POLITECNICO DI MILANO

School of Industrial and Information Engineering
Master of Science in Biomedical Engineering

Comparison of statistical methods for missing data imputation in MRI-radiomics

Supervisor: Prof. Luca Mainardi

Co-Surpervisor: Dott. Marco Bologna

Thesis of:

Jairo Pinedo Taquía

ID: 893884

ACADEMIC YEAR 2019 – 2020

ACKNOWLEDGEMENTS

Firstly, I would like to say ‘thank you, God’ who made the dream of studying at a world-class university become true.

Likewise, I want to show my eternal gratitude to my family who always shows their confidence in me regardless of the distance; to my mom Justa, dad Nicolas, brothers Aldeir, Adler and Gheraldin and my aunt Yohana who say to me, ‘you have to be perseverant’. Also, I want to mention especially to the mother of my son. Stefany, who is a great woman and has taken care of my son Erhen so much. He will always encourage me to overcome myself more.

A special thanks to my supervisor Luca Mainardi who give me his confidence to develop this work and to my tutor Marco Bologna who help me all the time by providing me with all his experience and knowledge.

Last but not least, I feel grateful to have many friends who were always presented during all this time as Valeryn La Rosa, Ketty C., Teresa, Alejandra, Angely and Sylvana. Moreover, to those ones who are far away from here as Anderson Benites and Mario Inga who helped me so much in difficult moments, also to Tony, Jacky, German, Gino, and Elsa.

Every moment that I was here, at Polimi, will be kept on my memories as an unforgettable part of my life.

Contents

Abstract

Radiomics is a new field of medical image analysis consisting in the extraction of a large quantity of features from non-invasive medical imaging. When performing radiomic features extraction on multiparametric magnetic resonance imaging (MRI), there is the possibility to have missing information since there are not defined protocols at hospitals to acquire data, and the MRI sequences acquired may vary among centers. Statistical methods to impute missing data exist, but it is not clear which method could be best for a radiomic application. The main objective of this thesis was to compare different imputation methods in terms of accuracy and time of imputation.

An initial dataset of 185 patients with affected by nasopharyngeal carcinoma was used for this thesis. Each patient had imaging available, with at least one of the following MRI sequences available: pre- and post-contrast T1-weighted (T1w) images; T2-weighted (T2) images; apparent diffusion coefficient (ADC) maps. Manually segmented regions of interest (ROI), representing the main tumor (T) and lymphnode (N) were used to extract the radiomic features. For each combination of image type, 536 different radiomic features were computed, for a total of 4288 features, but only but only the 2144 were considered for the analysis. Initially the information was represented in a matrix of 185 rows (representing the observations or patients) and 2144 columns (representing the radiomic features). However, in order to have a complete matrix to be used as a gold standard, the patients with missing data were removed, reducing the number to 115. From such matrices, derived matrices with missing columns for post-contrast T1w and ADC were generated using different level of missingness (1, 5, 10, 20, 30, 40 and 50%). For each level of missingness, 10 matrices with random missing data were generated. Having a new matrix of 115 x 2144 elements with missing information, we could apply different methods 6 different imputation methods. Imputation was performed on two sets of matrices, without and with normalization process (z-score was used). After the process of imputation, the performance of each method was assessed by root mean square error (NRMSE) and time of imputation. In reference to the first one, the comparison was between the values given by the imputation methods and the values of the original matrix. Friedman tests with post-hoc comparisons were used to detect significant differences in the errors obtained with the different methods. As for the second one, it was evaluated the dependence of time with the level of missing information.

All the metrics were compared amongst the used methods in two different groups the normalized and non-normalized. The results showed that the simple methods of imputation are the optimal trade-off between accuracy and time of imputation, especially for high levels of missingness. However, the decision to choose a method will be strongly linked to our priority that can be defined in terms of either time or accuracy or both.

List of figures

List of tables

List of abbreviations

3D-CRT	3D Conformal RadioTherapy
ADC	Apparent Diffusion Coefficient
AUC	Area Under the Curve
CE-T1w	Contrast Enhanced T1-weighted
CT	Computed tomography
DWI	Diffusion-Weighted Imaging
FOS	First Order Statistics
fSVT	fast Singular Value Thresholding
GLCM	Grey Level Co-occurrence Matrix
GLRLM	Grey Level Run Length Matrix
IMRT	Intensity Modulated RadioTherapy
KNN	K-Nearest Neighbor
LASSO	Least Absolute Shrinkage and Selection Operator
MRI	Magnetic Resonance Imaging
NPC	NasoPharyngeal Cancer
NRMSE	Normalized Root Mean Square Error
OAR	Organ At Risk
PET	Positron Emission Tomography
RF	Radio Frequency
ROI	Region of Interest
TNM	Tumor Node Metastasis
TE	Time of Echo

TR	Time of Repetition
T1w	T1-weighted
T2w	T2-weighted

Chapter 1: Introduction

In this part, all the background necessary to understand this project of thesis is provided. Background on nasopharyngeal cancer, magnetic resonance imaging (MRI), radiomic features extraction and post processing is given.

1.1 Aim of the project

This project has like main objective to develop a comparison amongst different methods of missing data imputation applied for radiomics features from MRI, assessing its performance in terms of both imputation error and time of imputation. The mathematical algorithms to fill the missing data are based on statistical theory. The imputation was performed on both normalized and non-normalized dataset to understand whether normalization could impact the performance of the different methods.

1.2 Clinical background

1.2.1 General introduction

Rhinopharynx cancer is also called 'nasopharyngeal carcinoma' (NPC). The nasopharynx is considered as a bridge for air passing from nasal cavity to throat [1]. Figure 1.1 where the nasopharynx is located.

NPC is a non-common tumour emerging from epithelial cell affecting the nasopharynx [2]. According to statistical data, of almost 18.1 million of new cancer cases and around 1.8 million deaths recorded in 2018, there were about 177 422 people detected with NPC and 94 711 registered deaths[3]. These values represent about 1% of cases and 5% of deaths. In 2012 the incidence rate in Europe was " 0.4 cases/100,000/year (in Spain, 0.5 cases/100,000/year) " [4].

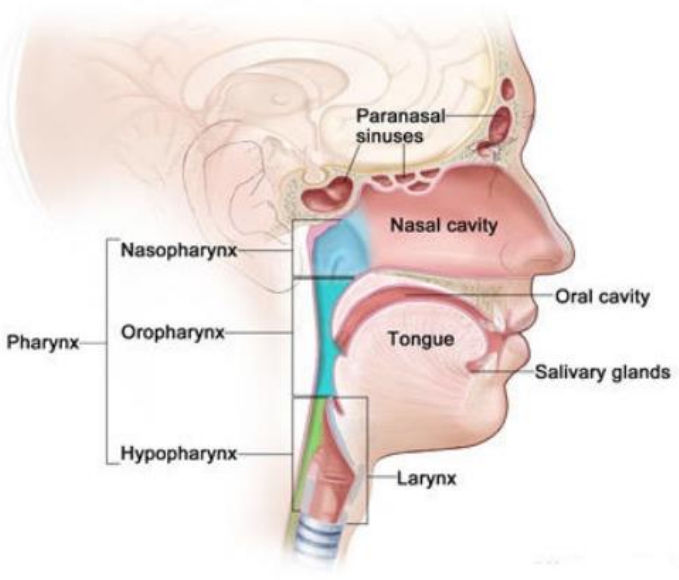

Figure 1.1 *Location of nasopharynx [5].*

There are three different histologies of NPC [1]:

- Non-keratinizing undifferentiated carcinoma (this is the most common type of NPC in the US.).
- Non-keratinizing differentiated carcinoma.
- Keratinizing squamous cell carcinoma.

1.2.2 Cancer staging

The common way to categorize cancer stage is the Tumour, Node, Metastasis (TNM) system, which is used worldwide. It describes the spread of the tumour cells and indicates to what extend tumour is affecting surrounding tissues. It is based on three component (stage T, N and M) that describes the extension of the pathology:

- **Stage T:** Is the substage that defines the size and extension of the primary tumour with appropriate notations to describe the tumour. There are different levels of T-staging, like [6]:

- T0: No evidence of a primary tumour.
- Tis: Carcinoma in situ.
- T1, T2, T3 or T3: Primary invasive tumour that the higher categories show increasing of size and/or local extension.

- **Stage N:** Is the substage associated with the extensions of the tumours to surrounding lymph nodes. There are different levels of N-staging, like [6]:

- N0: No regional lymph nose involvement with cancer.
- N1, N2 or N3: Evidence of regional node containing cancer.

- **Category M:** Is the substage that provides information related to distant metastases and their presence. The levels of M-staging are the following [6]:

- M0: No evidence of distant metastasis.
- M1: Distant metastasis.

1.2.3 Cancer treatment

There are different types of therapies to deal with cancer. The choice of the proper treatment is based on the severity, histology, location and other factors associated with the tumour. The physicians choose the most appropriated treatment considering

effectiveness and minimization of damage to healthy tissues. The most common therapies used in for NPC are chemotherapy, radiotherapy, immunotherapy or a combination of them, depending of the clinical situation.

- **Chemotherapy:** It is a type of treatment that is related to the use of drugs that may destroy cancer cells. It provides with some advantages like the systemic effect (affect every cancer cell independently on the location) and may be used to reduce tumor size before surgery (induction chemotherapy). However, it may affect also healthy cells and therefore is associated with side effect like nausea or hair loss.

- **Radiotherapy:** It is the most common way to deal with NPC. In order to perform the radiotherapy, it is necessary to develop a radiotherapy plan, which requires the segmentation of the affected region, which is done using medical imaging as a reference (magnetic resonance imaging or computed tomography). Intensity modulated radiotherapy (IMRT) is the most appropriated treatment [7] to deal with NPC. This therapy is a three-dimensional state-of-the-art clinical advance to treat oncology-related issues. It has a better precision and control of the targeted district, being safer and less toxic for surrounding tissues [8]. This therapy controls the intensity profile of each radiation beam, maximizing where it is required and decreasing the damage to organs at risks (OAR) [9]. In the figure 1.2, it can be clearly seen a better control of dose when using IMRT than with another type of therapy, three-dimensional conformal radiotherapy (3D-CRT). The main advantage is that, being radiotherapy a localized treatment (affect only one district of the body at a time) this therapy does not have systemic side effects like chemotherapy. However, the process can damage the healthy tissues surrounding the treated region.

- **Immunotherapy:** This biological treatment basically makes immunity system deal with cancer in a more efficient way. It uses material made from living organisms to help body to recognize some types of cancer cells [10]. The treatment has different ways to be applied such as targeted antibodies, cancer vaccines, tumour-infecting viruses and others [11].

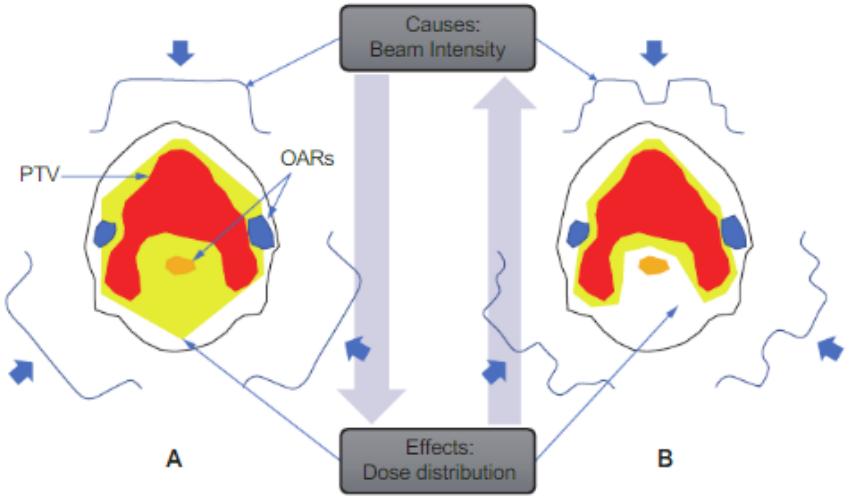

Figure 1.2 *Comparison of 3D-conformal radiotherapy (A) and intensity modulated radiotherapy (B) [9].*

1.3 Magnetic resonance imaging

1.3.1 Physical principle

Magnetic resonance imaging (MRI) is an imaging technique based on the phenomenon of spin resonance that happens when the hydrogen nuclei inside a uniform magnetic field are excited with radiofrequency wave. MRI is used widely in different fields of medicine, oncology in particular, where it is used for the detection of tumour and cancer. The advantages of this technique are the use of non-ionizing radiations (radio waves) which reduces the risk for health, and the excellent soft-tissue contrast, that can be controlled by properly adjusting some image acquisition parameters (contrast weighting).

