D I P L O M A T H E S I S

APPLICATION OF RADIOMICS PARAMETERS IN IN VIVO MEDICAL IMAGING

NÓRA JUHÁSZ

Thesis Supervisor: DR. KRISZTIÁN SZIGETI
Research Associate
Department of Biophysics and Radiation Biology
Semmelweis University-Faculty of Medicine

Co-Supervisor: DR. DÁVID LÉGRÁDY
Associate Professor
BME, Institute of Nuclear Techniques
Department of Nuclear Techniques

BME

2021

MŰEGYETEM 1782

Diplomamunka feladat a Fizikus mesterképzési szak hallgatói számára

A hallgató neve: Juhász Nóra	specializációja: Orvosi fizikus MSc.
A záróvizsgát szervező tanszék neve: NTI	

A témavezető neve: Dr. Szigeti Krisztián - munkahelye: Sememlweiss Egyetem Biofizikai Intézet - beosztása: Laborvezető - email címe: krisztian.szigeti@gmail.com	A konzulens neve: Dr. Légrády Dávid - tanszéke: NTI - beosztása: egy.docens - email címe: legrady@reak.bme.hu

A kidolgozandó feladat címe: Diagnosztikai célú invivo multimodális képalkotás parametrizációja

A téma rövid leírása, a megoldandó legfontosabb feladatok felsorolása:

Az elmúlt években az in vivo képalkotás jelentős teret nyert a preklinikai gyógyszerkísérletekben és az egyetemi élettani kutatásokban. A NanoSPECT/CTnek köszönhetően a kutatók az egerekben a molekuláris folyamatokat nanoliteres felbontással tudják vizsgálni – pontos anatómiai lokalizációval. A kisállatokon végzett kísérletek során kapott nagy mennyiségű képi információ feldolgozásában nagyon fontos a hatékony és gyors automatikus orvosdiagnosztikai célú adatkiértékelés. A feladat a képi 3D rekonstrukciók során kapott képek kiértékelése, és olyan geometriai parametrizációs eljárások kifejlesztése, amelyek elősegítik a hatékonyabb és pontosabb diagnosztikai döntéseket. A feladat során a hallgató megismeri a képalkotó diagnosztikai szakirodalomban használt módszereket, továbbá új és komplexebb, a jelenlegi technológiai követelményeknek megfelelő algoritmusok fejlesztésében vesz részt. A feladat során megismeri és alkalmazza ezen algoritmusok in vivo és in vitro validációját is.

A záróvizsga kijelölt tételei:

Dátum:

Hallgató aláírása:	Témavezető aláírása*:	Tanszéki konzulens aláírása:	A témakiírását jóváhagyom (tanszékvezető aláírása):

*A témavezető jelen feladatkiírás aláírásával tudomásul veszi, hogy a BME TVSZ 145. és 146.§ alapján az egyetem a képzési célok megvalósulása érdekében a szakdolgozatok, illetve diplomamunkák nyilvánosságát tartja elsődlegesnek. A hozzáférés korlátozása csak kivételes esetben, a dékán előzetes hozzájárulásával lehetséges.

PLÁGIUM – NYILATKOZAT

/Nyilatkozat a diplomamunka készítésére vonatkozó szabályok betartásáról/

Alulírott **Juhász Nóra (Neptun-kód: N7QHF5)** a Budapesti Műszaki és Gazdaságtudományi Egyetem fizikus MSc. szakos hallgatója kijelentem, hogy ezen diplomamunkát meg nem engedett segédeszközök nélkül, önállóan, a témavezető irányításával készítettem, és csak a megadott forrásokat használtam fel.

Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból vettem, a forrás megadásával jelöltem.

Budapest, 2021. 01.07.

aláírás

PLAGARISM STATEMENT

This project was written by me, **Nora Juhasz (Neptun-code: N7QHF5)** - as a student of the Physicist MSc. program on the Budapest University of Technology and Economics - and in my own words, except for quotations from published and unpublished sources which are clearly indicated and acknowledged as such. I am conscious that the incorporation of material from other works or a paraphrase of such material without acknowledgement will be treated as plagiarism, subject to the custom and usage of the subject, according to the University Regulations on Conduct of Examinations. The source of any picture, map or other illustration is also indicated, as is the source, published or unpublished, of any material not resulting from my own experimentation, observation or specimen-collecting.

Budapest, 07.01.2021

signature

Tartalom

1. Introduction

The technologies of medical imaging play a core role in the whole process of cancer management [1]. It can be accurately determined several important data such as tumor location, size, cancer metastasis, and whether treatment involves critical anatomical structures. However, the main advantage of the imaging technology is the ability to visualize tissue in non-invasive ways to avoid injury from invasive biopsy. [2]

There are many mature imaging technologies, such as Computed Tomography (CT) imaging, Positron-emission Tomography (PET) imaging, Magnetic Resonance Imaging (MRI) and medical ultrasound imaging. Different modalities of molecular imaging technology can observe different information. CT images can assess the cancer structural features, especially soft tissue organs such as the spinal cord, lung, liver, pancreas, etc., but it cannot describe the functional details of solid tumors and it is difficult to find initial symptoms of cancer. PET images can detect the presence of early cancer cells and their molecule metabolic activities but have a poorer ability to describing structural tissue information [2]. MRI images are superior to CT in soft tissue such as nerves, blood vessels, muscles, etc. However, the detection of lung, liver, pancreas, adrenal gland and prostate is worse than CT and more expensive [3].

Although these imaging techniques are widely implemented in hospitals for treating patients, the expression of information from imaging is limited due to the oversimplification of internal diagnostic criteria. Therefore, there is an urgent need for an approach for quantitatively mining more valuable information from imaging to diagnose, treat and monitor disease. With the rapid development of hardware devices and imaging agents in medical imaging technology, a computer-assisted standard quantitative extracted features method, known as radiomics, as the true transformative power of medical imaging analysis.

2. Literature

2.1 Radiomics

The quantitative analysis of tumor characteristics based on medical imaging in an emerging field of research [1] called radiomics. Radiomics, as its name already shows, is a process that converts medical images into high dimensional quantitative features. The features can be used as a training data for decision making in medicine. Radiomics shows important roles in precision medicine, thanks to its non- invasive characteristics.

Radiomics begins with acquisition of high-quality images, followed by segmentation of region or volume of interest (ROI/VOI) extraction of quantitative features from the ROI/VOI, which are analyzed along with clinical and genomic data to develop diagnostic, predictive, or prognostic models for decision support. Radiomic analysis exploits advanced image analysis tools and the rapid development and validation of medical imaging data that uses image/based signatures for precision diagnosis and treatment.

Radiomics has shown its ability in many areas in medicine, including prediction of patient survival, cancer recurrence, cancer stages, cancer risks and genetic features. The prediction ability of radiomics makes it a powerful tool for treatment assessment in medicine.[2]

Studies have shown that quantitative imaging features derived from computed tomography, positron emission tomography and magnetic resonance imaging scans could add value in the prediction of outcome parameters in oncology. [3]

For example, Nie et al. [5] evaluated multiparametric MRI features in predicting response after preoperative chemoradiation therapy for locally advanced rectal cancer and were able to build models with improved predictive value over conventional volume-based imaging metrics. [6]

2.2 MRI (FLAIR)

In this thesis I am going to present MRI examinations, so I would like to give a brief introduction of its theoretical foundation.

MRI is a non-invasive tomographic method based on the magnetic resonance of nuclei. Its advantage over CT is that it has better contrast resolution in areas of soft tissues and the patient does not have to be exposed to the damaging effects of ionizing radiation.

There is a structural MRI examination (sMRI), where the morphology of organs and tissues can be observed, and there is also a functional magnetic resonance imaging (fMRI), which can be used to obtain information about the function of examined organs and metabolic processes.

The body to be examined is placed in a strong external magnetic field. The vector quantity characterizing the strength of the magnetic field is the magnetic induction (B), the unit of which is

tesla (T). MRI involves the use of 1–3 tesla-strength magnetic fields [8]. Under the influence of the magnetic field, the elementary magnets (in our case the protons) are oriented. In addition to the magnetic field, we give radio frequency pulses, which, when eliminated, try to restore the nuclear spins to their original random arrangement.

With this method, the pixels are measured one by one. Protons can be characterized by three basic physical parameters for imaging: the density of protons and the two relaxation times associated with loss of orientation (i.e., T1 and T2 relaxation times) of nuclear spins oriented through the absorption of electromagnetic radiation. During T1 or longitudinal relaxation, the direction of macroscopic magnetization wants to return to the direction of the magnetic field. T1-weighted images are rich in detail, on them the fat has an increased signal intensity, the liquor has a reduced signal intensity, the gray matter can be well distinguished from the white matter. During T2 or transverse relaxation, the elementary magnets tend to move away from each other in two dimensions of space. In T2-weighted images, liquor gives an enhanced signal intensity, bordered by strong contrast, the gray matter gives a darker image, and the white matter gives a lighter image. [9]

After the T1 and T2 MRI sequences the third commonly used sequence is the fluid-attenuated inversion recovery, or as I mentioned above FLAIR, is an MRI sequence with an inversion recovery set to null fluids. For example, it can be used in brain imaging to suppress cerebrospinal fluid (CSF) effects on the image.[4] So, in other words, FLAIR is an MRI technique that shows areas of tissue T2 prolongation as bright while suppressing (darkening) cerebrospinal fluid (CSF) signal, thus clearly revealing lesions in proximity to CSF. [8-10]

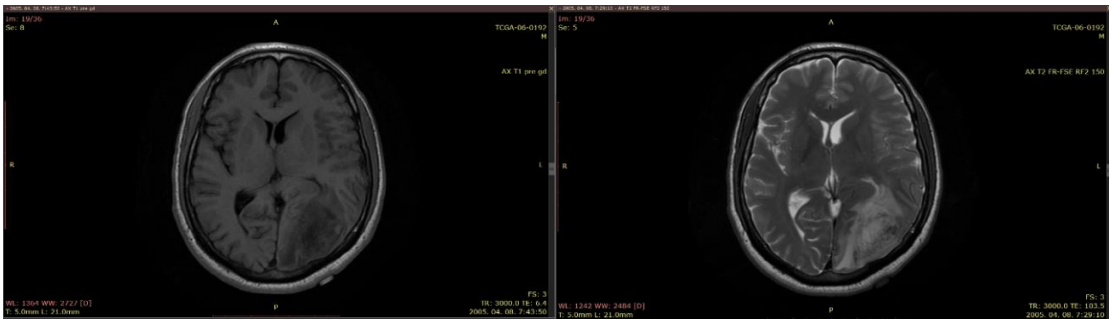


Figure 1.: The same glioblastoma multiforme tumor in T1- and T2- weighted MRI sequence

The Flair sequence is similar to a T2-weighted image except that the TE (Time to Echo) and TR (Repetition Time) times are very long. By doing so, abnormalities remain bright but normal CSF fluid is attenuated and made dark. This sequence is very sensitive to pathology and makes the differentiation between CSF and an abnormality much easier. [11]

	TR (msec)	TE (msec)
T1-Weighted (short TR and TE)	**500**	**14**
T2-Weighted (long TR and TE)	**4000**	**90**
Flair (very long TR and TE	**9000**	**114**

