation correction is the process of estimating and correcting for this attenuation.

- Image reconstruction: The 2D projection data acquired by the PET scanner is mathematically processed in a sinogram to create a 3D image of the radiotracer distribution in the body. This is typically done using iterative algorithms that iteratively refine the estimate of the radiotracer distribution until it converges into a stable solution.

- Post-processing: The reconstructed PET image is often post-processed to enhance the quality of the image and to extract quantitative information about the radiotracer distribution. This may involve filtering, smoothing, and segmentation techniques.

Multiple positron-emitting isotopes exist, but not all are useful. The International Atomic Energy Agency (IAEA) has summarized some of the most used radioisotopes for PET imaging as seen in Table 2.

Table 2: Most frequently used radioisotopes for PET [41].

Radionuclide	Source	Half-life (min)	Maximum (and mean) positron energies (keV)	Mean positron range in water (mm)
C-11	Cyclotron	20.4	970 (390)	1.1
N-13	Cyclotron	9.96	1190 (490)	1.3
O-15	Cyclotron	2.07	1720 (740)	2.5
F-18	Cyclotron	110	635 (250)	0.5
Ga-68	Generator	68	1899 (836)	0.8
Rb-82	Generator	1.25	3356 (1532)	1.5

Two main prerequisites are put forth by the IAEA for an isotope to be acceptable for PET:

- Readily available or (relatively) easy to produce, in adequate quantities, and with the required purity;

- Suitable for the synthesis of radiopharmaceuticals that allow the study of biochemical processes in vivo [41].

The radioisotope is then bound to a specific molecule that has an affinity for accumulating in specific locations. As explained in a previous chapter, glucose is used by cancerous cells at higher rates than normal cells. This means the glucose concentration will be higher in and around tumors [44]. By labeling glucose with a radioisotope, these higher concentrations can be pinpointed using PET. A commonly used glucose analog is fludeoxyglucose (FDG). This molecule is mostly marked with an ^{18}F atom substituting a hydroxyl group to form the [^{18}F]-2-deoxy-2-fluoro-D-glucose. The ^{18}F is produced in a cyclotron facility [45-46]. To assess glucose metabolism, the standardized uptake value (SUV) is used. This is a semi-quantitative method based on the activity concentration of the injected radiopharmaceutical. It can be calculated by Formula 1.

$$SUV = \frac{Activity\ Concentration\ At\ Time\ Of\ PET\ [Bq/ml]}{Initial\ Activity\ Injected\ [Bq]/\ patient\ weight\ [ml]} \quad (1)$$

4 Radiomics

4.1 What is radiomics?

A topic gaining popularity in cancer research is radiomics. It encompasses the extraction of features (higher dimensional data) from images. The image can be the volume of interest (VOI) of a tumor itself. These features contain quantitative information based on the intensity, shape, size or volume, and texture of tumor phenotype and its microenvironment. It is an extension of computer-aided diagnosis and detection (CAD) systems [7-8]. Several studies have shown that the aforementioned features extracted from computer tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), or nuclear medicine imaging correlate with underlying tumor biology changes as summarized by Ardakani et al. [47]. Moreover, some features have indicated prognostic capabilities for lung and head-and-neck cancer patients [48]. This suspected prognostic ability will be discussed in this research paper. It is important to note that the extracted data can be cancer-specific [49]. Therefore this study focuses on radiomics research concerning lung cancer and more specifically non-small cell lung cancer (NSCLC).

4.1.1 Radiomics features

Radiomics data consists of quantitative results of image features. First and foremost, radiomics features are descriptive data. They summarize the characteristics and distribution of a set of data values and include include “minimum,” “maximum,” “range,” “percentile,” “mean,” “median,” “mode,” “mean deviation,” “standard deviation,” “variance,” “skewness,” and “kurtosis.” This type of data represents information that can be used as the basis for comparing how data series differ [50]. The extracted features can be grouped into multiple classes: first-order, shape, second-order or texture, and higher-order features.

First-order features, also referred to as intensity histogram features and statistical features, provide information about the distribution of voxel intensities within the segmented region of the image. Features found in this class are the ones listed previously with the addition of energy (magnitude of voxel values).

Shape features, subdivided into 3-dimensional and 2-dimensional shape features, describe the shape and size of the VOI. These are also called morphology features. This class includes features such as volume, surface area, sphericity, diameter, elongation, and flatness.

Second-order or texture features supply information on the texture of the VOI. The quantification of the texture is based on the arrangement of voxel gay-level intensities in the VOI. These features are further subdivided in five groups. **Gray Level Co-occurrence**

Matrix (GLCM) features describe textural indices based on the arrangements of pairs of voxels. More specifically, the features describe how combinations of discretized gray levels of neighboring pixels of voxels are distributed along an image direction. **Gray Level Run Length Matrix (GLRLM) Features** represent the quantification of a length or number of consecutive pixels with the same gray level value, also known as a run length. **Gray Level Size Zone Matrix (GLSZM) Features** provide data on the gray level zones of the segmented area of the image. These gray level zones are areas of connected voxels sharing the same gray level intensity, thus indicating uniformity. **Neighboring Gray Tone Difference Matrix (NGTDM) Features** quantify the difference between a voxel gray level and the average gray level of its neighbors in all three dimensions within a given distance. Finally, **Gray Level Dependence Matrix (GLDM) Features** describe the gray level dependency of a number of connected voxels within a certain distance of the centermost voxel.

Higher-order features can be obtained after transformation or filtering of the original segmented region. These features, however, are not part of this study. [51-54]

This paper will use the nomenclature set forth in the Image Biomarker Standardisation Initiative (IBSI) [53].

4.2 Radiomics: the state of affairs

Radiomics is used in different ways for all different kinds of cancer, going from non-small cell lung cancer (NSCLC) to pancreas cancer. This research only looks at lung cancer tumors, but it can be helpful to look further than only lung cancer research and take a look at all kinds of previous radiomics research. The results of these previous studies, the implementation, and the difficulties will be discussed in this chapter.

4.2.1 Radiomics for lung cancer

Although radiomics is a relatively new method, the results found in multiple lung cancer studies are promising. Most research about lung cancer radiomics uses (^{18}F)FDG-PET/CT scans whereby a positron-emitting radionuclide (^{18}F) is bound to a glucose molecule, fludeoxyglucose (FDG), and injected into a patient. Cancer cells use sugar at higher rates than normal cells which increases the concentration of the ^{18}F in those regions. The positron emitting tomography (PET) camera detects the radiation born from the ^{18}F-FDG and produces an image detailing the areas with higher glucose uptake. When combining these images with computed tomography (CT) images giving anatomical information, possible cancerous cells can be pinpointed in the body [55].

In the radiomics lung cancer studies, the data is extracted by a professional who marks the region of interest. In general, the steps in radiomics research are segmentation, quantification,

extracting the different features, and performing statistical analyses on the radiomics datasets. There are different methods to perform these steps, and the data from the extracted features highly depends on them. There is still no standardized approach for radiomic research which is still an issue, but the results are very useful for predicting prognosis and response to different therapies [9]. What is essential for the standardization of radiomics studies is quality and reproducibility. Studies about data-driven gating and free-breathing PET-CT acquisitions show that the radiomic features derived from pulmonary lesions located inferior to the superior lobes, and pulmonary lesions of a smaller size, have a more significant variability [56]. Reconstruction and delineation are also two important factors when it comes to radiomic research. Studies show that many features have similar standardized uptake values (SUV), which is a simple way of determining activity in PET imaging and is used to measure the response of cancers to treatment. In general, the performance of radiomic studies depends more on the delineation method than on the applied reconstruction algorithm [57-58]. Finally, there are different types of segmentation. Research shows that reproductivity and reliability are better for semi-automatic segmented volumes than for manually segmented ones. It is suggested that there is a development needed for fully automatic segmentation tools. This will minimize the impact of contouring uncertainties as well as increase the repeatability and reproducibility of studies concerning radiomics [59].

4.2.2 Radiomics applications in other types of cancer

Radiomics is not only used for lung cancer research. Head-and-neck cancer radiomic studies show that radiomics can be used to identify patients with a high risk of local tumor recurrence in an early stage [60]. Radiomics is also used for liver fibrosis, which is a disease that results in liver failure, cirrhosis, and portal hypertension. Applying radiomics has shown promising results in staging liver fibrosis and characterizing hepatocellular carcinoma, but there is still no agreement on how to use these properly for specific applications [61-62].

The use of radiomics in breast cancer research is a relatively new topic. It is used to improve the diagnosis and characterization. Where lung cancer radiomic research uses mostly PET-CT, breast cancer research mostly involves magnetic resonance imaging (MRI). Since these studies are in an early stage, research on high-quality prospective and reproductivity is still needed. At the time of writing, there are still quality limitations [63].

At last, radiomics can also help in the research of prostate tumors, but previous studies have shown that the features vary greatly in their repeatability [64].

4.2.3 Prognosis

Other than retrieving detailed information about a tumor and its VOI, studies have been focussing on producing prognostic models for a wide range of cancer phenotypes. For example, radiomics can be used to predict the recurrence of acute pancreatitis (AP). Here the radiomics features are used as a biomarker and in combination with the radiomics model it could help predict the recurrence of AP as a quantitative analysis method [65].

Going back to the lungs, radiomic features capturing detailed information about the tumor phenotype can be used as a prognostic biomarker for distant metastasis in lung adenocarcinoma [66]. This type of cancer is found on the outside of the lungs compared to NSCLC which is typically an internal form of lung cancer. For this same type of cancer, another study demonstrated the link between imaging characteristics and patient survival [67].

For lung cancer, in general, Chen et al. were able to predict a prognosis in 80.7% of the cases using radiomics and high throughput data extraction [68]. More proof of the prognostic capabilities comes from Jiang et al. who put together a prognostic model by combining metabolomic and radiomic features of primary gastrointestinal diffuse large B cell lymphoma [69]. It is important to note that studies have indicated that the software platform version can affect feature reliability in CERR and LIFEx. Features identified as having a significant relationship to survival varied between these platforms [70].

4.3 Workflow of a radiomics study

The workflow of a radiomics study consists of several steps and is visualized in Figure 12. In general, these steps are the acquisition of the images, the segmentation of the volumes of interest, the feature extraction, the selection of the features, and finally the data analyses. However, the standardization of these steps is still a debated issue because different studies use different methods for these different steps [9].

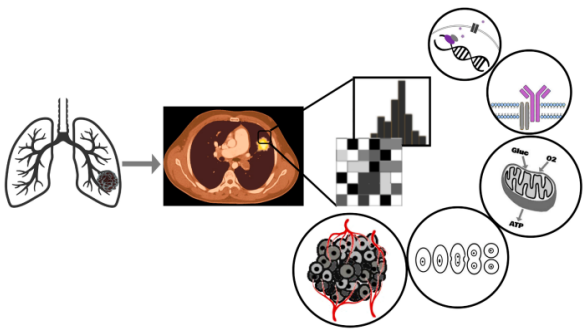


Figure 12: A simple scheme showing that ^{18}F-FDG PET/CT imaging may mirror the genotype of tumors [9].

The first step in the radiomics workflow is image acquisition. These are medical images, such as MRI, CT, and PET, of study patients. Image acquisition is an important step because the extracted features depend on the quality of the images. Therefore, to produce generalizable data, the process of image acquisition should be standardized [9]. In the case of NSCLC, PET/CT images are used to visualize the tumor. The images of this study are anonymized so that the identity of the different patients is unknown.

After the imaging is done, the next step is the segmentation of the two-dimensional region of interest (ROI) or the three-dimensional volume of interest (VOI). In the case of NSCLC, there has to be a segmentation of the tumor tissue. The segmentations have to be done by a professional, for example, a doctor who's specialized in lung cancer. There are also other techniques for segmentation, such as semi-automatically and fully automatic. In semi-automatic segmentation, standardized algorithms are used, while fully-automatic segmentation uses deep learning algorithms [71]. In this study, the image segmentation is done by a semi-automatically method with an SUV threshold of 0%. The VOI, created on the basis of this threshold is then checked and adjusted manually by a medical professional. After the segmentation, the radiomics features, which are discussed in *3.1.1*, can be extracted.

In general, feature extraction is the calculation and quantification of the characteristics of the gray levels within the segmented regions. It can be done by different programs, such as LifeX, Moddicom, and Pyradiomics.

5 Statistical analyses

5.1 Student t-test

When comparing two population groups and whether they express a significant or non-significant difference, a Student t-test can be used. A requirement for this statistical tool is that both datasets approximate a normal distribution. Generally, a t score is calculated and a t distribution is produced. The t score is defined by the ratio of the mean difference to the standard error as seen in Formula 2.

$$t\ score\ =\ \frac{Mean\ Difference}{Standard\ Error} \tag{2}$$

The t distribution is based on a probability distribution and parameterized by the degrees of freedom for the dataset. These degrees of freedom equal the total sample size minus two for a student t-test. Having more degrees of freedom correlates with the distribution increasingly favoring the mean. Based on this, a threshold value can be determined and the aforementioned t-score can be compared with this threshold [72].

A paired t-test is a form of t-test used on before-and-after observations on the same subjects [73]. For example, for comparing healthy and cancerous tissue in the same patient, a paired t-test could prove useful. Apart from a t-score, a paired t-test returns other significant results. First and foremost the null hypothesis H0 is tested. In the case of the healthy and cancerous tissue, H0 would be that both datasets are the same. When the statistical test is performed, an H value of 0 or 1 is returned, accepting or rejecting the null hypothesis respectively. A second result is the *p*-value. It can be described as a probability of how likely the similarity or difference between the two datasets is due to chance. P has a value ranging from 0 to 1. The closer to 0, the less likely the result is by chance and the more statistically significant the result is [74]. A third resulting parameter is the t score or t stat, described in this paragraph. Finally, the paired t-test returns the degrees of freedom (df) of the dataset. This is the number of variables that can be changed without breaking any restraints [75].

5.2 Machine learning

5.2.1 Classification trees

Tree classification is a fundamental task in the field of machine learning and pattern recognition, aimed at categorizing data samples into distinct classes or categories based on their features. It is an easy and fast way to interpret the data and to find a good fitting and prediction [76].

It is similar to a regression tree, but a classification tree is used to predict a qualitative response rather than a quantitative one. The 'most commonly occurring class' of each observation is predicted in the region to which it belongs. Not only the class prediction corresponding to a particular terminal node region is important, but also the class proportions among the training observations that fall into that region are interesting for interpreting the results [77].

5.2.2 Support vector machine

The support vector machine (SVM) is an objective of support that classifies the data points by finding a hyperplane in an N-dimensional space. There are many possible hyperplanes that could be chosen to separate two classes of the dataset. This is done by finding a plane that has a maximum distance between the data points of both classes as shown in Figure 13 [78].

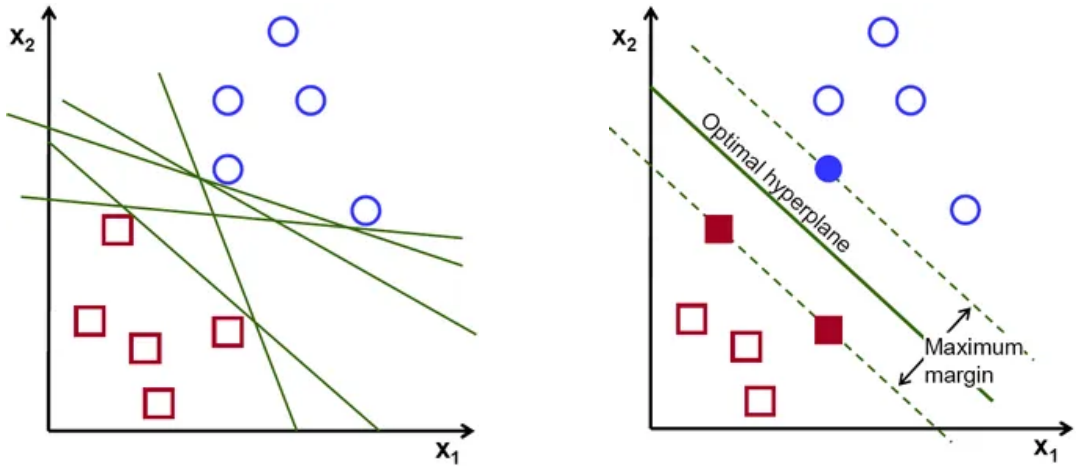


Figure 13: Possible hyperplanes [78].

By maximizing the margin distance between the data points, the better future data points can be classified. The data points that are closer to the hyperplane are the support vectors, and they influence the position and the orientation. The SVM that is used, uses these points to build the classifier [78].

5.2.3 K-nearest neighbors algorithm

The K-nearest neighbors algorithm is a supervised machine learning algorithm used for both classification and regression tasks. It operates by finding the k data points in the training set that are closest to a given test point, based on a chosen distance metric. KNN assigns the class label of the majority of the K-nearest patterns in the data space. The K stands for the neighborhood size and it defines the locality. Figure 14 shows the difference between a KNN for a small neighborhood (a) and a larger neighborhood (b). For a small K-value, the prediction is local, while a larger K-value ignores these small agglomerations of patterns [79].

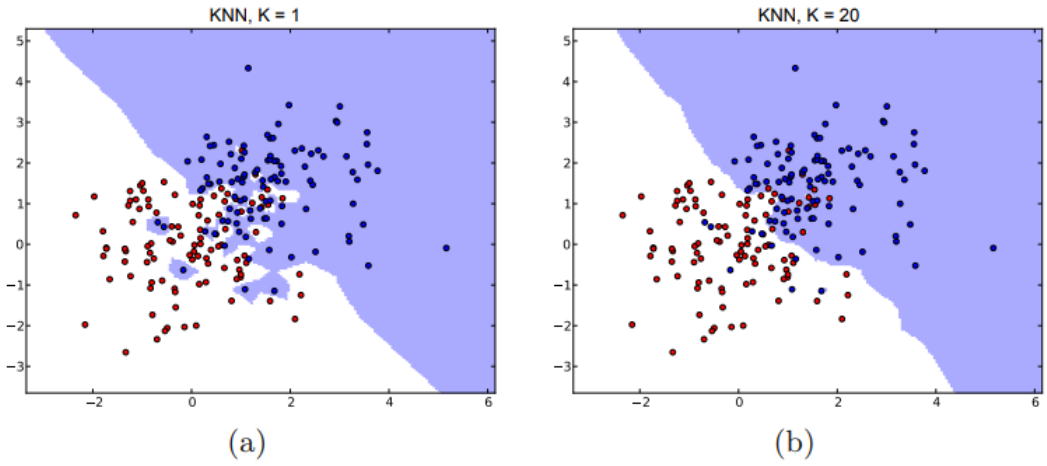


Figure 14: Visualization of the K-nearest neighbors algorithm for different K values [79].

5.2.4 Ensemble learning

Ensemble learning is a machine learning technique that combines the predictions from multiple models. There are an unlimited number of ensembles that can be developed, but there are three methods that are most commonly used:

- Bagging stands for bootstrap aggregating. For this method, many decision trees are fitted and the predictions are averaged for the final prediction.

- Stacking is a method that uses another model that learns how to combine the predictions of many different prediction models to make a final prediction

- Boosting is an iterative ensemble learning technique where base learners are trained sequentially, and each subsequent model focuses on correcting the mistakes made by the previous models.

Overall, ensemble learning is a powerful technique that can enhance the accuracy and reliability of machine learning models [80].

5.3 Cluster analysis

5.3.1 Dendrogram

A dendrogram is a tree-based representation of the hierarchical clustering of data. It visually displays the relationships between different objects or groups of objects based on their similarities. In a dendrogram, each object/group of objects is represented by a leaf/branch of the tree. The objects that are more similar to each other are clustered together and connected by branches. The height of the branches in the dendrogram represents the dissimilarity or distance between the clustered objects. The longer the branch, the greater the dissimilarity [77].

The visualization of a dendrogram is shown in Figure 15. It shows how the data can be clustered, where the dashed line indicates the cut-off height. The left side of Figure 15 shows the dendrogram with complete linkage and Euclidean distance. The center shows the dendrogram with a cut at a height of nine, which forms two clusters. The right side of Figure 15 shows the dendrogram with a cut at a height of 5, which forms three clusters [77].

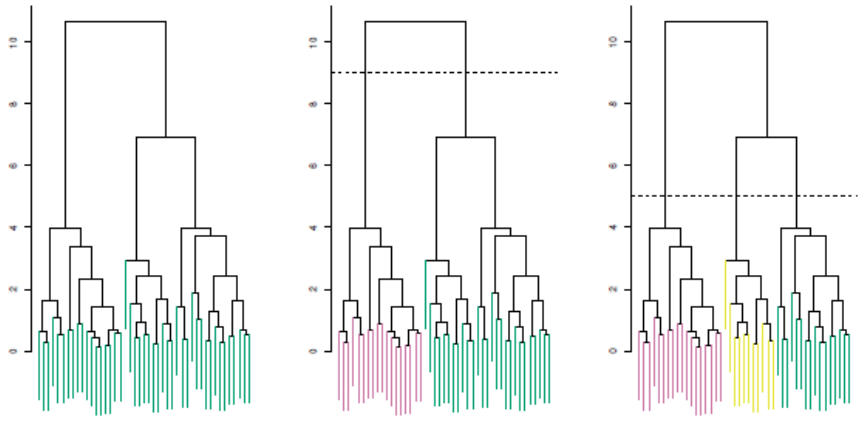


Figure 15: Hierarchical clustering with a dendrogram [77].

5.4 Principal Component Analysis (PCA)

The Singular Value Decomposition (SVD) provides a systematic way to determine a low-dimensional approximation to high-dimensional data in terms of dominant patterns. This technique is data-driven in that patterns are discovered purely from data, without the addition of expert knowledge or intuition. The SVD is numerically stable and provides a hierarchical representation of the data in terms of a new coordinate system defined by dominant correlations within the data. Principal component analysis (PCA) is one of the most important applications of SVD. PCA is based on reducing the dimensionality of a data set made up of a large number of interrelated variables. The important aspect of this method is that the variation present in the dataset is retained as much as possible To achieve this goal, the original variables are transformed into a new set of variables, the principal components. These principal components are vectors that are uncorrelated and ordered in such a way that the first few hold most of the variation originally present in the variables [81]. The first principal component is the direction along which the samples have the largest variation. The second principal component is the direction along which the samples show the largest variation while being uncorrelated to the first component [82].

PCA can also serve as a tool for data visualization. When handling datasets of multiple observations, each containing a large number of features or variables, it will produce a large number of scatterplots, each providing very little information about the dataset as a whole. PCA can be used to reduce this to a lower dimensional plot for example while retaining most of the variation. This can aid in visually determining if studied characteristics can be grouped as seen in Figure 16 [77].

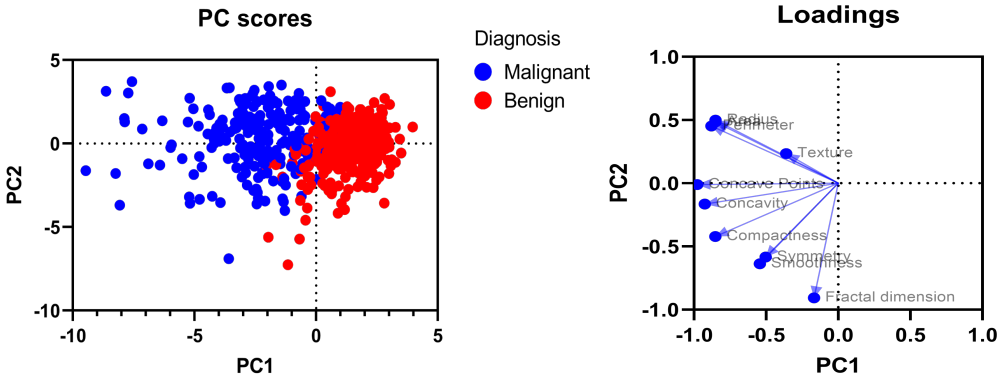


Figure 16: PCA results enabling grouping of benign and malignant tumors (left) and visualization of relevant features (right)[83].

Combining both graphs from Figure 16 creates a biplot. The cosine of the angle between a vector and an axis indicates the importance of the contribution of the corresponding variable to that principal component. The cosine of the angle between pairs of vectors indicates the correlation between the corresponding variables. Highly correlated variables point in similar directions; uncorrelated variables are nearly perpendicular to each other. Points that are close to each other in the biplot represent observations with similar values [84].

6 Objectives

The main objective of this Master's Thesis is to unearth possible distinguishing powers of radiomics, extracted from the ^{18}F-FDG PET images of a homogeneous group of 49 early-stage NSCLC patients. More specifically, can these features be used to discriminate healthy tissue from NSCLC tissue? The hypothesis is that the radiomics features, based on image characteristics, are a good tool for separating both tissue types. Hereto, this study aims to create an optimized discriminative model using different sets of radiomics features to distinguish healthy from NSCLC tissue. In the case of clear separation and a model with high accuracy values can be obtained, it is important to know which features are most important in this discrimination. Additionally, it will be investigated if these classifiers are able to make predictions on the tissue type. This is further expanded by training and testing models on other clinical parameters, as described in the patient demographics. These extra parameters include: tumor phenotype; tumor tissue location; diabetes; glycemia levels and packyears.

7 Materials and methods

7.1 Patient cohort

The ProLung study is a research project at Ziekenhuis Oost-Limburg (ZOL) located in Genk and is funded by 'Kom op tegen Kanker'. The study is made up of two branches, metabolomics, and radiomics applied to patients with primary lung cancer, specified type NSCLC. For this study, a patient cohort consisting of 53 patients is used. Each patient underwent a PET/CT scan in ZOL followed by an anatomopathological analysis test of the tissue of interest to determine the tumor phenotype. Four patients from the original patient cohort were removed from the study since it did not concern a malignant tumor, but inflammation. The remaining 49 patients were all diagnosed with NSCLC stage I-IIIA. No metastasis of the tumor had occurred and all patients underwent a lobectomy as part of their standard-of-care treatment plan. The treatment and imaging took place at ZOL. All patients provided written consent to be included in the ProLung study and they can choose to leave the study at any time.

7.2 PET-CT scanner

To obtain the PET images needed for the radiomics analyses, all patients had a PET-CT scan taken with a Biograph Horizon PET-CT scanner. This machine uses lutetium oxyorthosilicate ($Lu_2(SiO_4)O$, LSO) scintillation crystals. The technical specifications of the PET part of this scanner are shown in Table 3. The imaging procedure followed the guidelines of the European Association of Nuclear Medicine (EANM). This study focuses on the meta-data extraction of PET images, but the CT images were used by Prof. Dr. Mesotten to adjust the images for attenuation. All the images obtained from the scan are saved in the Picture Archiving and Communications Systems (PACS).

Table 3: Technical specifications of the PET imager [85].

Biograph Horizon PET	
Axial field of view	16.4, 22.1* cm
Crystal size	4 x 4 x 20 mm
Confidence window	4.1 ns
Effective sensitivity	14.9, 26.5*cps/kBq
Effective peak NEC rate	224, 336* kcps ≤26 kBq/cc

** Optional*

7.3 ^{18}F-FDG-PET scanning procedure

As laid out in Chapter 3.2, ^{18}F-FDG is used as the radiopharmaceutical biomarker for the imaging of the patients in this study. An autoinjector, Iris automated multidose injection by Comecer, is used to deliver the ^{18}F-FDG to the patient while minimizing the radiation dose to the medical personnel. An hour elapses between the administration of the radiopharmaceutical and the PET/CT imaging. Firstly, a CT image (25 mA, 130 kV) is obtained where the imaging field extends from the midthighs to the base of the skull resulting in a 512 x 512 matrix. Secondly, a PET scan covering the same area is performed for 15 - 20 minutes. Based on the mass of the patients, <50 kg, 50-80 kg- and >80 kg the emission time per bed position is one minute, one minute and a half, and two minutes respectively.

7.4 Data acquisition

During a previous study, the PET-CT images have been anonymized and tumor segmentation has been performed in the ACCURATE tool, developed by the research team of Prof. Dr. Boellaard (Amsterdam, UMC) as seen in Figure 17. This segmentation is semi-automatic where the tool suggests the tumor VOI coördinates based on an SUV threshold with a medical professional, Prof. Dr. Mesotten, performing any necessary corrections. The segmentation is performed solely on PET images and an SUV threshold of 0% is maintained. The CT images acquired during the PET/CT scan are used to correct the PET images for breathing artifacts. The created VOIs are saved as projects (.prj) containing the PET .dcm files and the VOI (.nii and .voi files).

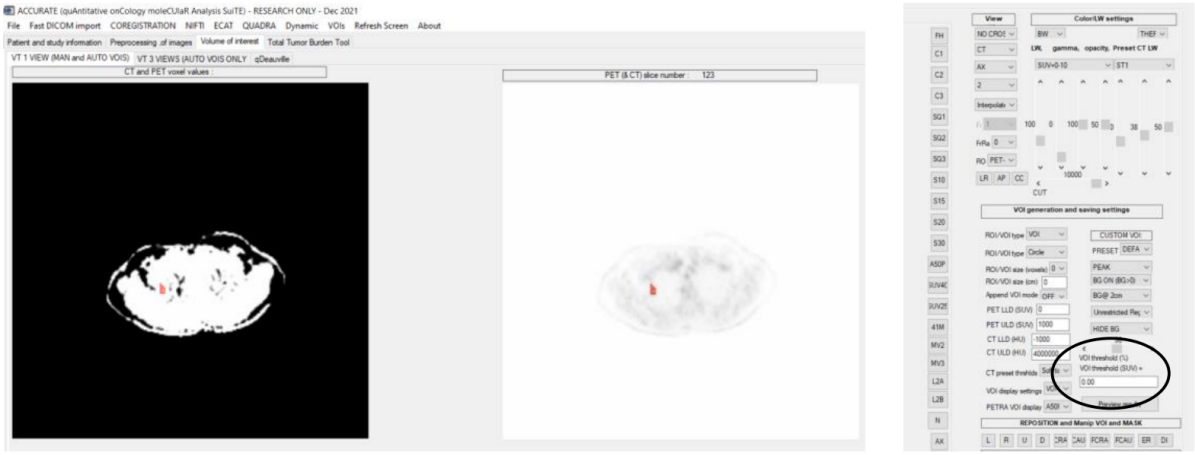


Figure 17: The ACCURATE tool showing a PET image.

Next, the projects of the VOIs are loaded in the RADIOMICS tool also developed by the research team of Prof. Dr. Boellaard (Amsterdam, UMC). The radiomics analysis is performed resulting in an Excel file containing 498 radiomics features and 6 PET Uptake Metrics per patient.

To be able to compare the aforementioned radiomics of tumor tissue with healthy tissue, VOIs of healthy tissue of the same patient have to be segmented. The shape, volume, and surface of this new VOI must match the tumor VOI exactly. The accurate tool can translate and rotate VOIs, making this comparison possible. The location of the healthy VOI is also important for the radiomics data [86]. This study by Trojani et al. suggests three possible options for segmenting healthy tissue for comparison with NSCLC tissue: the liver, the opposite lung at the same height, and the upper-right lung above the aortic arch. For the purposes of this research paper, a VOI in the opposite lung at the same height was chosen in agreement with the medical professional in charge of the tumor segmentation. Analogous to the NSCLC tissue, the RADIOMICS tool is used to extract the same 498 radiomics features and 6 PET Uptake Metrics per patient. A more detailed step-by-step plan of the data acquisition for both tumor and healthy tissue is added in Annex IV.