From a physical point of view, MRIs use the magnetic fields generated by a powerful magnet that enforces the hydrogen protons of the analysed body to be aligned in either parallel or antiparallel way to the direction of that field as we can see in the figure 1.3. Usually clinical MRIs devices come in different range of magnetic field, usually between 0.5 and 3 Tesla [12].

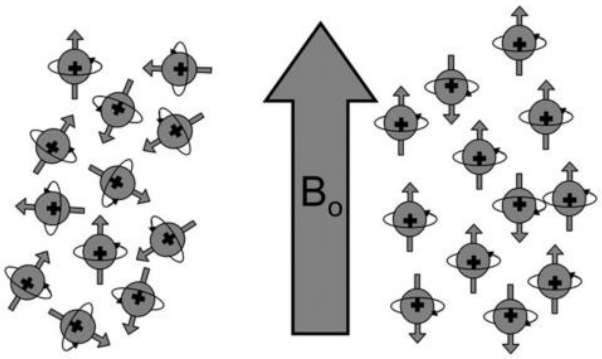

Figure 1.3 *Alignment of protons due to an external magnetic field (B_o). Left: There is no external magnetic field. Right: External magnetic field applied [13].*

In this equilibrium state, the protons are characterized by a motus of precession around their axis, which is like the movement of a spinning top (figure 1.4). It is important to state that the stronger magnetic field, the higher precession frequency. This relation can be mathematically described by the Larmor equation:

$$f = \gamma \times B_o \tag{1.1}$$

where f is the Larmor frequency [MHz], γ is the gyromagnetic ratio [MHz/Tesla] and B_o is the magnetic field [Tesla].

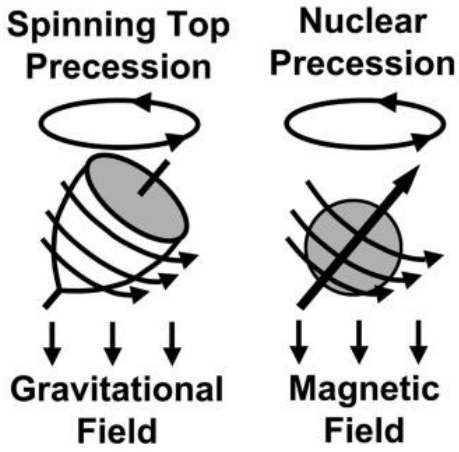

Figure 1.4 *Comparison between the precession of a spinning top (left) and the nuclear precession (right) [14].*

The net magnetic vector is the sum of all magnetic moments of the spins. Its component aligned in the direction of the magnetic field is called 'longitudinal magnetization' and is the result of the difference in the number of protons oriented parallel (low energy) and antiparallel (high energy) to the magnetic field. The longitudinal magnetization cannot be measured, because it is negligible compared to the external magnetic field B_o. To measure

the magnetization of the spin it is necessary to apply a radiofrequency (RF) in order to make the net magnetization vector to rotate 90° as it is seen in figure 1.5. If the RF pulse is applied for a longer time, longer rotations (e.g. 180°) of the longitudinal magnetization vector can be applied.

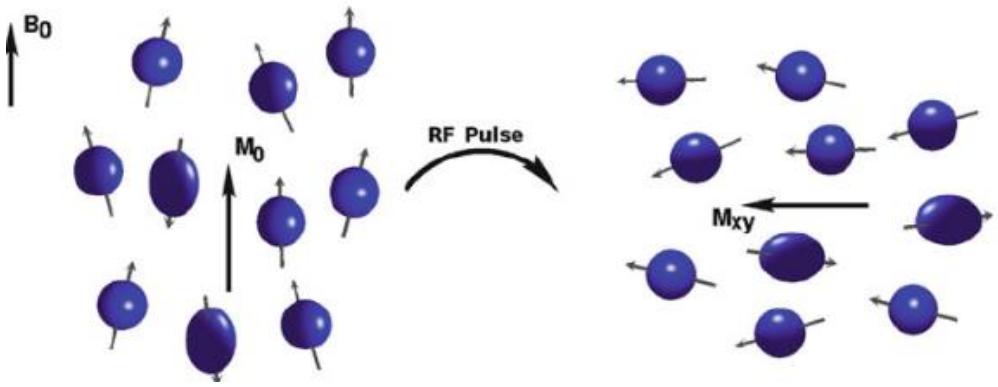

Figure 1.5 *Classical representation of the interaction of the spins with a 90° radiofrequency (RF) pulse to produce spin rotation. This leads to a macroscopic magnetization, M_{xy}, perpendicular to B_o [15].*

When the RF pulse is switched off, the spins return to their initial position, being realigned again with the magnetic field. Unlike the longitudinal magnetization, the transverse magnetization can be measured at any time after the end of the RF pulse. The decaying signal can be measured using sequences of multiple RF pulses and the information retrieved can be used to reconstruct tomographic images of the body (figure 1.6), where the signal depends on the quantity of spins in each element of volume (voxel) and the relaxation properties of the tissue in which the protons are in.

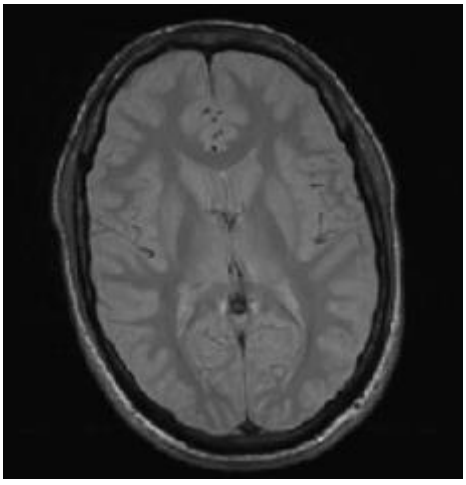

Figure 1.6 *Example of an image obtainable by magnetic resonance imaging [15].*

1.3.2 Magnetization relaxation and recovery

Figures 1.7-1.8 show what happens when the RF pulse is shut off. Two independent phenomena happen: the longitudinal magnetization quickly recovers from 0 to its original value; the transverse magnetization decays back to 0. The relaxation is the process that consist in the gradual reduction of the transverse magnetization and the corresponding recovery of the longitudinal magnetization.

Due to the phenomenon of relaxation, MRI has impressive ways to characterize a large number of tissues. The signals that are measured by the MRI and the contrasts between the signal from different regions of the body depend on both biological factors and imaging parameters. Among the biological factors, we can mention spin density, ρ, longitudinal relaxation time, T_1 and transverse relaxation times, $T_2$and T_2^* [15] . They are described below:

- **Spin density ρ**: is the density of the spin inside each voxel of the MRI image.

- **Longitudinal relaxation time T_1:** When defining two magnetic fields, B_0 and B_1 that are aligned in the z-axis and x-y plane respectively, we can assert that after removing the RF pulse generated by the transient magnetic field B_1 , the initial magnetic momentum M_z, return to its original value M_0 that is defined by a constant time also called 'the spin-lattice relaxation time constant (T_1)', as it is seen in figure 1.7 and equation (1.2).

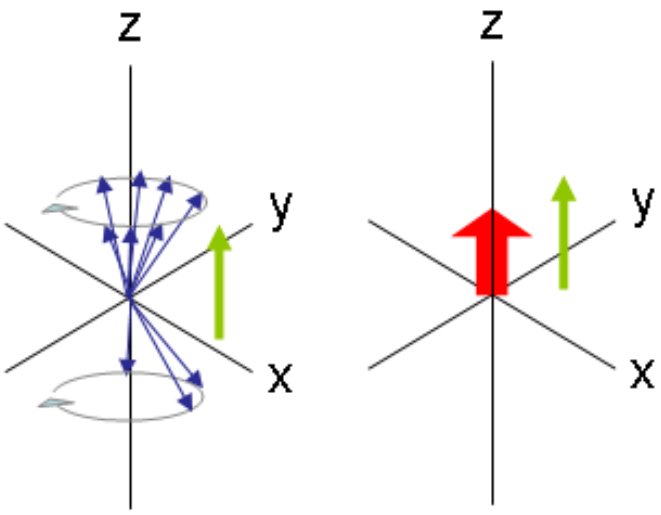

Figure 1.7 *Graphical representation of the recovery of the Mz magnetization to its original value [16].*

Such recovery can be described by the Bloch's Equation considering a certain flip angle α (the angle of rotation of the spin due to the RF pulse):

$$M_z(t) = M_0 \cos\alpha + (M_0 - M_0 \cos\alpha)(1 - e^{-\frac{t}{T1}}) \quad (1.2)$$

- **Transverse relaxation time T_2:** While the M_z comes to its initial state, there is a decay of the magnetization in the transverse plane in a constant time called 'the spin-spin relaxation time constant (T_2)'. This decay finish when reaching its initial value that is zero. The figure 1.8 represents graphically the decay of the magnetization in the transverse plane.

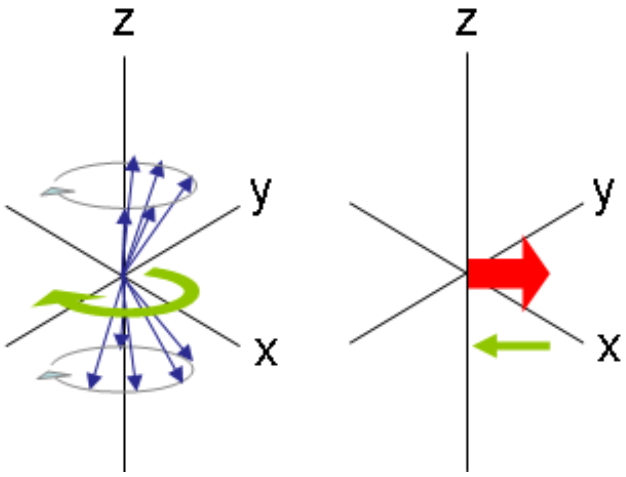

Figure 1.8 *Graphical representation of the decay of the magnetization in the transverse plane [16].*

The decay can mathematically be described by the following equation:

$$M_y(t) = M_0 \sin\alpha \, e^{-\frac{t}{T_2}} \quad (1.3)$$

The decay that is observed after shutting down the RF pulse is actually controlled by a smaller time constant (T_2*), which is smaller that T_2 and leads to a faster decay of the signal (figure 1.9). This faster decay is due to non-ideal acquisition condition in which the magnetic field is non-homogeneous inside the whole volume, causing a faster dephasing of the spins that leads to a faster decay.

1.3.3 Spin-echo imaging

The spin-echo is a pulse sequence used to acquire MRI images. It uses two different RF pulses, the first of 90° and the second of 180°, after a time interval τ after the end of the first one. This last pulse is added to recover the additional dephasing due to the magnetic field inhomogeneity that causes the T2*-decay. In the time interval [τ; 2τ], the dephasing that was accumulated during the period [0; τ] recovers and this cause a peak in the measured MRI signal that is called 'echo' [5]. The figure 1.9 display the spin echo process.

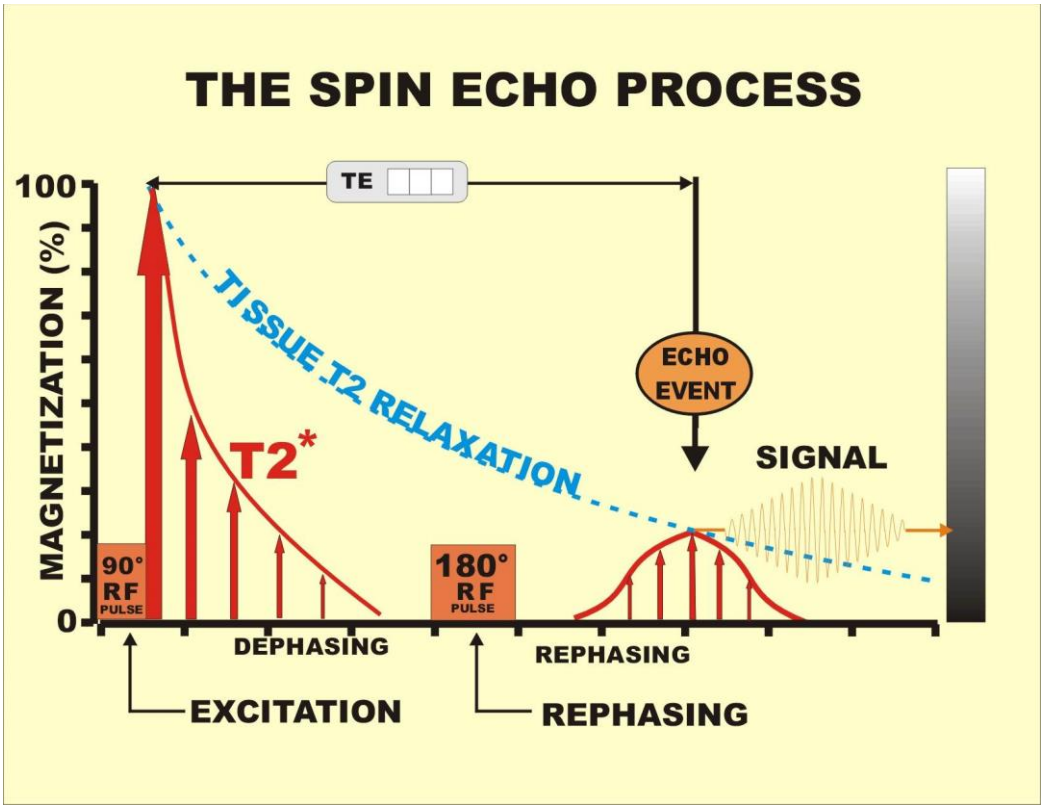

Figure 1.9 *Graphical representation of the spin echo process [17].*

The pulse sequence is characterized by two parameters (figure 1.10): the time of echo (TE), that represents the time between the middle of the 90°-pulse and the peak of spin-echo; the time of repetition (TR), representing the time between a 90° RF pulse and the next one.