Table 1.: Commonly used MRI Sequences and their approximate TR and TE times

Tissue	T1-Weighted	T2-Weighted	Flair
CSF	Dark	Bright	Dark
White Matter	Light	Dark Gray	Dark Gray
Cortex	Gray	Light Gray	Light Gray
Fat (within bone marrow)	Bright	Light	Light
Inflammation (infection, demyelination)	Dark	Bright	Bright

Table 2.: Comparison of T1 vs. T2 vs. Flair (Brain)

The aim of a FLAIR sequence is to suppress liquid signals by inversion-recovery at an adapted TI. Water has a long T1. Nulling of the water signal is seen at TI of 2000 milliseconds. As in the case of the other inversion-recovery sequences, an imaging sequence of the fast spin echo type is preferable to compensate the long acquisition time linked to long TR. [7]

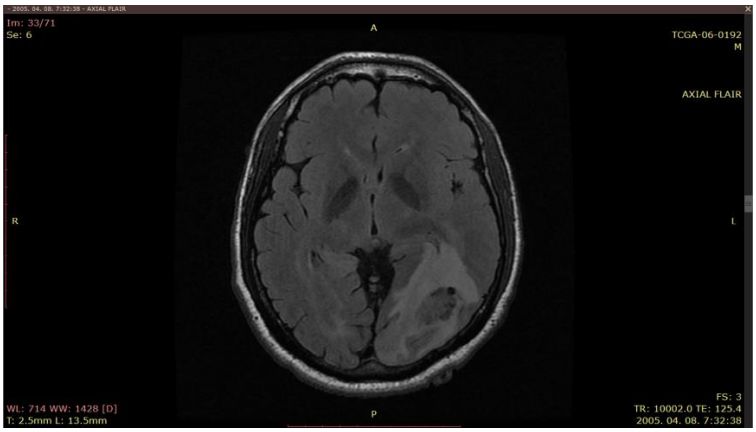


Figure 2.: Glioblastoma multiforme tumor in FLAIR MRI sequence

2.3 Glioblastoma multiforme

In the first stage of my thesis we had accessed a huge dataset from The Cancer Genome Atlas (TCGA) Glioblastoma Multiforme (GBM) and Low Grade Glioma (TCGA-LGG) collection, publicly available in the Cancer Imaging Archive [12] website.

After examining both types of cancer and their MRI scans, I had chosen the T2- Flair MRI scans of the glioblastoma multiforme (GBM) dataset due to the highest number of valid images for this study.

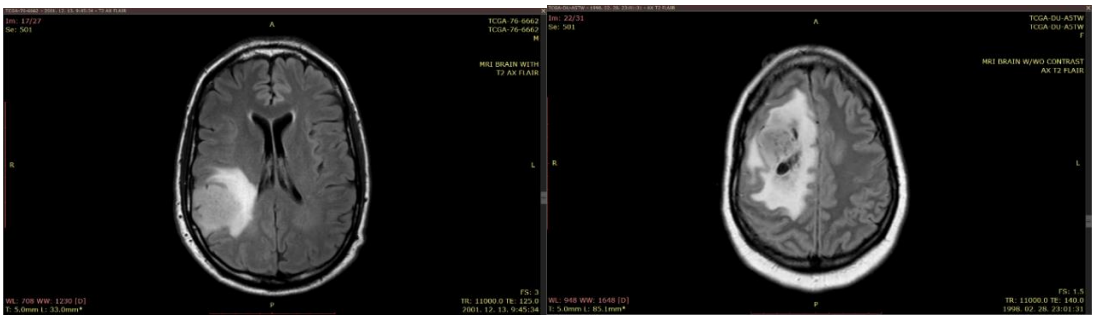


Figure 3.: left: Glioblastoma multiforme (T2-FLAIR); right: Low Grade Glioma (T2-FLAIR)

Glioblastoma multiforme (GBM) is the most aggressive and highly invasive high-grade glioma tumor with poor prognosis. It can occur in the brain or spinal cord. Glioblastoma develops from tar-shaped glial cells called astrocytes and oligodendrocytes that support the health of nerve cells within the brain. [13]

Because glioblastoma grows rapidly, its most common symptoms are due to a sudden increase in intracranial pressure. Symptoms include headache, dizziness, nausea, vomiting, memory loss, and personality change. Depending on the localization, different symptoms may appear, such as speech disturbances, vision problems, or unilateral paralysis [14].

Its diagnosis can be clarified with the help of biomarkers, primarily from radiological and biopsy samples, and possibly from blood. For example, immunohistochemical detection of the astrocyte-specific protein, GFAP (glial fibrillary acidic protein), helps to identify the tumor [15]. The exact root cause of the tumor is unknown.

2.3.1 The role of inflammation in glioblastoma

Inflammation is a natural response to any injury or damage to the body. A protective reaction also begins around glioblastomas, an aqueous, inflammatory yard develops around the tumor. Sometimes the tumor itself is quite small, one to two centimeters, but most of the brain is edematous and we detect this on the tumor with imaging. Acute inflammation can be transformed into chronic inflammation, which favors tumor growth [14]. We can conclude that, in terms of

both tumor growth and recognition, the brain tumor and the inflammation surrounding it can go hand in hand.

2.3.2 Treatment of glioblastoma multiforme

Treatment of glioblastoma is very difficult because the tumor is resistant to classical surgical therapy, certain brain functions may be impaired during surgery, and nervous system regeneration is limited, and because many drugs cannot cross the blood-brain barrier. Therapy consists of treating symptoms as well as reducing tumor size.[15]

Supportive care consists of improving the patient's neurological function, using corticosteroids and anticonvulsants. Steroids can reduce the edema around the tumor, their most commonly prescribed type being dexamethasone. Antiepileptic drugs inhibit the development of epileptic seizures, most notably levetiracetam. Brain-stimulating drugs can reduce fatigue; antiemetics and antidepressants may occasionally be considered as treatments [12].

Tumor reduction procedures do not cause complete recovery in the vast majority of cases, and the combined use of several methods is recommended.

Usually, the first step is surgical treatment, which aims to obtain a tumor sample for an accurate diagnosis and to remove as many tumor parts as safely as possible. The lesion is almost impossible to completely remove, especially if it is close to important brain centers (such as speech and coordination). Due to its protruding structure, it is difficult to find the boundary between cancerous and healthy tissue. Partial removal may reduce the rate of spread. Attempts are being made to reduce residual tumor growth with radiation therapy, chemotherapy, and biologic therapy. [13]

Radiation therapy usually lasts 5-6 weeks for 5 days a week, patients receive external radiation. Chemotherapy consists of 6 weeks of temozolomide treatment concomitantly with radiotherapy. The drug is an alkylating agent that crosses the blood-brain barrier and aids in the sensitivity of cells to radiation therapy [14].

The prognosis is poor, survival is around 2-3 years, 3 months without treatment. Not all glioblastomas are the same, so different patients respond better to different treatments.

3. Aim of the thesis

This thesis aims to explore the potential of radiomics in prognostic medical applications and to understand the relationships between different texture features. Because of the high dimensional data structure, we have proposed mature feature selection by classification models to select the most suitable radiomic pattern.

Moreover, the aims of this thesis are to show that radiomics has a great potential to improve the clinical decision support system. My purpose is to provide deeper understanding of fundamental technical and methodology aspects. I am going to use and introduce general and specific analysis methods of radiomics analysis.

My aim was, besides the understanding of the radiomics method, was to evaluate the potential application of textural analysis as used as a prognostic tool in glioblastoma, and to identify radiomic predictors to find the best subset one by the applied classifier.

4. Methods

The process of radiomics can be categorized in five steps as follows: image acquisition, image segmentation, feature extraction, feature selection and integrated analysis.

4.1 Image Acquisition

For any study, the first step is to acquire the appropriate images. As I have mentioned the MRI scans are publicly available for everyone on the Cancer Imaging Archive website.[2] Our purpose was to work with the largest dataset as possible, so I had chosen the Glioblastoma Multiforme (GBM) dataset with the T2- FLAIR scans. A total number of 138 patients were enrolled in this study (full dataset).

After examining 138 GBM T2-FLAIR MR sequenced images there was only evaluated the subset of cases, 77 patients. There were 29 scans where the tumor was on both sides of the brain, since we used one of the sides as a control group (the non-tumorous one), those scans couldn't be used in this study. 8 scans had artifacts, so the segmentation wouldn't have been accurate, 13 scans were obtained from the Sagittal plane, and 4 more from the Coronal plane. 7 scans were made in the 1990s thus the quality and number of slices of these images wouldn't allow us any further process. Thus, our restricted dataset contains 77 patients.

4.2 Segmentation

The Medical Imaging Interaction Toolkit (MITK Workbench, v.2018.04.2) is a free open-source software system for development of interactive medical image processing software. [17]. This software was used to display and process radiological images. The segmentation was manually accomplished by me. I had separated the brain to the left and right side one side with the tumor and one without.

Depending on which side has the GBM, the ROIs have to be marked. For ease of programming, I used the 1 and 0 values for differentiation in the evaluating program.

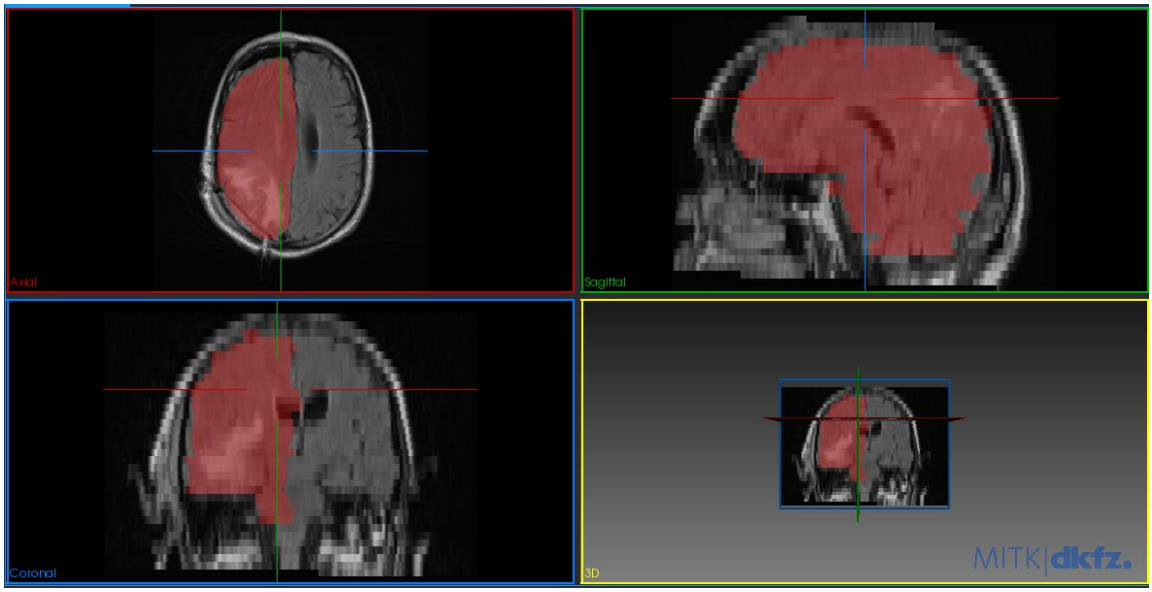


Figure 4.: Example: Left brain T2-FLAIR MRI sequence segmentation (with left side GBM)

The whole scans were in DICOM file format, however MITK made it possible to save the modified images, segmentations in DICOM and NifTi format as well. Both of the formats are easy to work with in MATLAB, because there are already existing built in functions to them, which helps in the evaluation.