Finally, the 504 obtained features for both the tumor and healthy lung tissue per patient are gathered in an Excel file with dimensions 106 x 504 for data analysis. The head of this file is shown in Figure 18.

	local inter	global inte	Original m	Original m	Original Tl	ExactVolu	Volume	approxim	Surface
PROLUNG002_tumor	10.9754	12.3966	14.2308	9.83963	532668	54135.4	48207.2	54504	9098.86
PROLING002_gezond	0.29307	0.416748	0.518437	0.258305	13983.4	54135.4	46593.7	54088	8902.79
PROLUNG004_tumor	1.85517	2.24605	4.31843	2.17688	5006.82	2300.96	1424.67	1984	537.171
PROLING004_gezond	0.475203	0.505545	0.574539	0.449072	1032.86	2300.96	1685.18	2080	660.661
PROLUNG009_tumor	16.2197	19.6004	27.0645	19.0972	89470.5	4685.6	3450.97	4560	1259.72
PROLING009_gezond	1.25291	1.30707	1.66441	1.11011	5200.85	4685.6	3455.66	4568	1259.52
PROLUNG010_tumor	6.0671	6.13819	9.59616	5.81687	41608.1	7153.9	5993.54	6640	2168.18
PROLING010_gezond	1.13319	1.15962	1.52519	0.944372	6755.1	7153.9	5995.68	6568	2194.66
PROLUNG011_tumor	0.756629	0.797352	1.17736	0.848976	2947.64	3472.36	2487.37	3328	1398.4
PROLING011_gezond	0.353467	0.400301	0.439736	0.321449	1116.07	3472.36	2779.22	3336	1368.49
PROLUNG012_tumor	5.55386	5.60374	11.6219	5.99932	18819.9	3137.68	2574.15	2928	1100.72
PROLING012_gezond	0.840482	0.840482	0.953779	0.589209	1848.35	3137.68	2603.72	2928	1118.46
PROLUNG013_tumor	2.65485	2.78948	5.8541	3.22871	7293.65	2259.13	1179.85	2096	708.294
PROLING013_gezond	0.586383	0.586383	0.734443	0.467573	1056.25	2259.13	1061.18	2104	631.292
PROLUNG015_tumor	2.63397	2.98138	8.54308	3.79327	5553.34	1464.25	784.447	1336	492.004
PROLING015_gezond	0.198238	0.20991	0.250666	0.174804	255.913	1464.25	787.719	1320	486.446

Figure 18: Head of the Excel file containing the features for both healthy and tumor tissue.

7.5 Data analysis

The first step is to filter out the features which are not useful in differentiating between the two types of tissue. This comparison is similar to a before-and-after observation of the same patient. A paired student t-test is therefore optimal for preprocessing the dataset. The null hypothesis H_0 indicates that a feature for healthy tissue is correlated with a feature for tumor tissue. Features in agreement with H_0 are not relevant in the differentiation and are removed from the dataset. A p-value of 0.05 or higher is used to determine the validity. This statistical test is performed in Matlab[1].

The second step is clustering the features by making a dendrogram in Matlab. This is done on the t-test corrected dataset. This dataset still has unknown values. The features that have unknown values are deleted before performing the linkage function which is used to form the tree that will form the dendrogram. To classify the different branches of the dendrogram, different cut-off heights are implemented because the clusters form on different levels in the dendrogram[1].

To further visualize the data, PCA is used since the dataset contains 504 features/dimensions. By reducing the features to principal components (PC) that explain most of the variance while being uncorrelated between themselves, a lower-dimensional dataset can be produced. Possible clustering of the data can be observed this way in a scatter plot. This plot can be further expanded by adding vectors of the most relevant features, creating a biplot[1].

7.6 Discriminative model

The final goal is to train different models in the Matlab app 'Classification learner'. The goal here is to train models to differentiate NSCLC tissue from healthy tissue. There is also demographic information available that can be used as classifiers. The demographic features used as classifiers are glycemia, tumor type, lung side (left or right), diabetes, and packyears. The models are trained on the first 30 patients. The model that gives the best accuracy is then used to make predictions for the other 19 patients of the dataset. The results are then visualized in a confusion matrix.

[1] Matlab scripts for the t-test and dendrogram as well as the Python code for the PCA are included in Annex III

8 Results

8.1 Demographics

For the purposes of this study, PET images of a cohort of 49 patients, each diagnosed with NSCLC stages I-IIIA, are used. Table 4 gives an overview of the most relevant demographic features of the patient group. The table includes packyears since there exists a strong correlation between smoking and lung cancer [16-17]. Furthermore, information on tumor location, diabetes status, and glycemia measurements before the scan are included. These characteristics are used as classifiers for the machine learning models. Figures 19 - 24 provide a more detailed overview of specific demographic characteristics of the patient cohort.

Table 4: Relevant demographic data of the patients included in this study.

Total patients		49
Sex (N, (%))	Men	31 (63.3)
	Women	18 (36.7)
Diabetes (N, (%))	Yes	11 (28.9)
	No	38 (71.1)
Packyears (y)	Median	35
	Average	40±27
	Range	0-139
	Unknown	1
Age (y)	Median	70
	Average	71±8
	Range	49-84
BMI (kg/m²)	Median	26
	Average	27±6
	Range	15-51
Lobe positioning	Right upper	15
	Right lower	8
	Left upper	14
	Left lower	12
Diameter (mm)	Median	20
	Average	28±21
	Range	7-120
Plasma glucose (mg% Glycemia)	Median	98
	Average	99±16
	Range	76-164

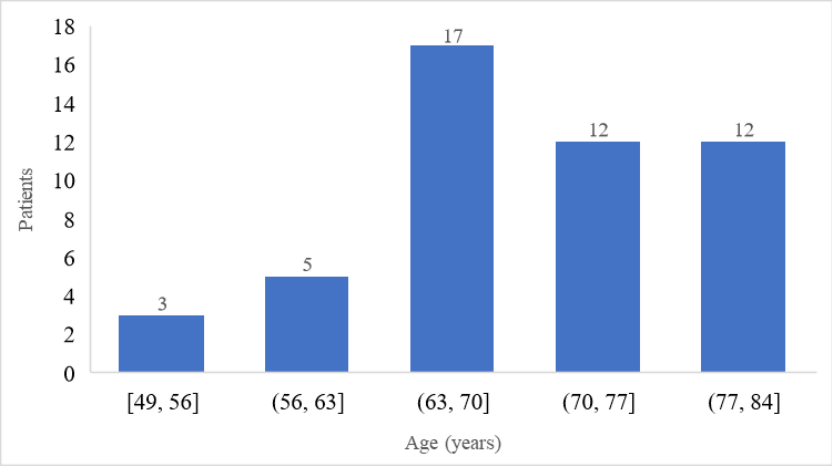


Figure 19: Bar chart of the ProLung patients' age.

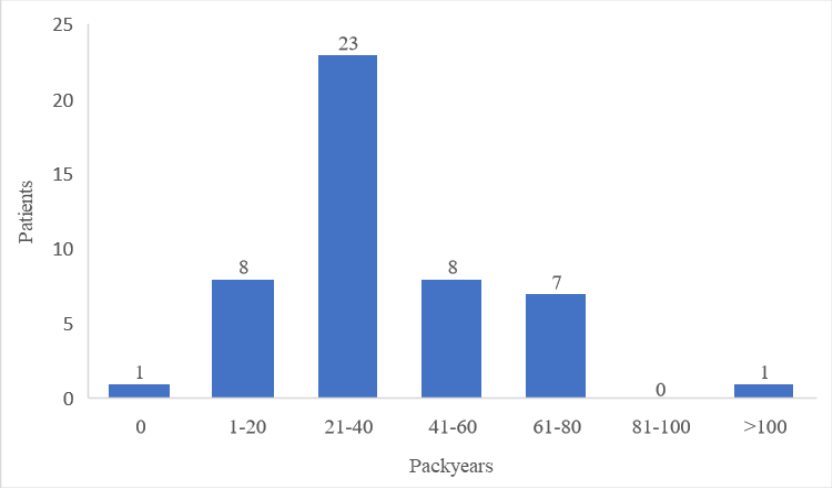


Figure 20: Bar chart of the ProLung patients' pack years.

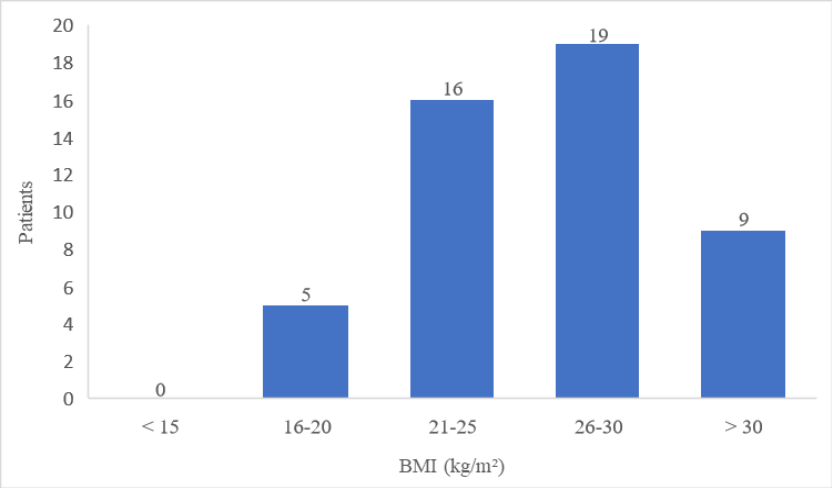


Figure 21: Bar chart of the ProLung patients' BMI.

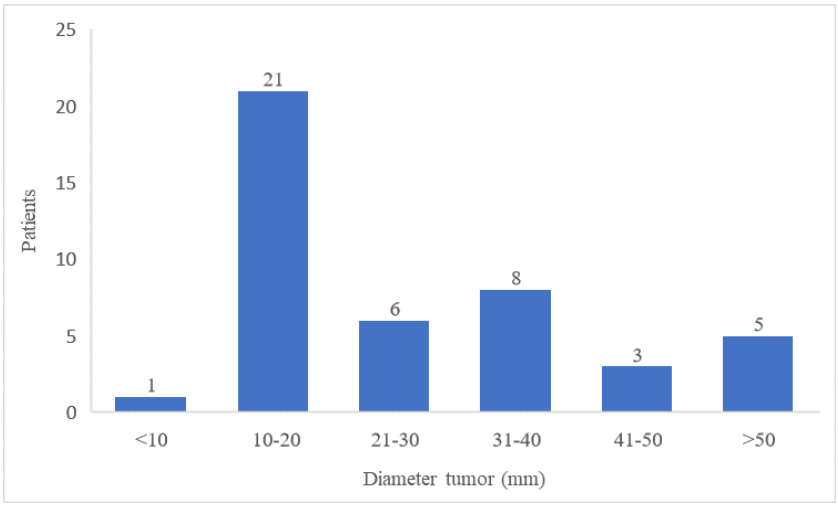


Figure 22: Bar chart of the diameter of the tumor of the ProLung patients.

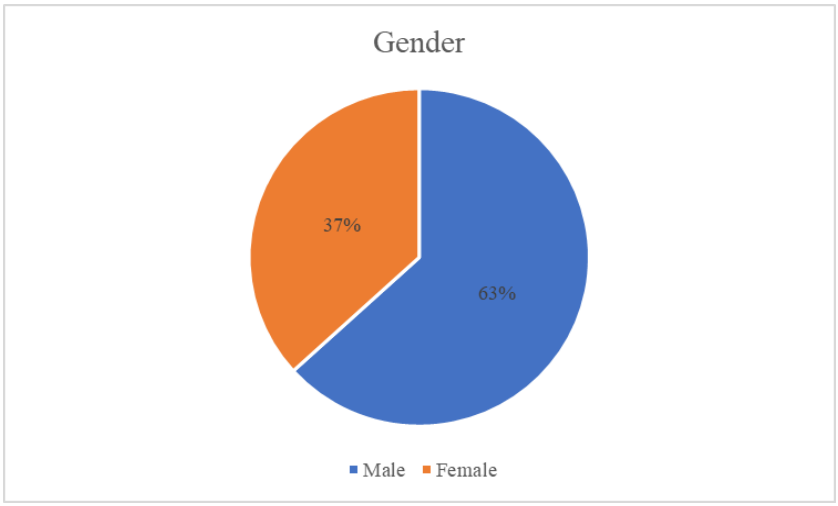


Figure 23: Pie chart of the gender of the ProLung patients.

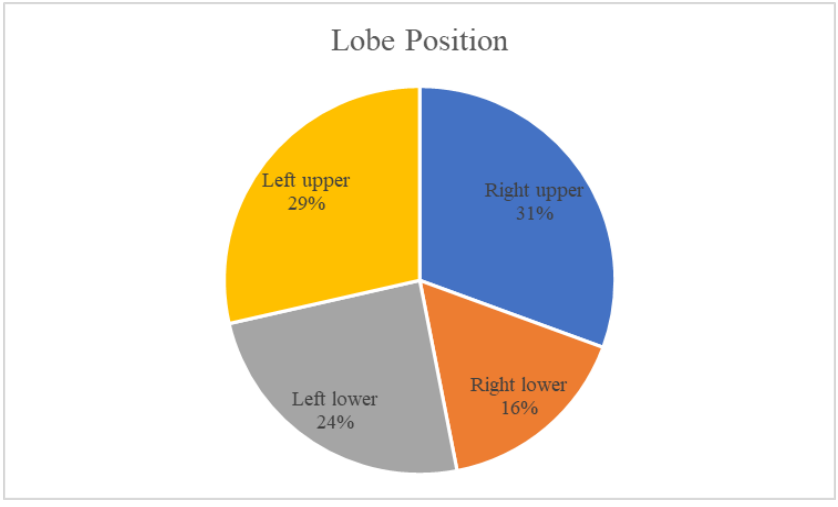


Figure 24: Pie chart of the lobe position of the ProLung patients.

8.2 Statistical results

8.2.1 Paired student t-test

A paired student t-test was performed on the radiomics dataset using Matlab to check which features are relevant in distinguishing healthy and NSCLC tissues. A null hypothesis of zero means a feature is similar in both types of tissues and therefore not useful for this study. A high *p*-value means the result is likely due to chance. Tabel 5 displays the radiomics features that are removed from the study based on their null hypothesis and/or a high *p*-value.

Table 5: Radiomics features not significant in differentiating healthy and NSCLC tissue.

	H	p-value	t-score	Standard deviation
PET Uptake Metrics - ExactVolume	0	0.321941	-1	68.959
Morphology - Volume	0	0.137958	1.506626	304.087
Morphology - Surface	0	0.608925	0.514727	73.744
Morphology - Surface to volume ratio	0	0.256654	1.146946	0.033
Morphology - Compactness1	0	0.862127	-0.17453	0.005
Morphology - Compactness2	0	0.999522	-0.0006	0.222
Morphology - Spherical disproportion	0	0.556949	0.591198	0.066
Morphology - sphericity	0	0.259347	-1.1404	0.025
Morphology - asphericity	0	0.875482	0.157473	0.023
Morphology - center of mass shift	0	0.219691	-1.30938	2970.070
Morphology - maximum 3D diameter	0	0.31811	-1.00803	0.752
Morphology - major axis length	0	0.572294	-0.56828	0.340
Morphology - minor axis length	0	0.590775	0.541063	0.352
Morphology - least axis length	0	0.824492	-0.22289	0.470
Morphology - elongation	0	0.30758	1.030424	0.013
Morphology - flatness	0	0.761051	0.305703	0.020
Morphology - vol density AABB	0	0.905052	-0.11987	0.023
Morphology - area density AABB	0	0.702015	0.384719	0.017
Morphology - vol density AEE	0	0.910132	0.113422	0.070
Intensity histogram - skewness	0	0.060643	-1.92552	3.633
glcmFeatures2Davg - cluster shade	0	0.23209	1.20911	985.639
glcmFeatures2DDmrg - first measure of information correlation	0	0.610693	0.51218	0.222
glcmFeatures2Dmrg - inverse difference moment normalised	0	0.18139	-1.35464	0.046
glcmFeatures2Dmrg - cluster shade	0	0.262703	1.132299	992.872
glcmFeatures3Davg - first measure of information correlation	0	0.770346	-0.29345	0.213
GLSZMFeatures2Davg - Zone size non uniformity normalized	0	0.39997	-0.84864	0.285
GLSZMFeatures3D - Zone size non uniformity normalized	0	0.259282	-1.14055	0.327
ngtdmFeatures2avg - coarseness	0	0.564438	-0.57997	0.257
ngtdmFeatures2avg - busyness	0	0.321571	-1.00077	25837.267
ngtdmFeatures2Dmrg - coarseness	0	0.175179	-1.37451	9.727
ngtdmFeatures3D - coarseness	0	0.308866	-1.02766	7.938
ngldmFeatures2Davg - Dependence count percentage	0	0.321941	1	0.007

As presented in Table 5, most of the features fall under the ‘Morphology’ or ‘Texture Features’ subgroup. Since the VOI for the healthy tissue and the Tumor tissue is identical and these features do not take into account the gray levels of the voxels in the image, these parameters are the same for both and can therefore not be the basis for a differentiating study. The same applies to the ‘PET Uptake Metrics -ExcatVolume’ feature.

Both the Gray Level Co-occurrence Matrix (GLCM) features, which describe textural indices based on the arrangements of pairs of voxels, and the Neighboring Gray Tone Difference Matrix (NGTDM) Features, which quantify the difference between a voxel gray level and the average gray level of its neighbors in all three dimensions within a given distance, have multiple parameters not relevant in making out healthy tissue form NSCLC tissue according to the paired t-test.

Gray Level Size Zone Matrix (GLSZM) Features provide data on the gray level zones of the segmented area of the image. These gray level zones are areas of connected voxels sharing the same gray level intensity, thus indicating uniformity. Both 2D and 3D ‘zone size non-uniformity normalized’ features, which measure the variability of size zone volumes in the image, with a lower value indicating more homogeneity in size zone volumes, are not included in this research paper.

The Final feature that will be excluded is the skewness trait of the ‘Intensity histogram (First order) features’. This represents the skewness of the histogram made of the intensities of the gray levels in the VOI.

8.2.2 Dendrogram

The corrected radiomics dataset exists out of 435 features. These features can be linked and clustered. To visualize these clusters, a dendrogram is made and the biggest clusters are marked. This is shown in Figure 25.

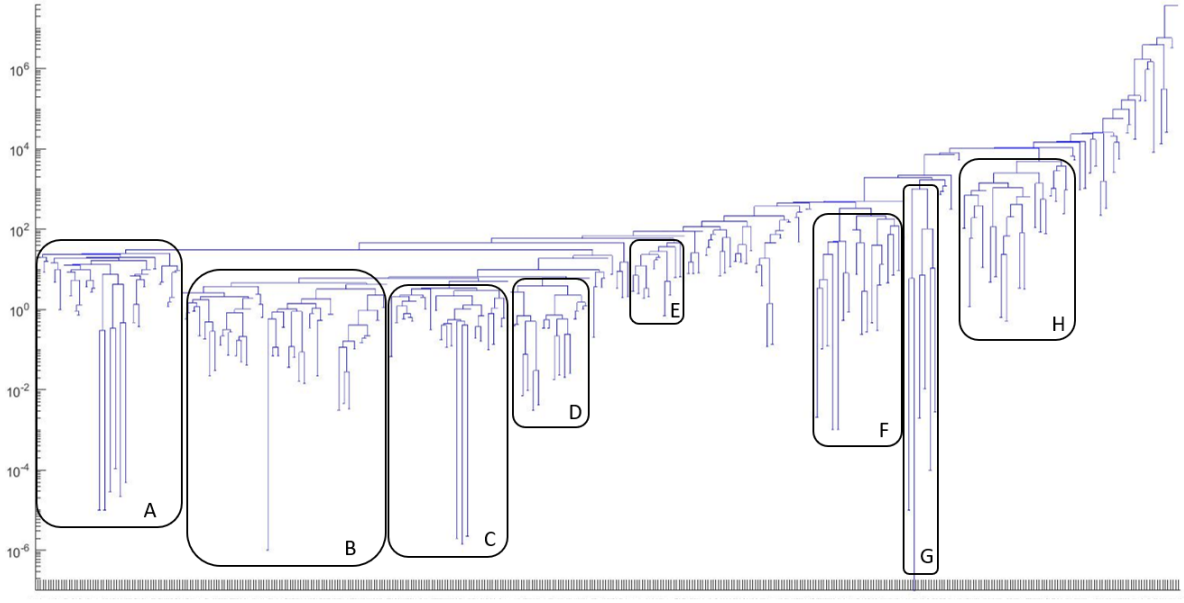


Figure 25: Dendrogram of the radiomics dataset.

There are 8 clusters marked in this dendrogram. The features that belong to each cluster are shown in Table 6. Also, the cutoff height is mentioned, this is the height of the position of the highest branch of the cluster.

Table 6: Information about the clusters in Figure 25.

Cluster	Information features
A (cutoff = 30)	Mostly the entropy, average, dissimilarity, emphasis, and uniformity features of the GLCM, GLRLM, GSZM, GLDZM, and NGLDM Also statistical features variance, minimum, 10th percentile and interquartile range, as intensity histogram features as mean absolute deviation and entropy
B (cutoff = 5)	Mostly joint maximum, angular second moment, inverse variance, first measure of information correlation and low gray level features of the GLCM, GLRLM, GSZM, GLDZM, and NGLDM Also morphology feature Moran's I, statistical features Coefficient of variation and Quartile coefficient and intensity features volume at int fraction 90, Coefficient of variation, Quartile coefficient and Uniformity
C (cutoff = 5)	Mostly correlation, second measure of information correlation, short run emphasis, Run length non uniformity normalized, run percentage and zone percentages of the GLCM, GLRLM, GSZM, GLDZM and NGLDM Also intensity features volume at int fraction 10 and difference vol at int fraction
D (cutoff = 5)	Mostly inverse difference features of the GLCM, GLRLM, GSZM, GLDZM and NGLDM Also morphology feature Geary's C
E (cut off = 48)	Pet Uptake metrics features as intensity peaks, statistical features as mean, maximum and range, GLCM features as difference variance and joint entropy
F (cutoff = 360)	Mostly joint variance, sum average, contrast features of GLCM and gray level variance features of GLRLM, GLSZM, GLDZMand NGLDM Also intensity histogram features variance and maximum
G (cutoff = 1200)	GLCM features sum variance and cluster tendency
H (cutoff = 5500)	GLCM autocorrection features and high gray level emphasis features of GLRLM, GLSZM, GLDZMand NGLDM

8.3 Principal component analysis

The same t-test corrected dataset of the cohort of 49 patients underwent a principal component analysis using Python. The script to perform the analyses and create the plots in this chapter can be found in Annex III. Firstly, all features were used to determine a set of principal components. The first two components, explaining the most variance while being uncorrelated among themselves, were then used to create a two-dimensional representation of the 435-dimensional original dataset. Since PCA is often used as a noise reduction operator, the most significant features are extracted after the first analysis. The 30 most important features resulting from this first PCA were used to perform a second PCA. Finally, a third PCA using only the five most significant features was then performed.

8.3.1 PCA using all 435 features

The first PCA is performed using all 435 features. For all the features of the 49 patients included in the study, the principal components and their explained variance were derived, as seen in Figure 26.

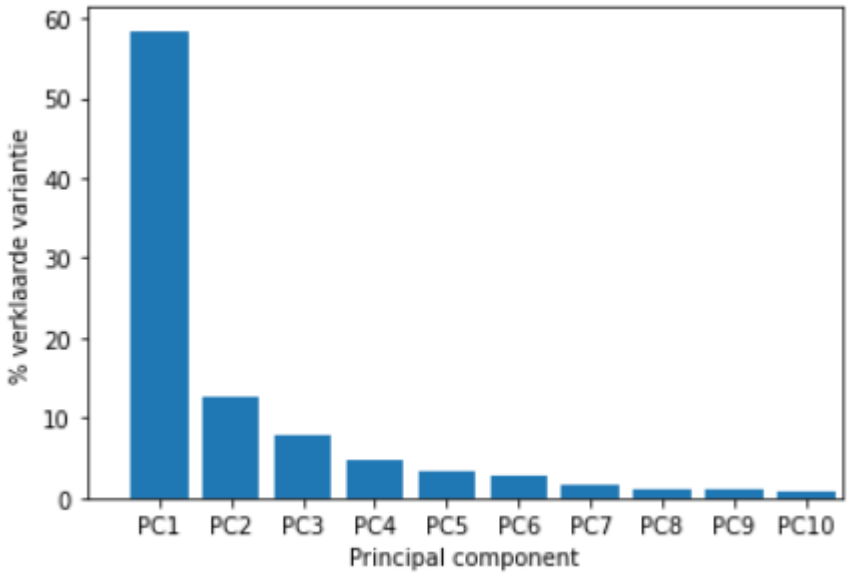


Figure 26: The first ten PCs showing the greatest variance.

Figure 26 shows the 10 most relevant principal components and their explained variance. These can be used to visualize the high-dimensional dataset in a lower-dimensional space. In this case, the first two components explaining 58.4% and 11.8% of the variance respectively, for a total of 70.6%, are used to make a 2D plot as seen in Figure 27.

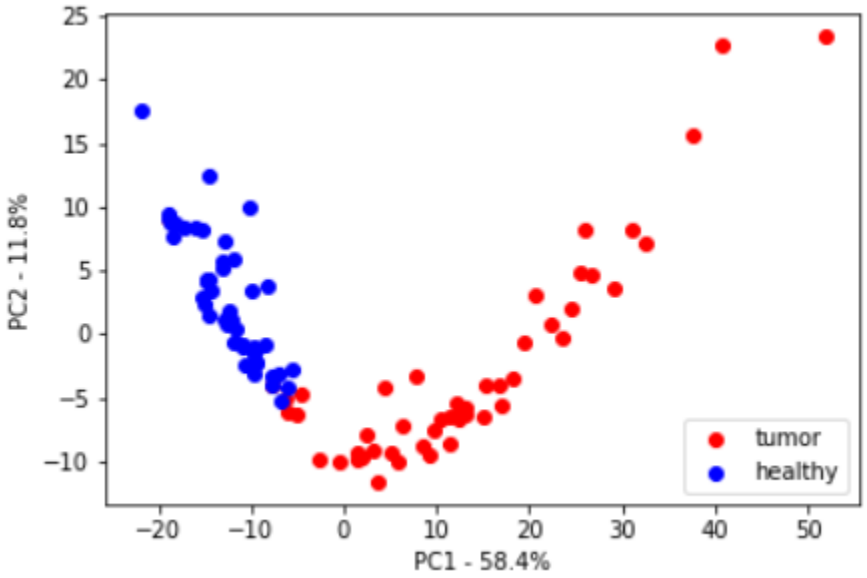


Figure 27: Scatterplot of PC1 and PC2 indicating clustering of healthy and tumor tissue.

Figure 27 displays the 2D scatterplot of the two most relevant PCs. The data points of healthy tissue VOIs are marked in blue, while the tumor tissue VOIs are marked in red. The combined variance of 70.6% is enough to almost completely distinguish between healthy and NSCLC tissue in the VOIs. The first principal component has a higher impact on the spread since it also contributes more to the explained variance. There is however some overlap around the coördinates (-7 , -6). Furthermore, one tumor VOI has been clustered with the healthy tissue at coördinate (-19 , 10). Figure 28 shows this more in detail.

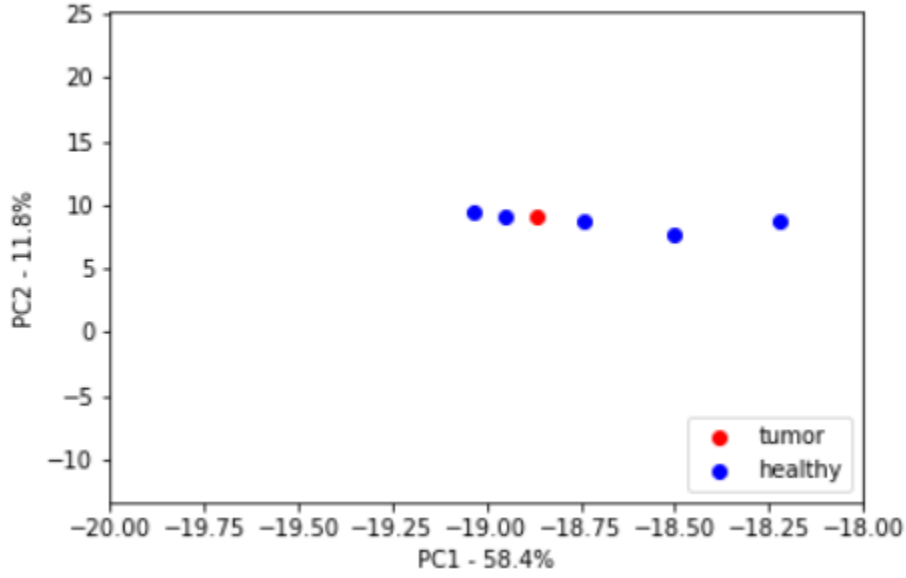


Figure 28: The scatterplot of PC1 and PC2, highlighting a wrongly clustered result.

From this PCA, the most important features contributing to the principal components can be identified. Table 7 shows the 30 most prevalent features and their loading scores. Similar scores indicate their weight in determining the principal components was alike.

Table 7: The 30 most significant features in differentiating between healthy and tumor tissue found using PCA.

(PET Uptake Metrics - Original max,)	0.061183
(Intensity histogram - maximum,)	0.061092
(Statistics - maximum,)	0.061091
(glcmFeatures2Davg - difference entropy,)	0.060996
(intensity volume - int at vol fraction 10,)	0.060943
(glcmFeatures2Dmrg - difference entropy,)	0.060928
(Intensity histogram - 90th percentile,)	0.060918
(Statistics - 90th percentile,)	0.060912
(Intensity histogram - Entropy,)	0.060567
(Intensity histogram - Robust mean absolute deviation,)	0.060521
(Intensity histogram - range,)	0.060439
(Statistics - range,)	0.060419
(Intensity histogram - Median absolut deviation,)	0.060402
(Intensity histogram - Mean absolut deviation,)	0.060346
(Statistics - Median absolute deviation,)	0.060309
(glcmFeatures3DWmrg - joint entropy,)	0.060272
(Statistics - Mean absolut deviation,)	0.060255
(glcmFeatures2Dmrg - sum entropy,)	0.060249
(glcmFeatures2Dvmrg - joint entropy,)	0.060245
(Statistics - Root mean,)	0.060159
(glcmFeatures3Davg - difference entropy,)	0.060110
(glcmFeatures2DDmrg - joint entropy,)	0.060080
(glcmFeatures2DDmrg - difference entropy,)	0.060057
(Intensity histogram - Interquartile range,)	0.060052
(glcmFeatures3Davg - sum average,)	0.060046
(glcmFeatures3Davg - joint average,)	0.060046
(glcmFeatures3DWmrg - joint average,)	0.060042
(glcmFeatures3DWmrg - sum average,)	0.060042
(glcmFeatures2DDmrg - sum average,)	0.060040
(glcmFeatures2DDmrg - joint average,)	0.060040

To further visualize the PCA, a biplot showing both the scatterplot and the vectors of the 10 most prevalent features is shown in Figure 29.