By controlling this TR and TE, a wide range of different images can be acquired, each with different contrasts among the tissues. The two main categories that were used for the thesis are T1-weighted and T2-weighted images (figure 1.11):

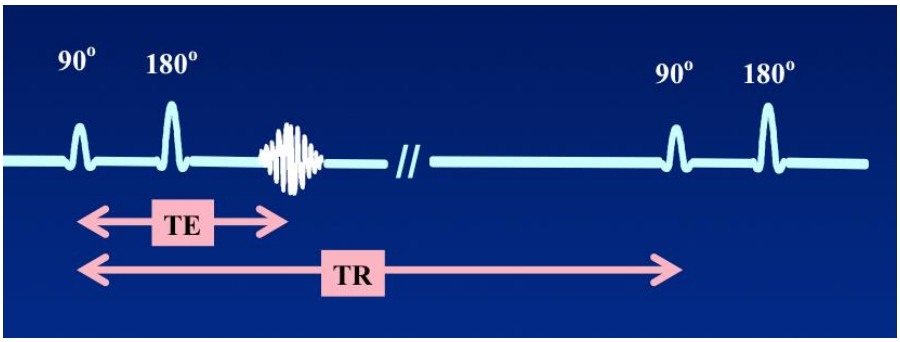

Figure 1.10 *Graphical representation of spin echo process [18].*

- **T1-weighted images (T1w):** In the case a short TR (<700 ms) and TE (<25 ms) are used, the contrast of the images is generated by T1 differences. The images associated with this process are called T1-weighted images. When removing the RF pulse, all the spins will go to the equilibrium. However not all of them will come back in the same time, tissues with shorter TR will appear brighter than tissues with longer TR. If it is necessary to improve the contrast, some types agents, such as gadolinium, can be used.

- **T2-weighted images (T2w):** when a long TR (>2000 ms) and TE (>60 ms) are used, we can obtain images called T2-weighted images, because they highlight contrasts based on T2 differences. Tissues with larger T2 appear brighter in this type of images.

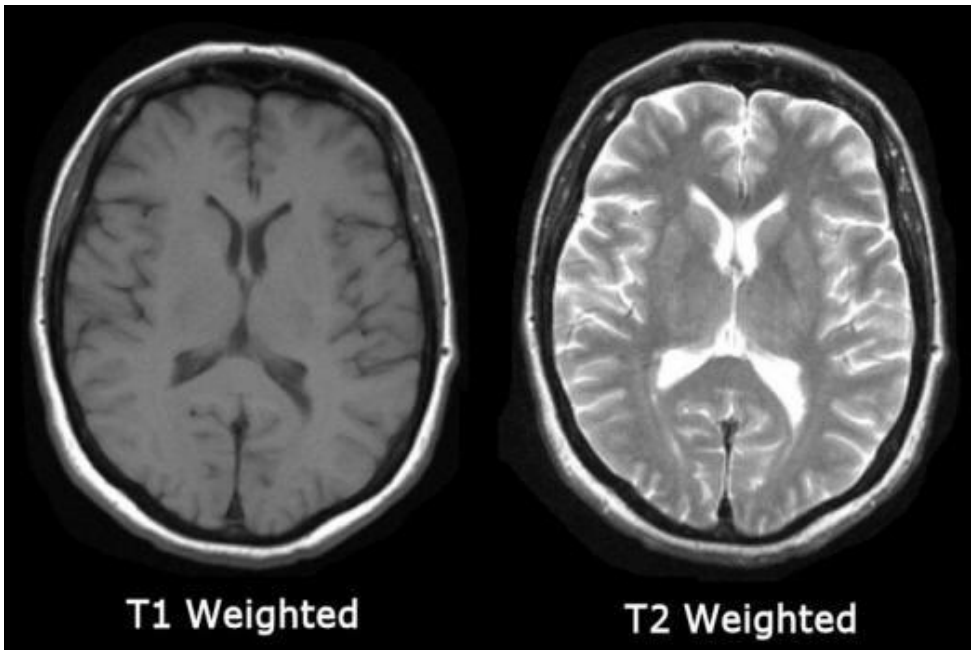

Figure 1.11 *Example of T1w and T2w images [19].*

1.3.4 Diffusion-weighted imaging (DWI)

Diffusion-weighted imaging (DWI) is an MRI technique that allows obtaining medical images that provide insight about the water molecules diffusion that is found in the tissues. It employs the difference in Brownian motion and let analyse in a depth way the district of human body. The motion of water molecules is limited by the interaction with the cell membrane and macromolecules. For example, the tumours have a high limitation of motion so the cellular density will be high [20]. The figure 1.12 shows districts of the brain with and without diffusion restriction. The process to generate diffusion weighted is to use two similar gradients around 180°-RF pulses, that is called the 'Stejskal-Tanner' approach [21]. The first gradient lets the spins to have its phase and the second one enforces an opposite phase but with the same value. When applying the second gradient two options can happen: on the one hand, there is no movement or position change of the spins, so in that case the conclusion is that there is no diffusion; on the other hand, there is change and also a loss of signal intensity, being concluded that the diffusion occurs [22].

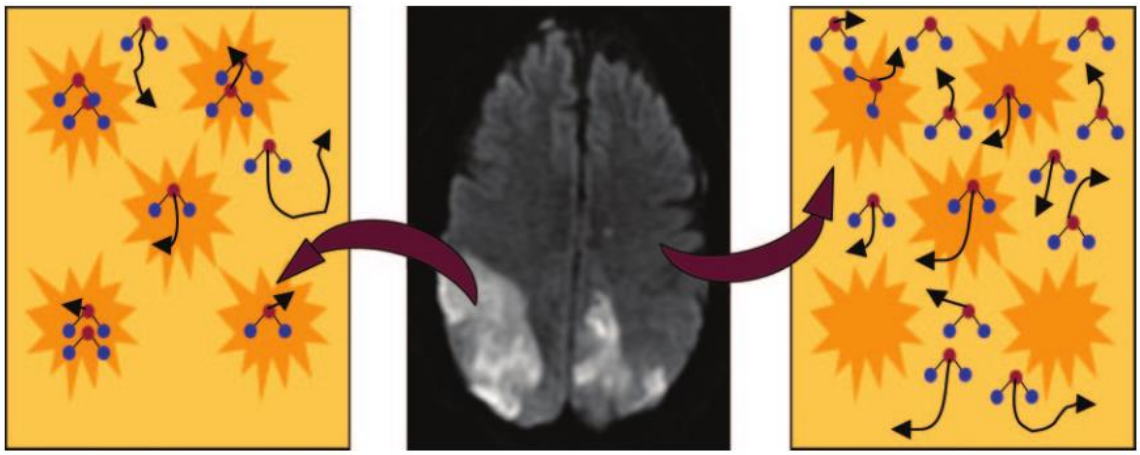

Figure 1.12 *Example of schematics and corresponding diffusion-weighted brain image show areas of restricted (left) and unrestricted (right) diffusion [22].*

The way to measure the sensitivity of DWI is through of 'b-value' parameter that is defined in s/mm^2 as unit of measurement.

$$b = \gamma^2 G^2 \delta^2 (\Delta - \frac{\delta}{3}) \tag{1.4}$$

The interpretation of each variable is: G, δ, Δ and γ represent the gradient pulse amplitude, the duration of the gradient pulse, diffusion time and the gyromagnetic ratio (defined by the ration between the magnetic moment and the angular momentum) respectively. Usually the parameter 'G' is variable and the other remains constant; and also we can comment that from a qualitative point of view, the higher b-value, the higher signal attenuation due to the diffusion [5]. As example, when b-value is 0 s/mm^2 is a T2w image. The b-value is strongly associated with the decay of the signal as it is shown below:

$$S(x, y, b) = S(x, y, 0)e^{-b.ADC(x,y)} \quad (1.5)$$

In this equation: S(x,y,0) is the signal measured in voxels (x,y) using a b-value of 0 s/mm^2, and ADC (x,y,z) represents the apparent diffusion coefficient that is a property of the tissue [5]. to obtain maps of ADC we need to have two or more DWI images calculated with different b-values, so the ADC can be computed pixel-wise. as represented in figure 1.13.

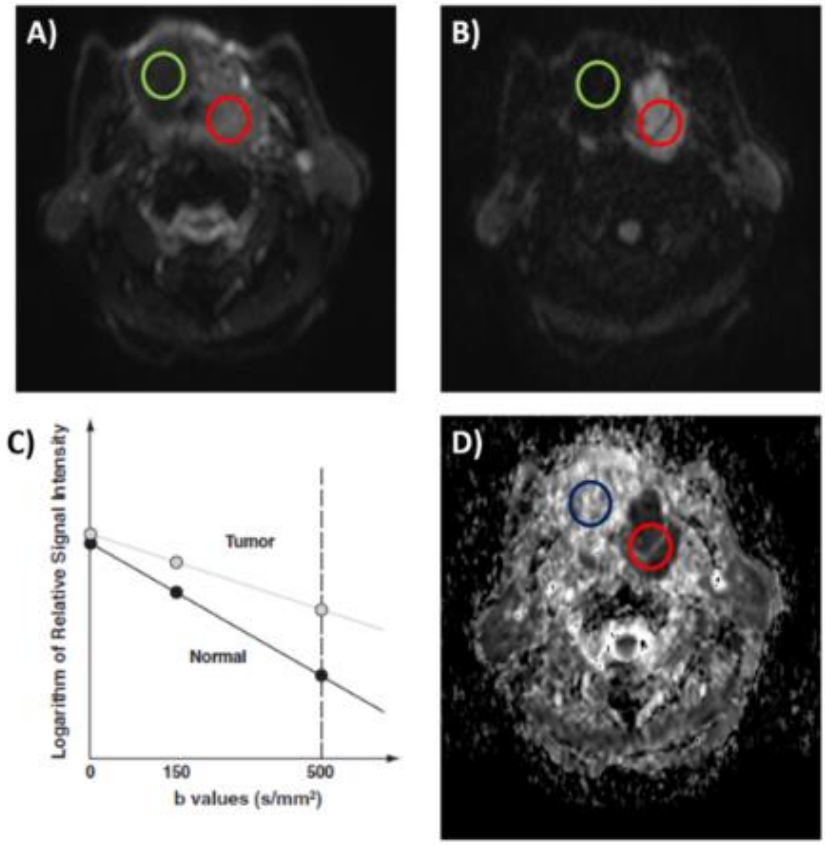

Figure 1.13 *Process of creation of an Apparent Diffusion Coefficient (ADC) map from diffusion-weighted images (DWI). A) DWI image with a b-value of 0 s/mm2. B) DWI image with a b-value of 500 s/mm2. C) Signal decay by b-value for tumor and normal tissue. D) Map of ADC [5].*

1.4 Radiomics

1.4.1 General introduction

It is widely known that all the tumours are different to each other and there is a wide number of available treatments to face them such as radiotherapy, chemotherapy, surgery, targeted drugs and immunotherapy. It is therefore necessary to identify biomarkers in order to distinguish biological differences among tumors. Also, there is a need to obtain those biomarkers that captures three-dimensional tumours complexity in a non-invasive and low-cost way. Radiomics, the technique of high-throughput features extraction from medical imaging, can fulfil all the requirements previously mentioned. It performs an analysis of standard clinical images from conventional scans such as computer tomography (CT), MRI or positron emission tomography (PET) by employing a specific software to compute features describing shape, intensity and texture of the tumour, thus capturing quantitative information that cannot be seen by eye, and which can be used as an independent source of biomarkers.

1.4.2 Workflow

There is a sequence of steps in the workflow of radiomics that must be followed in order to perform a complete analysis (see figure 1.14): 1) Image acquisition and processing; 2) Image segmentation; 3) Features extraction; 4) Data analysis (model building).