4.3 Feature Extraction

The third step is feature extraction, which computes hundreds of features from a given region of interest. There are plenty types of analysis what we are able to use on a (medical) images. *Figure 5.* shows some of the main kinds.

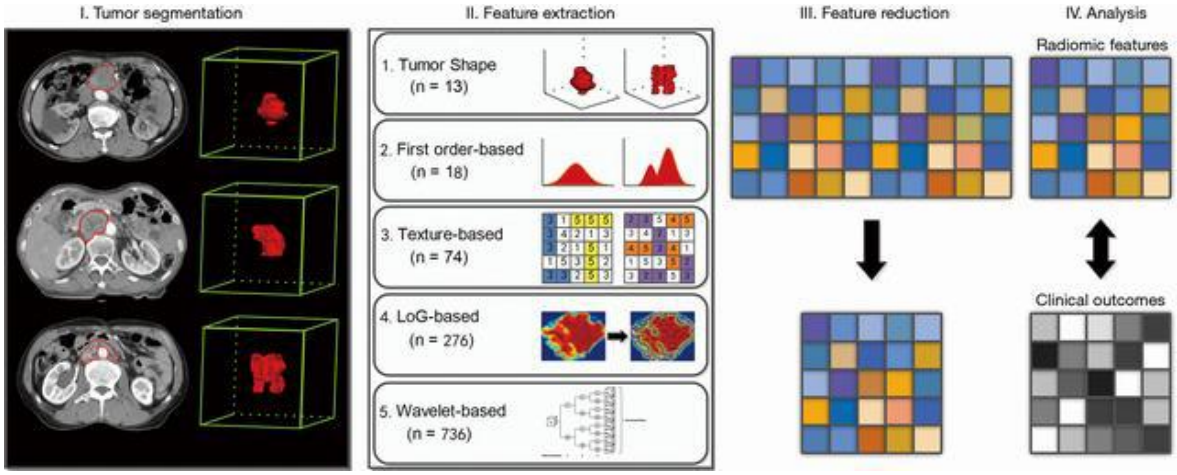


Figure 5.: Example of a workflow of radiomic analysis and different kind of feature extraction types [12]

The features are defined using mathematical formulas and are thus objective imaging features. The features are broadly classified into four categories: morphological (tumor shape), histogram- or

first order-based, textural, and transform-based (LoG- or Wavelet-based) features. In the case of the morphological process the features reflect the physical characteristics of the ROI such as the tumor area, volume, compactness. The first order based feature extraction and texture analysis exact information from the intensity of the pixels from the ROI. The main features in this case are main, median, entropy, etc. Texture-based analysis and texture features consider voxels and their neighbors grey-scale level. Transform-based features involve transforming the original image with a user-selected transform, such as low or high pass filtering, although these filtering make changes in the grey level.[39]

In this thesis the feature extraction is based on texture analysis. To calculate the various spatial parameters, a program called 'Radiomics Master' written in MATLAB (MathWorks, v.2020a) by Martin Vallieres [23] is available publicly. My program, calculations and results were built on this evaluation program. I had modified some of its algorithms to fit the files I had to work on and to gain specific evaluation outputs. I also had to implement a file management algorithm to evaluate the correct files in the folder system

Firstly, since the MRI scan slices were separated in every patient dataset, I wrote a merging loop which made it easier to handle with the large DICOM dataset.

Still working on the file management, the next step was to open up and merge the matching MRI scans, with its left and right sided ROIs and to get a mask and a volume data from the area.

With a small change in the algorithms (especially with the prepareVolume function), the program became suitable for parameterizing an image obtained with either functional (SPECT, PET) or morphological (MRI, CT) modality.

In general, the algorithm's process chart looks like:

N := length(folder system)	
i = 1 : N	
	files = dicomread[1] (list[2](i))

[1]dicomread: built-in function, reads the image data from the compliant Digital Imaging and Communications in Medicine (DICOM) file filename.[32]

[2]list: list of the dicom filenames

i,j,k ∈ **Z;**

```
i = 1 : length (files)
    [ROIonly, levels] = prepareVolume (…)
    nbins = length(histcounts[3](ROIonly))
    output.variance, output.skewness,…= getGlobalTextures(ROIonly, nbins)
group := [0 if the tumor is on the left side, 1 if the tumor is on the right side]
j = 1 : n_patient[4]
    k = 1 : length(files)
        featurematrix (j,k) = output.variance
        featurematrix (j,k) = output.skewness
        …
[B, fitinfo[6]] = Lassoglm[5](…)
LassoPlot(B, fitinfo,…)
```

[3]histcounts(X) partitions the X values into bins, and returns the count in each bin, as well as the bin edges. The histcounts function uses an automatic binning algorithm that returns bins with a uniform width, chosen to cover the range of elements in X and reveal the underlying shape of the distribution [32].

[4]n_patient: the number of the patients

[5]returns penalized, maximum-likelihood fitted coefficients for generalized linear models of the predictor data X and the response y, where the values in y are assumed to have a normal probability distribution. Each column of B corresponds to a particular regularization coefficient in Lambda. By default, Lassoglm performs Lasso regularization using a geometric sequence of Lambda values.

[6][B, FitInfo] = Lassoglm(___) returns the structure FitInfo, which contains information about the fit of the models, using any of the input arguments in the previous syntaxes.

The code's main function is the prepareVolume function, which uses our dataset properties, so let's take a deeper look at the in- and outputs of it.

$$[\boldsymbol{ROIonly}, \boldsymbol{levels}] = \boldsymbol{prepareVolume}(\boldsymbol{volume}, \boldsymbol{mask}, \boldsymbol{scanType}, \boldsymbol{pixelW}, \boldsymbol{sliceS}, \boldsymbol{R}, \boldsymbol{scale}, \boldsymbol{textType}, \boldsymbol{quantAlgo}, \boldsymbol{Ng})$$

where the inputs are the followings:

- *volume*: 2D (or 3D) array containing the medical images to analyze
- *mask*: 2D (or 3D) array of dimensions corresponding to 'volume'. The mask contains 1's in the region of interest (ROI), and 0's elsewhere

- *scanType*: String specifying the type of scan analyzed. Either 'PETscan', 'MRscan' or 'Other'. In this case we had chosen the ‘MRscan’ input.
- *pixelW*: Numerical value specifying the in-plane resolution (mm) of 'volume', which is 1.
- *sliceS*: Numerical value specifying the slice spacing (mm) of 'volume'. Put a random number for 2D analysis.
- *R*: Numerical value specifying the ratio of weight to band-pass coefficients over the weight of the rest of coefficients (HHH and LLL). Provide R=1 to not perform wavelet band-pass filtering. We used 1 as we didn’t want to perform wavelet band-pass filtering.
- *Scale*: Numerical value specifying the scale at which 'volume' is isotropically resampled (mm). If a string 'pixelW' is entered as input, the volume will be isotropically resampled at the initial in-plane resolution of 'volume' specified by 'pixelW'.
- *textType*: String specifying for which type of textures 'volume' is being prepared. Either 'Global' or 'Matrix'. If 'Global', the volume will be prepared for Global texture features computation. If 'Matrix', the volume will be prepared for matrix-based texture features computation (i.e. GLCM, GLRLM, GLSZM).[27-31]
- *quantAlgo*: String specifying the quantization algorithm to use on 'volume'. Either 'Equal' for equal-probability quantization, 'Lloyd' for Lloyd-Max quantization, or 'Uniform' for uniform quantization. Use only if textType is set to 'Matrix'.
The image first needs to be quantized to a reasonable bit depth prior to calculate the grey-level matrices. In the original paper, Haralick[26] proposes using an equal probability quantizer - in order for the extracted textures to be invariant under monotonic gray-tone transformations - thus we had chosen the Equal-probability quantization as well.[26].
- *Ng*: Integer specifying the number of gray levels in the quantization process. Use only if textType is set to 'Matrix'. [27-31]
All in all, the prepareVolume function and its inputs will look like:

$$[\boldsymbol{ROIonly}, \boldsymbol{levels}] = \boldsymbol{prepareVolume}(\boldsymbol{vol}, \boldsymbol{mask}, '\boldsymbol{MRscan}', \mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}, '\boldsymbol{Matrix}', '\boldsymbol{Equal}', \mathbf{32});$$

The outputs are:
- *ROIonly*: Smallest box containing the ROI, with the imaging data of the ready for texture analysis computations. Voxels outside the ROI are set to NaNs.
- *levels*: Vector containing the quantized gray-levels in the tumor region (or reconstruction levels of quantization).[23]

4.4 Texture Analysis

In this thesis I worked with texture- based radiomics analysis, since they are a central type of features that can be extracted from the region of interest. However, textures remain the core of radiomic feature computation, are given their higher-order characterization of spatial patterns in imaging volumes. [21]

In this thesis, texture features from four major categories were extracted*:*

I) Global features;

II) Gray-Level Co-occurrence Matrix (GLCM) features;

III) Gray-Level Run-Length Matrix (GLRLM) features;

IV) Gray-Level Size Zone Matrix (GLSZM) features.

To aim a deeper understanding of the background of the MATLAB code and the texture analysis, let's discuss firstly the 3 Grey-Level Matrix calculation methods.

4.4.1 Gray-Level Co-occurrence Matrix

Grey level Co-occurrence matrix is one of the earliest method feature extraction via texture analysis. It was proposed at first by Haralick at all [ref] in 1973. Grey Level Co-occurrence matrix or Co-occurrence distribution is defined over an image to be the distribution of co-occurring values at a given offset or represents the distance and angular spatial relationship over an image subregion of a specific size. As its name speaks GLCM is created from a gray scale image. In some simple words GLCM calculates the greyscale intensity or tone, so how often is a pixel with a specific grey level, thus it is based on the resolution of the image. GLCM has three main directions for the calculation. It is horizontal (0), vertical (90) and diagonal (-45 or 135).

If *P* defines the Gray Level Co-Occurrence Matrix of a quantized ROI imaging volume *V (x, y, z)* with isotropic voxel size, such as in this study the two sides of the brain, then each entry *P (i, j)* of *P* represents the number of times voxels of gray level *i* are neighbors with voxels of gray level *j* in *V*. [26]

The gray-level co-occurrence matrix is going to be a symmetric matrix with the size of the predefined number of quantized gray levels in the voxel.

4.4.2 Gray-Level Run-Length Matrix

The concept of GLRLM was proposed by Galloway in 1975. The Gray Level Run-Length matrix (GLRLM) quantifies how many consecutive pixels have the same value along a predefined direction. The rows of the matrix represent the discretized gray level and the columns the run—

length nonuniformity and the run-percentage – to emphasize different properties of these matrix [40].