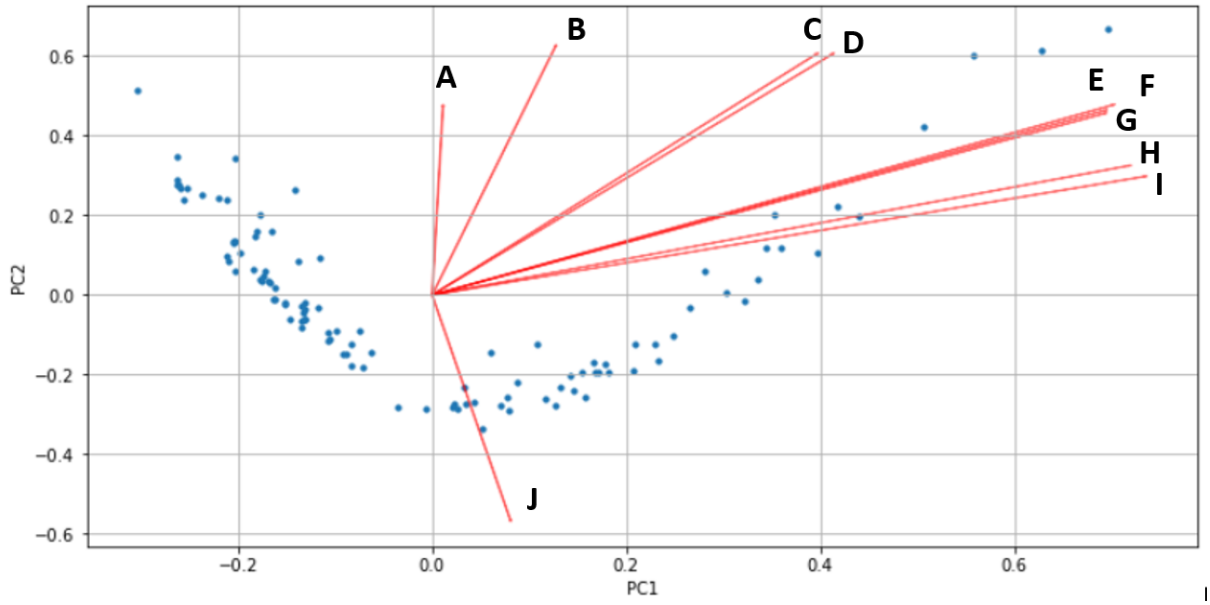


Figure 29: Biplot of the PC clusters and the vectors of the 10 most relevant features.

The vector length has been scaled to distinguish different feature vectors. The angle between a vector and the axis indicates its importance to the principal components on that axis. Angles between vectors indicate their underlying correlation. Table 8 serves as a legend for the letters in Figure 29.

Table 8: Legend for biplot Figure 29.

Letter	Radiomics feature
A	glcmFeatures2Dmrg - difference entropy
B	Statistics - 90th percentile
C	Intensity Histogram - 90th percentile
D	glcmFeatures2Davg - difference entropy
E	Statistics - maximum
F	Intensity histogram - range
G	PET Uptake Metrics - Original max
H	intensity volume - int at vol fraction 10
I	Intensity histogram - maximum
J	Intensity histogram - Robust mean absolute deviation

8.3.2 PCA with noise reduction (269 features)

PCA is a statistical test, often used to remove noise in large datasets. The 435 features and the absolute values of their loading scores are visualized in Figure 30. This shows that a drop-off occurs after the first 269 features, which indicates reduced importance of those features.

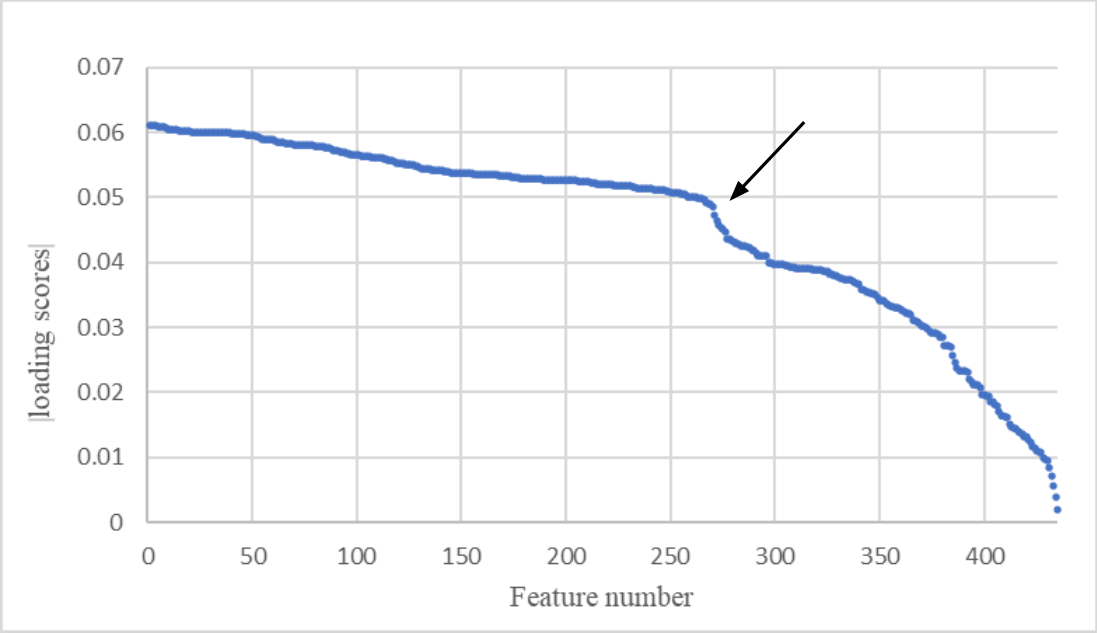


Figure 30: Plot of the loading- scores in descending order and their corresponding features.

As indicated by Figure 30, a drop in loading scores is present at the 270th feature. This in turn points to 166 features or 38.1% of the t-test corrected features being noise. The PCA is performed again using the 269 leftover features. This resulted in the bar chart of Figure 31 with the ten most relevant PCs, and the scatter plot in Figure 32.

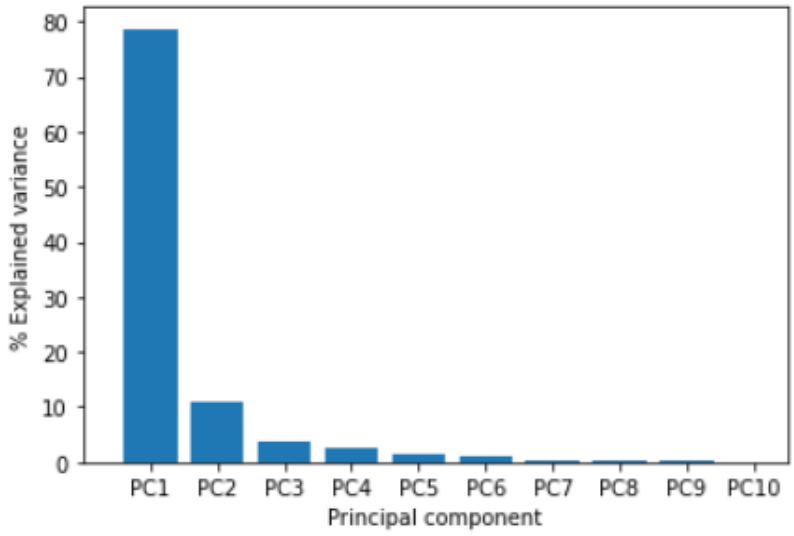


Figure 31: The first ten PCs showing the greatest variance after noise reduction.

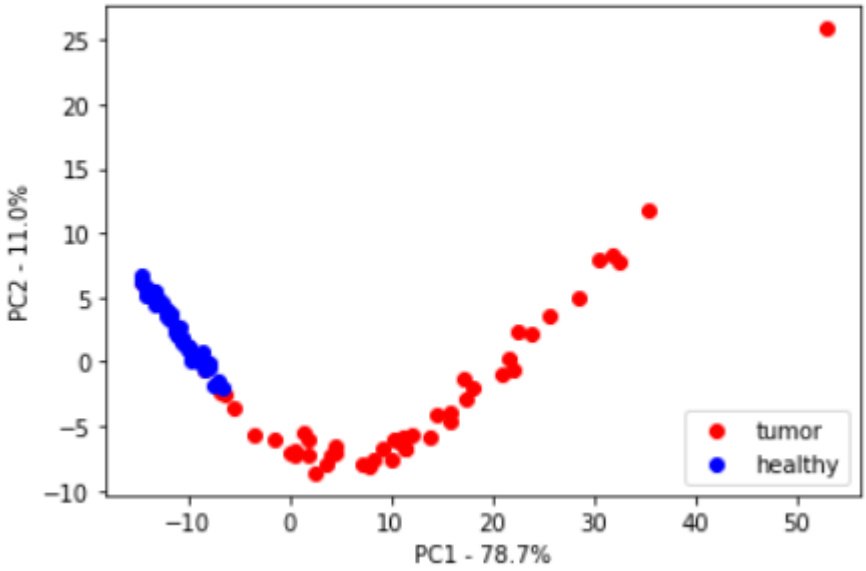


Figure 32: Scatterplot of PC1 and PC2 after noise reduction indicating clustering of healthy and tumor tissue.

Figure 32 shows both healthy and tumor tissue groups appearing more clustered after the noise reduction. As well as improved clustering, a higher explained variance of 89.7% by the first two PCs can be observed.

To observe the influence of the ten most influential features, a biplot is shown in Figure 33.

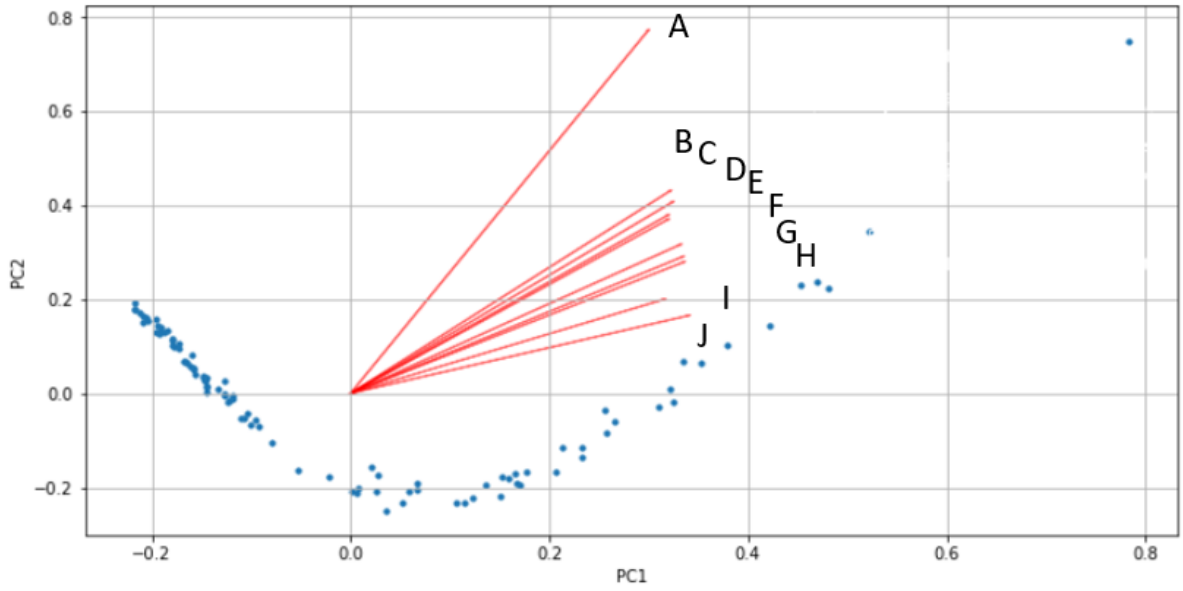


Figure 33: Biplot of the PC clusters and the vectors of the 10 most relevant features after noise reduction.

Table 9: Legend for biplot Figure 33.

Letter	Radiomics feature
A	Intensity histogram - Robust mean absolute deviation
B	Statistics - maximum
C	Statistics - 90th percentile
D	Intensity histogram - maximum
E	intensity volume - int at vol fraction 10
F	glcmFeatures3Davg - sum average
G	Statistics - root mean
H	PET Uptake Metrics - Original max
I	glcmFeatures3Davg - joint average
J	Intensity histogram - 90th percentile

Figure 33 shows higher clustering of the ten most relevant features in determining the spread of the data points compared to the biplot before noise reduction from Figure 32. Most of the features between these two plots are the same except for some 3D GLCM features being more prevalent than their 2D equivalents before noise reduction.

8.3.3 PCA using the 30 most significant features

Next, the 30 most significant features, as shown in Table 7, are used to perform a third PCA. The resulting graphs are shown in Figures 34-36, and in Table 10.

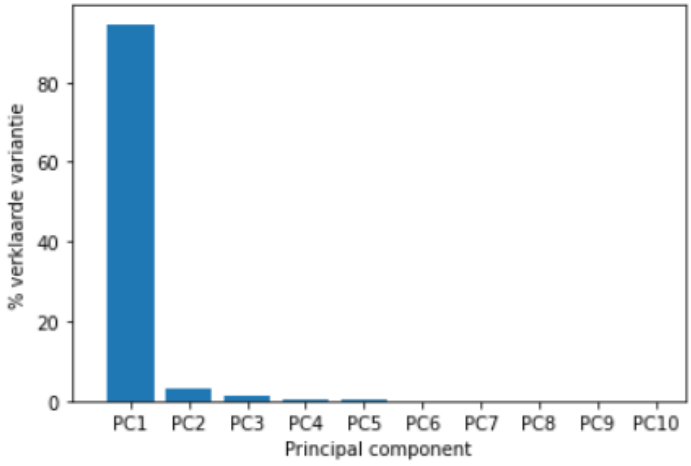


Figure 34: The five PCs showing the greatest variance for the 30 most significant features.

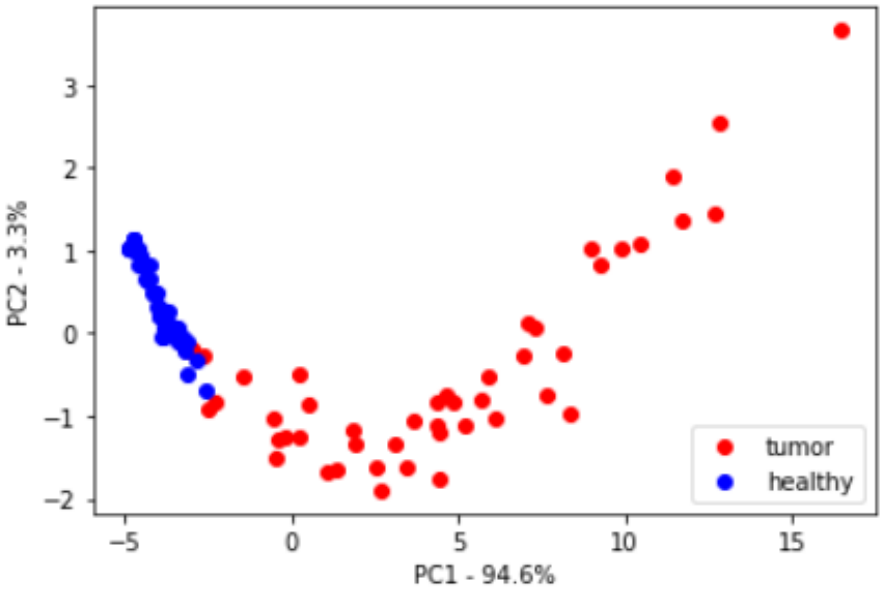


Figure 35: Scatterplot of PC1 and PC2 for the 30 most significant features.

As seen in Figures 34 and 35, the first two PCs can account for 97.7% of the total variance in the dataset. Healthy tissue clustering is improved compared to the PCA using all 435 features. Comparing the healthy tissue to the noise-reduced PCA results, fewer features reduce the overall cluster density. The tumor tissue clustering worsened compared to both the first (all features) and the second (noise-reduced) PCA.

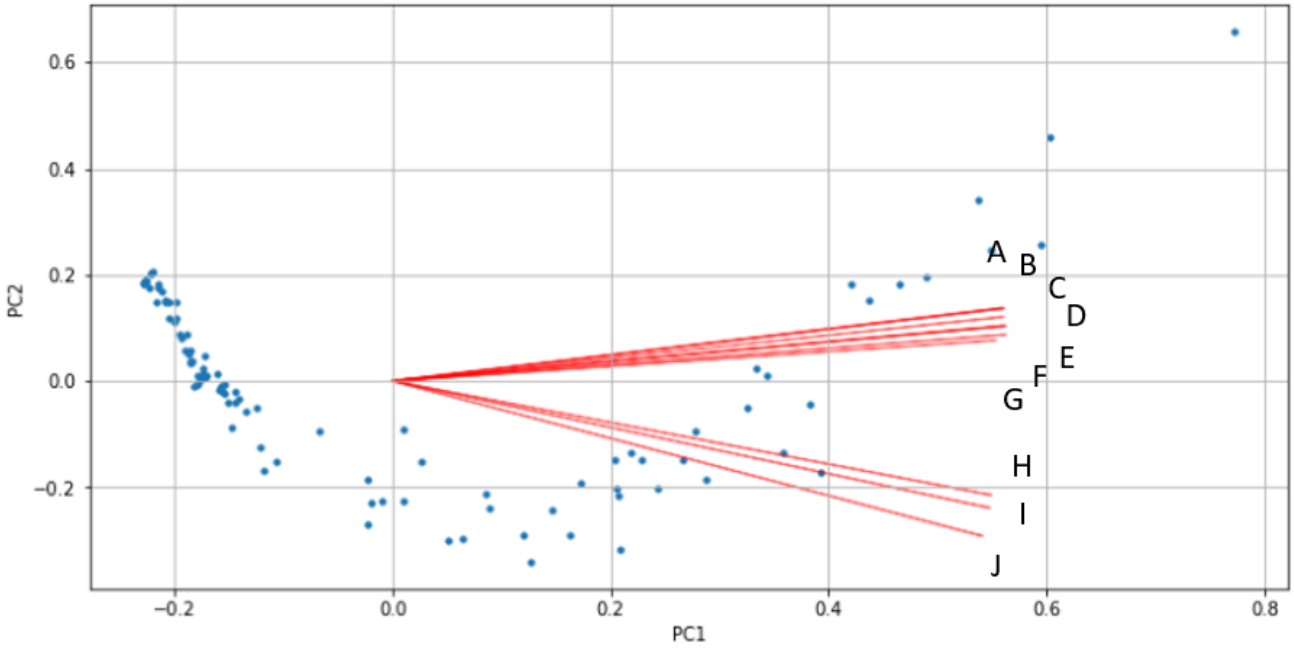


Figure 36: Biplot of the PC clusters and the vectors of the 30 most relevant features.

Table 10: Legend for biplot Figure 36.

Letter	Radiomics feature
A	Intensity histogram - range
B	Statistics - range
C	Intensity volume - int at vol fraction 10
D	Intensity histogram - maximum
E	Statistics - maximum
F	PET Uptake Metrics - Original max
G	Intensity histogram - Median absolute deviation
H	Statistics - 90th percentile
I	Intensity histogram - 90th percentile
J	Statistics - Root mean

8.3.4 PCA using the five most significant features

Finally, the five most significant features are extracted. The bar chart, scatterplot, and biplot resulting from the PCA as performed before are shown in Figure 37-39, and in Table 11.

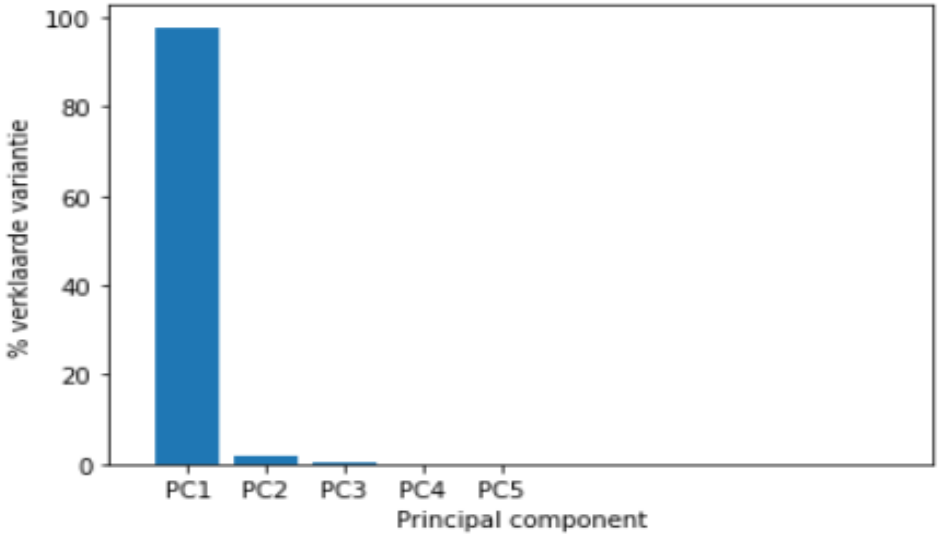


Figure 37: The five PCs showing the greatest variance.

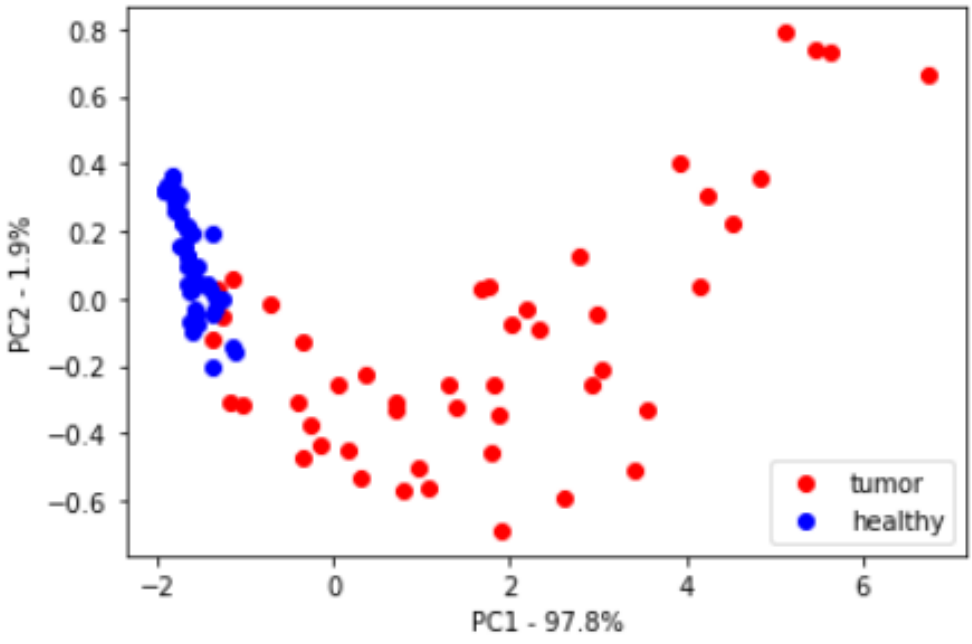


Figure 38: Scatterplot of PC1 and PC2 for the five most significant features.

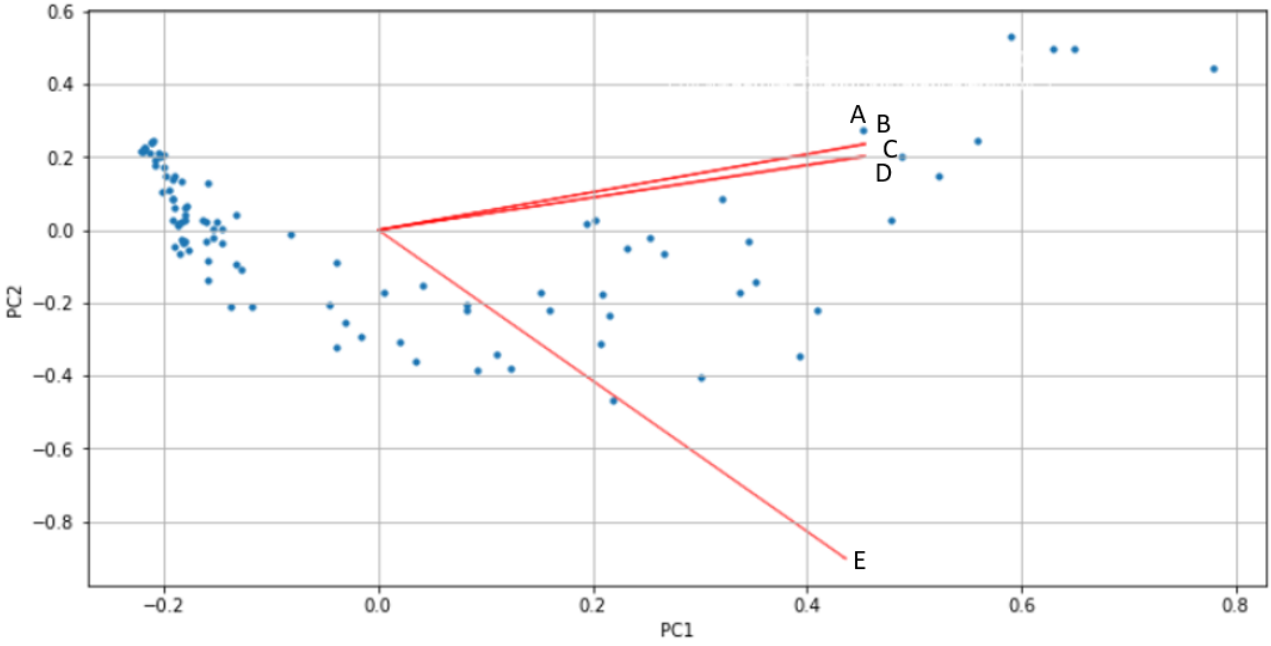


Figure 39: Biplot of the PC clusters and the vectors of the five most relevant features.

Table 11: Legend for biplot figure 39.

Letter	Radiomics feature
A	PET Uptake Metrics - Original max
B	Statistics - maximum
C	Intensity histogram - maximum
D	glcmFeatures2Davg - difference entropy
E	Intensity volume - int at vol fraction 10

Figure 38 shows that even based on five radiomics features, there still is a statistical difference between the healthy and NSCLC tissue. Enough difference to indicate the clustering of both datasets. To visualize this, the scatterplots of the PCA of all, the best 30, and the best five features are shown in Figure 40.

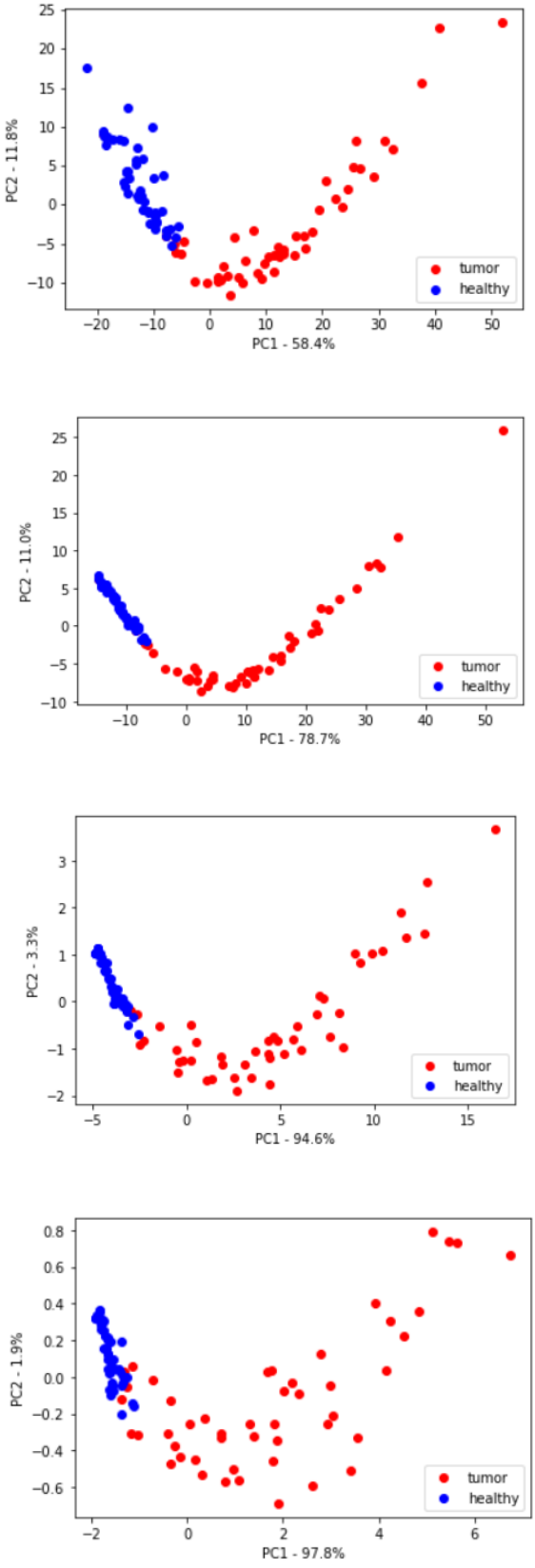


Figure 40: Top to bottom: Scatter Plot of PCA with all; noise reduced; 30 most significant and 5 most significant features.

Figure 40 indicates the clustering of the different tissues for all four PCAs. The cluster of the tumor tissue appears less cohesive when fewer features are taken into account. The healthy tissue data points on the other hand show tighter clustering when fewer features are used in the PCA. When using the 269 features after noise reduction, the healthy tissue cluster appears densest. Fewer features are relevant for determining healthy tissue compared to tumor tissue based on the density of those clusters with lower feature amounts. Another aspect to note is that the border between both clusters becomes less obvious when using a limited amount of features. Overall, the noise-reduced set of features indicates the best clustering for the data.

8.4 Classification Learning

To train the models, the patient cohort was subdivided into two groups. The first 30 patients, Group 1, are used to train the models in Matlab. The second group of 19 patients, Group 2, is used to test the model. For testing, the model with the highest accuracy is used for each classifier. The main classifier the models are trained on is 'tumor', to see whether the machine learning models can differentiate healthy and tumor tissue based on the radiomics output. There is also demographic information available that can be used as classifiers. The extra classifiers that will be investigated in this paper are diabetes, glycemia levels, tumor location (left or right lung), packyears, and tumor phenotype.

8.4.1 Tumor vs healthy tissue

After testing the first 30 patients on all the different models, the models in Table 12 give the best accuracy:

Table 12: The 10 models that give the best accuracy for predicting tumor and healthy tissue.

Model	Accuracy (%)
Fine Tree	96.7
Medium Tree	96.7
Coarse Tree	96.7
Linear Discriminant	96.7
Quadratic SVM	96.7
Cubic SVM	96.7
Fine KNN	**98.3**
Weighted KNN	**98.3**
Ensemble - Bagged Trees	**98.3**
Ensemble - Subspace Discriminant	**98.3**

The four last models give the best accuracy, the confusion matrices of these four models are shown in Figure 41.