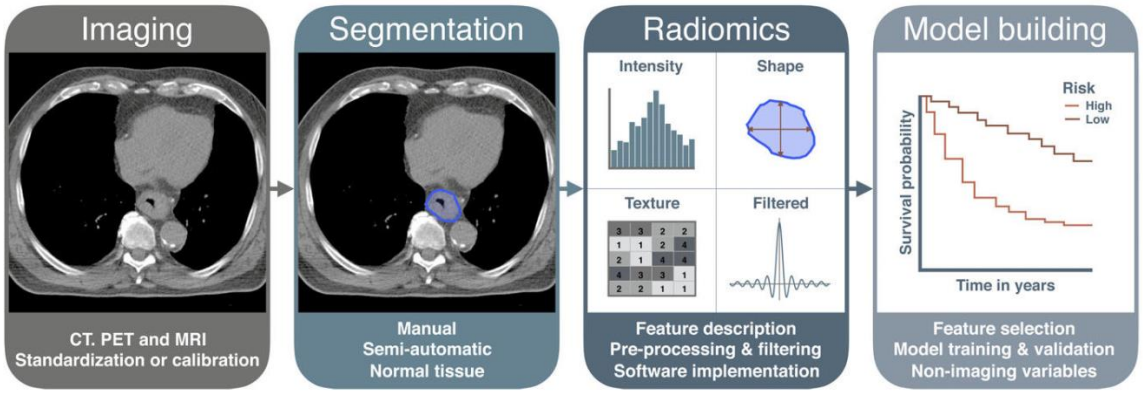

Figure 1.14 *Radiomics workflow. From left to right: Imaging acquisition and processing, Image segmentation, feature extraction and data analysis ADC [5].*

To have a better understanding, the single steps are herein described:

- **Image acquisition and processing:** The interaction of patient with clinical scan devices provides us with rich medical images in information.

The medical scan can be CT, PET and MRI usually. The images obtained in two-dimensions can be employed to build three-dimensional surfaces by using both different techniques reconstruction and a specific software. When the three-dimensional surfaces are reconstructed, it is necessary to smooth the surfaces and reduce all type of artifacts so that the quality would be improved. To obtain 3D surfaces, it is needed to process all the slices and performing a process like superposition.

- **Image segmentation:** This step concerns the delimitation of the tumour that is being separated from healthy tissues to be processed and analysed later. All the process must be carefully performed so that misleading information would not be obtained. Currently, there is not a standard way to segment the tumour given that there are numerous ways to carry out it such as manual, semiautomatic or automatic methods. Nonetheless, radiologists mostly prefer using manual segmentation, even if it is a time-consuming process, because for some districts, automatic or semiautomatic segmentation me [23].

- **Features extraction:** the phase of features extraction consists in the computation of a large number of quantitative characteristics. The number may vary from a few to several hundred or thousands, depending on the category of features considered and the software used. Usually, features can be divided in four main categories: shape, first order statistics, textural and filter-based features (see Subsection1.4.3).

- **Data analysis:** After the collection, data are analysed according to defined objectives. The application can be associated to machine learning process, survival prediction, regression process and other [24]. Data analysis may include some additional post-processing techniques such as features selection, normalization and missing data imputation.

1.4.3 Images features

There is a high number of features that can be extracted from medical images through of mathematical algorithms-based softwares. All these features can be divided mainly

in four categories: First order statistics; Shape and size features; textural features; filter-based (wavelet in the context of this thesis) features [24]. All the above-mentioned categories can be described as follows:

a. **First order statistics (FOS):** They are based on a histogram that has been derived of the segmented region of interest (ROI). The histogram of an image represents quantitatively the distribution of the pixels' intensity. Even though these features, such as intensity mean, intensity standard deviation and other, are not obtained from complex processing, they display the main properties of the tumour histogram. These features are sensitive to bias and scaling of the intensity histogram and therefore an adequate preprocessing is required to standardize the range of intensity values, in non-standardized images like MRI [24].

b. **Shaped and size features:** These features provide us with geometrical information associated with the shape of tumours or ROI that can be obtained from reconstructed three-dimensional surfaces as well as from two-dimensional structures [24].

c. **Textural features category:** This describe the spatial displacement of the different grey values inside the ROI. These features are a powerful tool when a non-homogeneous tissue is evaluated o analysed given that they extract information from the different levels of grey that are represented through matrix after a mathematical processing. The described matrixes mainly used to commonly extract data are the grey level co-occurrence matrix (GLCM) and the grey level run length matrix (GLRLM) [24].

   - **Grey level co-occurrence matrix:** This matrix of $N_g \; x \; N_g$ elements is used to obtain the distribution of intensities from a spatial point of view by considering two main parameters that are the angle (θ) , typically is 0°, 45°, 90° and 135°; and any integer distance (δ) admitted inside the image [25]. N_g is the number of discrete grey values in an image. This matrix represents how many times the element grey value *j* appears at a distance

δ from the grey value *i* in the direction *θ*. A graphical explanation of how to compute GLCM is given in figure 1.15.

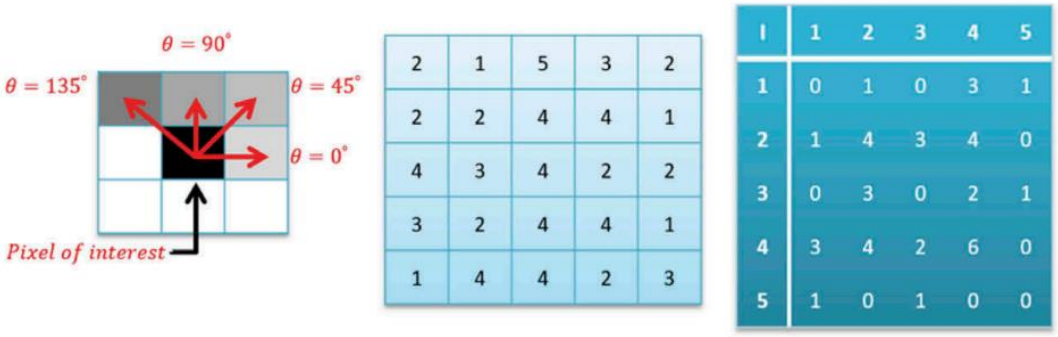

Figure 1.15 *(a) Illustration of the inter-pixel relationships characterized by the user defined parameter, θ (b) An example 5 × 5 matrix with gray values ranging from 1 to 5. (c) The resultant symmetric grey level co-occurrence matrix (GLCM) obtained by multiplying the asymmetric GLCM with its transpose [25].*

- **Grey level run length matrix:** This matrix represents quantitively the grey level runs that are described as the length of consecutive pixels that have the same grey level [5]. Many voxels have the same intensity of grey level being shown in continuous voxels through different directions (identified by the angle θ). Given a direction, the GLRLM [25] shows how many times the element grey value *i* appears consecutively for *j* times [5]. In the figure 1.16 can be found the schematic process.

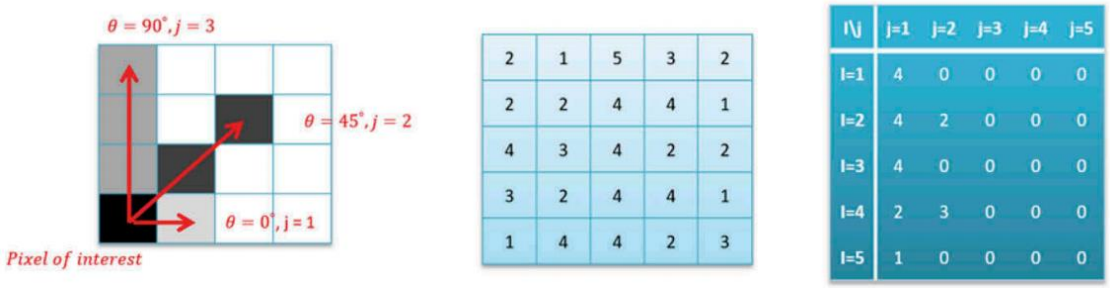

Figure 1.16 *Illustration of the inter-pixel relationships characterized by the user defined parameters, angle θ and run length j. (b) Example 5 × 5 matrix with values ranging from 1 to 5. (c) Resultant grey level run length matrix (GLRL) for run lengths of 1 to 5 and θ=0 [25].*

d. Wavelet features: Wavelet transforms decomposes the information from the original image to account for both frequency and spatial information. In this thesis, the 3D discrete wavelet transform (DWT) was used to obtain the radiomic features. In an image, the DWT is performed through a cascade tree

of low-pass and high-pass filters followed by down sampling by a factor of 2. In case of 2D images, for example *I(x,y)*, the decomposition is associated with a process of filtering and down-sampling in the x and y directions with a 1D low pass and high pass filter, being sub-bands obtained. An illustration of this process can be seen in the figure 1.17, where it is relevant to mention that its radiomic information gives new relevant information.

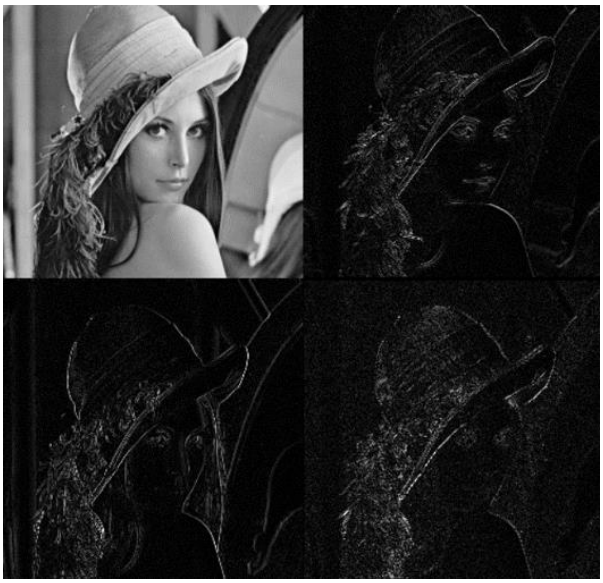

Figure 1.17 *Example of first level 2D discrete wavelet decomposition of an image. The upper-left corner represents the low-low pass sub-band (an approximation of the original image). The upper-right, lower-left and lower-right parts of the image represent the low-high pass, high-low and high-high sub-bands respectively [5].*

In case of 3D volumes such as the ones associated with MRI exams, there will be 8 possible combination as it is shown in the figure 1.18.

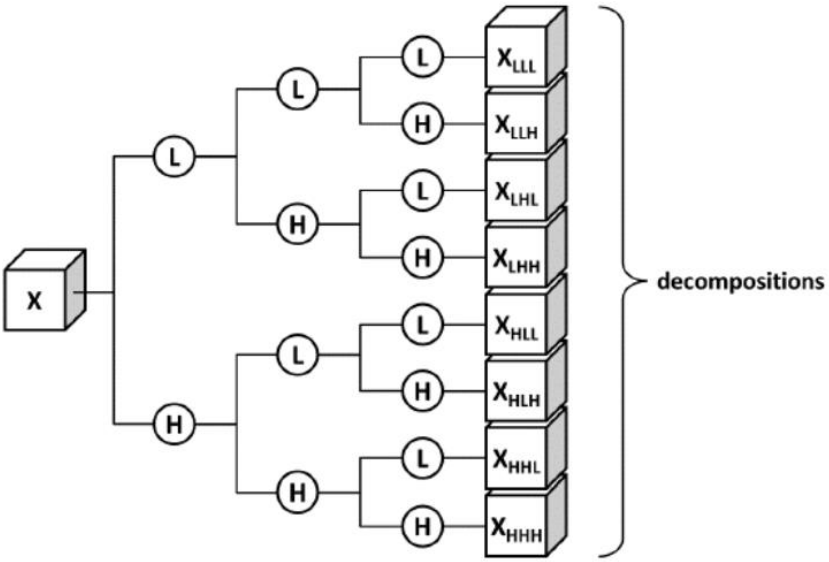

Figure 1.18 *Schematic representation of the 8 possible discrete wavelet decompositions of a 3D image [5].*

1.4.4 State-of-the-art

Radiomics is a new approach that promise benefits for people's health. There are many researches in all the world showing that its use can let to predict the pre-treatment of a tumour, risk of occurrence and survival analysis [26]. In the context of NPC, two studies can be cited:

- In the first study, an MRI-based radiomic nomogram for the assessment of the risk of recurrence was trained and validated on a dataset of 140 patients (80 and 60 patients for training and validation respectively). A total of 970 radiomic features extracted from pre-treatment MRI was used for the purpose. A Cox proportional hazard regression model was trained by selecting eight contrast-enhanced T1-weighted (CE-T1w) image features and seven T2-weighted (T2w) image features. As main result, it was concluded that radiomic nomogram was able to categorized patients in low and high risk groups [27].
- In the second study, radiomics was used to predict the risk of progression in NPC [28]. In this case, 113 patients were evaluated, being divided in 80 for training and 33 for validation. A total of 970 features were extracted and were select by using LASSO operator. The performance of the model was evaluated using area under the curve (AUC). This study demonstrated that a radiomic model based on the combination of CE-T1w and T2w images works better than radiomic models obtained from the individual imaging techniques.