Let P define the GLRLM of a quantized ROI imaging volume $V_Q(x, y, z)$ with isotropic voxel size. Each entry $P(i, j)$ of P represents the number of runs of gray level i and of length j in $V_Q(x, y, z)$. A run is a 1D line of connected voxels with an identical gray level. The GLRLM is a matrix of size $N_g \times L_r$, where Ng represents the pre-defined number of quantized gray levels set in $V_Q(x, y, z)$, and L_r the length of the longest run (of any gray level). The resulting GLRLM of that image is filled in by counting all the possible runs of connected pixels with identical gray levels for a given direction. [26] Similar to those derived from GLCMs, these statistics can be calculated in all four directions thus to obtain rotationally invariant results, they should be averaged.

4.4.3 Gray-Level Size Zone Matrix

The Grey Level Size Zone Matrix idea is based on the previous Grey Level Run Length Matrix. The concept of the GLSZM was proposed by Thibault et all in 2009. The matrix quantifies the number of continuous pixels with the same grey level. [41] A voxel is considered connected if the distance is 1 according to the infinity norm (26-connected region in a 3D, 8-connected region in 2D). Contrary to GLCM and GLRLM, the GLSZM is rotation independent, with only one matrix calculated for all directions in the ROI.

Let P define the GLSZM of a quantized ROI imaging volume $V_Q(x, y, z)$ with isotropic voxel size. In a gray level size zone matrix $P(i,j)$ the $(i,j)^{th}$ element equals the number of zones with gray level i and size j appear in image. A zone is a 2D region of connected voxels with an identical gray level. The GLSZM is a matrix of size $N_g \times L_z$, where N_g represents the pre-defined number of quantized gray levels set in $V_Q(x, y, z)$, and L_z the size of the largest zone (of any gray level). [37]

4.4.4 Example

I would like to represent an example for the three matrix calculations. Let's take a look at a specific example for all of them one by one. We can see on Fig. 6. a four different grey-scale level matrix. GLCM relies on pixel pairs (on *Fig.6.* interpixel distance are zero). On the *a*, matrix of *Fig. 6*. we can find the number 4 stands next to number 1 as an exact neighbor in a horizontal direction three times (highlighted in yellow) so this gives in the GLC matrix in 4th row and 1st column the value 3.

GLRLM relies on pixel runs. On *Fig. 6. (b)* the number 3 grey scale value runs 3 times long in a horizontal direction only in one case (highlighted in yellow) Thus the GLRL matrix in the 3rd row and 3rd column gives a 1 value.

GLSZM relies on areas of neighboring pixels with same gray-level. On *Fig. 6. (c)* the number 2 takes place 4 times right next to each other (reminder that GLSZM is a rotation independent calculation method) in all 4 directions (highlighted in yellow). Therefore, the Grey Level Size Zone Matrix's 2nd row and 4th column give us the value of 1.

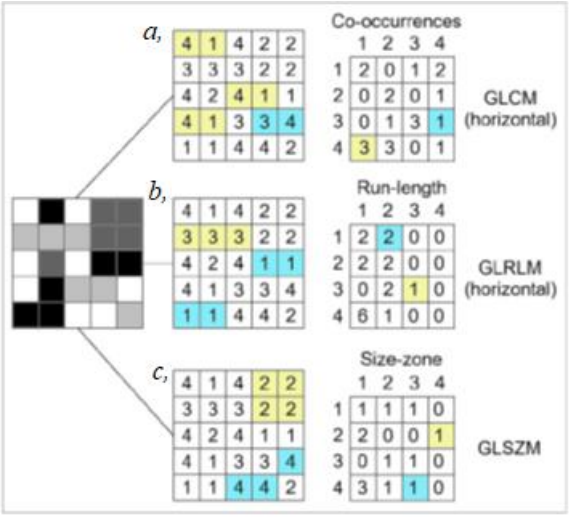


Figure 6.: Calculation of radiomic texture features.[38]

4.5 Texture features

In total, 36 texture features were extracted from the four separate groups divided by the left and right brain as well. *Table 3*. presents the list of texture features used in this thesis.

Global features are extracted from the intensity histogram of the ROI, whereas GLCM, GLRLM, and GLSZM textures are matrix-based features. In this work, histograms with 100 bins were used for the computation of Global features.

Texture Type	Texture name
Global	Variance Skewness Kurtosis
GLCM ***(Grey level cooccurrence matrix)***	Energy Contrast Correlation Homogeneity Variance Sum Average Entropy Autocorrelation Dissimilarity

GLRLM ***(Grey level run length matrix)***	Short-Run Emphasis (SRE) Long-Run Emphasis LRE Grey-Level Non-uniformity (GLN) Run Length Non-uniformity (RLN) Run Percentage (RP) Low Grey-Level Run Emphasis (LGRE) High Grey-Level Run Emphasis (HGRE) Short Run Low Grey-Level Emphasis (SRLGE) Short Run High Grey-Level Emphasis (SRHGE) Long Run Low Grey-Level Emphasis (LRLGE) Long Run High Grey-Level Emphasis (LRHGE) Grey-Level Variance (GLV) Run-Length Variance (RLV)
GLSZM ***(Grey level size zone matrix)***	Small Zone Emphasis (SZE) Large Zone Emphasis (LZE) Grey-Level Non-uniformity (GLN) Zone Size Non-uniformity (ZSN) Zone Percentage (ZP) Low Grey-Level Zone Emphasis (LGZE) High Grey-Level Zone Emphasis (HGZE) Small Zone Low Grey-Level Emphasis (SZLGE) Small Zone High Grey-Level Emphasis (SZHGE) Large Zone Low Grey-Level Emphasis (LZLGE) Large Zone High Grey-Level Emphasis (LZHGE) Grey-Level Variance (GLV) Zone Size Variance (ZSV)

Table 3: The list of texture features

4.6 Feature selection

When we reached this point in the study, we already have lots of data with lots of features of the images. Certainly, the following step in radiomics is to choose a variable or feature selection method to select the best subset of predictors. This way we can explain the data in the simplest way without any noise in the estimation of other quantities of interest caused by unnecessary predictors.

4.7 Correlation

The Pearson correlation coefficient (PCC) measures the linear relationship between two datasets. Strictly speaking, Pearson's correlation requires that each dataset be normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

So, in simpler words, significance of PCC is basically to show you how strongly correlated the two variables are. It is important to note that the PCC value ranges from -1 to 1. A value between 0 to 1 denotes a positive correlation. Value of 0 = highest variation (no correlation whatsoever). A value between -1 to 0 denotes a negative correlation.

From the asymmetric structure matrix shown in *Fig. 7*, we can read the correlations between the parameters. The proportion of the parameters are indicated in colors. The control group's texture features are in the columns and the rows contain the tumor side parameters. Based on the color scale on the right, we can tell what is the proportion among the tumor and control group.

Those close to 0 are indicated in green, the negative correlation is shown in blue and the positive correlation in yellow. The values of each parameter correlate well with each other, if it was found to be nearly the same color on the figure.

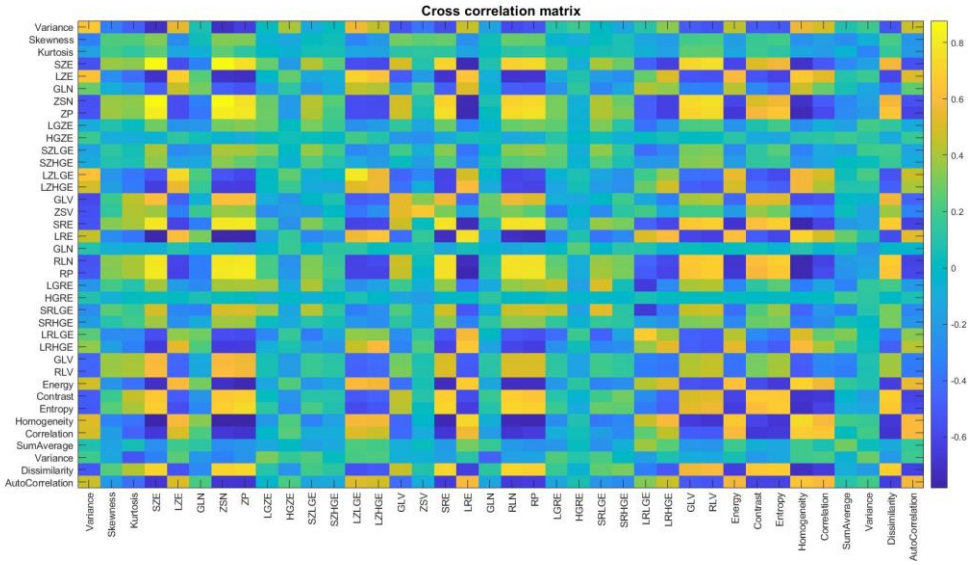


Figure 7: Cross correlation matrix. Numerical values correspond to Pearson correlation coefficient

The *Figure 7* shows the cross-correlation matrix, indicating that there are multiple and complex cross correlation among different covariates. The x-axis is the Control side, the non-tumorous side, and the y-axis is the tumor side. As we can see the least correlating parameters are the Size Zone High Grey-Level Emphasis (SZHGE), the High Grey-Level Run Emphasis (HGRE), the Grey Level Non-uniformity (GLN) and the Variance and Skewness, Zone Size Variance (ZSV).

The most positively correlating features are intensive yellow colored. Such as Run Length Non-uniformity (RLN), Run Percentage (RP), Zone Size Non-uniformity (ZSN) and Zone Percentage. The most negatively correlating features are deep blue colored. Such as Long Run Emphasis (LRE) and Homogeneity.

In a more detailed way, I would like to represent the correlation scatter graphs between the same features on the tumor side(x-axis) and the control side (y-axis).

4.7.1 Global texture features correlation

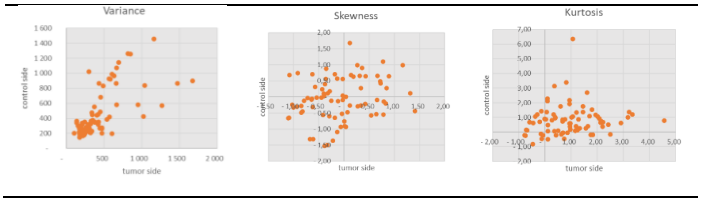


Figure 8 (a): Global features scatter plot

In the case of Global features, the scatter plot is shown in *Fig. 8 (a)*. The three features Variance, Skewness and Kurtosis were plotted to separately. It can be claimed that the most correlating one among the three features is the Variance. In the *Figure 9 (a)*, the exact Pearson correlation coefficients are represented in a heatmap matrix.