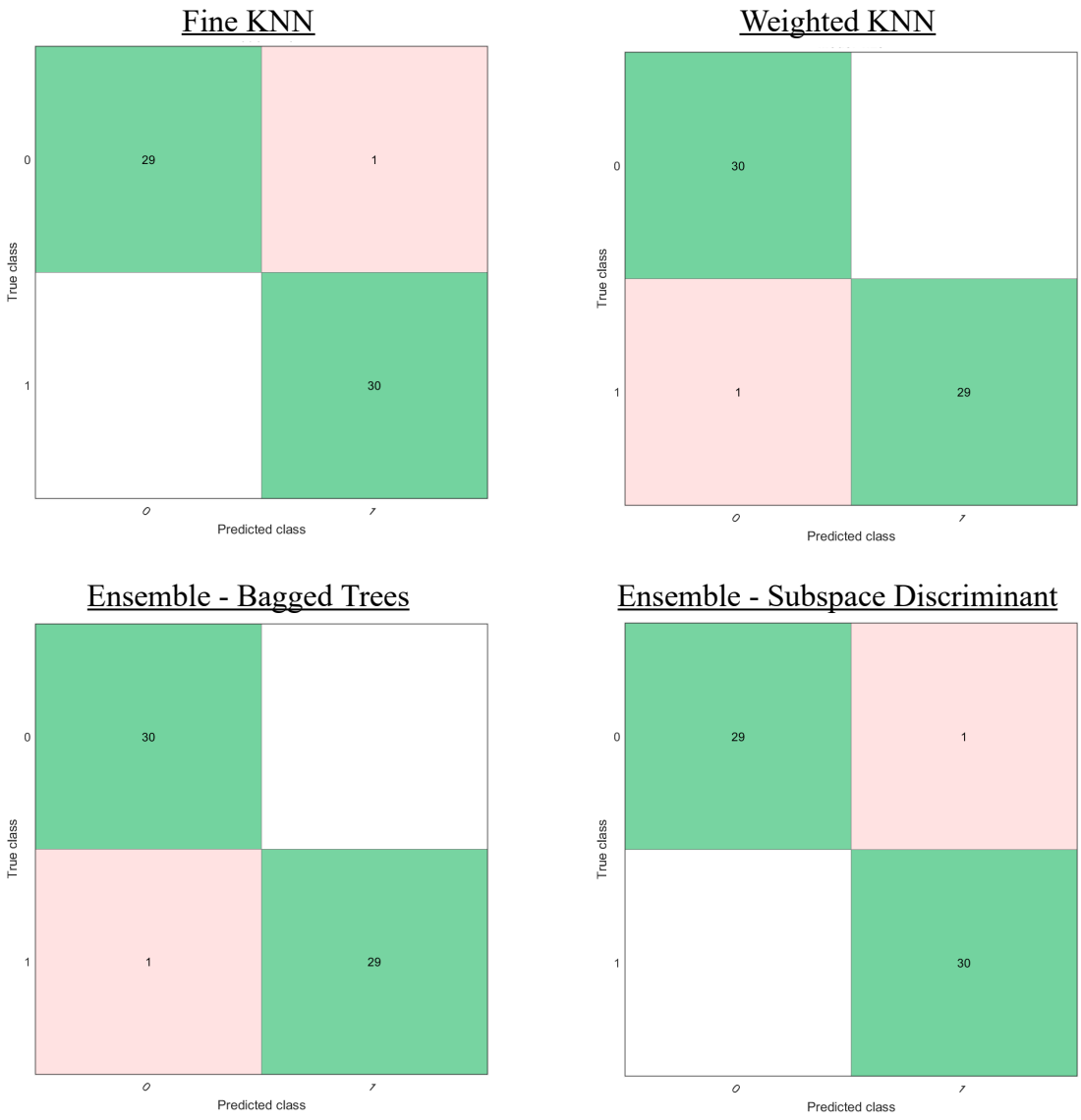
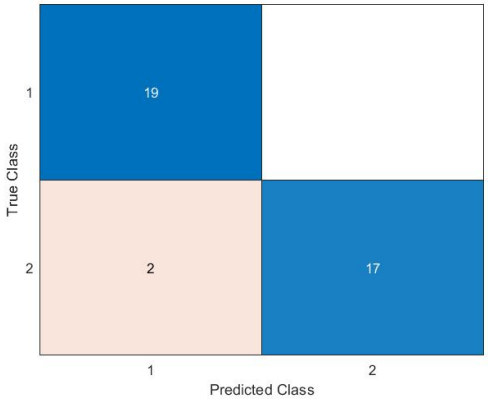


Figure 41: Confusion matrix for the Fine KNN, the Weighted KNN, Bagged Trees (Ensemble), and Subspace Discriminant (Ensemble) model for predicting tumor (1) or healthy (0) tissue of a cohort of 30 patients.

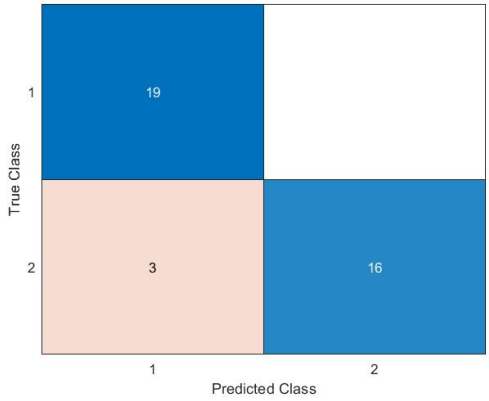
For the Fine KNN and Subspace Discriminant model, the healthy tissue of the ProLung63 patient is misclassified. For the Weighted KNN and Bagged trees model, the tumor tissue of the ProLung11 patient is misclassified. The scatter plots of the models for the two first features can be found in Annex I, Figures S1-S4. Here, the blue dots represent the healthy tissue and the orange dots represent the tumor tissues. The cross is the misclassified one.

These four models are then used to predict the other 19 patients. The confusion matrices are shown in Figure 42.

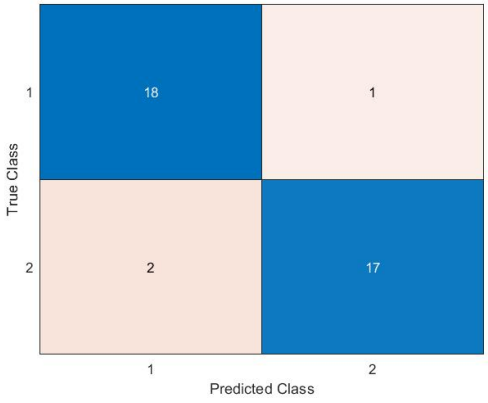
Fine KNN



Weighted KNN



Ensemble - Bagged Trees



Ensemble - Subspace Discriminant

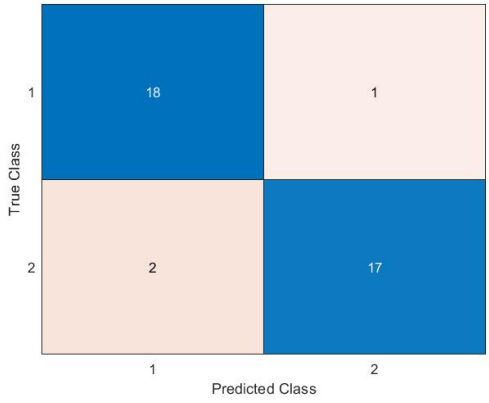


Figure 42: Confusion matrix for the Fine KNN, the Weighted KNN, Bagged Trees (Ensemble), and Subspace Discriminant (Ensemble) model for predicting tumor (2) or healthy (1) tissue of a cohort of 19 patients.

The patients that are misclassified in these models are shown in Table 13:

Table 13: Misclassified patients

Patient	Fine KNN	Weighted KNN	Bagged Trees	Subspace Disc.
ProLung098_tumor				x
ProLung101_gezond				x
ProLung102_tumor		x		
ProLung107_gezond			x	
ProLung112_tumor	x	x	x	
ProLung113_tumor	x	x	x	x

The model Fine Tree with an accuracy of 96.7% is also an interesting model because it gives a tree only one feature to classify the tumor and healthy tissues. The decision tree of the Fine Tree model that is trained on a cohort of 30 patients is shown in Figure 43.

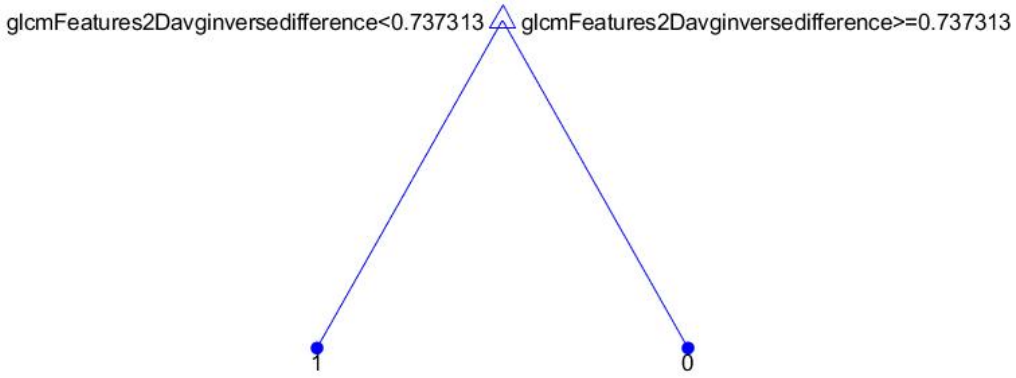


Figure 43: Decision tree of the Fine Tree model for predicting tumor (1) or healthy (0) tissue of a cohort of 30 patients.

8.4.2 Glycemia

The median of the glycemia is 98 mg%. The goal is to classify the patients to know if the amount of glycemia is larger or smaller than 98 mg%. After testing the first 30 patients on all the different models, the two models that give the best results are the Logistic Regression model with an accuracy of 66.7%, and the Ensemble - Subspace Discriminant model with an accuracy of 53.3%. The confusion matrices of these two models are shown in Figure 44.

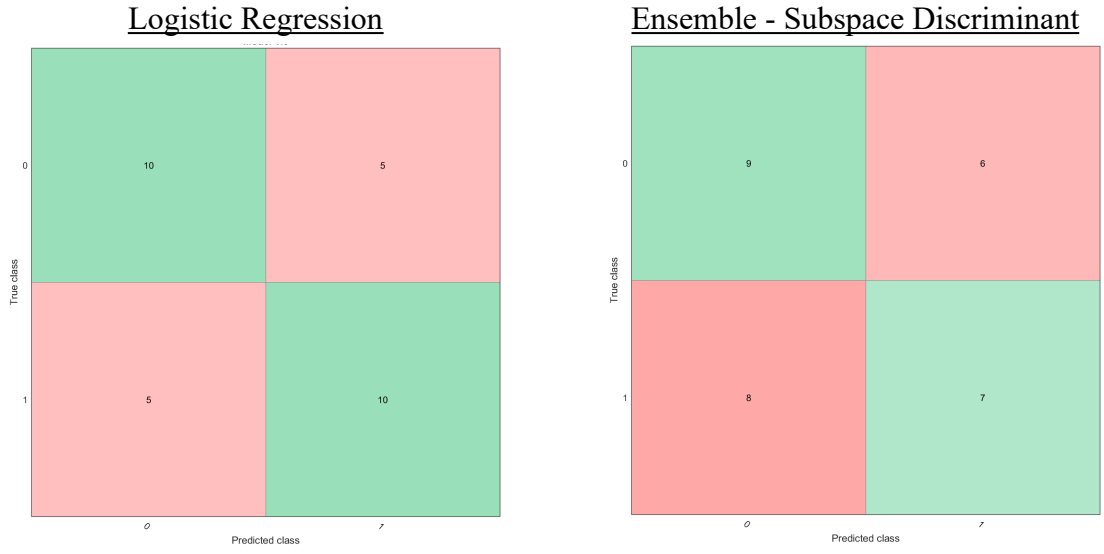


Figure 44: Confusion matrix for the Logistic Regression model and Subspace Discriminant (Ensemble) model for predicting if the glycemia is larger than 98 mg% (1) or smaller than 98 mg% (0) of a cohort of 30 patients.

These two models are then used to predict the other 19 patients. The confusion matrices are shown in Figure 45.

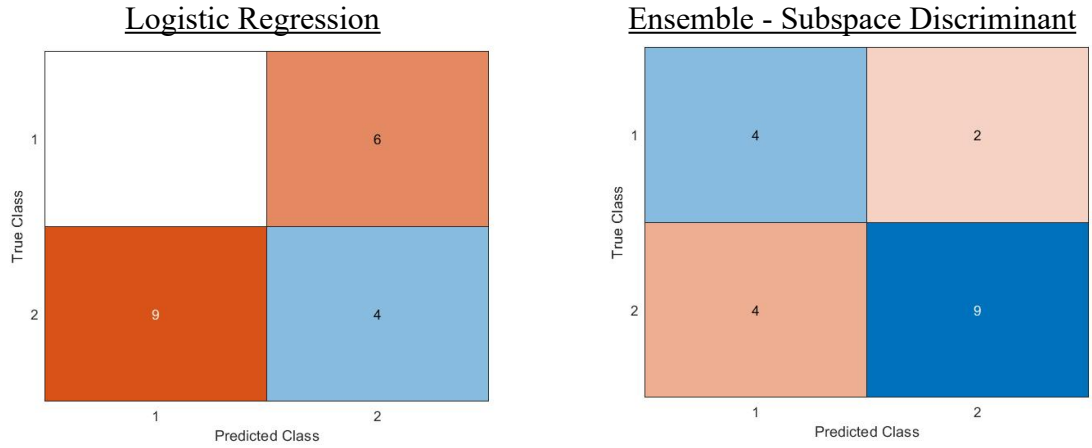


Figure 45: Confusion matrix for the Logistic Regression and Subspace Discriminant (Ensemble) model for predicting if the glycemia is larger than 98 mg% (2) or smaller than 98 mg% (1) of a cohort of 19 patients.

For the Logistic Regression model, only 4 out of 19 patients are classified correctly and for the Subspace Discriminant model classifies 13 out of 19 patients correctly.

8.4.3 Tumor type

In this study, the three classes of tumor types are neuroendocrine (NE), adenocarcinoma (AC), and squamous cell carcinoma (SCC). For this test, the three patients with neuroendocrine are not taken into account since neuroendocrine tumors use different metabolic pathways. Therefore, the results of comparing them to the other tumor phenotypes would be insignificant. Of the 30 patients that are used to train the different models, 23 patients have an AC and 7 have a SCC.

After testing the first 30 patients on all the different models, the Medium Gaussian SVM model gives the best accuracy (86.2%). The confusion matrix of this model is shown in Figure 46.

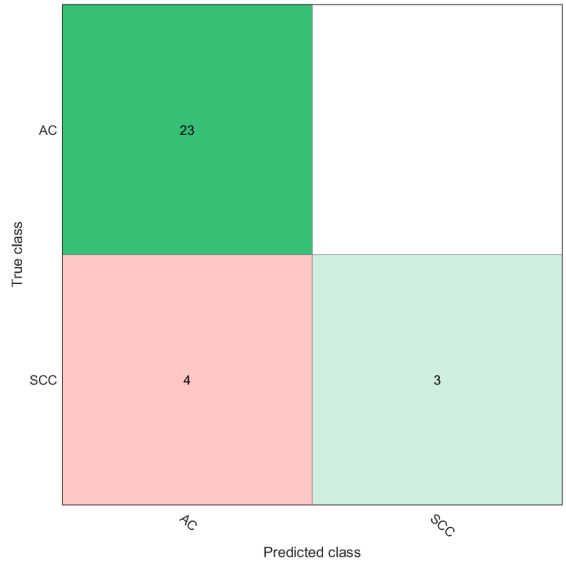


Figure 46: Confusion matrix for the Medium Gaussian SVM model for predicting the tumor type of a cohort of 30 patients.

This model is then used to predict the other 16 patients. Of these 16 patients, 12 patients have an AC and 4 have a SCC. The confusion matrices are shown in Figure 47.

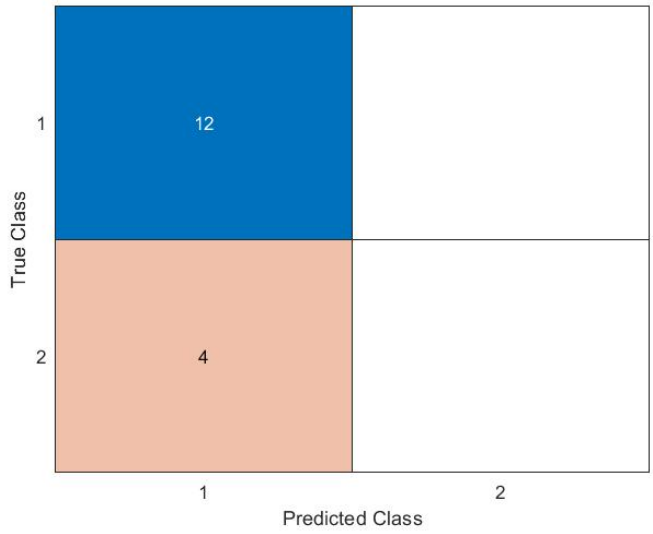


Figure 47: Confusion matrix for the Medium Gaussian SVM model for predicting the tumor type of a cohort of 16 patients.

This shows that 12 of the 16 are classified correctly, these are all AC tumors. All four of the SCC tumors are misclassified.

8.4.4 Lung side (left or right)

Next, the goal is to see if the models are capable of classifying the tumor's location, where L stands for the left lung and R for the right lung. After testing the first 30 patients on all the different models, the Logistic Regression model gives the best accuracy (73.3%). The confusion matrix of this model is shown in Figure 48.

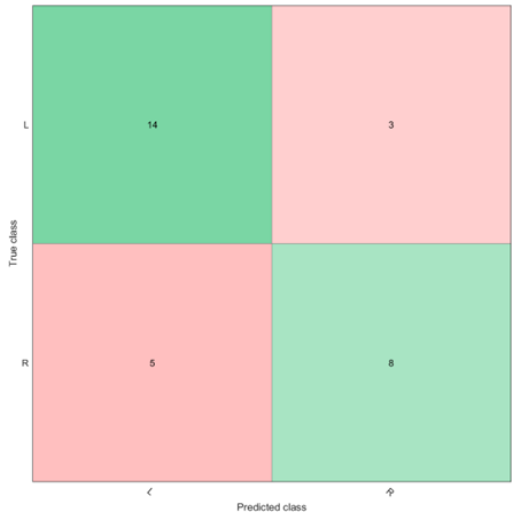


Figure 48: Confusion matrix for the Logistic Regression model for predicting the position of the tumor of a cohort of 30 patients

This model is then used to predict the other 19 patients. The confusion matrices are shown in Figure 49.

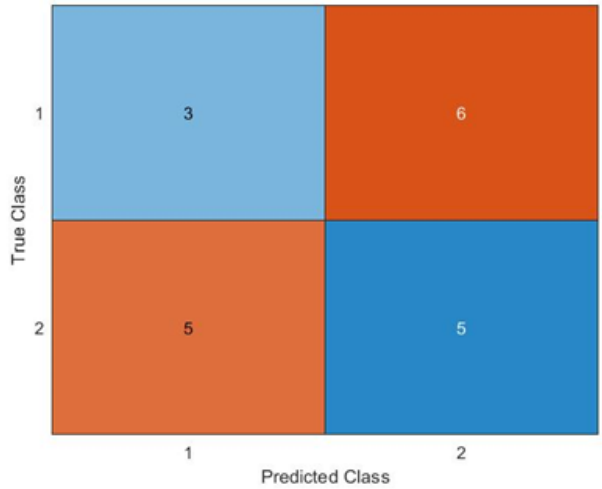


Figure 49: Confusion matrix for the Logistic Regression model for predicting the position of the tumor of a cohort of 19 patients.

This shows that only 8 out of 19 are classified correctly.

8.4.5 Diabetes

Next, the goal is to see if the models are capable of classifying the patients on having diabetes or not. After testing the first 30 patients on all the different models, the two models that give the best results are the Linear Discriminant model with an accuracy of 60.0%, and the Fine KNN model with an accuracy of 63.3%. The confusion matrices of these two models are shown in Figure 50.

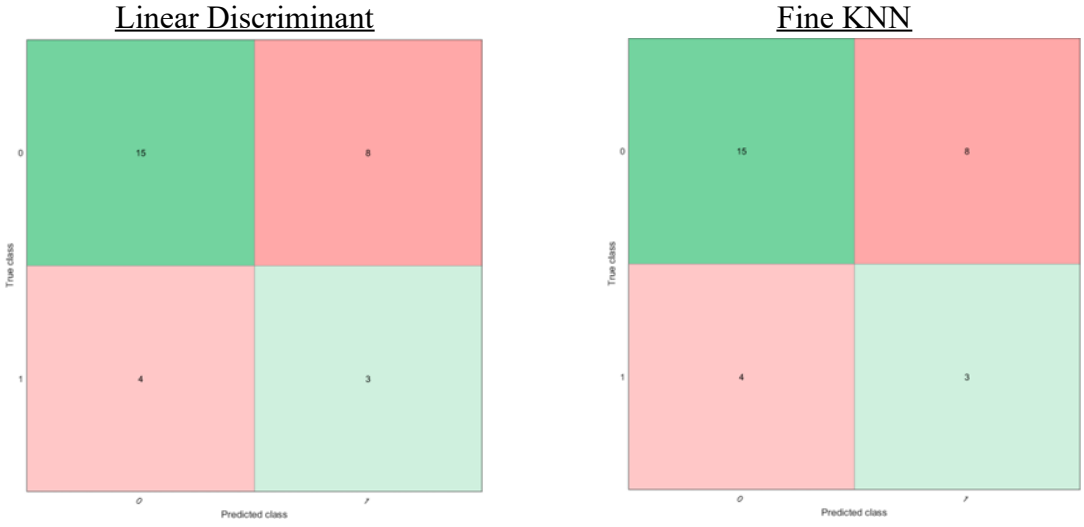


Figure 50: Confusion matrix for the Linear Discriminant model and Fine KNN model for predicting if the patient has diabetes (1) or not (0) of a cohort of 30 patients.

These models are then used to predict the other 19 patients. The confusion matrices are shown in Figure 51.

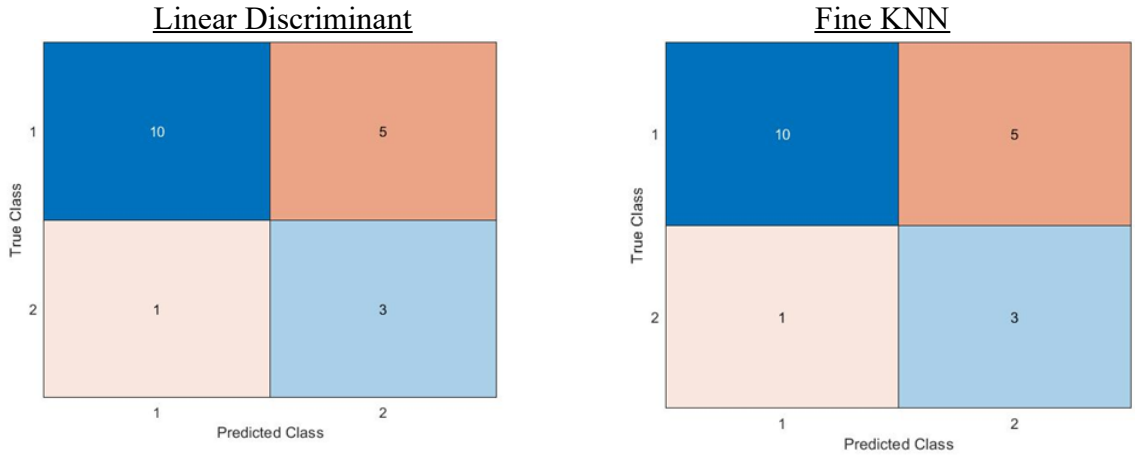


Figure 51: Confusion matrix for the Linear Discriminant and Fine KNN model for predicting if the patient has diabetes (2) or not (1) of a cohort of 19 patients.

For the linear discrimination model, 13 out of 19 predictions are correct. The Fine KNN model also predicted 13 out of 19 cases correctly.

8.4.6 Packyears

At last, the models are trained to study the number of the patients' packyears. Packyears can be defined as a number of years smoking one pack of cigarettes daily. The packyears of 1 patient are unknown, so these are not taken into account in this study. The median of the number of packyears is 35. The dataset is split into patients with more or equal than 35 packyears, and less than 35 packyears. The effect of the amount of packyears is tested on the dataset with the tumor tissue, and on the dataset with the healthy tissue.

Tumor tissue

First, all models are trained on the first 30 patients of the tumor tissue dataset. The RUS Boosted Trees (ensemble) model had the best accuracy, 60.0%. The confusion matrix of this model is shown in Figure 52.

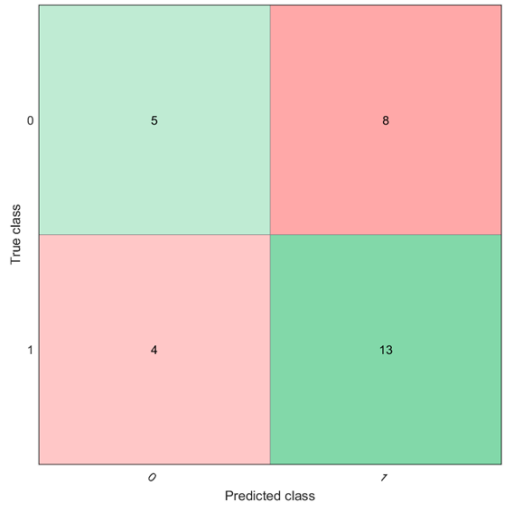


Figure 52: Confusion matrix for the RUS Boosted Trees (ensemble) mode for predicting if the amount of packyears is larger than 35 (1), or smaller (0) of the tumor tissue of a cohort of 30 patients.

This model is then used to predict the other 18 patients. The confusion matrices are shown in Figure 53.

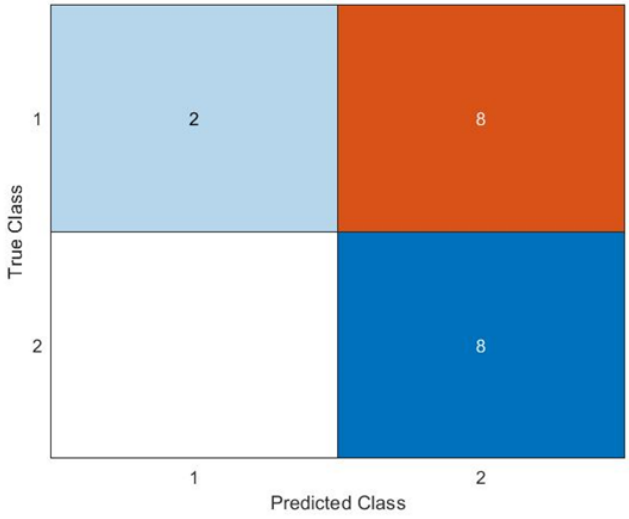


Figure 53: Confusion matrix for the RUS Boosted Trees (ensemble) model for predicting if the amount of packyears is larger than 38 (1), or smaller (0) of the tumor tissue of a cohort of 18 patients.

This shows that 10 out of 18 predictions are correct.

Healthy tissue

Now, all models are trained on the first 30 patients of the healthy tissue dataset. The Subspace KNN (ensemble) model has the highest accuracy, 56,7%. The confusion matrix of this model is shown in Figure 54.

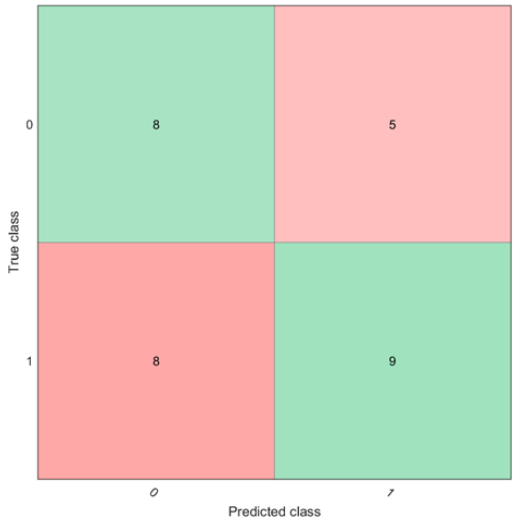


Figure 54: Confusion matrix for the Subspace KNN (ensemble) model for predicting if the amount of packyears is larger than 38 (1), or smaller (0) of the healthy tissue of a cohort of 30 patients.

This model is then used to predict the other 18 patients. The confusion matrices are shown in Figure 55.

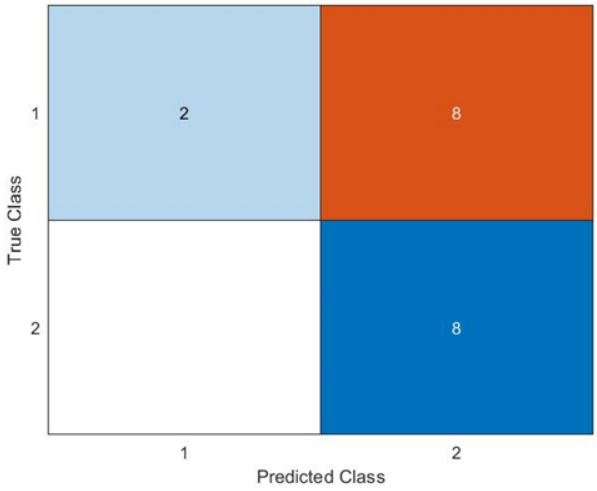


Figure 55: Confusion matrix for the Subspace KNN (ensemble) model for predicting if the amount of packyears is larger than 38 (1), or smaller (0) of the healthy tissue of a cohort of 18 patients.

This shows that 10 out of 18 predictions are correct.

To verify these results, a PCA is performed using 12 patients from the 25th percentile and 12 patients from the 75th percentile. The noise-reduced 269 features are used to perform the analysis resulting in Figure 56.

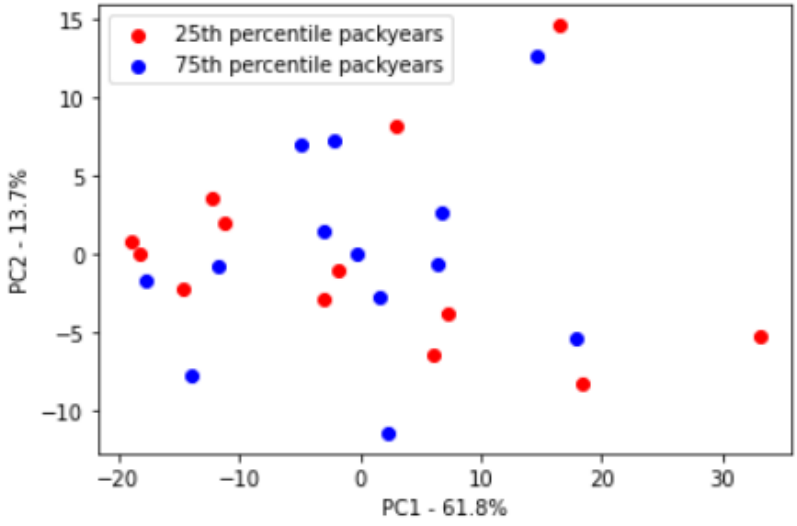


Figure 56: Scatter Plot of PCA of packyears.

Figure 56 shows no clear clustering of tissue of heavy smokers (more than 35 packyears) compared to patients who have less than 35 packyears.

This group of two times 12 patients is finally used to train and test the models again. 16 patients are randomly chosen to train the models, the best model is then used to make predictions about the remaining eight patients. The highest accuracy achieved by the training models was 43.8% by both the fine and coarse Gaussian SVM models. This accuracy is too low to meaningfully make predictions on the remaining eight patients.

8.5 Classification learning on the noise-reduced dataset

In the PCA test, the noise-reduced dataset of 269 radiomics features performed best in clustering the patients. These features are therefore now used for the classification learner to see if the results are better than the results on the t-test corrected dataset. The models are again trained to differentiate healthy and tumor tissue, and on the extra classifiers diabetes, glycemia levels, tumor location (left or right lung), packyears, and tumor phenotype. The first 30 patients, Group 1, are again used to train the models in Matlab. The second group of 19 patients, Group 2, is used to test the model. The confusion matrices of these tests can be found in Annex II (figures S5-S16).

8.5.1 Tumor vs healthy tissue

The tests on the noise-reduced dataset for differentiating healthy and tumor tissue return only two models with an accuracy of 98.3%. These are the Fine Gaussian SVM model and the Subspace Discriminant (ensemble) model. After testing these models on the other 19 patients, The Fine Gaussian SVM came out as the best by only misclassifying 2 out of 19 patients. These are the same patients as the Fine KNN model, which came out as the best in 8.4.1. The Subspace Discriminant (ensemble) model misclassified five patients.

8.5.2 Glycemia

Next, the models are trained to classify whether the amount of glycemia is larger or smaller than 98 mg%. The model that came out best is the Logistic Regression model with an accuracy of 80%, which is a better result than the model in 7.4.2. After testing these models on the other 19 patients, this model classified 10 out of 19 patients correctly. This is worse than the model in 8.4.2.

8.5.3 Tumor type

For the classification of tumor types, the three patients with neuroendocrine are again not taken into account. After testing the first 30 patients on all the different models, the Medium Gaussian SCM model came out as the best with an accuracy of 90%. This is a little more than the result in 8.4.3, where the accuracy was 86.2%. This model is then used to predict the

other 16 patients. 5 out of 16 patients are misclassified. This is one more than the model in 8.4.3. There, four patients with an AC were misclassified, and in this model four patients with an AC and one with a SCC.