However, it is important to emphasize that there are no previous studies about how to impute missing data when some features are missing for example to missing imaging technique. This is an important issue to be solved, since, in the clinical practice, it is not guaranteed that every image sequence will be available for any patient. A study comparing different missing data imputation techniques was found for DNA microarrays [29], but no equivalent study was found for radiomics. The purpose of this thesis was to try to fill this gap in knowledge by comparing different missing data imputation methods for radiomics.

1.5 Data analysis

1.5.1 Data normalization

The process of normalization implies to put different features of many observations in a same scale, and it is too useful due to fact that when analysing features with different range of values, some of them can be neglected. For example, there can be two types of features A and B, the first one can have a range of values from 0 to 5000 and the other from 1 to 3, so when process information feature 'A' may have more relevance for some algorithms. Normalization avoids this issue an may help the model trained on the normalized features to perform better [30].

One of the most used method for normalizing data is 'z-score normalization', it mathematically can be expressed in this way:

$$z = \frac{value - \mu}{\sigma} \tag{1.5}$$

Here, μ and σ represent mean value and standard deviation respectively.

1.5.2 Missing data imputation

When different types of data are obtained by different sources, they are stored one by one to in a database wherein usually some of them can be missing. In many cases, missing data reduce the quality of downstream analysis, since is not possible to exploit all the information. To deal with this issue, different approaches are employed such as:

- **KNN imputation:** This approach called 'the k-nearest neighbours' is very simple to understand. The idea is that having a set of data or observations with missing and non-

missing information, the ones with non-missing feature values can be used to impute data where necessary. It means that if there is a specific observation with some missing values, this algorithm will try to find the most similar case amongst the observations with non-missing data to assign proper values by using weighted [29].

- **Imputation by simple rules:** It is one the easiest way to fill missing data. If there are a range of observations listed in rows being characterized by features displayed columns, there are some statistics (for example. mean, median or random values) that can be computed from the column and use to impute the missing data where necessary.

- **Low-Rank Completion with Nuclear Norm Optimization:** It works under the assumption of an existing low-rank structure that will give rise to optimization-based problem for matrix filling [31].

- **Generalized Spectral Regularization (Hard imputation):** To address missingness data-related issues, this method uses the least absolute shrinkage and selection operator (LASSO) under this scenario: If the analysed model has enough zeros or is sparse, the LASSO overestimates quantitively the coefficients different from zero. It can be an inconvenient for data analysis, so low-rank assumption and hard thresholding idea are employed to face it [32].

- **Singular Value Thresholding for Nuclear Norm Optimization (fSVT imputation):** The algorithm of this approach generates matrices through of an iterative process. The most attractive features are two basically that are, first, soft-thresholding applied to disperse matrix and, second, the rank of iteration is not decreasing [33].

1.5.3 Statistical analysis

In this part, the statistical comparison used in this thesis (Friedman's test and post-hoc comparisons) are described.

- **Friedman's test**

  The Friedman's test a non-parametric statistical test that is used in the analysis of three or more variables or conditions of a group of related samples (null hypothesis). It evaluates the differences in the medians amongst the different conditions. The outcome gives information about the variations of the samples but do not offer information about the difference among the pairs. To illustrate, it can be cited that whether there is a list of patient with a specific disease, they can be treated with three or more different drugs in order to know their health improvement; after the treatment all the results may be quantified and in order to be able to distinguish statistically the difference between the effects, Friedman's test can be applied to the outcome values.

- **Post-hoc comparisons**

  In some cases, the null hypothesis can be rejected, thus it is concluded that there are differences amongst the conditions, treatments o variables. However, it is not possible to understand which are the pairs of groups that present the significant differences. A powerful tool to deal with that scenario are the post-hoc comparisons that are used to compare more than one pair of conditions or treatment at the same time. Since multiple comparisons are performed at once, the p-value of the test must be corrected in order to reduce the probability of false positives. There are different options for p-value correction, such as the Tukey method, Scheffé's test, Bonferroni method and so on [34].

- **Tukey method:** It tests all possible pairwise differences by employing a studentized range distribution in order to determine if at least one differences is significantly different from zero [34].
- **Scheffé's test:** It tests all possible joint pairwise simultaneously in order to determine if at least one is significantly different from zero [34].
- **Bonferroni method:** uses a lower threshold of significance compared to the single comparison; the threshold is the value traditionally used for significance ($p=0.05$) divided by the number of comparisons [34].

We can remember that whether the null hypothesis is validated, this analysis will not provide relevant information.

Chapter 2: Materials and methods

In this part, the dataset and the methods used to perform a comparison among missing data imputation method are described.

2.1 Patient data

The dataset was collected by the "Istituto Nazionale dei Tumori (INT)" of Milan, Italy as part of a project of collaboration with Politecnico di Milano. The initial dataset was composed of 185 patients affected by NPC. The main clinical and follow-up details of the data for the patients are presented in table 2.1.

CLINICAL DATASET	
Number of patients	185
Age (years)	49 [40-59]
Sex	55 females; 130 males
Overall stage	I-II: 11 III-IV: 174
Stage T	T1-T2: 104 T3-T4: 81
Stage N	N0: 10 N1-N2: 88 N3: 87
Follow-up time (months)	60 [45-66]
Number of deaths	23
Number of concurrences	51

Table 2.1 *Clinical and demographic characteristics of the patients. Numerical values are displayed as median and interquartile range.*

2.2 Image acquisition

For the 185 patients of interest, up to four types of images were acquired: pre-contrast T1w images, post-contrast T1w images, T2w images and DWI images. DWI images were acquires using different b-values in the range 0-1000 s/mm^2. ADC images were reconstructed using multiple DWI images as described in Subsection 1.3.4 All the images were acquired using a Siemens Magnetom Avanto 1.5 T. T1w and T2w images were acquired using turbo spin-echo pulse sequence, while DWI image were acquired using echo

planar imaging. It is needed to remark that four types of images were obtained under the following procedure: to have the availability T1w MRI (pre and post contrast) and T2W MRI where it was used a spin-echo pulse sequence; also to have DWI MRI images acquired with at least two b-values in range 0-1000 s/mm2 acquired by echo-planar imaging. Details of image acquisition are reported in table 2.2.

CLINICAL DATASET			
Image sequence	**T1w/T1wCont**	**T2W**	**ADC**
Scanner	Siemens Magnetom Avanto 1.5 T (163 patients) Philips Achieva 1.5 T (12 patients) Others (10 patients)		
Number of images	178/150	180	115
Pulse sequence	Turbo spin-echo	Turbo spin-echo	Echo planar imaging
Time repetition (ms)	528 [479-587]	4655 [3360-5300]	4800 [4200-5200]
Time of echo (ms)	12 [12-12]	109 [105-109]	79 [77-93]
Slide thickness (mm)	3 [3-3]	3 [3-3]	4 [3-4]
Slice spacing (mm)	3.9 [3.9-3.9]	3.9 [3.9-3.9]	5.2 [3.9-5.2]
Pixel spacing (mm)	0.57 [0.57-0.69]	0.51 [0.41-0.57]	1.89 [1.19-1.89]

Table 2.2 *Description of the image acquisition parameters for the patients of the dataset. Data are divided by image sequence. Numeric variables are displayed as median and interquartile range.*

2.3 Image segmentation

The segmentation was performed manually by radiologist with a wide experience employing appropriated software. The ROIs associated to this study are focused on primary tumour and main affected lymph nodes. In particular, T2w MRI was used to segment the ROIs and they

were employed to extract the features from T1w images and ADC maps. An example of segmentation from this study is illustrated in figure 2.1.

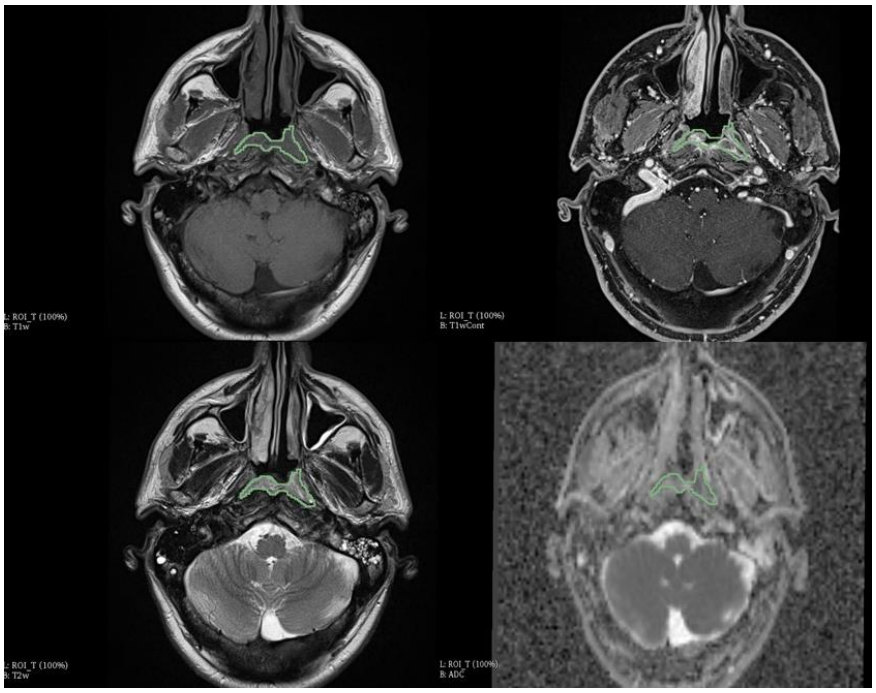

Figure 2.1 *Examples of our study images of a tumour segmentation. A) T1-weighted image. B) Contrast T1-weighted image. C) T2-weighted image. D) Apparent diffusion coefficient maps. The segmentation of the tumour is also shown.*

2.4 Image pre-processing

The medical images of each patient are processed one by one, developing a selection process according to the type of defined image. The way of pre-processing T1w and T2w follows this order:

- The removal of noise was done by using a 3D gaussian filter with 3x3 voxel kernel and σ=0.5.
- The intensity inhomogeneities were corrected by employing a N4ITKalgorithm.
- Apply z-score normalization to standardize the intensity.
- Resample of the voxel size to an isotropic resolution of 2 mm was performed using B-spline interpolation.

In case of ADC, intensity standardization and N4ITK were not performed, since ADC is a quantitative image so it already has a standardized range of intensity, and does not have

inhomogeneity effect, because it given by the differences between two images with identical inhomogeneities, so the net effect is null.

2.5 Radiomics feature extraction

The radiomic features are extracted using the open-source software Pyradiomics, and the full list can be found in the online documentation [35]. The feature obtained are associated with the following categories: FOS, shape and size, GLCM, GLRLM and wavelet.

After the images pre-processing, the extraction of radiomic features is performed with the use of a combination of appropriate software. The computational tools employed are PYRADIOMICS [36] and MATLAB 2018 [37]. In order to connect them, a script was developed in MATLAB wherein the application of the first one was included.

The radiomic features were extracted from each combination of image sequence (T1w pre- and post-contrast, T2w and ADC) and for each ROI type (tumor or lymph node) for a total of 8 subcategories. From each subcategory, 536 features were extracted, for a total 4288 features. The output of all this process is stored in a 'csv file' wherein the information of the image type, ROI type and radiomic features is shown and divided in column. An example of that can be seen in the figure 2.2.The script developed in MATLAB provided us with a file to each patient.

	1 ImageType	2 ROIType	3 ImgFile	4 ROIFile	5 original_shape_Elongation	6 original_shape_Flatness	7 original_shape_LeastAxisLength	8 original_shape_MajorAxisLength
1	'ADC'	'N'	'ADC.mha'	'ROI_N.nrrd'	0.4127	0.3272	20.4346	62.44
2	'T1w'	'N'	'T1w_2.mha'	'ROI_N.nrrd'	0.4357	0.3310	19.7376	59.62
3	'T1wCont'	'N'	'T1wCont_2...	'ROI_N.nrrd'	0.4453	0.3315	19.3858	58.47
4	'T2w'	'N'	'T2w_2.mha'	'ROI_N.nrrd'	0.4389	0.3339	19.6878	58.97
5	'ADC'	'N'	'ADC_3.mha'	'ROI_N_2.nrrd'	0.5236	0.4114	17.5542	42.66
6	'T1w'	'N'	'T1w.mha'	'ROI_N_2.nrrd'	0.4502	0.3662	17.5414	47.90
7	'T1wCont'	'N'	'T1wCont....	'ROI_N_2.nrrd'	0.4720	0.3820	18.0659	47.29
8	'T2w'	'N'	'T2w.mha'	'ROI_N_2.nrrd'	0.4466	0.3659	17.4802	47.77
9	'ADC'	'T'	'ADC_2.mha'	'ROI_T.nrrd'	0.8932	0.8225	29.2778	35.59
10	'T1w'	'T'	'T1w.mha'	'ROI_T.nrrd'	0.8854	0.8156	29.2229	35.82
11	'T1wCont'	'T'	'T1wCont....	'ROI_T.nrrd'	0.9061	0.8302	29.2992	35.28
12	'T2w'	'T'	'T2w.mha'	'ROI_T.nrrd'	0.8875	0.8104	29.0824	35.88

Figure 2.2 *Extract of characterization information of the medical image from a chosen randomly patient. It can be seen the way of information organization, however in the graph it can be seen only 04 from 536 features.*

2.6 Features processing

When obtaining the information of each patient in different tables, that information is reshaped in a more appropriate way. Known that there are two types of ROI and that each of them has four image types and each image type has 536 features, all the information is

organized in the way that is shown in the figure 2.3 that represents the complete database after characterization of each medical images.