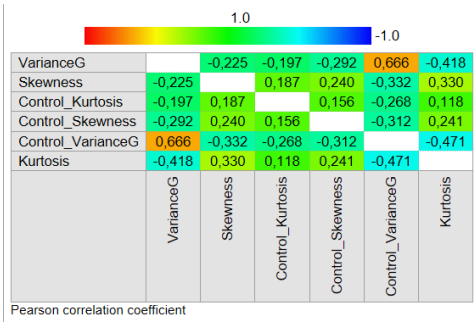
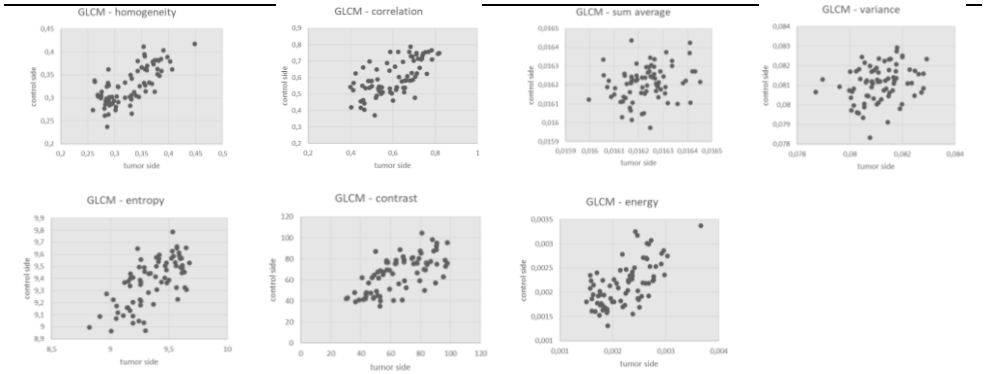


Figure 9 (a) Cross correlation matrix with Pearson correlation coefficients on Global texture analysis tumor and control features

4.7.2 Grey-Level Co-occurrence Matrix features correlation

As *Fig. 8 (b)* shows most of the control and tumor pair features from this type of analysis are mildly correlating with each other. The analysis of the cross correlation between the control and tumor features and also among them are showed several cross related covariates *(Fig. 9 (b))*. Uncorrelated features are the SumAverage and then Variance (also Control group SumAverage and Variance). GLCM features has, among all the features and all the other texture analysis methods, the highest and lowest Pearson correlation coefficient values *(Fig. 9 (b))*.



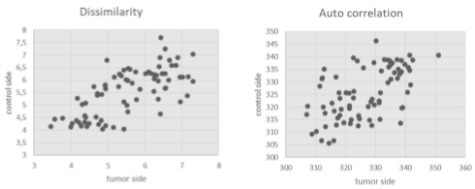


Figure. 8 (b): GLCM features scatter plots

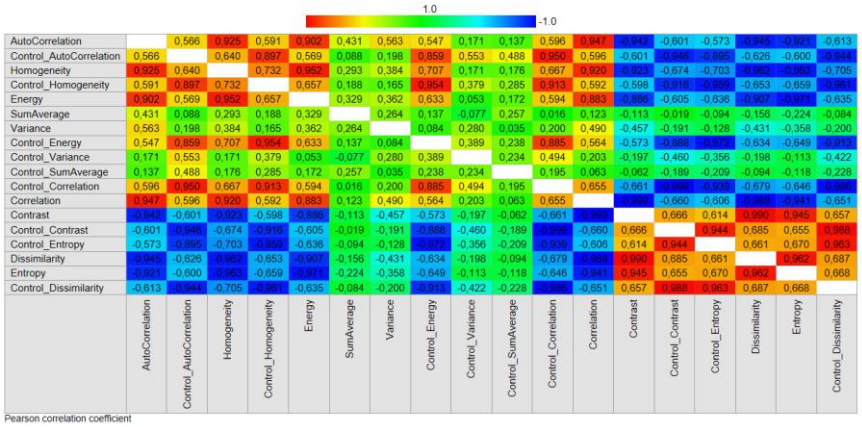
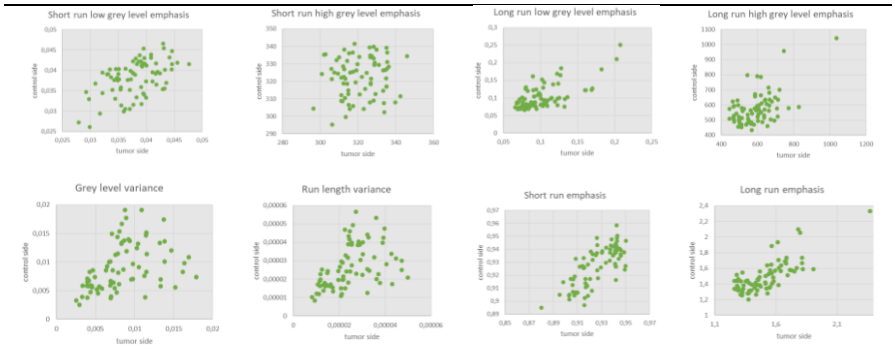


Figure 9 (b): Cross correlation matrix with Pearson correlation coefficients on GLCM control and tumor features

4.7.3 Grey-Level Run-Length Matrix features correlation

On *Fig. 8 (c)* it is conspicuous that the most correlating feature with its control group is the Grey-level Non-uniformity feature, however we find that on the heatmap (*Fig. 9 (c)*) GLN is the least correlating feature with any other features from this texture analysis. The most positively correlating feature with even a number 1 Pearson coefficient is the Run Percentage and the most negatively correlating one is the Long-Run Emphasis.



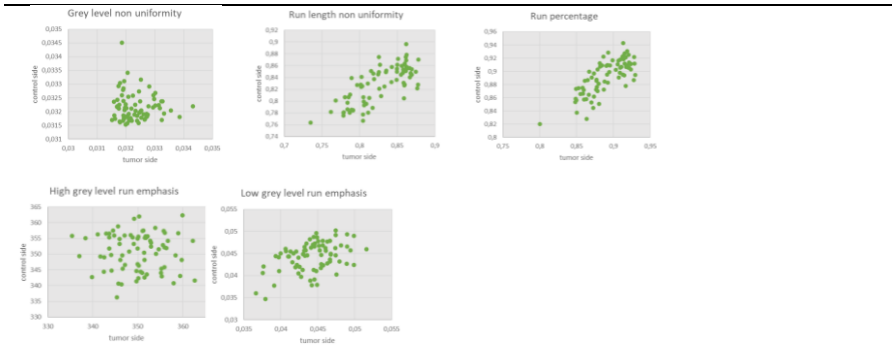


Figure. 8 (c): GLRLM features scatter plots

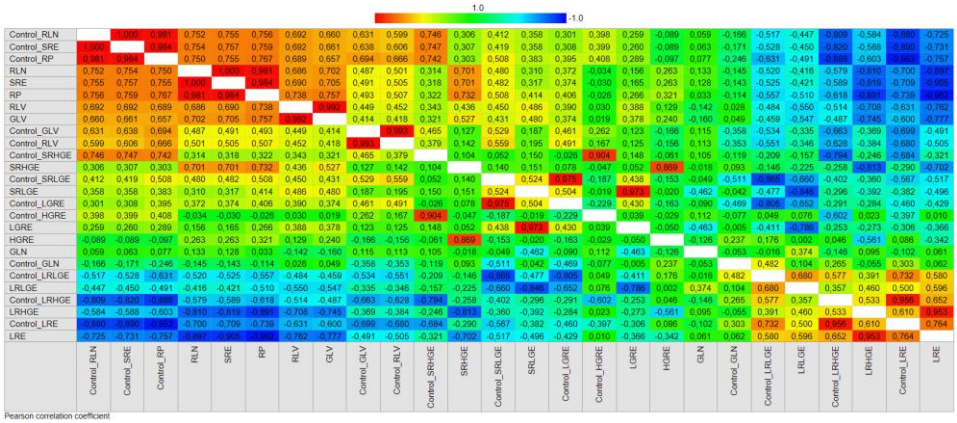
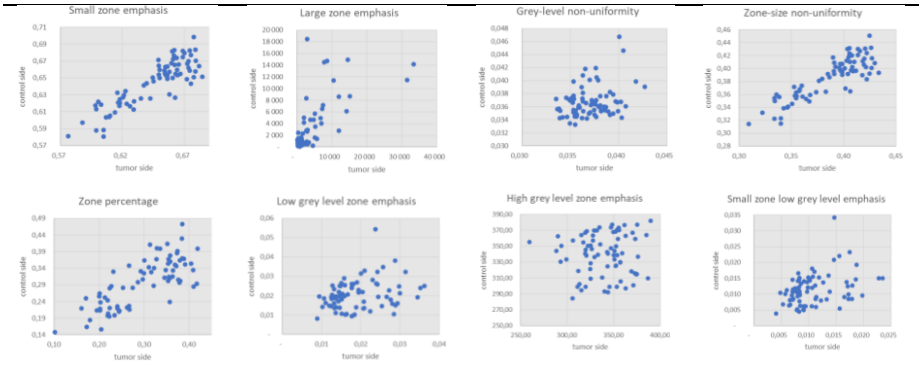


Figure 9 (c): Cross correlation matrix with Pearson correlation coefficients on GLRLM control and tumor features

4.7.4 Gray-Level Size Zone Matrix features correlation

Figure 8 (d) shows all the 13 data set correlation between control and tumor group. The most correlating one with each other are the Zone Size Variance, the Large Zone Low and High Grey-Level Emphasis. When we take a look at the heatmap as well on *Figure 9 (d)* we can find that Zone Size Variance doesn't really correlate with any other features in the matrix.



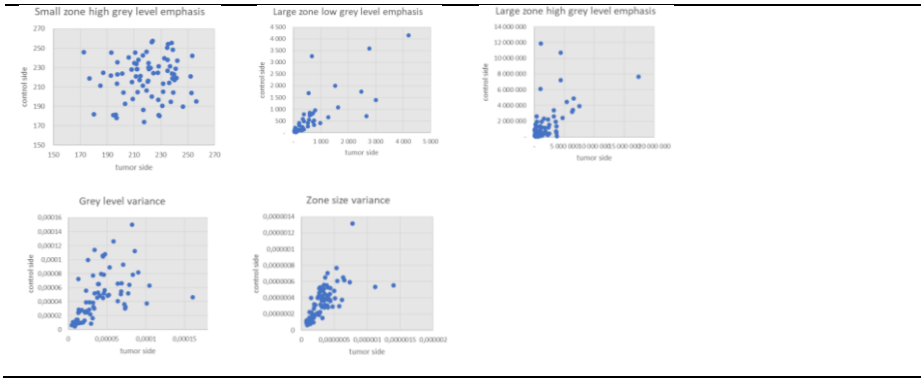


Figure 8 (d): GLSZM features scatter plots

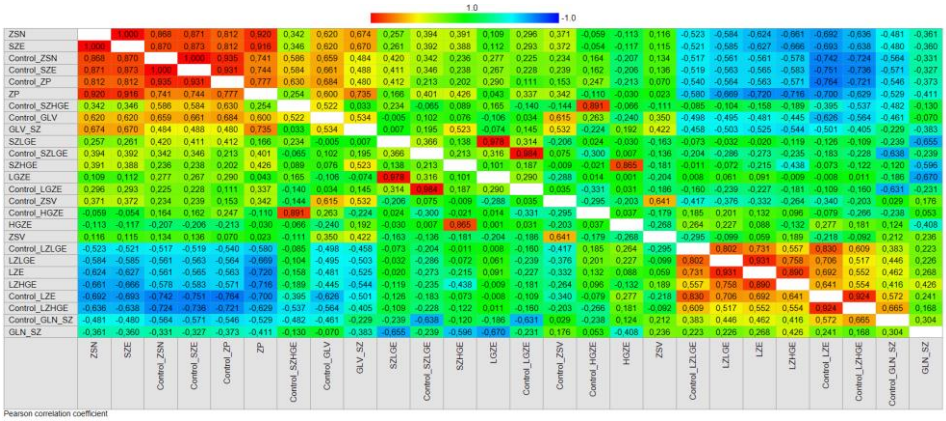


Figure 9. (d): Correlation matrix with Pearson correlation coefficients

To sum up, covariate with Pearson's correlation test ($p>0.05$; the P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional 5% ($P<0.05$) the correlation coefficient is called statistically significant.) non correlating features are great for further multivariate analysis (for example: in two different logistic models selecting two different groups of uncorrelated features).