8.5.4 Lung side (left or right)

Next, the goal is to see if the models are capable of classifying the tumor's location, where L stands for the left lung and R for the right lung. After testing the first 30 patients on all the different models, the Logistic Regression model gives the best accuracy (70.0%). After using this model to make predictions on the other 19 patients, only 7 out of 19 patients were classified correctly. This is one less than in 8.4.4.

8.5.5 Diabetes

Next, the goal is to see if the models are capable of classifying the patients on having diabetes or not using the tumor VOI. After testing the first 30 patients on all the different models, nine models returned the same accuracy (76.7%) and confusion matrix. All of these models however predicted none of the patients had diabetes. The models in question are: fine, medium, and coarse Gaussian SVM; fine, medium, coarse, cosine, and cubic KNN, and the boosted trees ensemble learning model. The next best models were all the Tree models with an accuracy of 73.3%. These models did make diabetes predictions. To make predictions for the other 19 patients, the Fine Tree model and the Fine KNN returned the best results. The Fine Tree predicted 12 out of 19 correctly, and the Fine KNN 14 out of 19. In 8.4.5, 13 out of 19 patients were classified correctly.

8.5.6 Packyears

At last, all models are trained on the first 30 patients of the healthy tissue dataset. The Kernel naive Bayes model has the highest accuracy, 70.0%. After using this model to make predictions on the other 18 patients, only 14 out of 18 patients were classified correctly. This is four patients more than in 8.4.6.

9 Relevant radiomics features

From the entire patient cohort, 498 radiomics features and 6 PET Uptake Metrics were collected per patient using the Radiomics tool. These features are subdivided into groups as described in Chapter 4.1.1.

From the statistical tests performed, 30 parameters appeared to be more relevant. These features originate from the following subgroups: gray level co-occurrence matrix (GLCM), intensity histogram (first order features), intensity volume (3D-shape features), and statistics (first order features). All features are aggregated using the 3D VOI except for some features from the GLCM. These features will be elucidated in this chapter.

9.1 Gray level co-occurrence matrix (GLCM) features

To quantify combinations of discretized gray levels of neighboring pixels (2D) or voxels (3D), distributed along an image direction, GLCMs can be used. By aggregating information from the different underlying directional matrices, GLCM feature values are computed with improved rotational invariance. The following aggregation methods can be used as described by Zwanenburg A. et al [53].:

- Features are computed from each 2D directional matrix and averaged over 2D directions and slices (2Davg).
- Features are computed from a single matrix after merging 2D directional matrices per slice and then averaged over slices (2Dmrg).
- Features are computed from a single matrix after merging 2D directional matrices per direction and then averaged over directions (2DDmrg).
- The feature is computed from a single matrix after merging all 2D directional matrices (2Dvmrg).
- Features are computed from each 3D directional matrix and averaged over the 3D directions (3Davg).
- The feature is computed from a single matrix after merging all 3D directional matrices (3DWmrg)

9.1.1 Difference entropy

The difference entropy feature is defined as a measure of the randomness/variability in neighborhood intensity value differences [51]. The 2Davg, 2Dmrg, 2DDmrg, and 3Davg aggregation methods are of interest, for a total of four GLCM difference entropy features. To calculate the difference entropy, Formula 3 can be used:

$$difference\ entropy\ =\ \sum_{k=0}^{N_g-1} p_{x-y}(k) log_2(p_{x-y}(k)\ +\ \epsilon) \quad (3)$$

Where

- ϵ is an arbitrarily small positive number ($\approx 2.2\times10^{-16}$),
- N_g is the number of discrete intensity levels in the image
- $p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g}$ p(i,j), where |i − j| = k, and k = 0,1,…,N_g − 1
- *p(i,j)* is the normalized co-occurrence matrix and equal to $\frac{P(i,j)}{\sum P(i,j)}$

9.1.2 Joint average

Joint average is the gray level weighted sum of joint probabilities [53]. Three features stemming from 2DDmrg, 3Davg, and 3DWmrg aggregation methods showed significance in this study. The average gray level intensity of the *i* distribution can be found using Formula 3 [51].

$$joint\ average\ =\ \mu_x = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)i \quad (4)$$

Where

- μ_x is the mean gray level intensity of p_x

9.1.3 Joint entropy

Joint entropy is similar to difference entropy in that it measures the randomness/variability in neighboring intensity values, but it looks at similarities instead of differences [51]. Three GLCM joint entropy features from 2DDmrg, 2Dvmrg, and 3DWmrg aggregation techniques were found to be significant. The joint entropy feature can be quantified by Formula (5).

$$joint\ entropy\ =\ -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) log_2(p(i,j)\ +\ \epsilon) \quad (5)$$

9.1.4 Sum average

The sum average feature of the GLCM class radiomics features is a measurement of the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values. Mathematically, the sum average is twice the joint average from chapter 8.1.2 [51]. Three features aggregated and computed via 2DDmrg, 3Davg, and 3DWmrg showed relevance for the purposes of this study. Using Formula 6, the sum average can be found.

$$sum\ average\ =\ \sum_{k=2}^{2N_g} p_{x+y}(k)k\ =\ 2\cdot joint\ average \tag{6}$$

Where

- $p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)$, where i + j = k, and k = 2,3,…,2N_g

9.1.5 Sum entropy

The sum entropy feature is defined as the sum of neighboring intensity value differences [51]. One feature using 2Dmrg aggregation showed significance. This feature can be calculated using Formula 7.

$$sum\ entropy\ =\ \sum_{k=2}^{2N_g} p_{x+y}(k)log_2(p_{x+y}(k)\ +\ \epsilon) \tag{7}$$

9.1.6 Inverse difference

GLCM - Inverse difference is a measure of localized homogeneity in an image. Co-occurrences in the matrix with significant differences in gray levels are weighed less. When the gray levels are equal, this feature is maximal [53]. This feature was the basis for differentiating healthy and tumor tissue in the Tree model in chapter 8.3.1 when aggregated using 2Davg. The calculation of this feature can be done using Formula 8 [51].

$$inverse\ difference\ =\ \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k} \tag{8}$$

9.2 Intensity histogram

By discretizing the original intensity distribution of the VOI into intensity bins, an intensity histogram is made. [53]. All features describe a 3D volume.

9.2.1 Entropy

The intensity histogram - entropy feature explains the uncertainty/randomness in the image values by measuring the average amount of information needed to encode those values [51]. Entropy being an information-theoretic concept is here defined as Shannon entropy and therefore is determined by Formula 9.[53].

$$entropy = -\sum_{i=1}^{N_g} p(i) log_2(p(i) + \epsilon) \quad (9)$$

Where

- p(i) is the normalized first order histogram and equal to $\frac{P(i)}{N_p}$

9.2.2 Statistical intensity histogram features

The intensity histogram class of radiomics features contained six more interesting features as indicated by the results in this study. These are statistical discretized features of the created histogram and their formulas are briefly summarized here [51],[53].

Interquartile range of the gray level values in the image array.

$$interquartile\ range = P_{75} - P_{25} \quad (10)$$

Range of occurring gray levels in the VOI.

$$range = max(X) - min(X) \quad (11)$$

Where

- X is a set of N_p voxels included in the VOI

Maximum gray level intensity in the segmented VOI.

$$maximum = max(X) \quad (12)$$

Median absolute deviation.

$$Median\ absolute\ deviation = \frac{1}{N_p}\sum_{i=1}^{N_p} |X(i) - M| \quad (13)$$

Mean absolute deviation, the mean distance of all intensity values from the Mean Value of the image array.

$$Mean\ absolute\ deviation = \frac{1}{N_p}\sum_{i=1}^{N_p} |X(i) - \overline{X}| \quad (14)$$

Robust mean absolute deviation, the mean absolute deviation calculated on the subset of the image array with gray levels in between, or equal to the 10th and 90th percentile.

$$Robust\ mean\ absolute\ deviation = \frac{1}{N_{10-90}}\sum_{i=1}^{N_{10-90}} \left|X_{10-90}(i) - \overline{X_{10-90}}\right| \quad (15)$$

9.3 Intensity volume

The intensity volume class of features expresses the relationship between a discretized intensity *i* and the fraction of the volume containing this intensity level or more [51]. One feature from this class showed promise in the ability to differentiate healthy lung tissue from NSCLC tissue.

9.3.1 Intensity at volume fraction 10

The intensity at volume fraction 10 is the minimum intensity present in a maximum of 10% of the VOI. Generally, only fractions 10 and 90 are used as radiomics parameters [87].

9.4 Statistical features

To describe the general distribution of the gray level intensities within a VOI, statistical features are used. Unlike the statistical features concerning the intensity histogram (chapter 9.2), these features are not discretized and can therefore be used to describe continuous intensity distributions [51]. The six relevant statistical features for this study are maximum, 90th percentile, mean absolute deviation, median absolute deviation, range, and RMS. The calculations for the median and mean absolute deviation can be done using formulas 13 and 14. The RMS is found using Formula 15.

$$RMS = \sqrt{\frac{\sum_{k=1}^{N_g} X_k^2}{N_g}} \tag{15}$$

9.5 PET Uptake Metrics - Original maximum

The PET Uptake Metrics consists of five features that were also extracted for the entire patient cohort. These metrics provide a quantization of the SUV in the VOI. One of these parameters appeared useful in differentiating between healthy and tumor tissue, namely: Original maximum. This feature corresponds to the more commonly used SUV_{Max} parameter which measures tumor glucose metabolism. It is also a frequently used feature to quantify tumor FDG uptake [88]. Formula 1 from Chapter 3.2 can be used to calculate the SUV.

10 Discussion

The focal point and the primary hypothesis of this research paper proposed the ability to distinguish healthy from NSCLC tissue using ^{18}F-FDG-imaging and radiomics. This has already shown promising results for several organs, including the liver, lungs, and prostate.[86],[89-90].

In this study, the radiomics extraction returned two sets of 504 features concerning the tumor and healthy VOI. Not all features are relevant and overfitting is a common issue in radiomics analyses [91-92]. A paired t-test was used to remove 69 features from the dataset. Furthermore, a PCA is a useful technique to reduce noise and is commonly used in radiomics studies [93-95] This analysis uncovered 166 features or 38.1% of the remaining features which proved less relevant. The clustering of the data points after this noise reduction improved greatly, indicating a clear difference between healthy tissue and tumor tissue radiomics features. When using the 30 most significant features, as indicated by the noise-reduced PCA, good clustering could still be observed. Only incorporating the top five features however generated clusters that appear less dense. In general, the cluster of the tumor tissue is more spread out than the cluster of the healthy tissue. This is a clear example of the Anna Karenina principle since the healthy tissue of the patients is more homogeneous and consistent compared to tumor tissue which comes in all shapes and sizes [96]. The tumor VOI of ProLung113 was clustered in the healthy tissue group for all PCAs. After reviewing the PET image and the tumor nodule, most of the VOI consisted of central necrosis which affects the SUV and therefore the gray levels in the image. This could explain the consistent miss clustering.

The dendrogram showed 8 clusters. The reason that features are in the same cluster is that they have the same order of magnitude. The dataset consists of different groups, such as GLCM, GLRLM, etc. The clusters that were formed are not linked to a specific group but to the quantities. For example, all the entropy values of the different groups are in one cluster. The 30 most significant features are spread over the different clusters. Therefore, there is no prominent cluster in this dendrogram. The PCA biplots showed grouping vectors indicating mutual correlation. These clusters did not however match up with the clusters found in the dendrogram.

The second objective was to find a discriminative model for healthy and tumor lung tissue using machine learning techniques. Machine learning in high dimensional datasets such as radiomics has gained popularity in recent years, specifically for the purposes of cancer staging, segmentation, and general -omics studies [97-98]. Other clinical parameters were also tested using the same machine-learning methods. These were glycemia, tumor phenotype, VOI location, diabetes, and packyears. The patient cohort was subdivided into a group of 30 patients used for training the models and a second group of 19 patients to then test the best model(s).

The first classifier was based on whether the VOI concerned ‘tumor’ or ‘healthy’ tissue. After training all different models for distinguishing the healthy tissue from the NSCLC tissue on the first 30 patients, the Fine KNN, Weighted KNN, Bagged trees, and Subspace Discriminant came out as the best models with an accuracy of 98.3%. They only classified one patient tissue wrong. After using these four models to make predictions on the other 19 patients, the KNN model gave the best result by only classifying two of the 19 patients wrong, corresponding to an accuracy of 89.5%. All four models classified ProLung113 wrong. This is the same patient that came out of the PCA test as misclassified. The other misclassified tissues are different for the different models. The Fine Tree model gave an accuracy of 96.7%. The tree model only uses the ‘glcmFeatures2Davg-inverse difference’. When this value is smaller than 0.737313, it is classified as a tumor. The inverse difference is a measure of homogeneity. A tumor tissue in general is more heterogeneous than healthy tissue, which makes this feature a good indication for discriminating between the two tissue types. In general, the results of the classification models are good, compared to previous studies where the accuracy has a value of 97.3%. For CT images, radiomics features have shown to be a good method in differentiating healthy from tumor tissue in pancreatic cancer [99]. For PET, radiomics coarseness features showed the ability to distinguish carcinoma from healthy tissue [100].

Next, glycemia was used as a differentiating parameter. The Logistic Regression model with an accuracy of 66.7%, and the Ensemble - Subspace Discriminant model with an accuracy of 53.3% came out on top. These models correctly predicted 4 and 13 out of 19 patients respectively. Similar results are found in other studies. In a study by M. Eskian et al., using 8380 patients, glycemia was significantly correlated with decreased SUV in brain and muscle tissue and increased SUV in liver tissue. Contrarily, no correlation was found between SUV and glycemia in tumor tissue [101]. This was also concluded in a study by K. A. Büsing et al., which found no impact of diabetes and glycemia levels on SUV [102]. The findings in this study lean towards the latter and the created models are not sufficient in distinguishing tumor VOIs based on the patients’ glycemia levels.

Tumor phenotype was the next parameter. The two types of nodules studied in this paper are adenocarcinoma (AC) and squamous cell carcinoma (SCC). The noise-reduced set of radiomics features produced the best results. The Medium Gaussian SVM model achieved an accuracy of 90%. The same model predicted 11 out of 16 correctly. Using radiomics and MRI on different brain tumor types, a study by F.J. Diaz-Pernas et al. reported increased performance in classifying the types into meningioma, glioma, and pituitary tumors. [103]. Prognostic capabilities and tumor phenotype determination of radiomics features have also been found for lung and head-and-neck cancer[48]. The high accuracy of the trained Medium Gaussian SVM model combined with positive findings in previous studies suggest further research should be done to predict lung tumor phenotype using radiomics features and machine learning.

Models able to determine the location of the tumor VOI were tested next. The Logistic Regression model gave the highest accuracy (73.3%). When predicting the location in the

second patient group, 8 out of 19 predictions were correct. The location of the VOI has been shown to affect feature quantification. [86]. This study was not able to build a model capable of consistently predicting VOI location between the left and the right lung.

The results from the classifier learner for diabetes indicated that the Linear Discriminant model with an accuracy of 60.0%, and the Fine KNN model with an accuracy of 63.3% were the best models. Both models predicted 13 out of 19 cases correctly in the test group. Linked with glycemia, no relevant correlations have been found between diabetes and SUV uptake in tumor tissue. [102]. This was extended to lung cancer specifically in a study by Gorenberg M. et al. [104]. The accuracy of the models in this study is too low to reasonably predict diabetes based on radiomics features.

The final clinical parameter used for training the models was packyears. The median of packyears (35) for the patient cohort was used to make the two groups since there was only one non-smoker in this study. The RUS Boosted Trees (ensemble learning) model achieved the highest accuracy for the tumor VOIs, with 60.0%. This corresponds to 12 out of 30 patients being misclassified. When making predictions on the remaining 18 patients, the model correctly predicted ten cases. For healthy tissue, the Subspace KNN (ensemble learning) model has the highest accuracy, 56,7%. This model predicted 10 out of 18 patients correctly from the test group. To create a more significant difference between both groups in healthy tissue, the patients under the 25th percentile and above the 75th percentile were used. to make two new groups. The models attained a maximum efficiency of 43.8% and the PCA did not reveal any data clustering even though the difference in tobacco burden was more significant. It has been shown that SUV using FDG correlates with tobacco burden in current smokers. This effect diminishes after the cessation of smoking [105]. Another study by Schroeder et al. demonstrated increased uptake of FDG in smoke-exposed lungs compared to control lungs [106]. The combination of smoking and diabetes has been shown to increase FDG uptake as well [107]. A study by D.A. Torigian et al. on the other hand found no relation between smoking and FDG uptake [108]. Due to the lack of non-smokers in this study, the median of packyears was used to divide the patient cohort into two groups. The models are not sufficient in determining the difference between both groups of (mostly) smokers. Refining the tests by making the gap between both groups bigger did not yield better results.

In summary, this study found separated clustering of healthy tissue and tumor tissue following a PCA. This clustering was optimal after removing noise and still apparent when using only 30 features with the highest loading scores. The discriminative models were capable of classifying the two tissue types with an accuracy of 89.5%, based on a training model with an accuracy of 98.3%.

The method of segmentation has a substantial effect on these results. In this study, the segmentation of the VOIs was exclusively conducted on ^{18}F-FDG PET images. A study by Lu et al. found that the selected segmentation method had an impact on the quantification of radiomics features. Similarly, differences were observed in the radiomics data obtained from

^{18}F-FDG and 11C-choline. The study observed the influence of the chosen imaging technique and the used radiopharmaceutical [109]. It is also important to emphasize the consistency of the radiomics data, as the delineation of tumors on the PET-CT images was semi-automatically performed by a single radiologist. A previous study by Zhao et al. demonstrated inconsistencies in VOI delineation when three different radiologists were involved [110]. These possible pitfalls in radiomics research have also been summarized by Somasundaramet al., who also listed pre-processing and noise as possible challenges [111]. Generally, reproducibility is found to be dependent on the scanner or acquisition and reconstruction settings making validation of radiomics studies difficult [112-115]. The specific methodologies and conditions are detailed in this paper. Nevertheless, we acknowledge that the reproducibility of this data using different technologies may lead to differing outcomes, as discussed by Gillies RJ et al. [8]. In their report, they enumerate various factors contributing to this phenomenon in radiomics data, including technical intricacies, data overfitting, incomplete result reporting, and unidentified confounding variables in the utilized databases.

This study concerns data from one healthy and one tumor VOI per patient and can therefore be labeled as pre/post intervention data. Consequently, the data is intrinsically paired which was not taken into account. A possible novel analysis technique for this datatype is proposed by P. Jonsson et al., namely orthogonal partial least squares-effect projections (OPLS-EP). This is a multivariate statistical analysis strategy allowing paired or dependent analysis of individual effects [116]. OPLS-discriminant analysis (OPLS-DA) has also been advised over PCA for similar studies since it uses separation based models instead of variance based models [117].

The sample size for training classifiers affects the models' accuracy. A smaller sample size can increase variance but removing outliers can decrease variance. The effects of sample size on machine learning is not studied enough to draw immediate conclusions as summarized by D. Rajput et al. [118]. Furthermore, guidelines and methodological conduct for machine learning clinical prediction models are limited [119].

Looking forward, it could prove useful comparing healthy tissue of smokers and non-smokers. This study included only patients with a smoking history, bar one, making it difficult to draw conclusions. Additionally, the same analyses can be done using CT segmented VOIs of the same patient cohort since these images were taken concurrently. This could expand the findings of this study to another modality as well as allowing for cross-modality comparisons. Furthermore, the methodology used in this study could be applied to tissues from other organs. To determine the validity of the machine learning results, sample-size calculations can be performed ex post facto by fitting a learning curve [120]. A larger patient cohort could prove useful in validating the results and allowing conclusions from the other clinical classifiers. We hypothesize that the models will be able to differentiate tumor from healthy lung tissue and smokers from non-smokers. Based on our findings and current findings in the literature, we expect no significantly different results for the other classifiers used in this study.

In the long term, this preliminary study can form the basis for automated detection and diagnosis of NSCLC. The results indicate a statistical difference in the radiomics data between healthy lung tissue and lung tumor tissue. Future studies can verify if this also goes for other tissues compared to an NSCLC VOI. A possible implementation could be a system that automatically subdivides a PET image into small cubic areas/VOIs and runs a radiomics analysis on each of the segmented cubes. From this data, the system can then decide if and where the patient has a tumor nodule by checking the grid elements corresponding to tumor tissue. This system can be AI-driven.

11 Conclusion

This study aimed to test the hypothesis that tumor and healthy lung tissue can be differentiated solely using ^{18}F-FDG PET-based radiomics data. Hereto, different discriminating classifiers were created with full and reduced datasets of radiomics features. Firstly, the PCA based on the full dataset showed separated clustering of the tumor and the healthy tissue. This differentiation was even more clear after noise-suppression, reducing the dataset to the 269 features with the highest loading scores. A set of 30 features still performed adequately. The second major finding was that four different machine learning models attained an accuracy of 98.3% in predicting the tissue type: Fine KNN, Weighted KNN, Bagged trees, and Subspace Discriminant models. The Fine Tree model achieved an accuracy 96.7% using only one radiomics feature: glcmFeatures2Davg-inverse difference. Machine learning models based on radiomics were not able to determine other clinical parameters, i.e. diabetes, glycemia levels, tumor location (left or right lung), number of packyears, and tumor phenotype (adeno- or squamous cell carcinoma).

This study adds promising results to the rapidly expanding field of radiomics. The identified models may assist in the further development of (semi)automated tools and provide a basis for further research in ^{18}F-FDG PET- based lung cancer diagnosis.

References

[1] “Cancer today,” Iarc.fr. [Online]. Available: https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=total&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=7&group_cancer=1&include_nmsc=0&include_nmsc_other=1&half_pie=0&donut=0. [Accessed 30 Mar 2023].

[2] G. Luo et al., “Projections of lung Cancer Incidence by 2035 in 40 countries worldwide: Population-based study,” JMIR Public Health Surveill., vol. 9, p. e43651, 2023.

[3] C. Allemani et al., “Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries,” Lancet, vol. 391, no. 10125, pp. 1023–1075, 2018.

[4] “Why is Lung Cancer So Deadly?,” NFCR, 30-Jul-2020. [Online]. Available: https://www.nfcr.org/blog/why-is-lung-cancer-so-deadly/. [Accessed: April 11 2023].

[5] B. Zhang, “Lung cancer,” NFCR, 10-Feb-2017. [Online]. Available: https://www.nfcr.org/cancer-types/lung-cancer/. [Accessed: March 11 2023].

[6] “Cancer,” Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed: March 10 2023].

[7] P. Lambin et al., “Radiomics: extracting more information from medical images using advanced feature analysis,” Eur. J. Cancer, vol. 48, no. 4, pp. 441–446, 2012.

[8] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images are more than pictures, they are data,” Radiology, vol. 278, no. 2, pp. 563–577, 2016.

[9] R. Manafi-Farid, N. Karamzade-Ziarati, R. Vali, F. M. Mottaghy, and M. Beheshti, “2-[18F]FDG PET/CT radiomics in lung cancer: An overview of the technical aspect and its emerging role in management of the disease,” Methods, vol. 188, pp. 84–97, 2021.

[10] C. Compton, Cancer: The enemy from within: A comprehensive textbook of cancer’s causes, complexities and consequences. Cham: Springer International Publishing, pp 26-31, 2020.

[11] “Stanford health care,” Stanfordhealthcare.org. [Online]. Available: https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html.
[Accessed: 10-Apr-2023]

[12] “Benign vs malignant tumors: What’s the difference?,” Cancer Treatment Centers of America, 21-Nov-2022. [Online]. Available: https://www.cancercenter.com/community/blog/2023/01/whats-the-difference-benign-vs-malignant-tumors. [Accessed: 10-Apr-2023]

[13] D. Hanahan, “Hallmarks of cancer: New dimensions,” Cancer Discov., vol. 12, no. 1, pp. 31–46, 2022.

[14] O. Warburg, “On the origin of cancer cells,” Science, vol. 123, no. 3191, pp. 309–314, 1956.

[15] R. J. Deberardinis, N. Sayed, D. Ditsworth, and C. B. Thompson, “Brick by brick: metabolism and tumor cell growth,” Curr. Opin. Genet. Dev., vol. 18, no. 1, pp. 54–61, 2008.

[16] R. Peto, S. Darby, H. Deo, P. Silcocks, E. Whitley, and R. Doll, “Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies,” BMJ, vol. 321, no. 7257, pp. 323–329, 2000.

[17] J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, and P. Boffetta, “Risk factors for lung cancer worldwide,” Eur. Respir. J., vol. 48, no. 3, pp. 889–902, 2016.

[18] A. J. Alberg, M. V. Brock, J. G. Ford, J. M. Samet, and S. D. Spivack, “Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines,” Chest, vol. 143, no. 5 Suppl, pp. e1S-e29S, 2013.

[19] G. Luo et al., “Projections of lung Cancer Incidence by 2035 in 40 countries worldwide: Population-based study,” JMIR Public Health Surveill., vol. 9, p. e43651, 2023.

[20] C. F. Mountain, “Revisions in the International System for Staging Lung Cancer,” Chest, vol. 111, no. 6, pp. 1710–1717, 1997.

[21] “Lung cancer stages,” cancer treatment centres of America, [Online]. Available: Cancer IAfRo. Cancer Today 2020 [Available from: https://gco.iarc.fr/today/home. [Accessed 6 May 2022].

[22] S. Abedi et al., “Estimating the survival of patients with lung cancer: What is the best statistical model?,” J. Prev. Med. Public Health, vol. 52, no. 2, pp. 140–144, 2019.

[23] Cancer research UK,[Online]. Available: https://www.cancerresearchuk.org/about-cancer/lung-cancer/survival. [Accessed 30 March 2023].

[24] “SURVMARK-2 - viz7,” Iarc.fr. [Online]. Available: https://gco.iarc.fr/survival/survmark/visualizations/viz7/?mode=%22circle%22&groupby=%22country%22&period=%221%22&cancer=%22NSLUNG%22&country=%22Australia%22&gender=0&stage=%22TNM%22&age_group=%2215-99%22&show_ci=true. [Accessed 30 Mar 2023].

[25] E. Lee and E. A. Kazerooni, “Lung cancer screening,” Semin. Respir. Crit. Care Med., vol. 43, no. 6, pp. 839–850, 2022.

[26] D. Lardinois et al., “Staging of non-small-cell lung cancer with integrated positron-emission tomography and computed tomography,” N. Engl. J. Med., vol. 348, no. 25, pp. 2500–2507, 2003.

[27] R. J. Downey et al., “Preoperative F-18 fluorodeoxyglucose-positron emission tomography maximal standardized uptake value predicts survival after lung cancer resection,” J. Clin. Oncol., vol. 22, no. 16, pp. 3255–3260, 2004.

[28] J. A. Zagzebski, Essentials of ultrasound physics. St. Louis, MO: Mosby, 1996.

[29] R. R. Gharieb, “X-rays and computed tomography scan imaging: Instrumentation and medical applications,” in Computed-Tomography (CT) Scan, R. R. Gharieb, Ed. London, England: IntechOpen, 2022.

[30] J. A. Verschakelen, J. Bogaert, and W. De Wever, “Computed tomography in staging for lung cancer,” Eur. Respir. J. Suppl., vol. 35, pp. 40s–48s, 2002.

[31] S. Abdulla, “CT equipment,” Radiology Cafe, 22-Mar-2017. [Online]. Available: https://www.radiologycafe.com/frcr-physics-notes/ct-imaging/ct-equipment/. [Accessed: 11-Apr-2023].

[32] The Editors of Encyclopedia Britannica, “X-ray tube,” Encyclopedia Britannica. 20-Apr-2017.

[33] Graham Lloyd-Jones BA MBBS MRCP FRCR-Consultant Radiologist, “Basics of X-ray physics,” Radiologymasterclass.co.uk. [Online]. Available: https://www.radiologymasterclass.co.uk/tutorials/physics/x-ray_physics_production. [Accessed: 11-Apr-2023].

[34] “Positron emission tomography (PET),” Hopkinsmedicine.org, 20-Aug-2021. [Online]. Available: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/positron-emission-tomography-pet. [Accessed: 17-Apr-2023]

[35] “11.4: Positron emission,” Chemistry LibreTexts, 15-Mar-2017. [Online]. Available: https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_Online_(Young)/11%3A_Nuclear_Chemistry/11.4%3A_Positron_Emission. [Accessed: 18-Apr-2023]

[36] SYNTHESIS AND APPLICATION OF TARGETED MOLECULAR IMAGING AGENTS FOR ENHANCED DISEASE IMAGING AND THERAPY - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Positron-emission-tomography_fig2_265232417 [accessed 18-Apr-2023]

[37] Positron emission tomography (PET),” Hopkinsmedicine.org, 20-Aug-2021. [Online]. Available: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/positron-emission-tomography-pet. [Accessed: 18-Apr-2023]

[38] Reniers B. Les PET-SPECT. 2021.

[39] FPGA based data acquistion system and digital pulse processing for PET and SPECT - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Block-detector-configuration-for-PET_fig7_266887992 [accessed 19 Apr, 2023]

[40] Rcr.ac.uk. [Online]. Available: https://www.rcr.ac.uk/system/files/publication/field_publication_files/evidence-based_indications_for_the_use_of_pet-ct_in_the_united_kingdom_2022.pdf. [Accessed: 19-Apr-2023].

[41] INTERNATIONAL ATOMIC ENERGY AGENCY, “Quality assurance for PET and PET/CT systems,” in Quality Assurance for PET and PET/CT Systems, pp. 1–145, 2019.

[42] R. H. Bijwaard, “Inventarisatie van ontwikkelingen van PET-CT,” Rivm.nl. [Online]. Available: https://www.rivm.nl/bibliotheek/rapporten/300080008.pdf. [Accessed: 18-Apr-2023].

[43] S. Tong, A. M. Alessio, and P. E. Kinahan, “Image reconstruction for PET/CT scanners: past achievements and future challenges,” Imaging Med., vol. 2, no. 5, pp. 529–545, 2010.

[44] S. Stroobants, J. Verschakelen, and J. Vansteenkiste, “Value of FDG-PET in the management of non-small cell lung cancer,” Eur. J. Radiol., vol. 45, no. 1, pp. 49–59, 2003.

[45] M. A. Ashraf and A. Goyal, Fludeoxyglucose (18F). StatPearls Publishing, 2022.

[46] M. Reivich et al., “The [18F]fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man,” Circ. Res., vol. 44, no. 1, pp. 127–137, 1979.

[47] A. Abbasian Ardakani, N. J. Bureau, E. J. Ciaccio, and U. R. Acharya, “Interpretation of radiomics features-A pictorial review,” Comput. Methods Programs Biomed., vol. 215, no. 106609, p. 106609, 2022.

[48] H. J. W. L. Aerts et al., “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” Nat. Commun., vol. 5, no. 1, p. 4006, 2014.

[49] C. Parmar et al., “Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer,” Sci. Rep., vol. 5, no. 1, p. 11044, 2015.

[50] J. Lee, “Statistics, Descriptive,” in International Encyclopedia of Human Geography, Elsevier, pp. 13–20, 2020.

[51] “Radiomic Features — pyradiomics v3.1.0rc2.post5+g6a761c4 documentation,” Readthedocs.io. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/features.html. [Accessed: 22-Apr-2023].

[52] C. Nioche, F. Orlhac, and I. Buvat, “Texture — User Guide Local Image Features Extraction — LIFEx —,” Lifexsoft.org. [Online]. Available: https://www.lifexsoft.org/images/phocagallery/documentation/ProtocolTexture/UserGuide/TextureUserGuide.pdf. [Accessed: 24-Apr-2023].

[53] A. Zwanenburg et al., “The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping,” Radiology, vol. 295, no. 2, pp. 328–338, 2020.