PATIENT	T				N			
	T1w (536 features)	T1wCont (536 features)	T2w (536 features)	ADC (536 features)	T1w (536 features)	T1wCont (536 features)	T2w (536 features)	ADC (536 features)
Patient 1								
Patient 2								
...	...	...	...	...	...	...	...	...
Patient 184								
Patient 185								

Figure 2.3 *Format of the data base obtained after the reshape of information.*

For this study we chose to use only the features related to the main tumour (indicated with T) and we excluded the features referred to the largest lymph node (indicated with N). This was done in order to reduce the dimensionality of the dataset and make the following computation easier. The N-based features were therefore removed. As consequence the database is reduced as is shown in the figure 2.4.

PATIENT	T			
	T1w (536 features)	T1wCont (536 features)	T2w (536 features)	ADC (536 features)
Patient 1				
...				
...				
...				
...				
Patient N				

Figure 2.4 *Reduction of the database due to staging selection.*

Then, in the table reduced, we explored and is observed that there were some missing values (indicated with 'nan') due to the lack of availability of some imaging sequences. The figure 2.5 displays as these 'nan' values are distributed.

	1 Exam_ID	2 T_T1w_original_shape_Elongation	3 T_T1w_original_shape_Flatness	4 T_T1w_original_shape_LeastAxisLengt	5 T_T1w_original_shape_MajorAxisLengt	T_T1v
13	'179_00'	0.9550	0.7444	30.7767	41.3469	
14	'178_00'	0.6166	0.2551	11.7917	46.2179	
15	'66_00'	0.7220	0.4999	26.2657	52.5403	
16	'177_00'	0.8221	0.4743	25.0823	52.8881	
17	'176_00'	0.8018	0.5104	19.8448	38.8779	
18	'173_00'	0.6244	0.3208	15.9926	49.8599	
19	'166_00'	0.4236	0.2617	7.1172	27.1986	
20	'171_00'	0.5919	0.3832	13.3516	34.8410	
21	'172_00'	0.8557	0.4675	17.9956	38.4902	
22	'167_00'	NaN	NaN	NaN	NaN	
23	'71_00'	0.8254	0.4953	16.7745	33.8642	
24	'180_00'	0.7957	0.6556	25.2311	38.4881	
25	'174_00'	0.9805	0.8287	27.9844	33.7671	
26	'169_00'	0.6725	0.4418	23.7220	53.6981	

Figure 2.5 *Observation detected with missing information.*

Thus, it was necessary to remove all the observations with missing information in order to consolidate a non-missing complete database. After the removal of non-complete patients, the observation number was reduced from 185 to 115 like is shown in the figure 2.6.

	1 Exam_ID	2 T_T1w_original_shape_Elongation	3 T_T1w_original_shape_Flatness	4 T_T1w_original_shape_LeastAxisLengt	5 T_T1w_original_shape_MajorAxisLengt	6 T_T1w_original_shape_
104	'113_00'	0.6976	0.4820	27.3503	56.7492	
105	'114_00'	0.6684	0.4276	19.5912	45.8182	
106	'122_00'	0.7261	0.5845	36.7000	62.7856	
107	'120_00'	0.4980	0.3945	17.7405	44.9665	
108	'121_00'	0.6980	0.4247	14.1124	33.2281	
109	'124_00'	0.8336	0.4495	17.0946	38.0343	
110	'123_00'	0.7809	0.4312	16.5897	38.4767	
111	'126_00'	0.4755	0.2888	15.6477	54.1909	
112	'127_00'	0.9536	0.5915	30.4023	51.4017	
113	'05_00'	0.8914	0.6417	17.8870	27.8753	
114	'159_00'	0.6012	0.3353	20.4703	61.0546	
115	'04_00'	0.6385	0.3806	18.4993	48.6013	

Figure 2.6 *Extraction where it is seen the reduction of observations due to information missingness.*

Once processed, the information is processed additionally in a normalized scale made column by column by using the z-score method; that process was performed in order to analyse the performance of a normalization process in radiomic application. As stated in section 1.5.1, normalizing the features ranges can be useful to help the following data analysis part. For our case, z-score [38] was chosen as a normalization algorithm since is the one that has been used the most in studies of radiomics.

The subsequent analysis (Section 2.7-2.9) were performed on both the normalized dataset and a non-normalized dataset. This was done because, in general, previous studies of literature used both non-normalized and normalized dataset.

2.7 Generation of missing data

The complete database was used to perform a simulation of missing information. The simulation consists in replacing some random values with empty spots (not a number or nan). Then, trying to simulate a real clinical situation, the images features coming from post-contrast T1w and ADC images are the ones for which 'nan values' are assigned randomly. For the randomly selected patients, all the features related to types mentioned obtained 'nan values' as it is shown in the figure 2.7. Pre-contrast T1w and T2w images were always kept because it is reasonable to assume that they are always available in the clinical practice.

PATIENT	T			
	T1w (536 features)	T1wCont (536 features)	T2w (536 features)	ADC (536 features)
Patient 1				
…		nan		nan
…				
…		nan		nan
…				
Patient N				

Figure 2.7 *Format of assignment of nan values for types of images T1WCont and ADC.*

The way to impute nan values is under the following procedure:

- Define the percentage of missing data. For this case, the defined percentages are 1, 5, 10, 20, 30, 40 and 50%.
- Starting from 1%, randomly select the observations and replace the numeric values with nan in the type of images T1wCont and ADC, row wise, i.e. there are 536 nan assignment by each image type.
- Following with 5%, keep the assigned nan imputation of the previous step and increase the missing data level until complete the desired percentage (5 % in this case).
- Repeat the process up to obtain 50% missing data.
- Perform and repeat all the procedure 10 times in order to have 10 iterations.

2.8 Missing data imputation

At this point, there are 7 different tables with missing data (one for each level of missingness) repeated 10 times (iterations), i.e. 70 tables. Knowing that the statistical methods of imputation were applied for missing information of the database. Two types of dataset were processed, one is the data properly obtained after missing imputation and the other that consists in a pre-processing step. A process of normalization is carried out to each table in order to have a better distribution of the information.

A series of missing data imputation are used to each table in order to fill the nan values properly. The used methods are listed below and were previously explained in the Subsection 1.5.2:

- KNN imputation
- Simple method by mean imputation
- Simple method by median imputation
- Simple method by random imputation
- Singular Value Thresholding for Nuclear Norm Optimization
- Generalized Spectral Regularization (Hard imputation)

Each method was used to impute each dataset. In other words, there were 6 output tables for each dataset. The implementation of these algorithms were developed in a free software tool called 'R 3.6.2' [39] with the *filling* package [40]. It is necessary to report each dataset that was normalized by z-score method prior to the missing data imputation, was denormalized immediately after. It is to say that the results obtained from imputation for normalized data must come back to its original scale in order to be compared with the true or original scale.

2.9 Evaluation of the imputation methods

In this part, a statistical process will be performed with the information obtained from the previous step. It was focused on two metrics:

- First, the normalized root mean square error (NRMSE) is computed for each method and each level of missingness.

In order to statistically compare the NRMSE from different methods we constructed a matrix with the values of the mean of NRMSE in 10 different rows and 6 columns representing the number of iteration and methods respectively (figure 2.8). Based on this matrix, the Friedman's test also was carried out in order to find whether there are significant differences amongst the errors in the different imputation methods.

Iteration	Method used					
	FSVT	Hard	K-NN	Mean	Median	Random
Iteration 1	NRMSE					
Iteration 2						
Iteration 3						
Iteration 4						
Iteration 5						
Iteration 6						
Iteration 7						
Iteration 8						
Iteration 9						
Iteration 10						

Figure 2.8 *Format of input to perform Friedman's test.*

A Friedman test was performed for each level of missingness. In case significant differences were detected, post-hoc comparisons with Tukey correction were performed to identify the pairs of significant different methods.

- Second, all the processes must have a further analysis that is related to the time of imputation. Our data was processed in a computer with these features:

- Computer processing unit: Intel (R) Core (TM) i5-7200U CPU @ 2.5 GHz 2.71GHz
- RAM: 8 GB

Chapter 3: Results and discussion

In this chapter, the results of the analysis are reported together with a brief discussion.

3.1 Introduction

The results of this chapters are expressed both using plots and statistical test. Here is a list with the acronyms of the evaluated methods:

- fSVT: Singular Value Thresholding for Nuclear Norm Optimization.
- hard: Generalized Spectral Regularization.
- knn: KNN imputation.
- mean: Simple method by mean imputation
- median: Simple method by median imputation
- random: Simple method by random imputation

Additionally, it is mentioned that in the legends of the figures the following syntaxis is used: "Method of imputation – Way of normalization – Method of normalization", to explain better this part, a brief description is listed:

- Method of imputation: The six ones mentioned above.
- Way of normalization: 'Column' for this study or 'none' if there is no normalization.
- Method of normalization: 'Z-score' or 'none' if there is no normalization.

The registration of error is divided in two groups, normalized and non-normalized, since in general both type of data may be used for radiomic studies and it is important to understand how the performance of statistical imputation is affected by normalization. To each of them, the Friedman's test was performed to all the considered percentages. The results are shown in the table 3.1:

Friedman's test	Percentage of missing data						
	1%	5%	10%	20%	30%	40%	50%
Non-normalized	2.06E-04	3.58E-07	7.53E-09	4.17E-09	4.90E-09	4.29E-09	3.75E-09
Normalized - Z scorre	1.27E-06	2.20E-08	2.79E-08	2.13E-09	1.77E-09	5.04E-09	2.13E-09

Table 3.1 *P-values of the Friedman test as a function, divided by level of missingness and normalization of the dataset.*

The results obtained shown that all the values are less than 0.05, so the null hypothesis is rejected, and it can be concluded that all the medians compared are not equal.

3.2 Missing data imputation: Original data

3.2.1 Imputation error

The boxplots related to non-normalized information are shown below, where the dependence of NRMSE on missing information quantity and on the different methods of imputation is visible. Figure 3.1 shows the range of error for each method according to level of missingness.

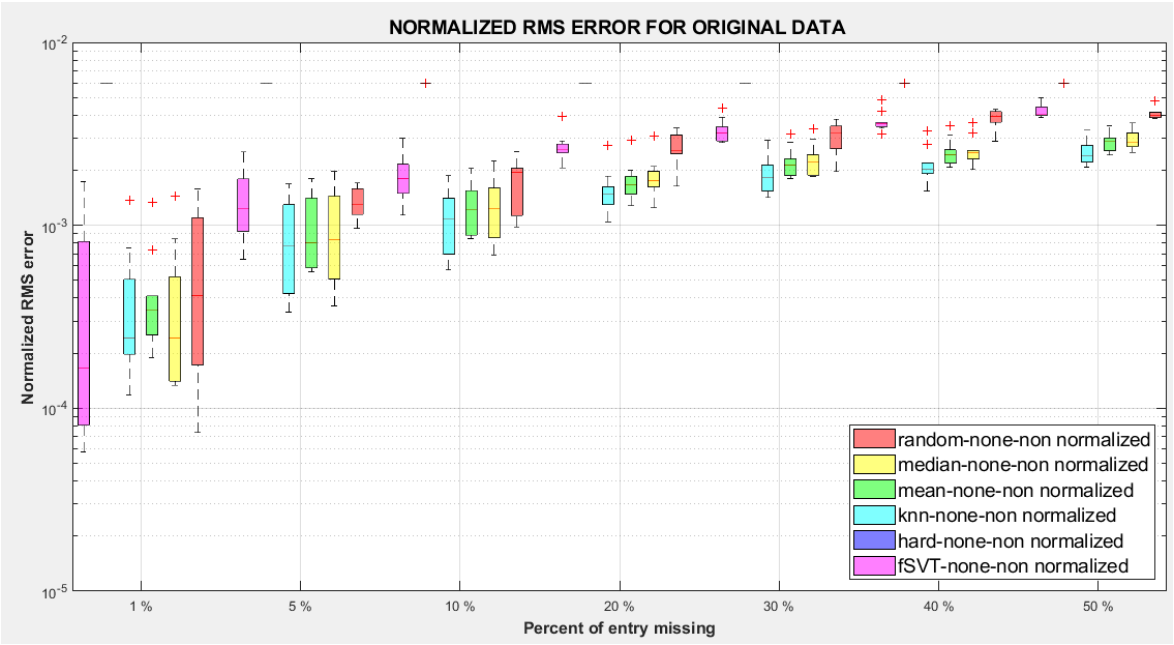

Figure 3.1 *Boxplots representing the normalized root mean square (RMS) error for non-normalized data for different combinations of missingness and imputation method.*

On the other hand, the mean of NRMSE from each method is displayed in the figure 3.2. On that, its evolution with respect to missing data percentage is displayed so that the performance comparison can be analysed. The data of this graph is shown in the table 3.2.