4.8 Regression

We can find in published articles, literature and studies that the most commonly used analyses methods and classifications for radiomics dataset are: logistic regression (L1 or L2), elastic net. They are preferred because they are considered as supervised learning methods. It means it is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.

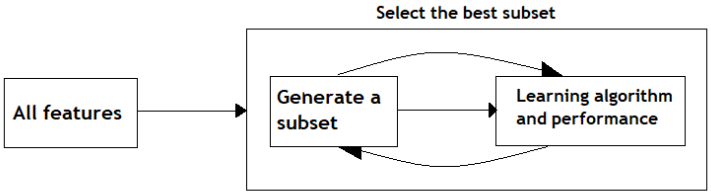


Figure 10.: The method of the best subset selection

First, let's clarify that linear models are one of the simplest ways to predict output using a linear function of input features. However, overfitting (when we have large dataset) and underfitting (when we have small dataset) can cause easily a problem.[19]
Linear model with n features for output prediction:

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n + b \qquad \textit{Eq. 1.}$$

In the 3.5.1 equation shows that β_0 will be slope and *b* will represent intercept. Linear regression looks for optimizing β_0 and *b* such that it minimizes the cost function.
Cost function for simple linear model:

$$\sum_{i=1}^{M} (y_i - y_i)^2 = \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} \beta_j \cdot x_{ij} \right)^2 \qquad \textit{Eq. 2.}$$

where the dataset has *M* instances and *p* features.

Two special linear regression model which are able to reduce complexity and prevent over-fitting as the result of a linear regression, are the Lasso and Ridge regression. They are very similar methods yet not so similar, let me explain why and which one is the better choice for my thesis. [20]

4.8.1 Lasso Regression

One of the most well-known powerful methods that helps regularization and feature selection of the given data is Least Absolute Shrinkage and Selection Operator (LASSO). The Lasso method puts a limitation/restrictions on the sum of the values of the model parameters. The sum has to be

less than the specific fixed value. This Shrinks some of the coefficients to zero, Indicating that a certain predictor or certain features will be multiplied by zero to estimate the target.[42] During this process the variables that have non-zero co-efficient after shrinking are selected to be the part of the model. It also adds a penalty term to the cost function of a model, with a lambda value that must be tuned.[20]

The cost function of Lasso is:

$$\sum_{i=1}^{M}(\mathrm{y_i}-\mathrm{y_i})^2 = \sum_{i=1}^{M}\left(\mathrm{y_i}-\sum_{j=0}^{p}\beta_j\cdot x_{ij}\right)^2 + \lambda\sum_{j=0}^{p}|\beta_j| \qquad \textit{Eq. 3.}$$

where β is the coefficient and λ is the shrinkage parameter.

When λ lambda is 0, the equation is reduced and this leads to no elimination of the parameters. Increase in λ causes the increase in bias, decrease in λ causes the increase in variance.[42]

Although Lasso regression seems to be a great choice as a classifier, it still has some limitations.[18] It sometimes struggles with some types of data. If the number of predictors (p) is greater than the number of observations (n), Lasso will pick at most n predictors as non-zero, even if all predictors are relevant. If there are two or more highly collinear variables then Lasso regression select one of them randomly which is not good for the interpretation of data. [36]

4.8.2 Ridge Regression

In Ridge or L2 regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients

$$\sum_{i=1}^{M}(\mathrm{y_i}-\mathrm{y_i})^2 = \sum_{i=1}^{M}\left(\mathrm{y_i}-\sum_{j=0}^{p}\beta_j\cdot x_{ij}\right)^2 + \lambda\sum_{j=0}^{p}\beta_j^2 \qquad \textit{Eq. 4.}$$

where β is the coefficient and λ is the shrinkage parameter.

Ridge regression penalizes the β coefficients for being too large, but it doesn't shrinks the coefficient to zero only close to it. It helps to reduce the model complexity and multi-collinearity. Multicollinearity is a situation that occurs when independent variables are highly correlated. This is the case when we apply Ridge regression to our data. Compared to Lasso this regularization term will decrease the values of coefficients but unable to force them to zero. If the number of

predictors is greater than the number observations it is capable of selecting more than n relevant predictors. [35] Ridge regression isn't preferable when the data contains huge number of features out of which only few are actually important, as it might make the model simpler but the model built will have poor accuracy. Hence, this model is not good for feature reduction. [42]

4.8.3 Elastic net

Lasso regression is a great algorithm for variable selection with high dimensional data, however sometimes it over regularize the data. The solution for the problem is a third type of regression method is the Elastic net. It includes both L1 and L2 norm regularization terms and combine the penalties of both Ridge and Lasso Regression. The elastic net method improves on Lasso's limitations, i.e., where Lasso takes a few samples for high dimensional data, the elastic net procedure provides the inclusion of "n" number of variables until saturation. [43]

Take a look at equation Eq. 6. [44] We multiply the L2 norm by 1 - α, multiply the L1 norm by α, and add these values both up. We multiply this value by lambda and add it to the sum of squares. Alpha here can take any value between zero and one:

- when alpha is zero, the L1 norm becomes zero, and we get ridge regression
- when alpha is one, the L2 norm becomes zero, and we get LASSO
- when alpha is between zero and one, we get a mixture of ridge regression and LASSO.[44]

Elastic net's cost function with the L1 and L2 loss:

$$\sum_{i=1}^{M}(\mathrm{y_i}-\mathrm{y_i})^2=\sum_{i=1}^{M}\left(\mathrm{y_i}-\sum_{j=0}^{p}\beta_j\cdot x_{ij}\right)^2+\lambda_1\sum_{j=0}^{p}|\beta_j|+\lambda_2\sum_{j=0}^{p}\beta_j^2 \qquad \textit{Eq. 5.}$$

or as the following format:

$$\sum_{i=1}^{M}(\mathrm{y_i}-\mathrm{y_i})^2=\sum_{i=1}^{M}\left(\mathrm{y_i}-\sum_{j=0}^{p}\beta_j\cdot x_{ij}\right)^2+\lambda\left(\alpha\sum_{j=0}^{p}|\beta_j|+\frac{1-\alpha}{2}\sum_{j=0}^{p}\beta_j^2\right) \qquad \textit{Eq. 6.}$$

,where β is the coefficient and λ_1 and λ_2 is the Lasso and Ridge shrinkage parameter and α is the mixing parameter between ridge and Lasso.

Groupings and variables selection are the key roles of the elastic net technique. Elastic Net Regression encourages group effect in case of highly correlated variables. The elastic net technique is most appropriate where the dimensional data is greater than the number of samples used.

4.8.4 Lasso – Ridge – Elastic net Comparison

Ridge, Lasso, and elastic net regularization are all methods for estimating the coefficients of a linear model while penalizing large coefficients. The following illustration will clarify why Lasso Regularization leads to feature selection, why Ridge only reduces the coefficients close to zero but never exactly zero and why Elastic Net Regularization is the best of both words.

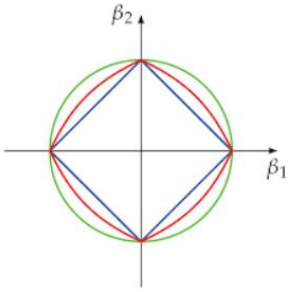


Figure 11.: Compering penalty of the LASSO(blue) and Ridge(green) and Elastic net(red) (β penalty)[45]

On the *Figure 11* the plotted constraint regions of three cost function of the Lasso and the Ridge regression and Elastic net. The illustration compares the shapes of the ridge, LASSO and elastic net penalties. As the elastic net penalty is somewhere between the ridge and LASSO penalties, it looks like a square with rounded sides.

5. Results and evaluation

5.1 Clinical characteristics

The restricted dataset contains 77 patient's data. All the MR images were obtained from the Axial plane. Thus, a total of 77 patients were enrolled in this thesis, including 50 males and 27 females (minimum age years, maximum age years, median age 60 years.) 47 of them were diagnosed with Glioblastoma Multiforme on the right hemisphere if the brain and 30 of them were diagnosed with Glioblastoma Multiforme on the left hemisphere of the brain. All the patients were treated with External Beam Therapy.

Sex	Female: 27 (35.1%) Male: 50 (64.9%)
Age	Mean: 58 Median: 60 Minimum: 17 Maximum: 84
Tumor location	Right: 47 (61.0%) Left: 30 (39%)
Radiation Therapy Type	External Beam
Radiation Therapy Site	Primary Tumor Field
MR plane	Axial

Table 4.: Clinical characteristics of the cohort of the patients (Restricted dataset)

5.2 LASSO classification results

In this subchapter I am going to represent the results of my analysis with Lasso Classification for each of the texture analysis group separately. Each curve on the following graphs corresponds to a variable. The feature variables are separated by sides. The response variable is the Tumor (1) and Control (0) vector depending on which side of the brain has the Glioblastoma Multiforme. Speaking in general, the plot shows the nonzero coefficients in the regression for various values of the Lambda regularization parameter. Larger values of Lambda appear on the left side of the graph, which means more regularization, resulting in fewer nonzero regression coefficients.[32] The dashed vertical lines represent the Lambda value with minimal mean squared error. The upper part of the plot shows the degrees of freedom (df), meaning the number of nonzero coefficients in the regression, as a function of Lambda.

For small values of Lambda (toward the right in the plot), the coefficient values are close to the least-squares estimate.

The plot shows the path of its coefficient against the L1-norm of the whole coefficient vector at as λ varies. The axis below indicates the number of nonzero coefficients at the current λ, which is the effective degrees of freedom (df) for the Lasso.

5.2.1 Global features

From the feature extraction, in the Global function case, we get 6 output values from the MATLAB program: Variance, Skewness and Kurtosis for both left and right sides. As the classification results show us *(Fig. 12 (a))*, from these 6 values 3 are nonzero values which include 2 features the Skewness of the right side (green) and left side (orange) of the brain and the Kurtosis only for the right side (blue).

To concretize the above mentioned general description of the graph, *Fig. 12 (a)* shows on the upper part 6 degress of freedom however only 3 appreciable from the zero line. Also we can find that a larger value of Lambda resulting a fewer nonzero coeffiecient which in this case is the right brain side Skewness.