[54] Y. Cui and F.-F. Yin, “Impact of image quality on radiomics applications,” Phys. Med. Biol., vol. 67, no. 15, p. 15TR03, 2022.

[55] “PET scan,” nhs.uk. [Online]. Available: https://www.nhs.uk/conditions/pet-scan/. [Accessed: 10-Mar-2023].

[56] D. Faist et al., “Reproducibility of lung cancer radiomics features extracted from data-driven respiratory gating and free-breathing flow imaging in [18F]-FDG PET/CT,” Eur. J. Hybrid Imaging, vol. 6, no. 1, p. 33, 2022.

[57] D. Bell and K. Yap, “Standard uptake value,” Radiopaedia.org. Radiopaedia.org, 27-Jul-2011.

[58] F. H. P. van Velden et al., “Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: Impact of reconstruction and delineation,” Mol. Imaging Biol., vol. 18, no. 5, pp. 788–795, 2016.

[59] C. A. Owens et al., “Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer,” PLoS One, vol. 13, no. 10, p. e0205003, 2018.

[60] M. Bogowicz et al., “Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models,” Radiother. Oncol., vol. 125, no. 3, pp. 385–391, 2017.

[61] R. Bataller and D. A. Brenner, “Liver fibrosis,” J. Clin. Invest., vol. 115, no. 2, pp. 209–218, 2005.

[62] X. Zhu, “Editorial for ‘radiomics approaches for predicting liver fibrosis with nonenhanced T1-weighted imaging: Comparison of different radiomics models,’” J. Magn. Reson. Imaging, vol. 53, no. 4, pp. 1090–1091, 2021.

[63] F. Valdora, N. Houssami, F. Rossi, M. Calabrese, and A. S. Tagliafico, “Rapid review: radiomics and breast cancer,” Breast Cancer Res. Treat., vol. 169, no. 2, pp. 217–229, 2018.

[64] M. Schwier et al., “Repeatability of multiparametric prostate MRI radiomics features,” Sci. Rep., vol. 9, no. 1, p. 9441, 2019.

[65] Y. Hu et al., “Three-dimensional radiomics features of magnetic resonance T2-weighted imaging combined with clinical characteristics to predict the recurrence of acute pancreatitis,” Front. Med. (Lausanne), vol. 9, p. 777368, 2022.

[66] T. P. Coroller et al., “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma,” Radiother. Oncol., vol. 114, no. 3, pp. 345–350, 2015.

[67] O. Grove et al., “Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma,” PLoS One, vol. 10, no. 3, p. e0118261, 2015.

[68] J. Chen et al., “Lung cancer diagnosis using deep attention-based multiple instance learning and radiomics,” Med. Phys., vol. 49, no. 5, pp. 3134–3143, 2022.

[69] C. Jiang et al., “Radiomics signature from [18F]FDG PET images for prognosis predication of primary gastrointestinal diffuse large B cell lymphoma,” Eur. Radiol., vol. 32, no. 8, pp. 5730–5741, 2022.

[70] I. Fornacon-Wood et al., “Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform,” Eur. Radiol., vol. 30, no. 11, pp. 6241–6250, 2020.

[71] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, “Radiomics in medical imaging-"how-to" guide and critical reflection,” Insights Imaging, vol. 11, no. 1, p. 91, 2020.

[72] W. Haynes, “Student’s t-Test,” in Encyclopedia of Systems Biology, New York, NY: Springer New York, pp. 2023–2025, 2013.

[73] “Statistics: 1.1 Paired t-tests,” Statstutor.ac.uk. [Online]. Available: https://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf. [Accessed: 04-May-2023]

[74] T. Dahiru, “P - value, a true test of statistical significance? A cautionary note,” Ann. Ib. Postgrad. Med., vol. 6, no. 1, pp. 21–26, 2008.

[75] Statisticsbyjim.com. [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/. [Accessed: 04-May-2023]

[76] The MathWorks, Inc. (2022). MATLAB version: 9 (R2019b)

[77] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning: With applications in R, 2013th ed. New York, NY: Springer, 2013.

[78] R. Gandhi, “Support vector machine — introduction to machine learning algorithms,” Towards Data Science, 07-Jun-2018.

[79] O. Kramer, “K-Nearest Neighbors,” in Dimensionality Reduction with Unsupervised Nearest Neighbors, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 13–23, 2013.

[80] Jason Brwnlee, A Gentle Introduction to Ensemble Learning Algorithms, Ensemble Learning, April 19, 2021.

[81] I. T. Jolliffe, “Introduction,” in Principal Component Analysis, New York, NY: Springer New York, pp. 1–7, 1986.

[82] M. Ringnér, “What is principal component analysis?,” Nat. Biotechnol., vol. 26, no. 3, pp. 303–304, 2008.

[83] GraphPad Software, LLC, “GraphPad prism 9 statistics guide - graphs for principal component analysis,” Graphpad.com. [Online]. Available: https://www.graphpad.com/guides/prism/latest/statistics/stat_pca_graphs_tab.htm. [Accessed: 04-May-2023]

[84] R. Wicklin, “What are biplots?,” The DO Loop, 06-Nov-2019. [Online]. Available: https://blogs.sas.com/content/iml/2019/11/06/what-are-biplots.html. [Accessed: 04-May-2023].

[85] Healthineers S. Biograph Horizon 2020 [Available from: https://www.siemens-healthineers.com/molecular-imaging/pet-ct/biograph-horizon. [Accessed: 18-May-2023].

[86] V. Trojani et al., “Radiomic features characterization in healthy and NSCLC tissues,” Phys. Med., vol. 92, p. S187, 2021.

[87] I. El Naqa et al., “Exploring feature-based approaches in PET images for predicting cancer treatment outcomes,” Pattern Recognit., vol. 42, no. 6, pp. 1162–1171, 2009.

[88] Z. A. Kohutek et al., “FDG-PET maximum standardized uptake value is prognostic for recurrence and survival after stereotactic body radiotherapy for non-small cell lung cancer,” Lung Cancer, vol. 89, no. 2, pp. 115–120, 2015.

[89] S. Lysdahlgaard, “Comparing Radiomics features of tumor and healthy liver tissue in a limited CT dataset: A machine learning study,” Radiography (Lond.), vol. 28, no. 3, pp. 718–724, 2022.

[90] S. Ghezzo et al., “State of the art of radiomic analysis in the clinical management of prostate cancer: A systematic review,” Crit. Rev. Oncol. Hematol., vol. 169, no. 103544, p. 103544, 2022.

[91] Z. Jin, T. Yuan, Y. Tokuda, Y. Naoi, N. Tomiyama, and K. Suzuki, “Radiomics: Approach to precision medicine,” in Intelligent Systems Reference Library, Cham: Springer International Publishing, 2023, pp. 17–29

[92] S. Roy et al., “Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging,” EBioMedicine, vol. 59, no. 102963, p. 102963, 2020.

[93] P. Bulens et al., “Predicting the tumor response to chemoradiotherapy for rectal cancer: Model development and external validation using MRI radiomics,” Radiother. Oncol., vol. 142, pp. 246–252, 2020.

[94] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider, and F. Khalvati, “Radiomics-based prognosis analysis for non-small cell lung cancer,” Sci. Rep., vol. 7, no. 1, pp. 1–8, 2017.

[95] S. Ibrahim, S. Nazir, and S. A. Velastin, “Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis,” Journal of Imaging, vol. 7, no. 11, p. 225, Oct. 2021.

[96] D. G. Baur, “The Anna Karenina principle and stock prices,” J. Behav. Exp. Finance, vol. 33, no. 100602, p. 100602, 2022.

[97] L. Rundo, C. Militello, V. Conti, F. Zaccagna, and C. Han, “Advanced Computational Methods for Oncological Image Analysis,” Journal of Imaging, vol. 7, no. 11, p. 237, Nov. 2021.

[98] K. Marias, “The Constantly Evolving Role of Medical Image Processing in Oncology: From Traditional Medical Image Processing to Imaging Biomarkers and Radiomics,” Journal of Imaging, vol. 7, no. 8, p. 124, Jul. 2021.

[99] L. C. Chu et al., “Utility of CT radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue,” AJR Am. J. Roentgenol., vol. 213, no. 2, pp. 349–357, 2019.

[100] D. Markel et al., “Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT,” Int. J. Mol. Imaging, vol. 2013, p. 980769, 2013.

[101] M. Eskian et al., “Effect of blood glucose level on standardized uptake value (SUV) in 18F- FDG PET-scan: a systematic review and meta-analysis of 20,807 individual SUV measurements,” Eur. J. Nucl. Med. Mol. Imaging, vol. 46, no. 1, pp. 224–237, 2019.

[102] K. A. Büsing, S. O. Schönberg, J. Brade, and K. Wasser, “Impact of blood glucose, diabetes, insulin, and obesity on standardized uptake values in tumors and healthy organs on 18F-FDG PET/CT,” Nucl. Med. Biol., vol. 40, no. 2, pp. 206–213, 2013.

[103] Díaz-Pernas, F.J.; Martínez-Zarzuela, M.; Antón-Rodríguez, M.; González-Ortega, D. A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. Healthcare,9, 153, 2021.

[104] M. Gorenberg, W. A. Hallett, and M. J. O’Doherty, “Does diabetes affect [(18)F]FDG standardized uptake values in lung cancer?,” Eur. J. Nucl. Med. Mol. Imaging, vol. 29, no. 10, pp. 1324–1327, 2002.

[105] L.-Q. Tong, Y.-F. Sui, S.-N. Jiang, and Y.-H. Yin, “The association between lung fluorodeoxyglucose metabolism and smoking history in 347 healthy adults,” J. Asthma Allergy, vol. 14, pp. 301–308, 2021.

[106] T. Schroeder, M. F. Vidal Melo, G. Musch, R. S. Harris, T. Winkler, and J. G. Venegas, “PET imaging of regional 18F-FDG uptake and lung function after cigarette smoke inhalation,” J. Nucl. Med., vol. 48, no. 3, pp. 413–419, 2007.

[107] J. Bucerius et al., “Impact of noninsulin-dependent type 2 diabetes on carotid wall 18F-fluorodeoxyglucose positron emission tomography uptake,” J. Am. Coll. Cardiol., vol. 59, no. 23, pp. 2080–2088, 2012

[108] D. A. Torigian et al., “A study of the feasibility of FDG-PET/CT to systematically detect and quantify differential metabolic effects of chronic tobacco use in organs of the whole body-A prospective pilot study,” Acad. Radiol., vol. 24, no. 8, pp. 930–940, 2017.

[109] L. Lu et al., “Robustness of radiomic features in [11C]choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: Impact of segmentation and discretization,” Mol. Imaging Biol., vol. 18, no. 6, pp. 935–945, 2016.

[110] B. Zhao et al., “Reproducibility of radiomics for deciphering tumor phenotype with imaging,” Sci. Rep., vol. 6, no. 1, p. 23428, 2016.

[111] A. Somasundaram et al., “Mitigation of noise-induced bias of PET radiomic features,” PLoS One, vol. 17, no. 8, p. e0272643, 2022.

[112] Y. Balagurunathan et al., “Test-retest reproducibility analysis of lung CT image features,” J. Digit. Imaging, vol. 27, no. 6, pp. 805–823, 2014.

[113] M. G. Lubner, A. D. Smith, K. Sandrasegaran, D. V. Sahani, and P. J. Pickhardt, “CT texture analysis: Definitions, applications, biologic correlates, and challenges,” Radiographics, vol. 37, no. 5, pp. 1483–1503, 2017.

[114] D. Mackin et al., “Measuring computed tomography scanner variability of radiomics features,” Invest. Radiol., vol. 50, no. 11, pp. 757–765, 2015.

[115] F. A. Shaikh et al., “Technical challenges in the clinical application of radiomics,” JCO Clin. Cancer Inform., vol. 1, no. 1, pp. 1–8, 2017.

[116] P. Jonsson et al., “Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples,” Metabolomics, vol. 11, no. 6, pp. 1667–1678, 2015.

[117] “OPLS vs PCA: Explaining differences or grouping data?,” Sartorius. [Online]. Available: https://www.sartorius.com/en/knowledge/science-snippets/explaining-differences-or-grouping-data-opls-da-vs-pca-data-analysis-507204. [Accessed: 09-Jun-2023].

[118] D. Rajput, W.-J. Wang, and C.-C. Chen, “Evaluation of a decided sample size in machine learning applications,” BMC Bioinformatics, vol. 24, no. 1, 2023.

[119] P. Dhiman et al., “Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review,” BMC Med. Res. Methodol., vol. 22, no. 1, p. 101, 2022

[120] R. L. Melvin, “Sample size in machine learning and artificial intelligence,” Uab.edu. [Online]. Available: https://sites.uab.edu/periop-datascience/2021/06/28/sample-size-in-machine-learning-and-artificial-intelligence/. [Accessed: 01-Jun-2023].

Annex I: Scatter plots healthy vs tumor tissue

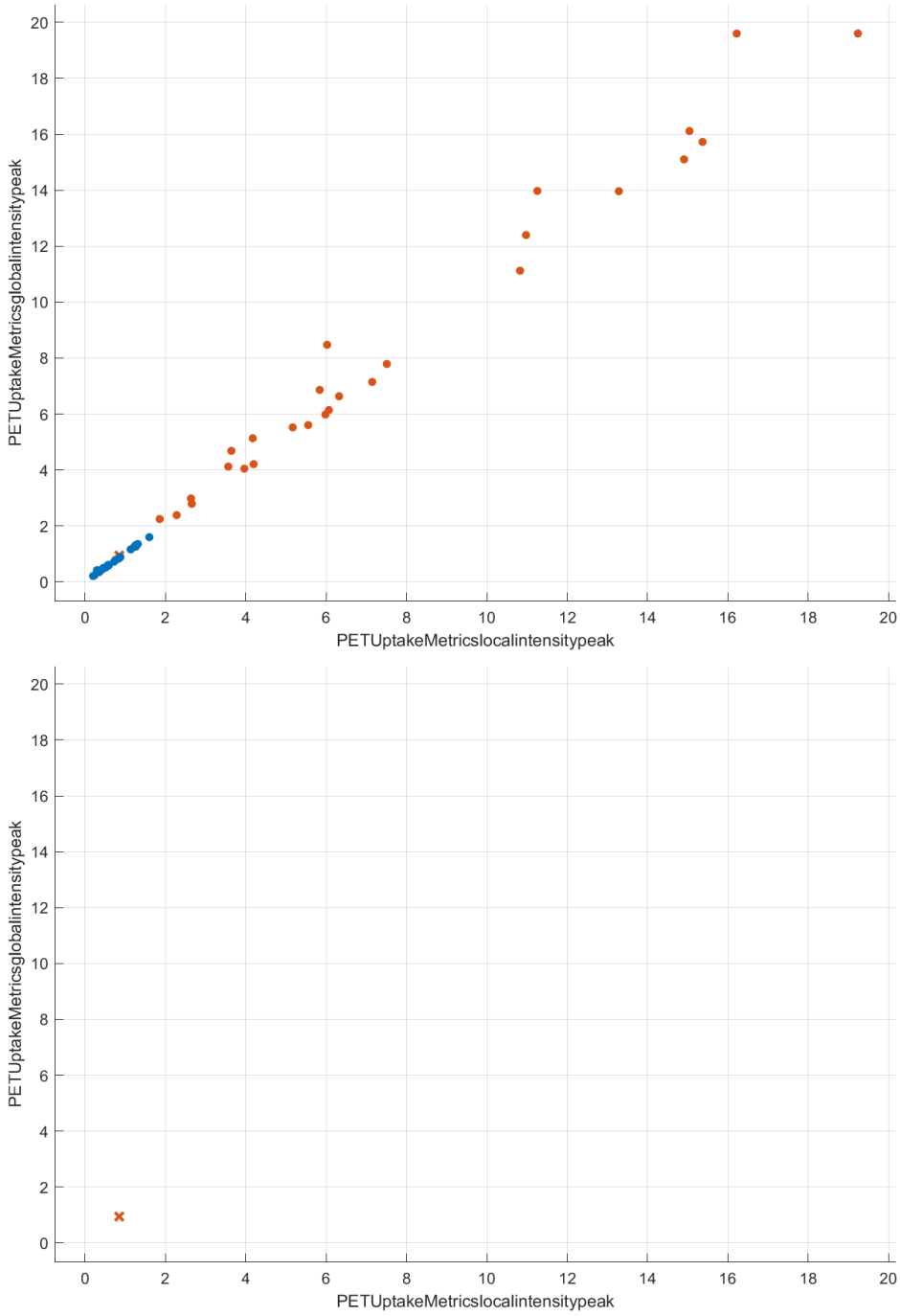


Figure S1: The scatter plot of the Fine KNN model on a cohort of 30 patients, where the blue dots represent the correct classified healthy tissues, the orange dots the correct classified tumor tissues, and the orange cross the incorrect classified tumor tissue. This for the first two radiomics features 'PET Uptake Metrics - local intensity peak' (x-axis) and 'PET Uptake Metrics - global intensity peak' (y-axis).

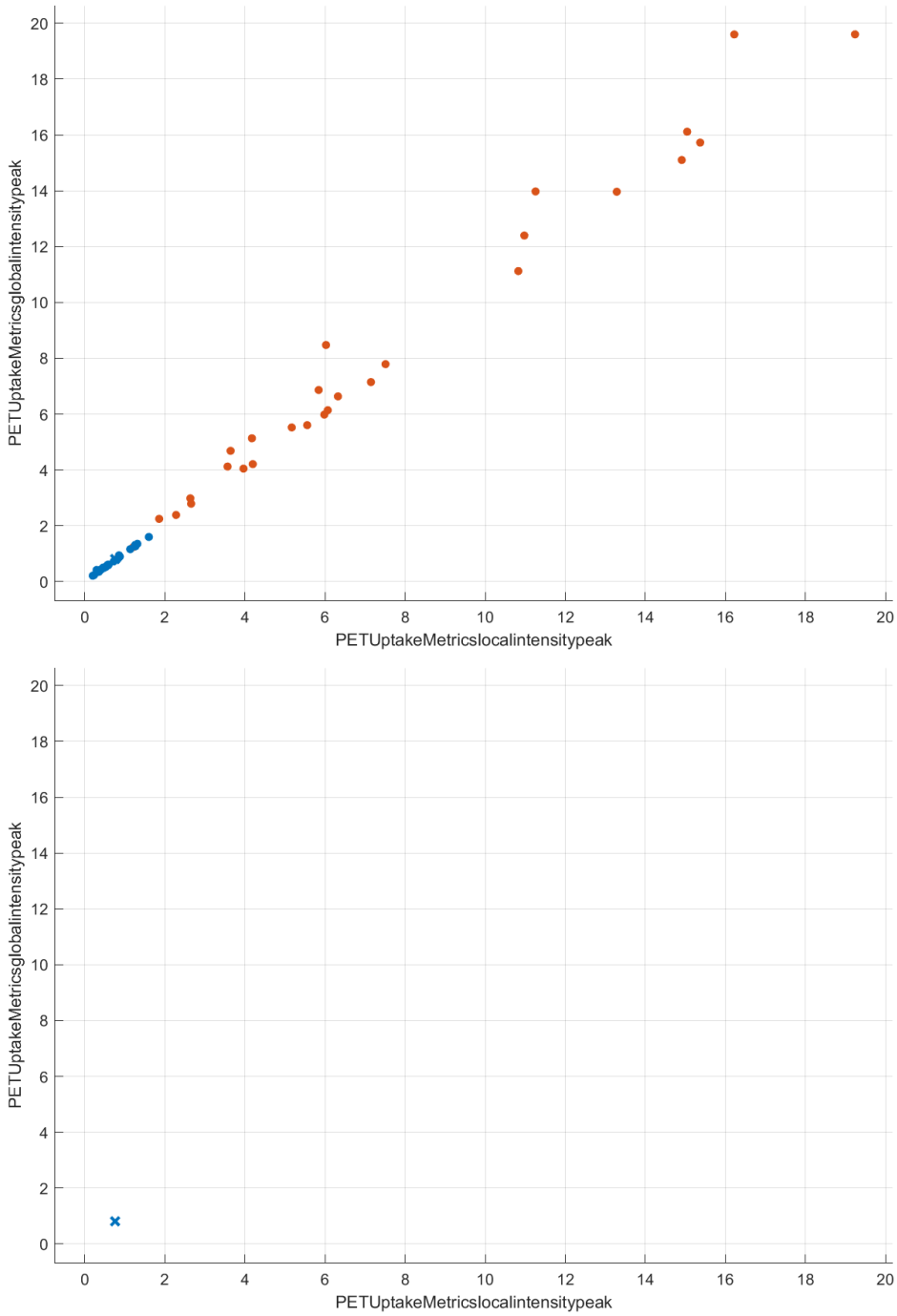


Figure S2: The scatter plot of the Weighted KNN model on a cohort of 30 patients, where the blue dots represent the correct classified healthy tissues, the orange dots the correct classified tumor tissues, and the blue cross the incorrect classified healthy tissue. This for the first two radiomics features 'PET Uptake Metrics - local intensity peak' (x-axis) and 'PET Uptake Metrics - global intensity peak' (y-axis).

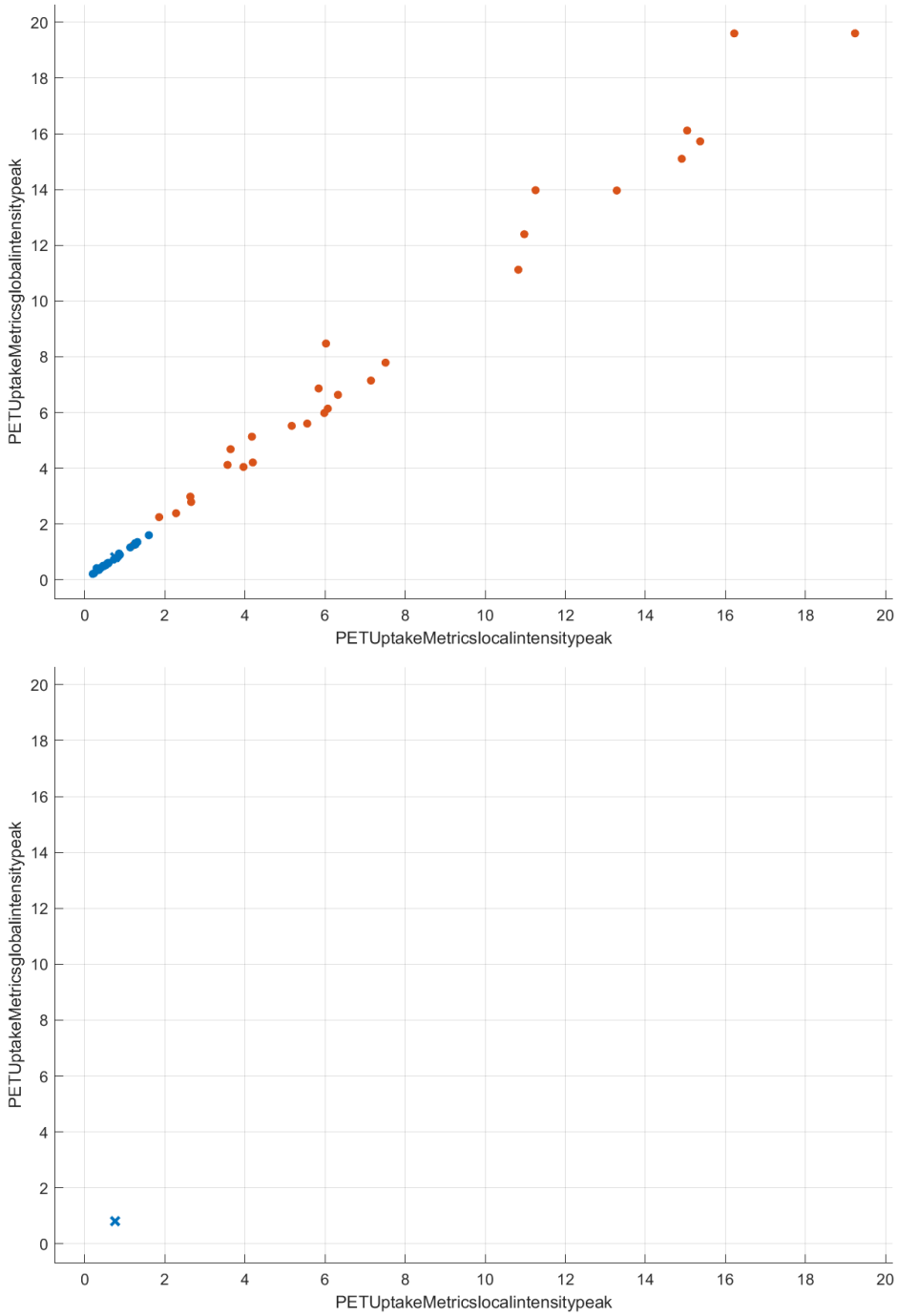


Figure S3: The scatter plot of the Ensemble - Bagged Trees model on a cohort of 30 patients, where the blue dots represent the correct classified healthy tissues, the orange dots the correct classified tumor tissues, and the blue cross the incorrect classified healthy tissue. This for the first two radiomics features 'PET Uptake Metrics - local intensity peak' (x-axis) and 'PET Uptake Metrics - global intensity peak' (y-axis).

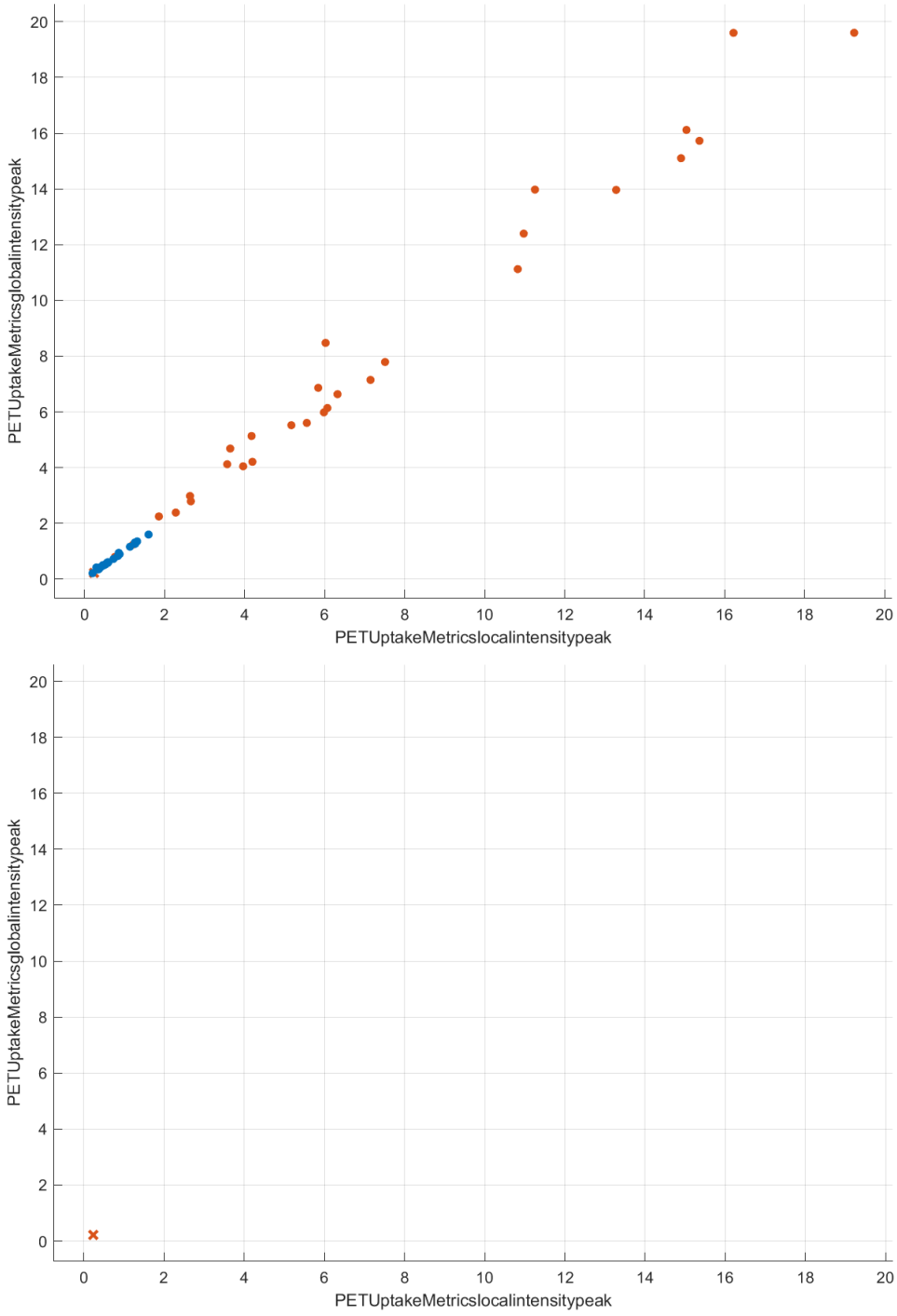


Figure S4: The scatter plot of the Ensemble - Subspace Discriminant model on a cohort of 30 patients, where the blue dots represent the correct classified healthy tissues, the orange dots the correct classified tumor tissues, and the orange cross the incorrect classified tumor tissue. This for the first two radiomics features 'PET Uptake Metrics - local intensity peak' (x-axis) and 'PET Uptake Metrics - global intensity peak' (y-axis).

Annex II: Classification learner results on the noise-reduced dataset

Tumor vs healthy tissue

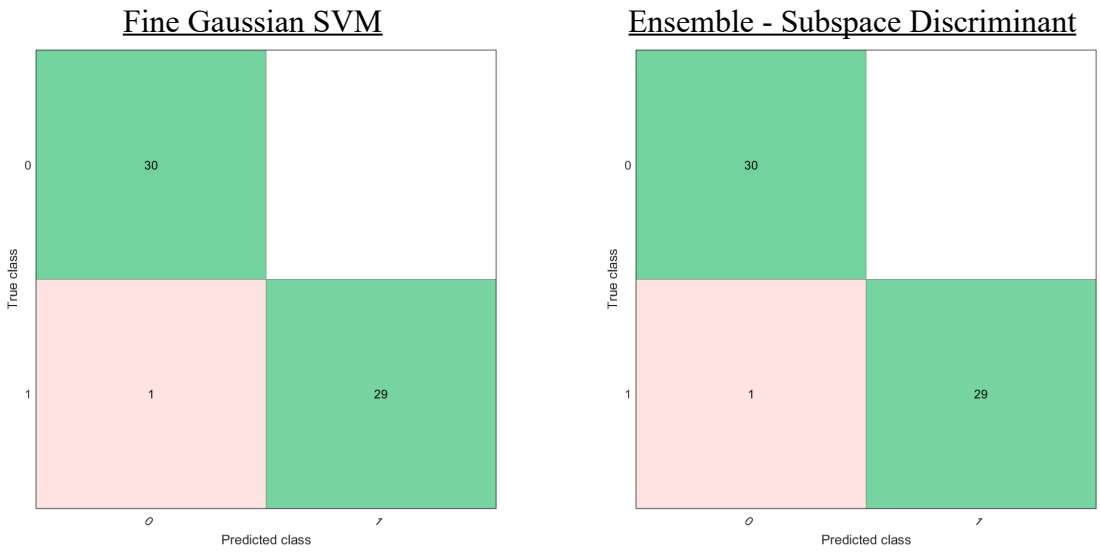


Figure S5: Confusion matrix for the Fine Gaussian SVM and Subspace Discriminant (Ensemble) model for predicting tumor (1) or healthy (0) tissue of a cohort of 30 patients, with an accuracy of 98.3%.