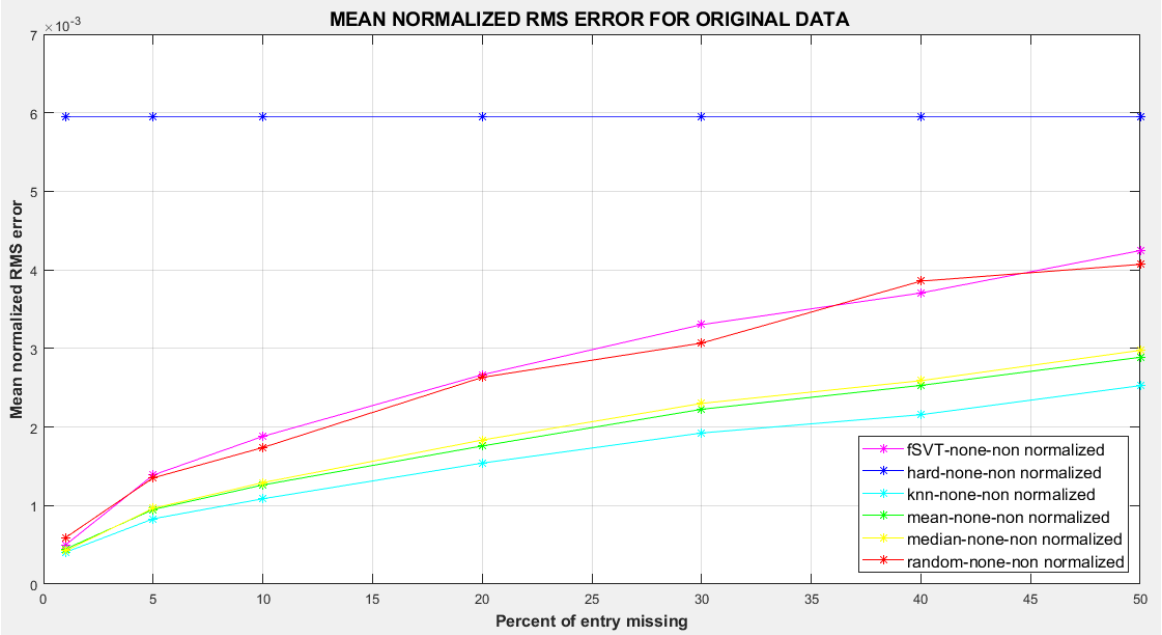

Figure 3.2 *Plots representing normalized root mean squared (RMS) error as a function of level of missingness for different levels of missingness and imputation method for non-normalized data.*

Imputation method	Percentage of missing data						
	1%	5%	10%	20%	30%	40%	50%
fSVT-non normalized	0.0005	0.0014	0.0019	0.0027	0.0033	0.0037	0.0042
hard-non normalized	0.0059	0.0059	0.0059	0.0059	0.0059	0.0059	0.0059
knn-non normalized	0.0004	0.0008	0.0011	0.0015	0.0019	0.0022	0.0025
mean-non normalized	0.0004	0.0009	0.0013	0.0018	0.0022	0.0025	0.0029
median-non normalized	0.0004	0.001	0.0013	0.0018	0.0023	0.0026	0.003
random-non normalized	0.0006	0.0014	0.0017	0.0026	0.0031	0.0039	0.0041

Table 3.2 *Level of normalized root mean squared error (NRMSE) as a function of the level of missingness and the imputation method, for the non-normalized dataset.*

3.2.2 Statistical comparison of imputation errors

The test post-hoc comparisons were performed to each percentage associated with missing data. Figure 3.3 shows the results for the multiple comparison in case of 1% of missing data. All the methods analysed can be superposed, except ‘hard’ method, which had a significantly higher NMRSE.

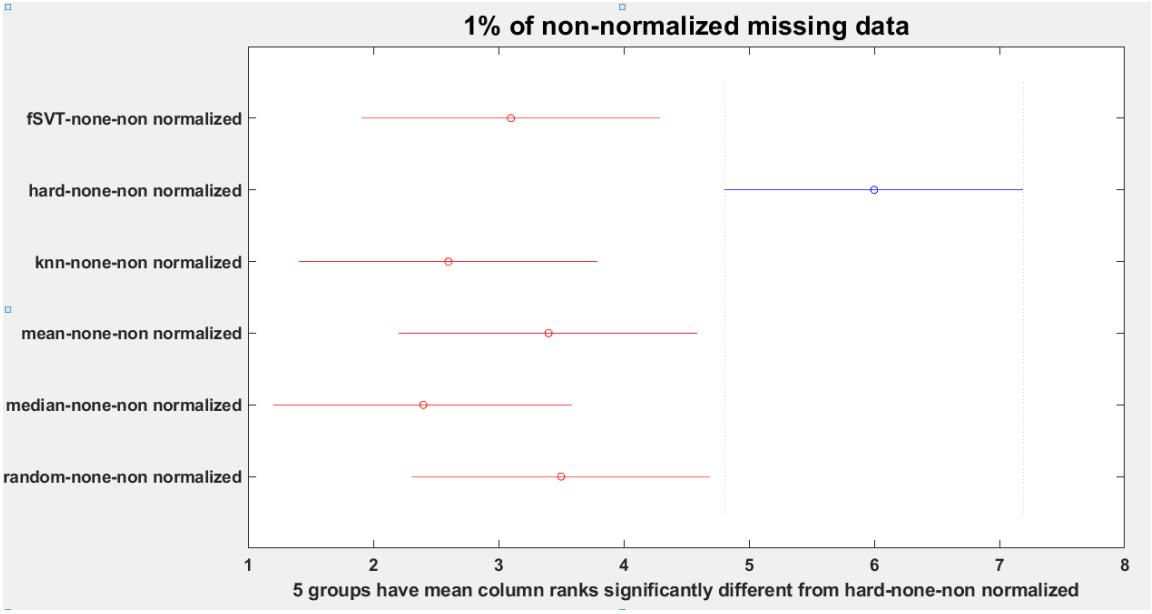

Figure 3.3 *Multicomparison analysis for 1% of non-normalized missing data.*

Figure 3.4 shows the results for the multiple comparison in case of 5% of missing data. In this case, 'knn' method provided the lowest error and it was significantly lower than 'hard' method, 'fSVT' and 'random' imputation, but not significantly lower than mean and median imputation.

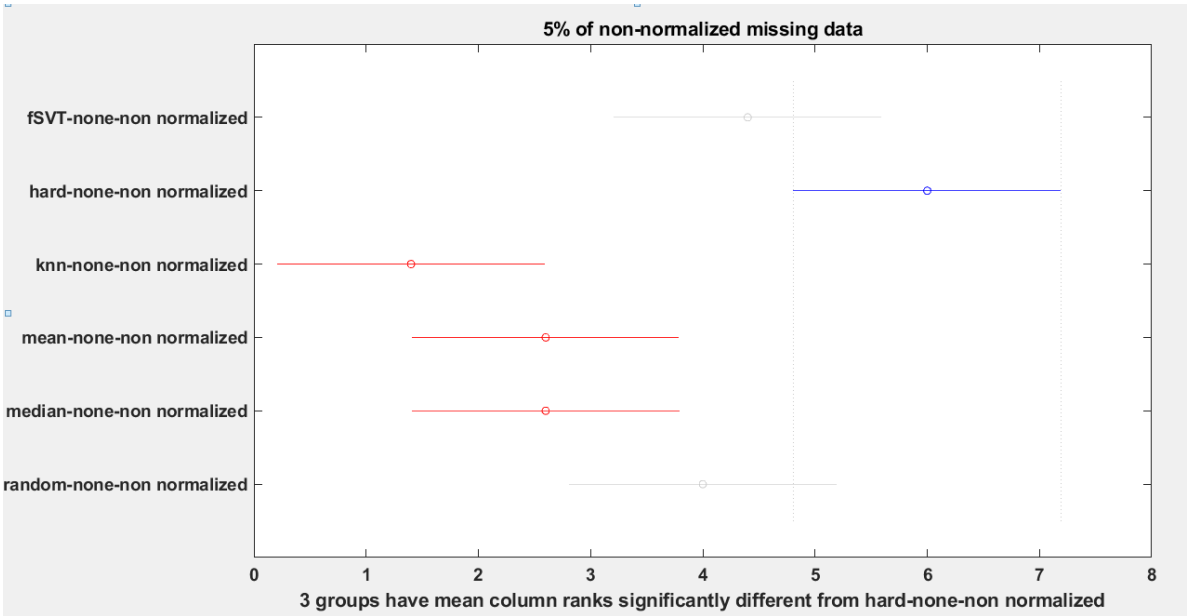

Figure 3.4 *Multicomparison analysis for 5% of non-normalized missing data.*

A similar pattent can be observed for all the other percentages of missing data (Figures 3.5-3.9): 'knn' was the method with the lower NRMSE, which was statistically lower than 'hard', 'fSVT' and 'random' imputation method.