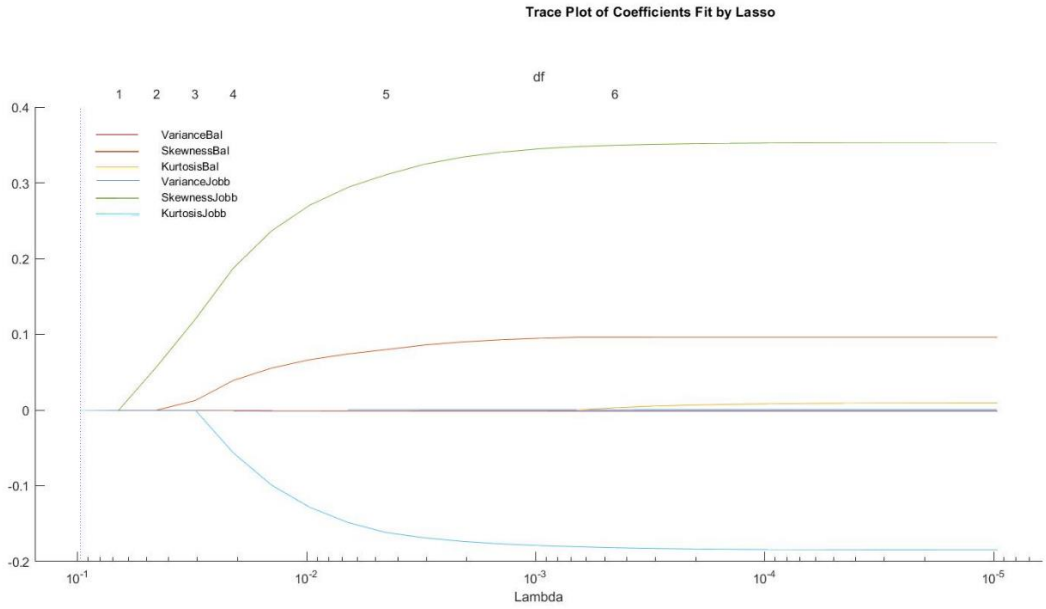


Figure 12 (a): LASSO classification result from Global feature extraction

The Skewness defined as (adapted from [3]):

$$s = \sigma^{-3} \sum_{i=1}^{N_g} (i - \mu)^3 p(i)$$

$$p(i) = \frac{P(i)}{\sum_{i=1}^{N_g} P(i)}$$

Eq. 7.

,where N_g represents the number of gray-level bins set for *P*, and *P(i)* represent the number of voxels with gray-level *i*.

5.2.2 GLCM features

In the case of the Gray Level Co-Occurrence Matrix (GLCM) texture analysis, we gain 18 features from the calculation and there are two features with nonzero coefficient value, the 'Sum Average' for both left and right side of the brain and the Variance for the right side as well *(Fig. 12 (b)).*
On *Figure 12 (b)* we can see that Lasso regression retains two nonzero coefficients as Lambda increases (toward the left of the plot), and these two coefficients (SumAvarage, Variance) reach 0 at about the same Lambda value. The Lasso plot shows two of the coefficients becoming 0 at the same value of Lambda, while another coefficient remains nonzero for higher values of Lambda. In general, Lasso tends to drop smaller groups, or even individual predictors, this is a general pattern for L1 regression.

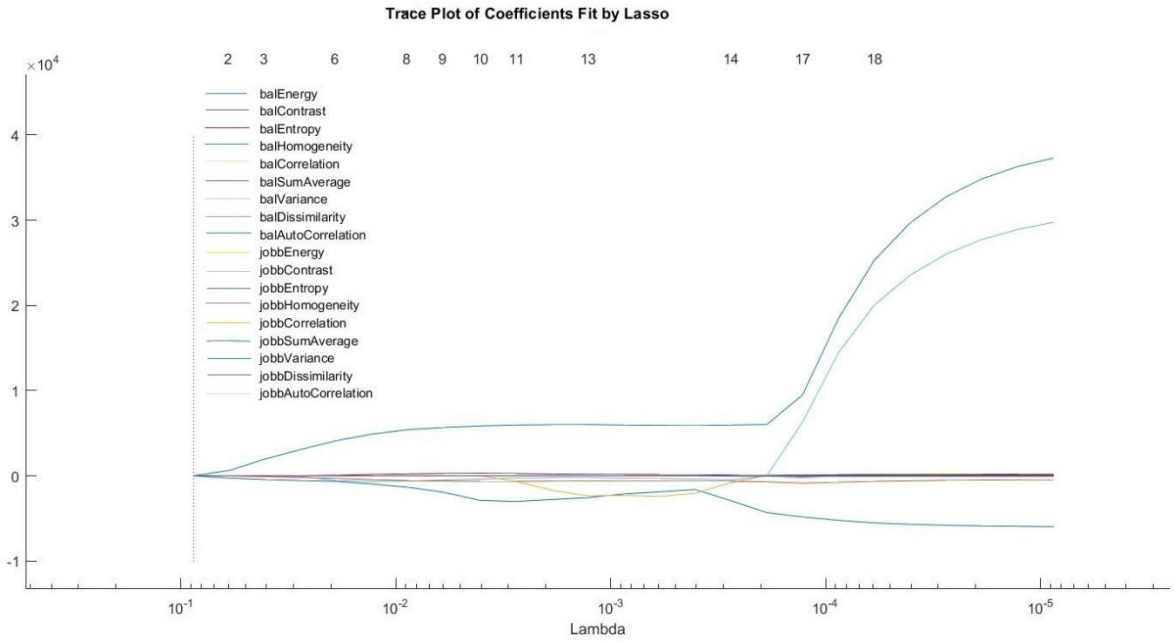


Figure 12 (b): LASSO classification result from GLCM feature extraction

The Sum Average texture feature is defined as (adapted from [3]):

$$\text{sum average } = \frac{1}{N_g \times N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [ip(i,j) + jp(i,j)] \qquad \textit{Eq. 8.}$$

,where N_g represents pre-defined the number of quantized gray-level bins set in *V*, and *P(i,j)* represents the number of 3D zones of gray-levels *i* and of size *j* in the voxel.

5.2.3 GLRLM features

In the case of the Gray Level Run-Length Matrix, the RLV, the Run- Length Variance texture feature for both sides was the only non-zero coefficient *(Fig. 12 (c))* The graphs shows that the number of nonzero coefficients in the regression increases with the Lambda values reduction.

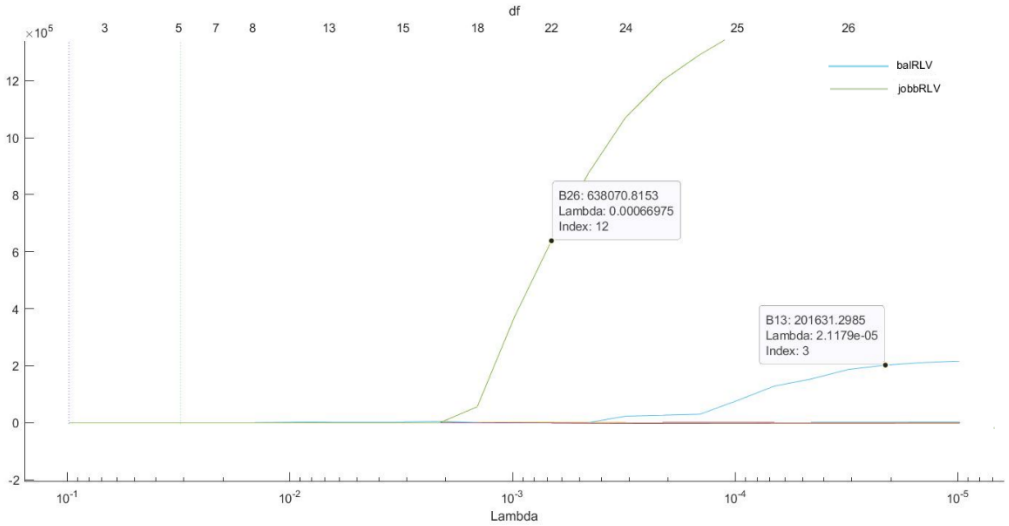


Figure 12 (c): LASSO classification result from GLRLM feature extraction

The Run-Length Variance (RLV) defined as (adapted from [3]):

$$\mathrm{RLV} = \frac{1}{N_g \times L_r} \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} (jp(i,j) - \mu_j)^2$$

$$\mu_j = \sum_{j=1}^{L_z} j \sum_{i=1}^{N_g} p(i,j) \qquad \textit{Eq. 9.}$$

,where N_g represents pre-defined the number of quantized gray-level bins set in *V*, and *P(i; j)* represents the number of 3D zones of gray-levels *i* and of size *j* in the voxel, L_r represents the length of the longest run of any grey level in *V*.

5.2.4 GLSZM features

In case of the Gray-Level Size Zone Matrix just like in the previous case we got one non-zero coefficient. It is the ZSV also known as Zone-Size Variance for both sides. *(Fig. 12 (d))*

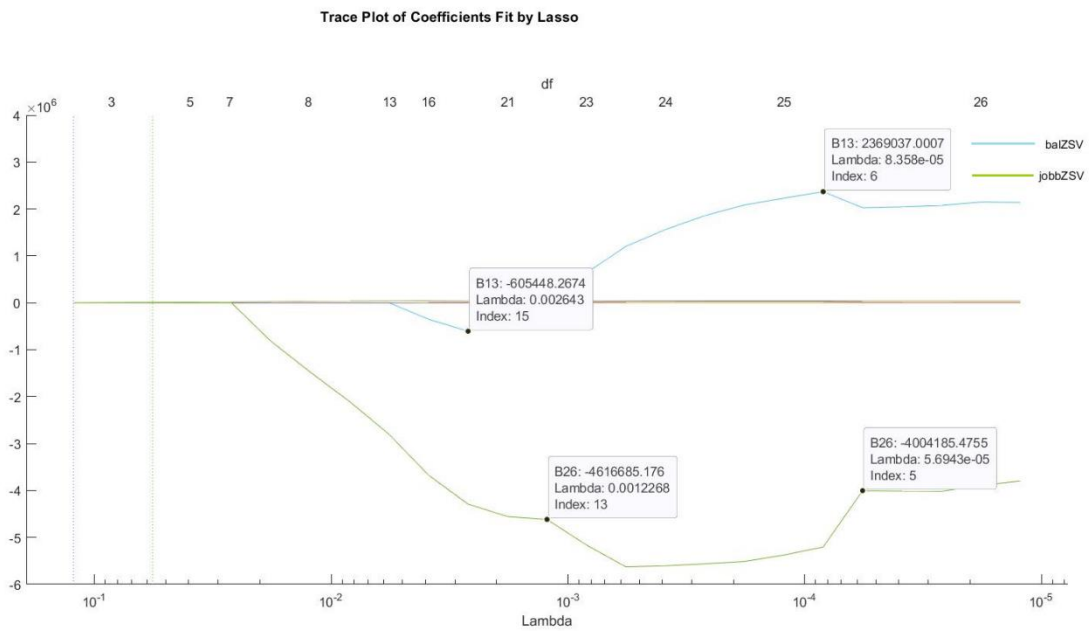


Figure 12 (d): LASSO classification result from GLSZM feature extraction

The Zone-Size Variance (ZSV) defined as (adapted from [3]):

$$\mathrm{ZSV} = \frac{1}{N_g \times L_z} \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} (jp(i,j) - \mu_j)^2$$

$$\mu_j = \sum_{j=1}^{L_z} j \sum_{i=1}^{N_g} p(i,j)$$

Eq. 10.