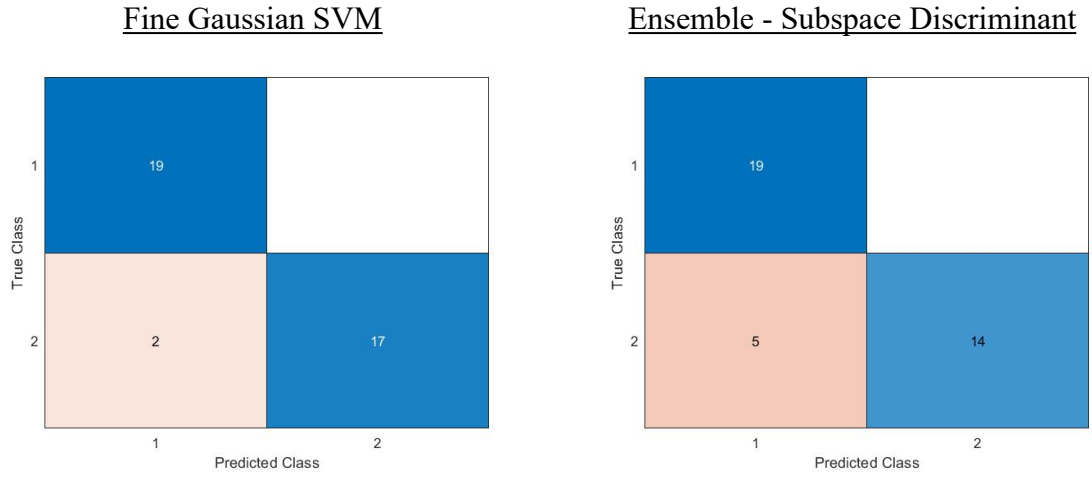


Figure S6: Confusion matrix for the Fine Gaussian SVM and Subspace Discriminant (Ensemble) model for predicting tumor (2) or healthy (1) tissue of a cohort of 19 patients.

Glycemia

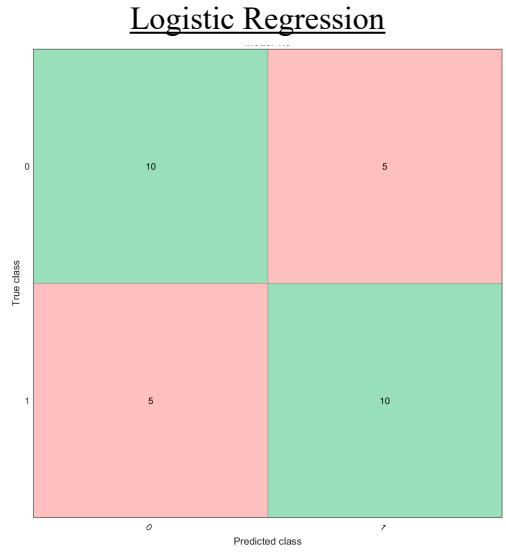


Figure S7: Confusion matrix for the Logistic Regression model and Subspace Discriminant (Ensemble) model for predicting if the glycemia is larger than 98 mg% (1) or smaller than 98 mg% (0) of a cohort of 30 patients, with an accuracy of 80.0%.

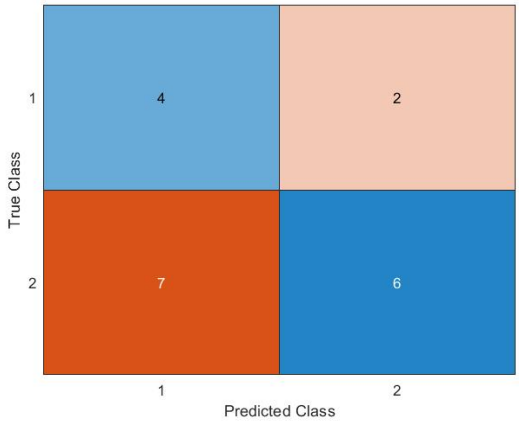


Figure S8: Confusion matrix for the Logistic Regression and Subspace Discriminant (Ensemble) model for predicting if the glycemia is larger than 98 mg% (2) or smaller than 98 mg% (1) of a cohort of 19 patients.

Tumor type

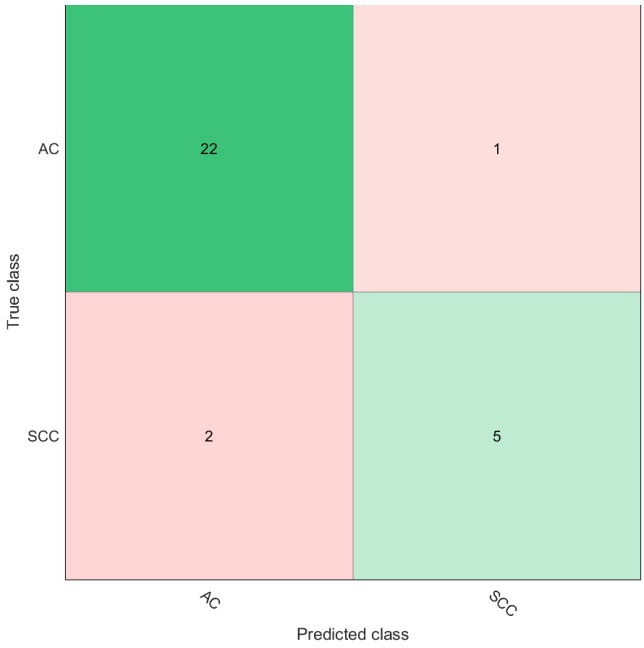


Figure S9: Confusion matrix for the Medium Gaussian SVM model for predicting the tumor type of a cohort of 30 patients, with an accuracy of 90.0%.

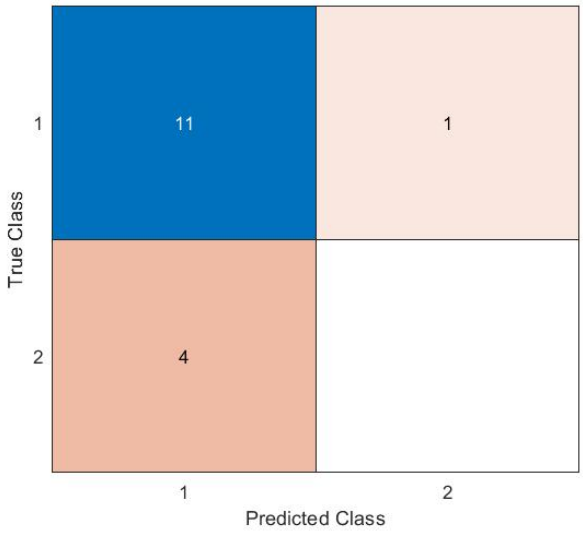


Figure S10: Confusion matrix for the Medium Gaussian SVM model for predicting the tumor type of a cohort of 16 patients.

Lung side (left or right)

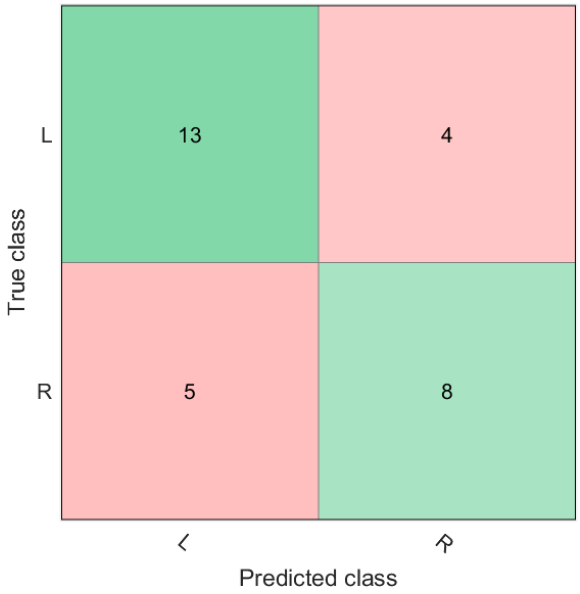


Figure S11: Confusion matrix for the Logistic Regression model for predicting the position of the tumor of a cohort of 30 patients, with an accuracy of 70.0%..

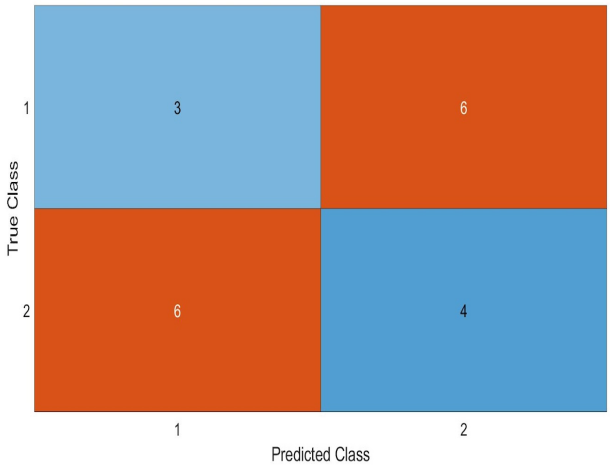


Figure S12: Confusion matrix for the Logistic Regression model for predicting the position of the tumor of a cohort of 19 patients.

Diabetes

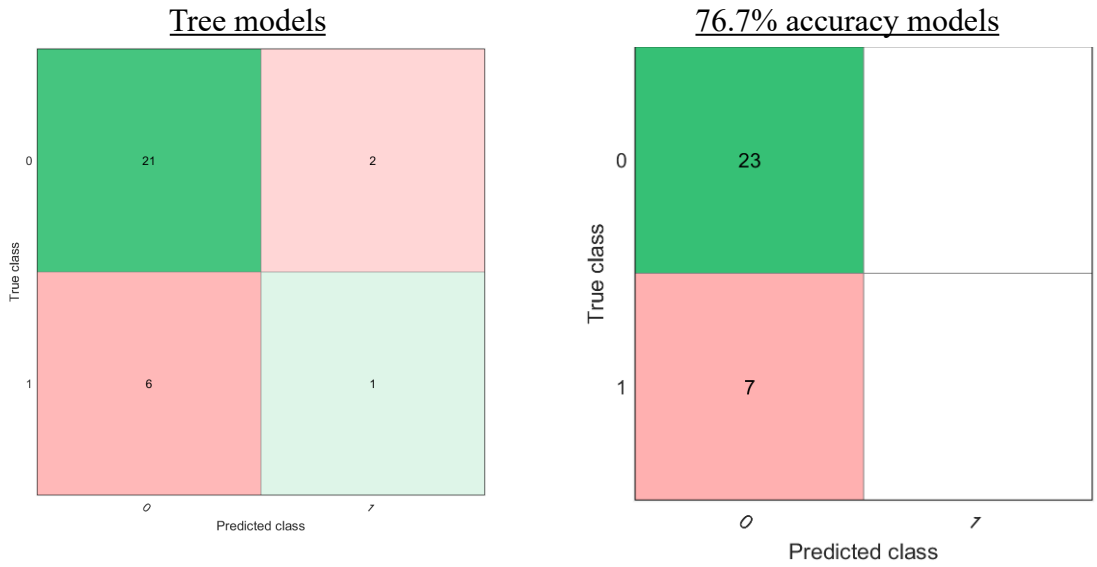


Figure S13: Confusion matrix for the tree model and the nine highest scoring models for predicting if the patient has diabetes (1) or not (0) of a cohort of 30 patients.

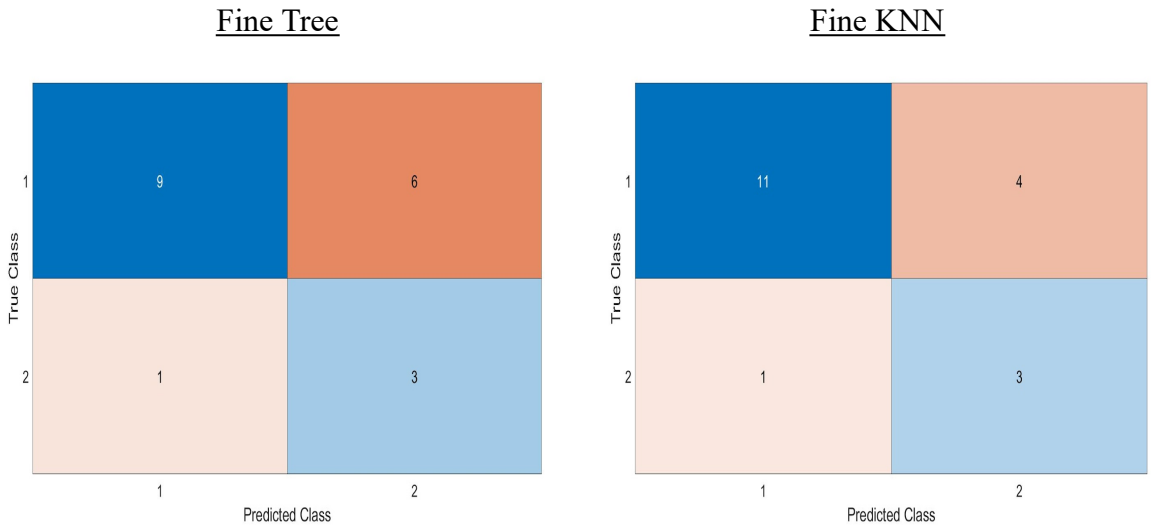


Figure S14: Confusion matrix for the Linear Discriminant and Fine KNN model for predicting if the patient has diabetes (2) or not (1) of a cohort of 19 patients.

Packyears

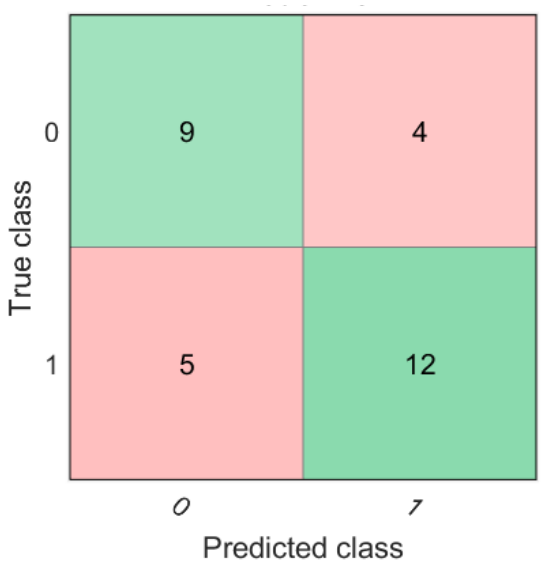


Figure S15: Confusion matrix for the Kernel Naive Bayes model for predicting if the amount of packyears is larger than 35 (1), or smaller (0) of the healthy tissue of a cohort of 30 patients.

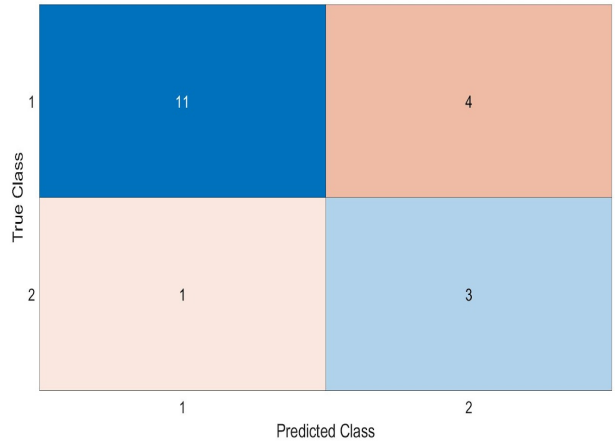


Figure S16: Confusion matrix for the SKernel Naive Bayes model for predicting if the amount of packyears is larger than 35 (1), or smaller (0) of the healthy tissue of a cohort of 18 patients.

Annex III: Codes

Python code for PCA

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.decomposition import PCA
from sklearn import preprocessing
import matplotlib.pyplot as plt

quantificatie=FILE_WITH_FEATURES_DATA
features=FILE_WITH_FEATURE_NAMES
index=FILE_WITH_TUMOR(1)_HEALTHY(0)

scaled_data = preprocessing.scale(quantificatie)

pca = PCA() #make PCA object
pca.fit(scaled_data) #Calculation of the PCs
pca_data = pca.transform(scaled_data) #generate coordinates


per_var = np.round(pca.explained_variance_ratio_*100, decimals=1) # % of the variance per PC
labels = ['PC' + str(x) for x in range(1, len(per_var)+1)]
plt.bar(x=range(1,len(per_var)+1), height=per_var, tick_label=labels)
plt.ylabel('% Explained variance')
plt.xlabel('Principal component')
plt.xlim(0,10.5)
plt.show() #graph with the ten most relevant PCs



pca_df = pd.DataFrame(pca_data, columns=labels)
tumor_array = pca_df[::2]
healthy_array = pca_df[1::2]
plt.scatter(tumor_array.PC1, tumor_array.PC2,c='red', label='tumor')
```

```
plt.scatter(healthy_array.PC1, healthy_array.PC2,c='blue', label='healthy')
plt.xlabel('PC1 - {0}%'.format(per_var[0]))
plt.ylabel('PC2 - {0}%'.format(per_var[1]))
plt.legend(loc='lower right')
plt.show() #scatterplot with the first two PC's as axis and the data plotted out accordingly

loading_scores = pd.Series(pca.components_[0], index=features)
sorted_loading_scores = loading_scores.abs().sort_values(ascending=False)
top_10_features = sorted_loading_scores[0:10].index.values
print(loading_scores[top_10_features]) #show the 10 most relevant features and their loading scores


def biplot(score,coeff,labels=None): #structure for the biplot with the 10 most relevant features
    xs = score[:,0]
    ys = score[:,1]
    n = coeff.shape[0]
    scalex = 1.0/(xs.max() - xs.min())
    scaley = 1.0/(ys.max() - ys.min())
    plt.scatter(xs * scalex,ys * scaley,s=10)
    for i in range(10):
    plt.arrow(0, 0, coeff[i,0]*1, coeff[i,1]*12,color = 'r',alpha = 0.5)
    if labels is None:
    plt.text(coeff[i,0]* 1, coeff[i,1] * 1, "Var"+str(i+1), color = 'green', ha = 'center', va = 'center')
    else:
    plt.text(coeff[i,0]* 1, coeff[i,1] * 1, labels[i], color = 'g', ha = 'center', va = 'center')


    plt.xlabel("PC{}".format(1))
    plt.ylabel("PC{}".format(2))
plt.figure(figsize=(12, 6))
biplot(pca_data, np.transpose(pca.components_[0:2]), list(loading_scores[top_10_features].index))
plt.grid()
```

Matlab code

t-test

```
data = xlsread('Dataset_Boellaard_t_2.xlsx');          % Load data
[rows,cols] = size(data);                              % Define number of rows and columns
features = (cols-1)/2;                                 % Define amount of features

results=zeros(features, 5);                            % Empty matrix to store results

for i = 1:features
   tumor = data(:,i);                                  % First column
   gezond = data(:,i+features+1);
   [h,p,ci,stats] = ttest(tumor, gezond);

   MyFieldNames = fieldnames(stats);                   % Field to values
   for j=1:3
      stat(j,1) = getfield(stats,MyFieldNames{j});
   end
   results(i,1) = h;                                   % Store results
   results(i,2) = p;
   results(i,3) = stat(1,1);
   results(i,4) = stat(2,1);
   results(i,5) = stat(3,1);
end
```

Dendrogram

```
data = xlsread('Dataset_dendrogram.xlsx');   % Load the data
data_trans=transpose(data);                  % Transpose data matrix (for linkage function)
[rows,cols] = size(data_trans);              % Define rows and cols
tree = linkage(data_trans,'average');        % Create tree with the linkage function
cutoff = 5;                                  % Define cutoff height
[H, T, outperm] = dendrogram(tree,0,'ColorThreshold',cutoff);        % Create dendrogram
outperm=transpose(outperm);
linesColor = cell2mat(get(H,'Color'));                                % Get lines color;
colorList = unique(linesColor, 'rows');
data_trans_color     = zeros(rows,3);        % Create zero vectors for saving data later
data_trans_cluster   = zeros(rows,1);
for iLeaf = 1:rows
   [iRow, ~] = find(tree==iLeaf);
   color = linesColor(iRow,:);               % Assign color to each observation
   data_trans_color(iLeaf,:) = color;        % Assign cluster number to each observation
   data_trans_cluster(iLeaf,:) = find(ismember(colorList, color, 'rows'));
end

set(gca,'YScale','log');                     % Set the plot in logarithmic scale for better
view
```

Classification learner

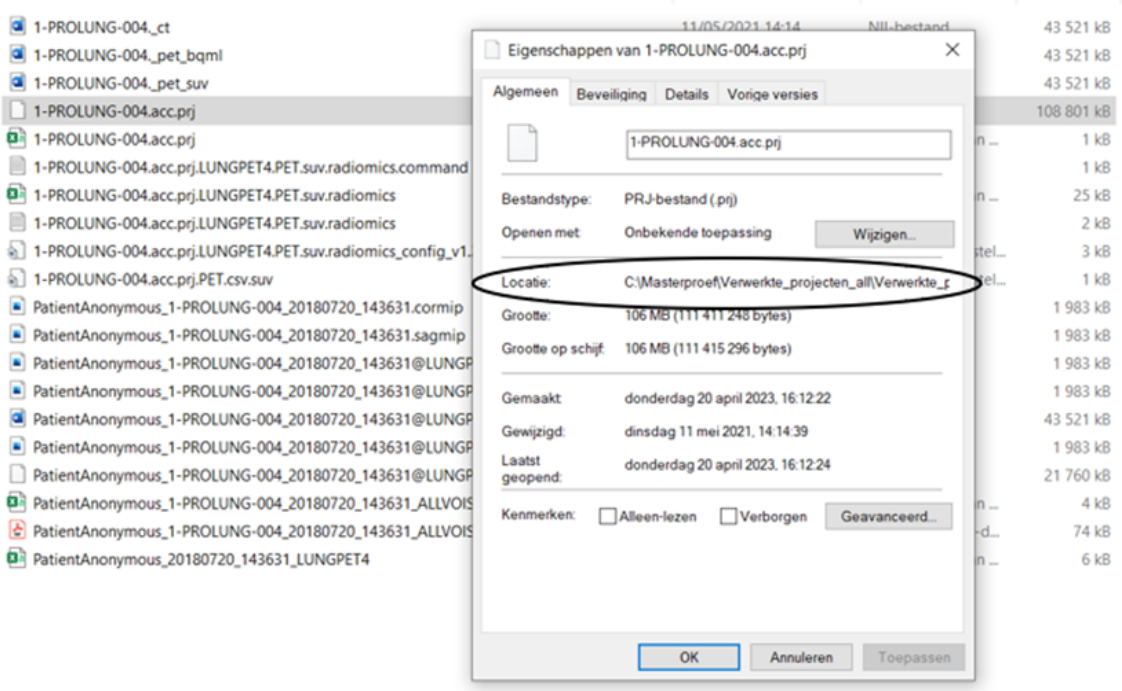
```
yfit = Glycemie_LogReg.predictFcn(DATASET.xls);          % Load prediction function
yfit = transpose(yfit);
predictions = table(yfit);
predicted_class = table2array(predictions);
true_class = A;                                          % Correct values of the input data
C = confusionmat(true_class, predicted_class);           % Create confusion matrix
confusionchart(C);
```

Annex IV: Guidelines for radiomics data extraction of healthy and tumor tissue in the Accurate and Radiomics tools (Prof. Dr. Boellaard, Amsterdam UMC)

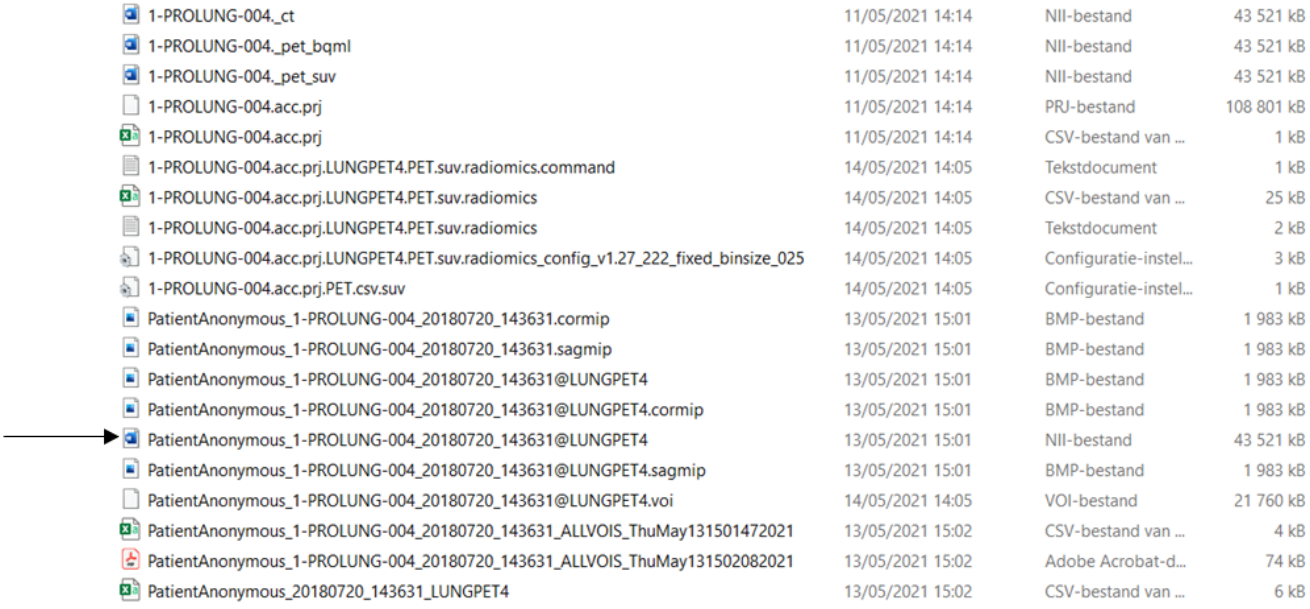
Step 1: First of all, you have to check whether all necessary files are at hand. The following image shows what a normal patient file looks like. The most important files for the data extraction are .prj, .nii and .voi

1-PROLUNG-004._ct	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_bqml	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_suv	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	PRJ-bestand	108 801 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	CSV-bestand van ...	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics.command	14/05/2021 14:05	Tekstdocument	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	CSV-bestand van ...	25 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	Tekstdocument	2 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics_config_v1.27_222_fixed_binsize_025	14/05/2021 14:05	Configuratie-instel...	3 kB
1-PROLUNG-004.acc.prj.PET.csv.suv	14/05/2021 14:05	Configuratie-instel...	1 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.voi	14/05/2021 14:05	VOI-bestand	21 760 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131501472021	13/05/2021 15:02	CSV-bestand van ...	4 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131502082021	13/05/2021 15:02	Adobe Acrobat-d...	74 kB
PatientAnonymous_20180720_143631_LUNGPET4	13/05/2021 15:02	CSV-bestand van ...	6 kB

Important note: make sure that the path of the patient file does not contain spaces! Verify this by right-clicking a file in the patient folder, go to properties and search for the location. The entire path of the file should be displayed as seen in the image below.

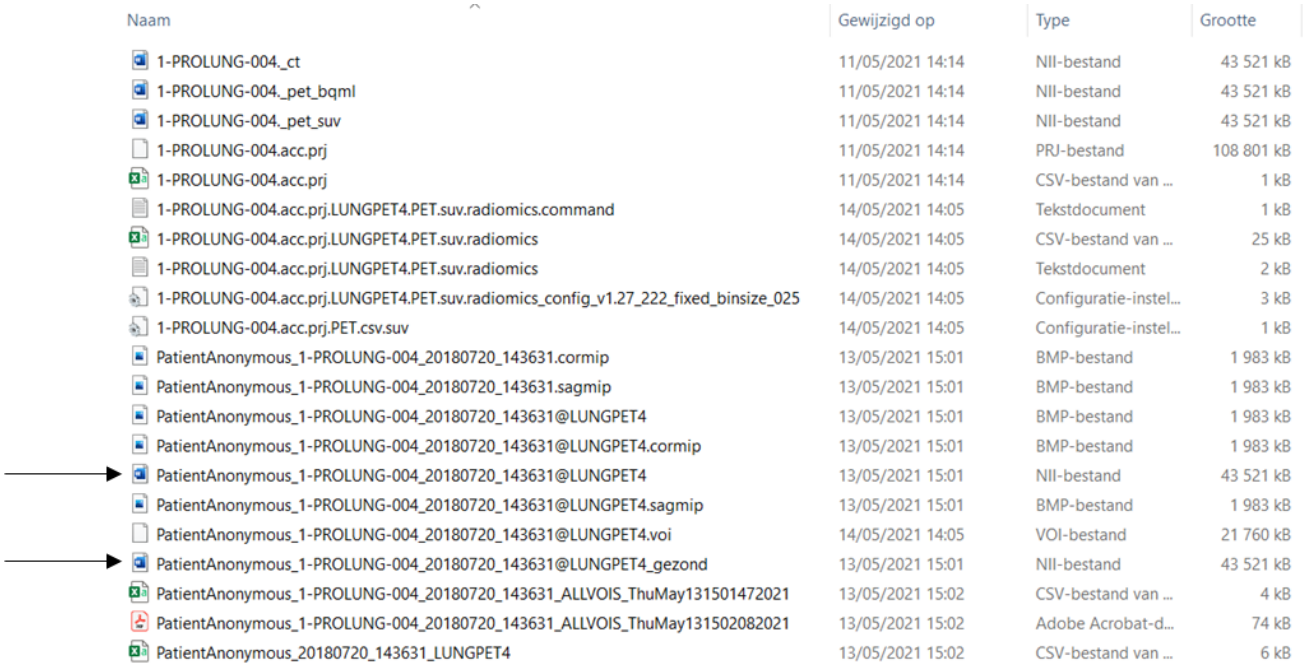


Step 2: To create a new volume of interest (VOI) for the healthy tissue, copy the NII-file of the VOI.



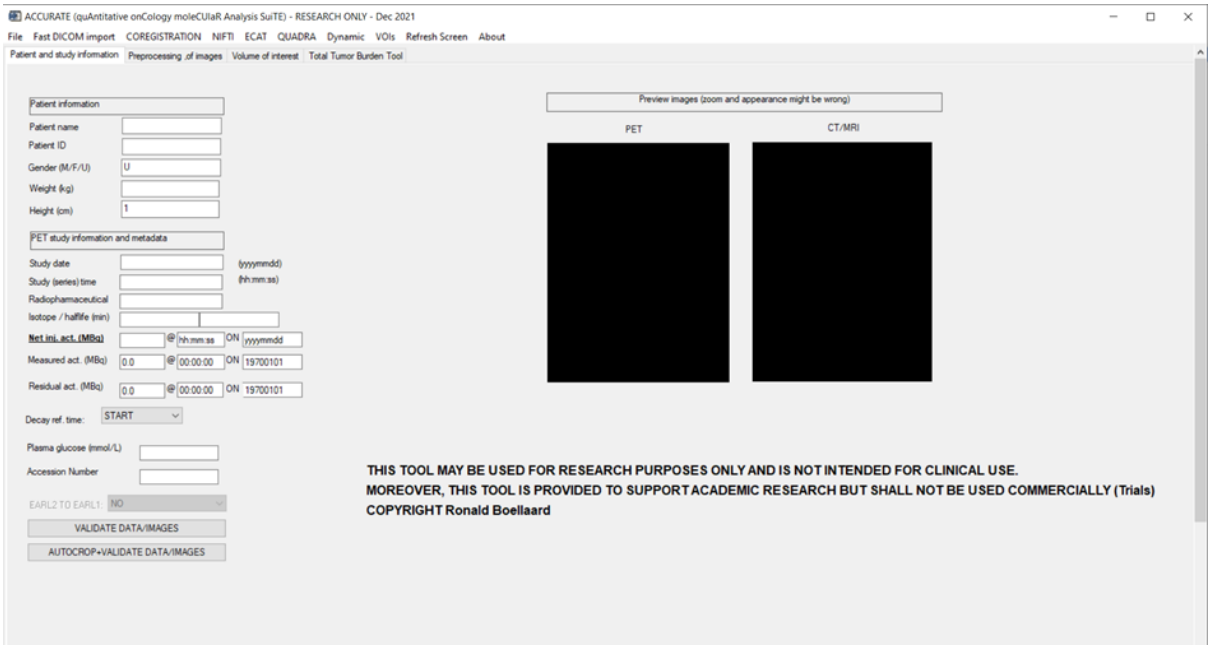
1-PROLUNG-004._ct	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_bqml	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_suv	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	PRJ-bestand	108 801 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	CSV-bestand van ...	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics.command	14/05/2021 14:05	Tekstdocument	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	CSV-bestand van ...	25 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	Tekstdocument	2 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics_config_v1.27_222_fixed_binsize_025	14/05/2021 14:05	Configuratie-instel...	3 kB
1-PROLUNG-004.acc.prj.PET.csv.suv	14/05/2021 14:05	Configuratie-instel...	1 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.voi	14/05/2021 14:05	VOI-bestand	21 760 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131501472021	13/05/2021 15:02	CSV-bestand van ...	4 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131502082021	13/05/2021 15:02	Adobe Acrobat-d...	74 kB
PatientAnonymous_20180720_143631_LUNGPET4	13/05/2021 15:02	CSV-bestand van ...	6 kB

Add _healthy to his newly created NII-file. Now there should be 2 files for the VOI. One which contains the tumor tissue, and one which will be used to create a volume of healthy tissue.



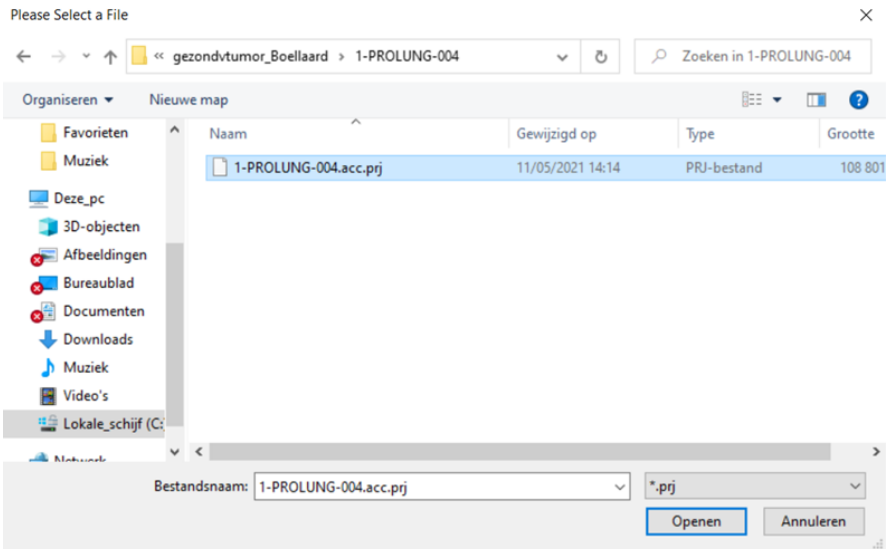
Naam	Gewijzigd op	Type	Grootte
1-PROLUNG-004._ct	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_bqml	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_suv	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	PRJ-bestand	108 801 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	CSV-bestand van ...	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics.command	14/05/2021 14:05	Tekstdocument	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	CSV-bestand van ...	25 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	Tekstdocument	2 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics_config_v1.27_222_fixed_binsize_025	14/05/2021 14:05	Configuratie-instel...	3 kB
1-PROLUNG-004.acc.prj.PET.csv.suv	14/05/2021 14:05	Configuratie-instel...	1 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.voi	14/05/2021 14:05	VOI-bestand	21 760 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4_gezond	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131501472021	13/05/2021 15:02	CSV-bestand van ...	4 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131502082021	13/05/2021 15:02	Adobe Acrobat-d...	74 kB
PatientAnonymous_20180720_143631_LUNGPET4	13/05/2021 15:02	CSV-bestand van ...	6 kB

Step 3: Open the Accurate tool (developed by a research team at Amsterdam UMC under Prof. Dr. Boellaard) by launching 'accurate4petct_v06022022', press 'Click to continue' and click yes for the scrollable version. The application should look like this:



Note that the tool can crash when you miss click or when an error occurs. When this happens, go back to Step 3.

Step 4: Next, the patient project has to be loaded into the tool. This can be done by clicking 'File' in the toolbar and 'LoadProject'. Access the folder of the patient you want to analyse and open the .prj file.

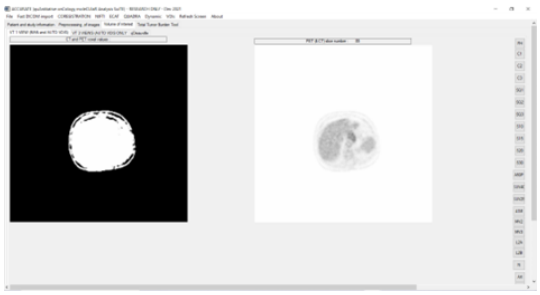


When done correctly, the PET image, patient info and radiopharmaceutical info should appear.

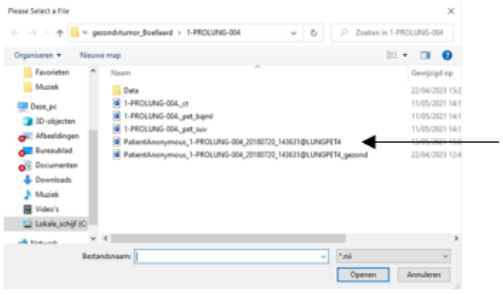
Step 5: Check if the patient weight, height and the plasma glucose fields are filled in correctly and click 'VALIDATE DATA/IMAGES'. A pop-up will appear saying you can continue the analysis, click 'ok'.

Open the 'Volume of interest' view by clicking the third tab under the toolbar.

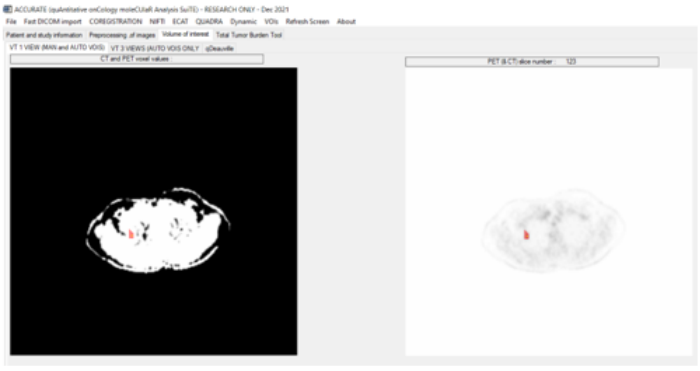
This view should appear:



Step 6: In the taskbar, click NIFTI -> Open VOI (Nifti). Now load in the VOI of the tumor.

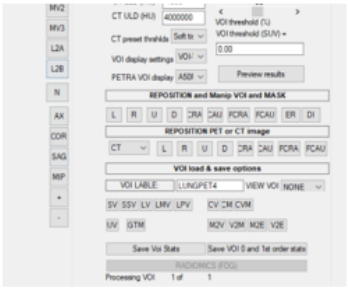


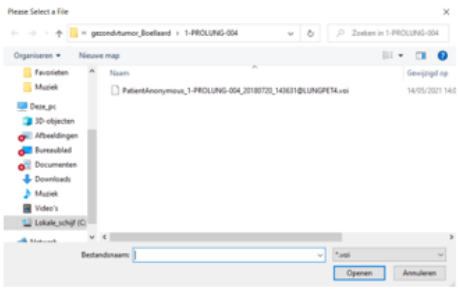
The VOI should appear marked in red.



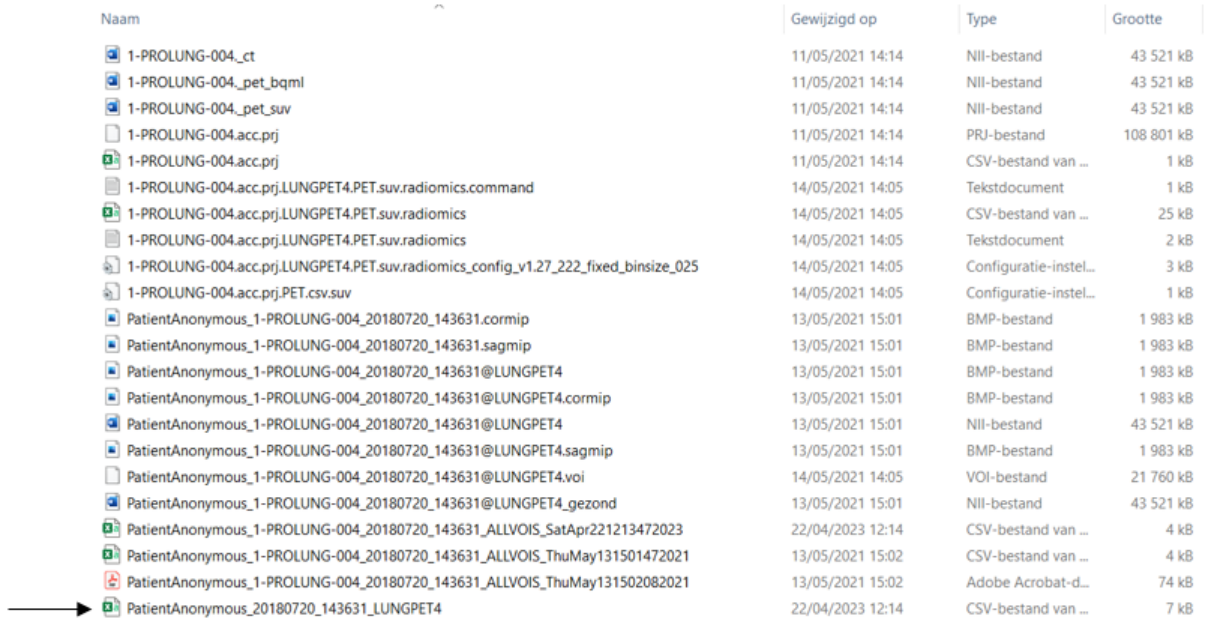
By scrolling right in the tool, the VOI information will become visible. Take note of the Threshold. This should be the same for all images in the study. In this case it is set at 0.00%.

Step 7: Press save Voi Stats -> select the VOI-file and click ‘open’.





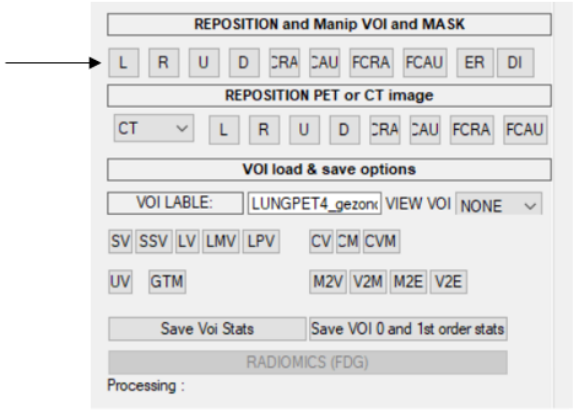
A new Excel file should appear in the patient folder containing the statistical data of the VOI.



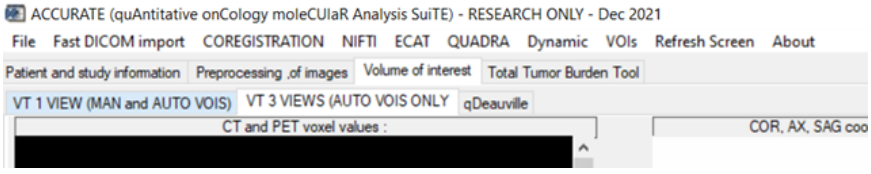
Naam	Gewijzigd op	Type	Grootte
1-PROLUNG-004._ct	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_bqml	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004._pet_suv	11/05/2021 14:14	NII-bestand	43 521 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	PRJ-bestand	108 801 kB
1-PROLUNG-004.acc.prj	11/05/2021 14:14	CSV-bestand van ...	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics.command	14/05/2021 14:05	Tekstdocument	1 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	CSV-bestand van ...	25 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics	14/05/2021 14:05	Tekstdocument	2 kB
1-PROLUNG-004.acc.prj.LUNGPET4.PET.suv.radiomics_config_v1.27_222_fixed_binsize_025	14/05/2021 14:05	Configuratie-instel...	3 kB
1-PROLUNG-004.acc.prj.PET.csv.suv	14/05/2021 14:05	Configuratie-instel...	1 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.cormip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.sagmip	13/05/2021 15:01	BMP-bestand	1 983 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4.voi	14/05/2021 14:05	VOI-bestand	21 760 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631@LUNGPET4_gezond	13/05/2021 15:01	NII-bestand	43 521 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_SatApr221213472023	22/04/2023 12:14	CSV-bestand van ...	4 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131501472021	13/05/2021 15:02	CSV-bestand van ...	4 kB
PatientAnonymous_1-PROLUNG-004_20180720_143631_ALLVOIS_ThuMay131502082021	13/05/2021 15:02	Adobe Acrobat-d...	74 kB
PatientAnonymous_20180720_143631_LUNGPET4	22/04/2023 12:14	CSV-bestand van ...	7 kB

Step 8: Open the VOI you want to use to create the healthy tissue VOI by clicking NIFTI -> Open VOI (Nifti) -> open the file with _healthy created earlier.

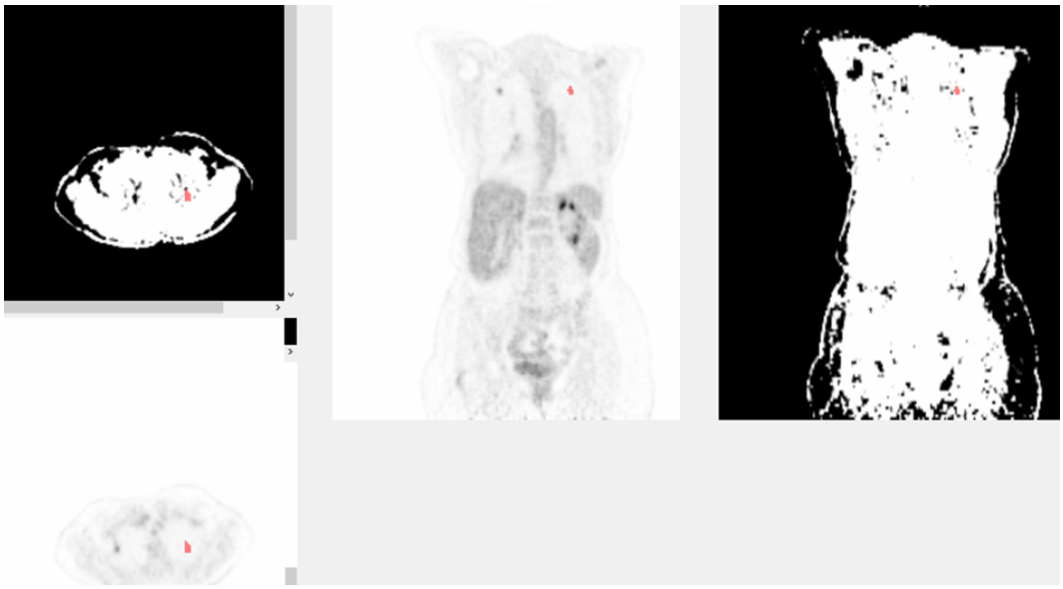
Step 9: Since this is just a copy of the tumor VOI, the Red marked volume will still be on the tumor. By using the reposition buttons as indicated below, the healthy tissue can be segmented.



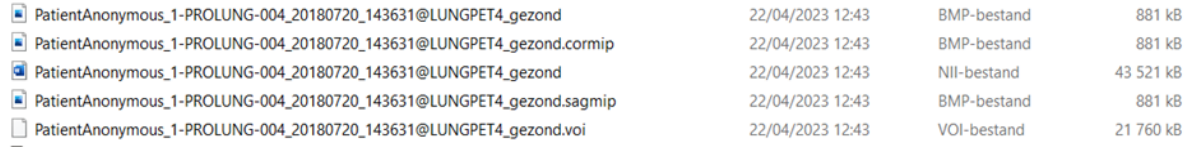
Important to note that the healthy VOI can only contain lung tissue. To get a better view of the images, click 'VT 3 VIEWS' .



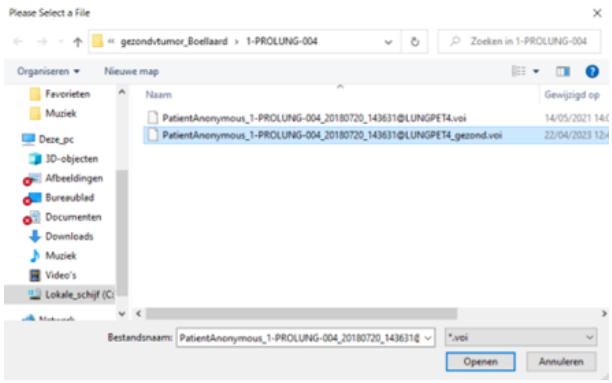
This is an example of a translated VOI to the opposite lung of where the tumor is.



Step 10: Save the newly created VOI by going to VOI in the taskbar -> Save VOI. The patient folder should now contain the following new files for the healthy tissue.

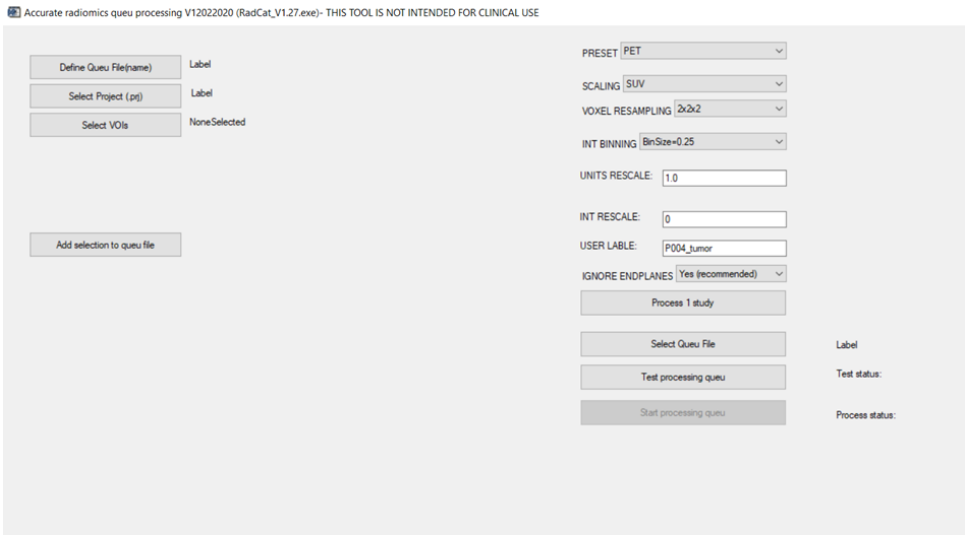


Step 11: Click save Voi Stats -> select the VOI-file of the healthy tissue and press ‘open’.



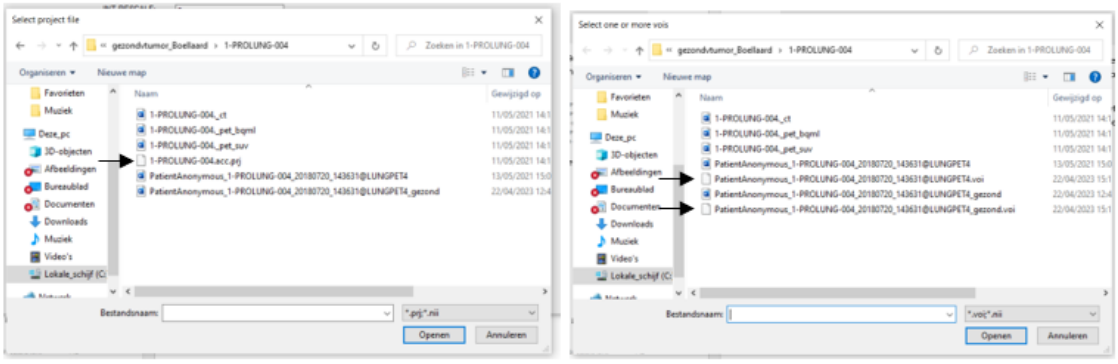
Now another Excel file should appear in the patient folder containing the statistical data of the healthy VOI.

Step 12: Close the Accurate tool and open the Radiomics tool. This should look as follows:



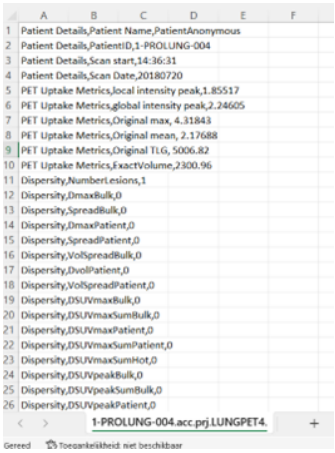
Step 13: Verify if the 'preset' and 'scaling fields' are correct. Next, change the 'USER LABLE'. Use appropriate naming for both tumor and healthy tissue as well as the patient number.

Step 14: Click 'Process 1 Study' -> select the project (.prj) -> select the VOI-file (.voi). Do step 13 and 14 twice per patient, once for the tumor radiomics extraction and once for the healthy tissue radiomics extraction.



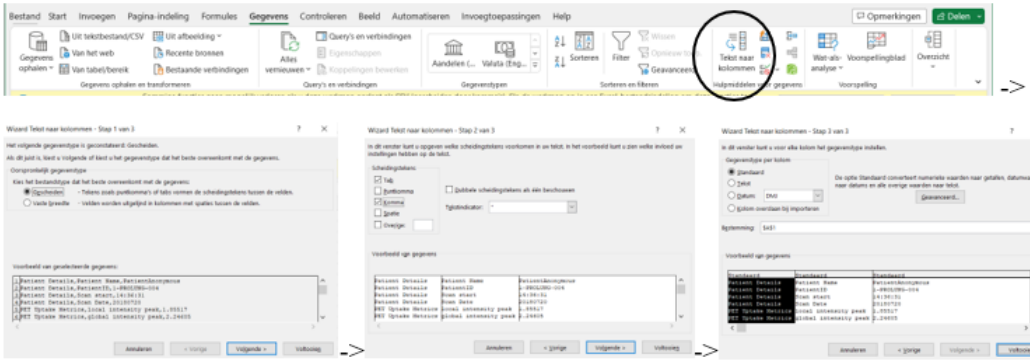
The following Excel files should be visible in the patient folder. These contain the extracted radiomics features.

Step 15: Open the Excel files. The data will look like this, in one column:



	A
1	Patient Details,Patient Name,PatientAnonymous
2	Patient Details,PatientID,1-PROLUNG-004
3	Patient Details,Scan start,14:36:31
4	Patient Details,Scan Date,20180720
5	PET Uptake Metrics,local intensity peak,1.85517
6	PET Uptake Metrics,global intensity peak,2.24605
7	PET Uptake Metrics,Original max, 4.31843
8	PET Uptake Metrics,Original mean, 2.17688
9	PET Uptake Metrics,Original TLG, 5006.82
10	PET Uptake Metrics,ExactVolume,2300.96
11	Dispersity,NumberLesions,1
12	Dispersity,DmaxBulk,0
13	Dispersity,SpreadBulk,0
14	Dispersity,DmaxPatient,0
15	Dispersity,SpreadPatient,0
16	Dispersity,VolSpreadBulk,0
17	Dispersity,DvolPatient,0
18	Dispersity,VolSpreadPatient,0
19	Dispersity,DSUVmaxBulk,0
20	Dispersity,DSUVmaxSumBulk,0
21	Dispersity,DSUVmaxPatient,0
22	Dispersity,DSUVmaxSumPatient,0
23	Dispersity,DSUVmaxSumHot,0
24	Dispersity,DSUVpeakBulk,0
25	Dispersity,DSUVpeakSumBulk,0
26	Dispersity,DSUVpeakPatient,0

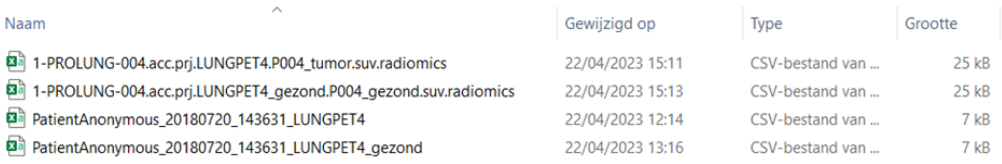
This can be solved by selecting the column containing the data and using the text to columns function under the data tab. Select comma as dividing symbol.



The dataset will look as follows:

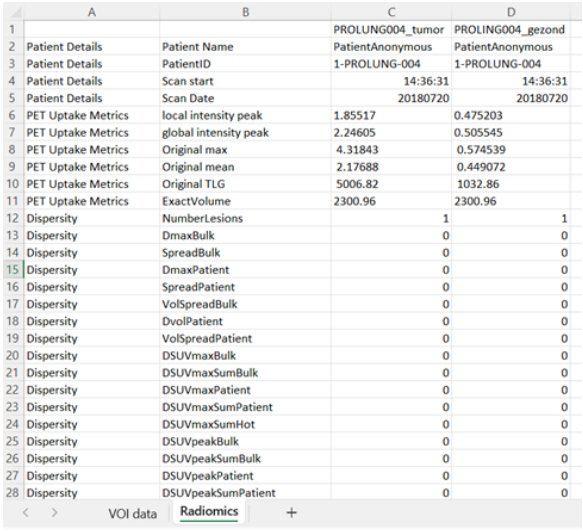
	A	B	C
1	Patient Details	Patient Name	PatientAnonymous
2	Patient Details	PatientID	1-PROLUNG-004
3	Patient Details	Scan start	14:36:31
4	Patient Details	Scan Date	20180720
5	PET Uptake Metrics	local intensity peak	1.85517
6	PET Uptake Metrics	global intensity peak	2.24605
7	PET Uptake Metrics	Original max	4.31843
8	PET Uptake Metrics	Original mean	2.17688
9	PET Uptake Metrics	Original TLG	5006.82
10	PET Uptake Metrics	ExactVolume	2300.96
11	Dispersity	NumberLesions	1
12	Dispersity	DmaxBulk	0
13	Dispersity	SpreadBulk	0
14	Dispersity	DmaxPatient	0
15	Dispersity	SpreadPatient	0
16	Dispersity	VolSpreadBulk	0
17	Dispersity	DvolPatient	0
18	Dispersity	VolSpreadPatient	0
19	Dispersity	DSUVmaxBulk	0

Repeat this process for the following four datasets per patient:



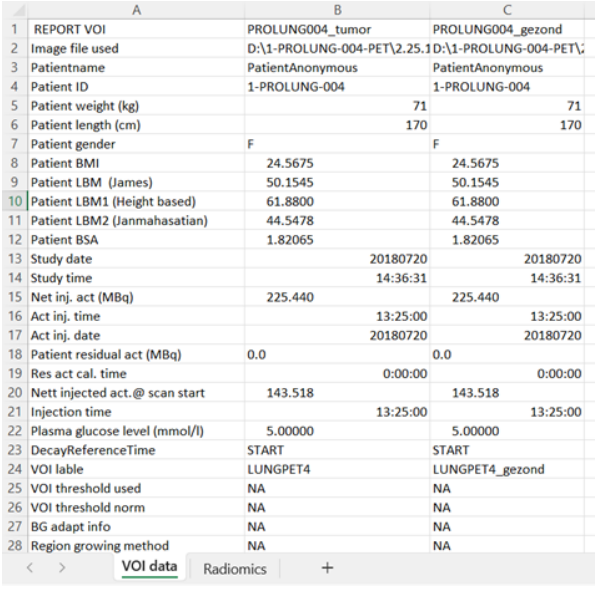
Naam	Gewijzigd op	Type	Grootte
1-PROLUNG-004.acc.prj.LUNGPET4.P004_tumor.suv.radiomics	22/04/2023 15:11	CSV-bestand van ...	25 kB
1-PROLUNG-004.acc.prj.LUNGPET4_gezond.P004_gezond.suv.radiomics	22/04/2023 15:13	CSV-bestand van ...	25 kB
PatientAnonymous_20180720_143631_LUNGPET4	22/04/2023 12:14	CSV-bestand van ...	7 kB
PatientAnonymous_20180720_143631_LUNGPET4_gezond	22/04/2023 13:16	CSV-bestand van ...	7 kB

Step 16: Combine all the extracted radiomics data per patient in one file. Row one contains the patients (two sets per patient). Do the same fort he VOI stats in a separate file/tab.



	A	B	C	D
1			PROLUNG004_tumor	PROLING004_gezond
2	Patient Details	Patient Name	PatientAnonymous	PatientAnonymous
3	Patient Details	PatientID	1-PROLUNG-004	1-PROLUNG-004
4	Patient Details	Scan start	14:36:31	14:36:31
5	Patient Details	Scan Date	20180720	20180720
6	PET Uptake Metrics	local intensity peak	1.85517	0.475203
7	PET Uptake Metrics	global intensity peak	2.24605	0.505545
8	PET Uptake Metrics	Original max	4.31843	0.574539
9	PET Uptake Metrics	Original mean	2.17688	0.449072
10	PET Uptake Metrics	Original TLG	5006.82	1032.86
11	PET Uptake Metrics	ExactVolume	2300.96	2300.96
12	Dispersity	NumberLesions	1	1
13	Dispersity	DmaxBulk	0	0
14	Dispersity	SpreadBulk	0	0
15	Dispersity	DmaxPatient	0	0
16	Dispersity	SpreadPatient	0	0
17	Dispersity	VolSpreadBulk	0	0
18	Dispersity	DvolPatient	0	0
19	Dispersity	VolSpreadPatient	0	0
20	Dispersity	DSUVmaxBulk	0	0
21	Dispersity	DSUVmaxSumBulk	0	0
22	Dispersity	DSUVmaxPatient	0	0
23	Dispersity	DSUVmaxSumPatient	0	0
24	Dispersity	DSUVmaxSumHot	0	0
25	Dispersity	DSUVpeakBulk	0	0
26	Dispersity	DSUVpeakSumBulk	0	0
27	Dispersity	DSUVpeakPatient	0	0
28	Dispersity	DSUVpeakSumPatient	0	0

VOI data | Radiomics | +



	A	B	C
1	REPORT VOI	PROLUNG004_tumor	PROLUNG004_gezond
2	Image file used	D:\1-PROLUNG-004-PET\2.25.1	D:\1-PROLUNG-004-PET\
3	Patientname	PatientAnonymous	PatientAnonymous
4	Patient ID	1-PROLUNG-004	1-PROLUNG-004
5	Patient weight (kg)	71	71
6	Patient length (cm)	170	170
7	Patient gender	F	F
8	Patient BMI	24.5675	24.5675
9	Patient LBM (James)	50.1545	50.1545
10	Patient LBM1 (Height based)	61.8800	61.8800
11	Patient LBM2 (Janmahasatian)	44.5478	44.5478
12	Patient BSA	1.82065	1.82065
13	Study date	20180720	20180720
14	Study time	14:36:31	14:36:31
15	Net inj. act (MBq)	225.440	225.440
16	Act inj. time	13:25:00	13:25:00
17	Act inj. date	20180720	20180720
18	Patient residual act (MBq)	0.0	0.0
19	Res act cal. time	0:00:00	0:00:00
20	Nett injected act.@ scan start	143.518	143.518
21	Injection time	13:25:00	13:25:00
22	Plasma glucose level (mmol/l)	5.00000	5.00000
23	DecayReferenceTime	START	START
24	VOI lable	LUNGPET4	LUNGPET4_gezond
25	VOI threshold used	NA	NA
26	VOI threshold norm	NA	NA
27	BG adapt info	NA	NA
28	Region growing method	NA	NA

VOI data | Radiomics | +

Step 17: Repeat these steps for each patient in the study and add the datasets to the file from step 16.