,where N_g represents pre-defined the number of quantized gray-level bins set in *V*, and *P(i; j)* represents the number of 3D zones of gray-levels *i* and of size *j* in the voxel, L_z represents the size of the largest zone of any gray-level in *V*.

5.2.5 All features

At last, I would like to interpret a plot. In this case the response dependent variable of the Lasso regression was the vector of vital status, and the independent variables were all the feature from all the four texture analysis.

Figure 12 (d) shows 4 main feature with a nonzero coefficient, these four are: Run-length Variance and Zone Size Variance for both the control and tumor group.

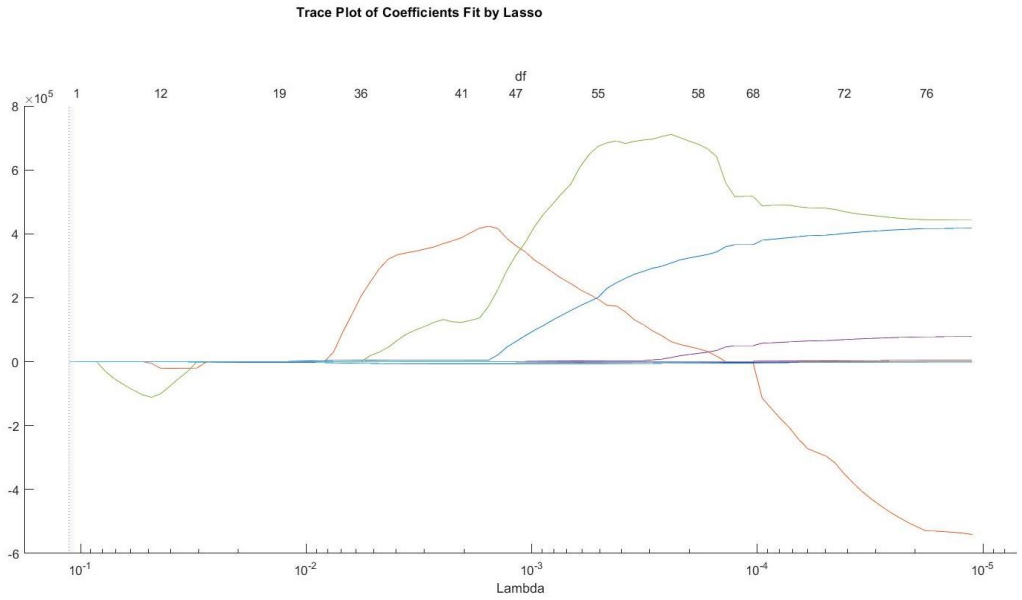


Fig. 12 (d): LASSO classification result for all the features

All in all, we tried to determine the strongest predictors, with the interpretation of the plots as evidence that variables that enter the model early are the most predictive and variables that enter the model later are less important.

We have four major texture features which are both non-zero for both side of the brain, therefore comparable their outcomes: the Skewness, the Sum Average, the Run-length Variance and the Zone -Size Variance. These features can be used for further analysis and evaluation for survival model such as Random Forest or Cox model, or to calculate the receiver operating characteristic curve (ROC) and AUC curves.

6. Conclusion

This thesis shows that glioblastoma multiforme brain one-side-tumor and the normal -side- brain as a control-group differences can be observed using non-invasive diagnostic imaging through textural analysis.

In the beginning of my thesis I separated and segmented the MRI scans manually. Nowadays there are a lot of semiautomated and automated methods for supervised segmentation, which can lead to improvement in the evaluation. Reducing the need for manual user input in the workflow is also an importance in the development of computer-aided diagnosis through machine learning or AI. Also, I was only working with the texture analysis for feature extraction, however as I mentioned above, there are several other ways for radiomics signature.

This thesis is focusing on to introduce texture analysis and 3 kinds of regression for feature selection. To purpose was to show how the Lasso classification model works for feature selection as well. From the applied technique, the results give us two features, Zone Size Variance and Run-Length Variance, which are probably to best to use in further evaluation, survival models etc. A well-chosen classifier which leads to the best subset for the large dataset can predict very effectively the overall survival.

Radiomics analysis requires a huge dataset, however, the number of patients included in the thesis cohorts did not provide sufficient statistical power to enable evaluation of the relationship between textural parameters and overall survival.

7. Acknowledgements

Here, I would like to express my sincere appreciation to all of those who have helped me during my thesis.

I would like to firstly thank my supervisor, Dr. Krisztian Szigeti for providing me the chance to join in his research group. He was a guide on the research road and always gave me professional advice.

I would like to express how thankful I am to Dávid Szőllösi for motivating me and supporting me. His research capabilities and enthusiasm always inspire me to better myself.

I would like to thank also to Balázs Fekete and Dániel Bercsényi for their patience, advices, but most importantly for finding the time from their busy schedule to support and take care of me.

Last but not least, I would like to thank to my twin, Miss Tamara Juhasz, who pushed me through difficult times. Even when I did, she *never* lost belief in me. Also, to my parents for unconditionally supporting and accommodating all of my shortcomings.

Thank you.

8. References:

[1]: Yiming Meng et al.: Application of Radiomics for Personalized Treatment of Cancer Patients, 10.2147/CMAR.S232473

[2]: Yi-Hua Zhang: Radiomics in cancer prognosis: Applications and Limitions of quantitative texture analysis, Karolinska Institute

[3]: Philippe Lambin: Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev.Clin Oncol. 2017;

[4]: Larue RT et al.: Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures, The British Journal of Radiology, 12 Dec 2016

[5] Ke Nie et al.: Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI, 10.1158/1078-0432.CCR-15-2997 Published November 2016

[6]: David C Preston, MD: Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics /https://www.imaios.com/en/e-Courses/e-MRI/MRI-Sequences/inversion-recovery-stir-flair/

[7]: Rohit Bakshi, MD et al.: Fluid-Attenuated Inversion Recovery Magnetic Resonance Imaging Detects Cortical and Juxtacortical Multiple Sclerosis Lesions Arch Neurol. 2001;58(5):742-748. doi:10.1001/archneur.58.5.742

[8]: Pooley, R.A., Fundamental physics of MR imaging. Radiographics, 2005. 25(4): p. 1087-1099.

[9]: Damjanovich, S., J. Fidy, and J. Szöllősi, Orvosi biofizika. Medicina, Budapest, 2007.

[11]: Prekeges, J., Nuclear medicine instrumentation. 2012: Jones & Bartlett Publishers.

[12]: https://wiki.cancerimagingarchive.net/display/Public/TCGA-GBM
https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-treatments/Glioblastoma-Multiforme

[13]: Lacroix, M., et al., A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. Journal of neurosurgery, 2001. 95(2): p. 190-198.

[14]: Stupp, R., et al., Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine, 2005. 352(10): p. 987-996.

[15]: Alex Lobera, M., Imaging in Glioblastoma Multiforme. 2017.

[16]: Wikipedia: Fluid-attenuated inversion recovery (https://en.wikipedia.org/wiki/Fluid-attenuated_inversion_recovery)

[17]: Wikipedia: MITK (https://de.wikipedia.org/wiki/MITK)

[18]: Lasso vs Ridge vs Elastic Net | ML :

(https://www.geeksforgeeks.org/Lasso-vs-ridge-vs-elastic-net-ml/#:~:text=Lasso%20regression%20stands%20for%20Least,term%20to%20the%20cost%20function.&text=The%20difference%20between%20ridge%20and,of%20coefficient%20to%20absolute%20zero.)

[19]: J. Friedman et.al.: The element of statistical learning; Springer, pages- 79-91, 2008.

[20]: Andreas Muller: Machine Learning with Python

[21]: Zhen Hou et al.: Radiomic analysis using contrast-enhanced CT: predict treatment response to pulsed low dose rate radiotherapy in gastric carcinoma with abdominal cavity metastasis doi: 10.21037/qims.2018.05.01

[22]: Ralph Theodoor Hubertina Leijenaar. Radiomics: Images are more than meets the eye

[23]: Vallières, M. et al. (2015). A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Physics in Medicine and Biology, 60(14), 5471-5496. doi:10.1088/0031-9155/60/14/5471

[24]: Zhou, H., Vallières, M., Bai, H.X. et al. (2017). MRI features predict survival and molecular markers in diffuse lower-grade gliomas. Neuro-Oncology, 19(6), 862-870. doi:10.1093/neuonc/now256

[25]: Vallière, M. et al. (2017). Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific Reports, 7:10117. doi:10.1038/s41598-017-10371-5

[26]: Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973). Textural features for image classication. IEEE Transactions on Systems, Man, and Cybernetics, smc 3(6), 610-621.

[27]: Hastie et al. Linear Regression in High Dimension and/or for Correlated Inputs EAS Publications Series 66:149-165 · January 2015

[28]: Thibault, G. (2009). Indices de formes et de textures: de la 2D vers la 3D. Applicationau classement de noyaux de cellules. PhD Thesis, Universite AIX-Marseille: p.172.

[29]: Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H. et al. (2014). Decoding tumour phe-no type by noninvasive imaging using a quantitative radiomics approach. Nat. Commun.5:4006, doi: 10.1038/ncomms5006.

[30]: Wei, X. (2007). Gray Level Run Length Matrix Toolbox v1.0, computer software. Beijing Aeronautical Technology Research Center.

[31]: Galloway, M.M. (1975). Texture analysis using gray level run lengths. Computer Graphicsand Image Processing, 4(2), 172-179.

[32]: Lasso and elastic net with Cross Validation, MathWorks; www.mathworks.com

[33]: Elastic net regression; /www.i2tutorials.com/machine-learning-tutorial/

[34]: Lasso, MathWorks; www.mathworks.com

[35]: Ridge and Lasso regression a complete guide with phyton scikit learn /towardsdatascience.com/

[36]: Bias variance and regularization in linear regression Lasso ridge and elastic net /towardsdatascience.com/

[37]: Guillaume Thibault; Bernard Fertil; Claire Navarro; Sandrine Pereira; Pierre Cau; Nicolas Levy; Jean Sequeira; Jean-Luc Mari (2009). “Texture Indexes and Gray Level Size Zone Matrix. Application to Cell Nuclei Classification”. Pattern Recognition and Information Processing (PRIP): 140-145.]

[38]: Marius E. et. all (2020): Introduction to radiomics; DOI: 10.2967/jnumed.118.222893

[39]: Seung-Hak Lee et all (2020): Radiomics in Breast Imaging from Techniques to Clinical Applications: A Review

[40]: Galloway et. all (1975): Machine Learning in Cardiovascular Medicine

[41]: Elastic net, /corporatefinanceinstitute.com/

[42]: From linear regression to ridge regression the lasso and the elastic net. /towardsdatascience.com/

[43]: Tutorial ridge lasso elastic-net /www.datacamp.com/

[44]: Hefin I. Rhys (2020): Machine Learning with R, the tidy verse, and mlr; ISBN 9781617296574

[45]: Frank Emmert -Streib et. all(2019): High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection; DOI: 10.3390/make1010021