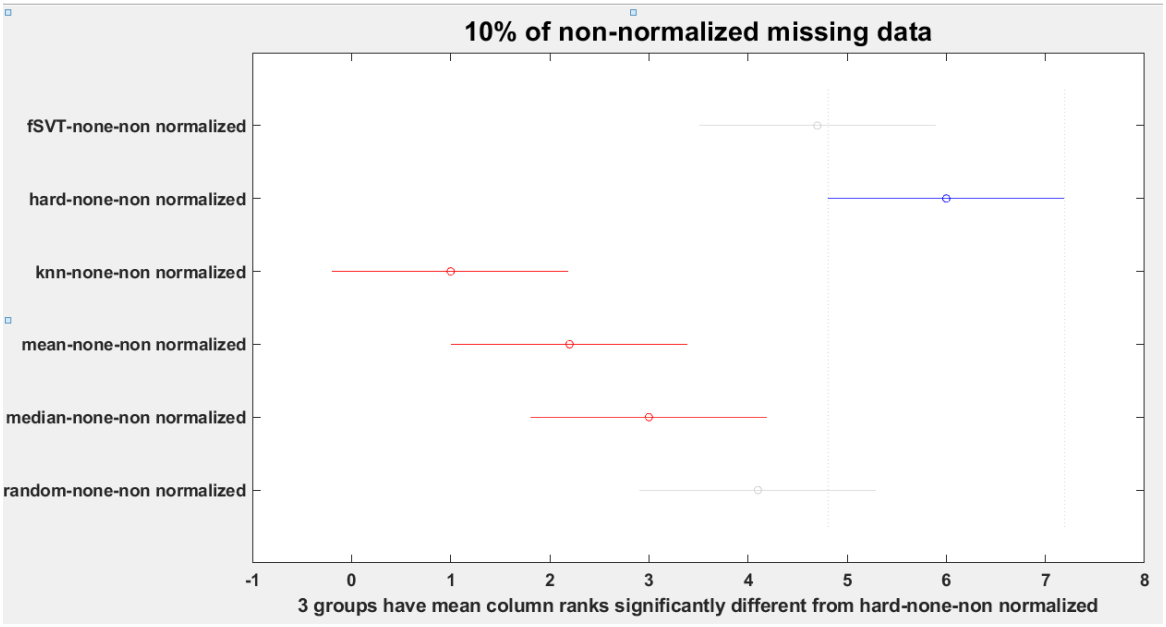

Figure 3.5 *Multicomparison analysis for 10% of non-normalized missing data.*

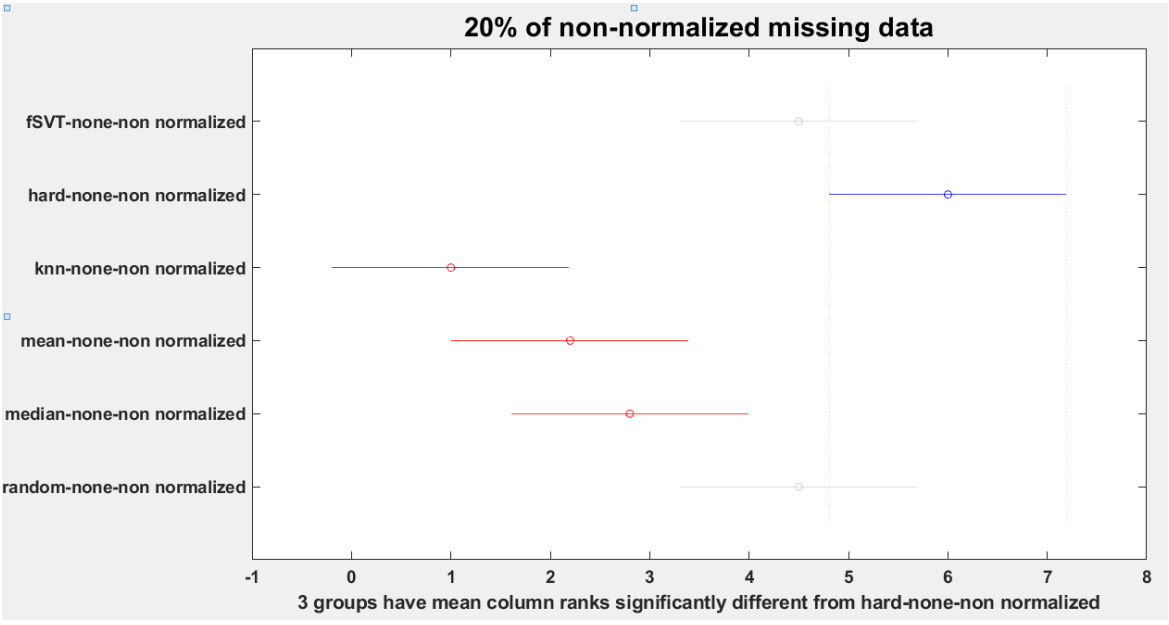

Figure 3.6 *Multicomparison analysis for 20% of non-normalized missing data.*

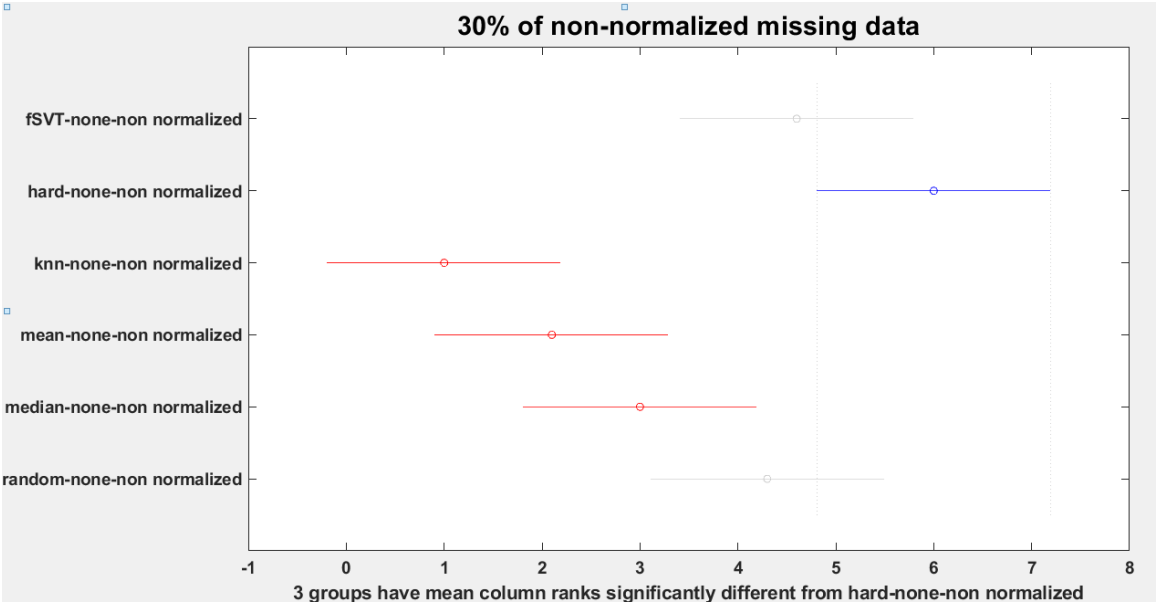

Figure 3.7 *Multicomparison analysis for 30% of non-normalized missing data.*

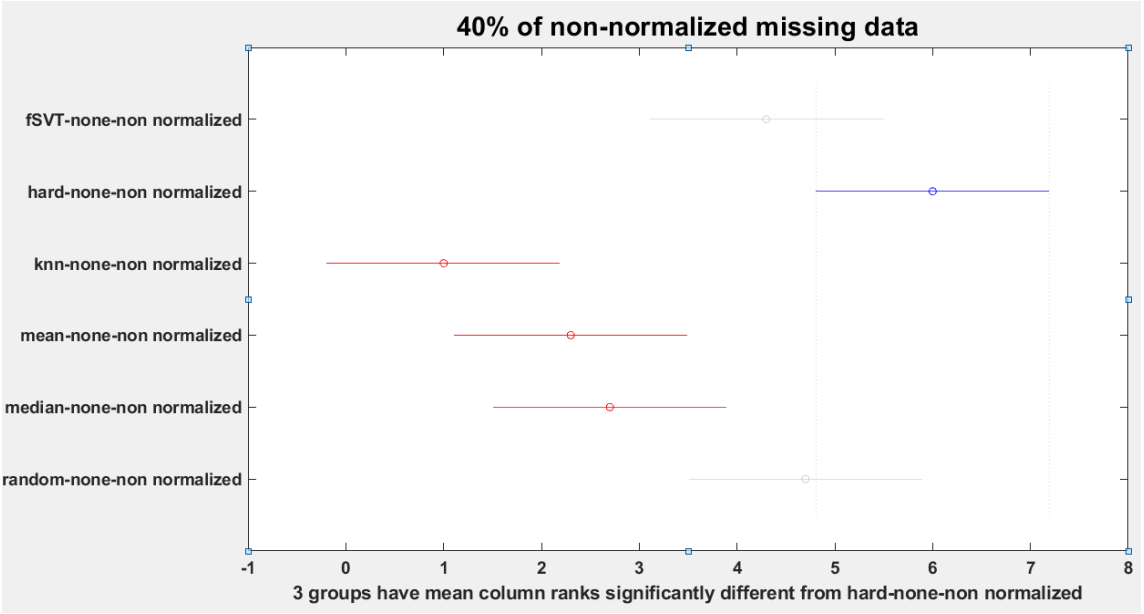

Figure 3.8 *Multicomparison analysis for 40% of non-normalized missing data.*

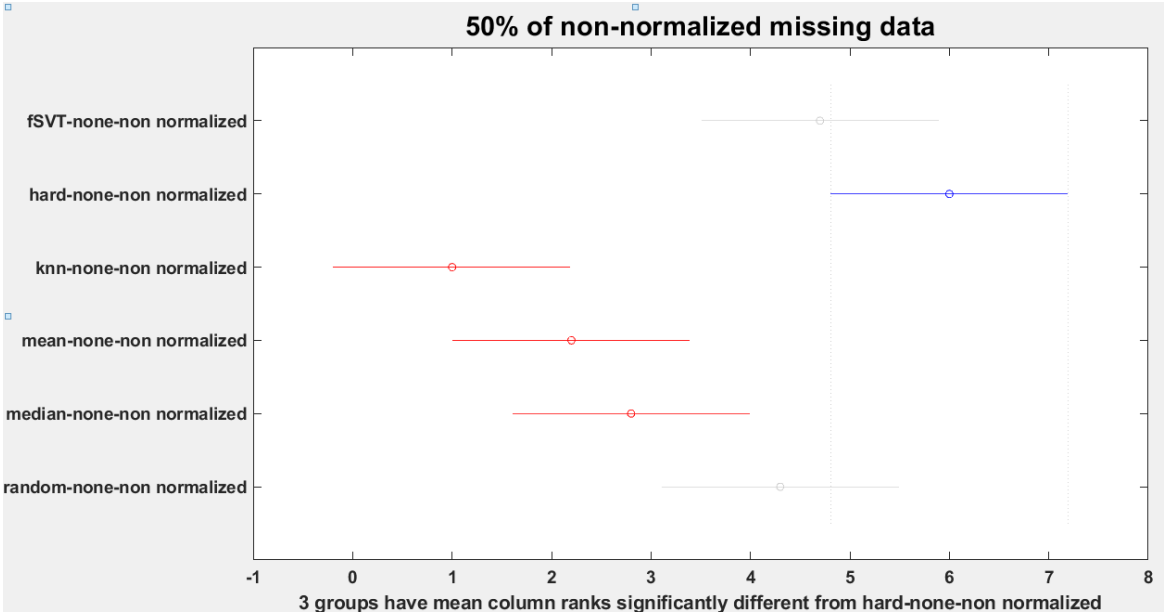

Figure 3.9 *Multicomparison analysis for 50% of non-normalized missing data.*

3.2.3 Time of imputation

The figure 3.10 shows the dependence of time imputation to each level of missingness. Clearly, it is seen that there is no dependence except for knn-imputation. The other methods display a trend regardless the quantity of missing data.

The table 3.3 show numerically the needed time to impute data, this table is related to the figure 3.2, wherein we can appreciate that all the methods, except for 'knn', do not have dependence with the level of missingness.

Method	percent_1	percent_5	percent_10	percent_20	percent_30	percent_40	percent_50
'random_imputation'	12.514	11.784	12.577	12.473	11.752	11.993	11.589
'median_imputation'	11.89	11.937	11.645	11.852	12.551	11.708	11.704
'mean_imputation'	11.89	11.681	11.653	11.137	11.266	11.27	11.531
'knn_imputation'	17.069	76.197	170.17	357.69	560.8	728.63	900.59
'hard_imputation'	39.996	41.211	40.276	40.1	40.54	41.376	41.136
'fSVT_imputation'	189.83	188.78	187.58	186.97	187.7	186.74	189.36

Table 3.3 *Times of imputation for the different imputation methods and the different levels of missingness.*

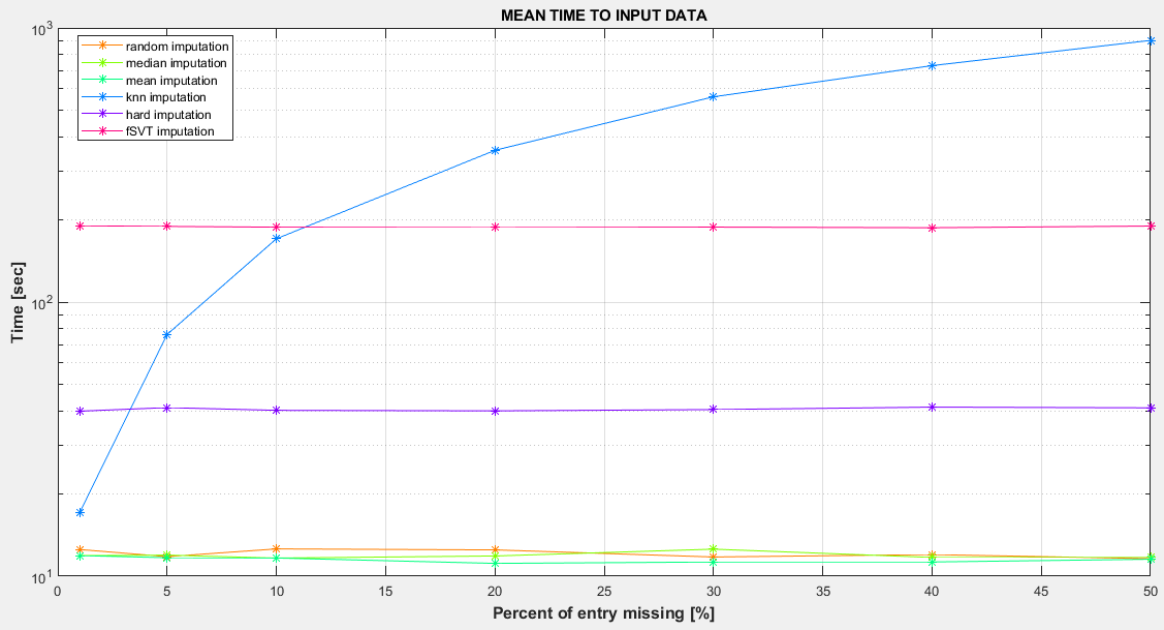

Figure 3.10 *Graphical representation of time to impute missing data for each method.*

3.3 Missing data imputation: Normalized data

3.3.1 Imputation error

The boxplots related to normalized data are shown in figure 3.11-3.12 and in table 3.4. It is possible to see that there has been a shift in performance for some methods. For example, 'knn' is no more the best performing method. The increasing trend of imputation error with missing data percentage is observed also for the normalized data.

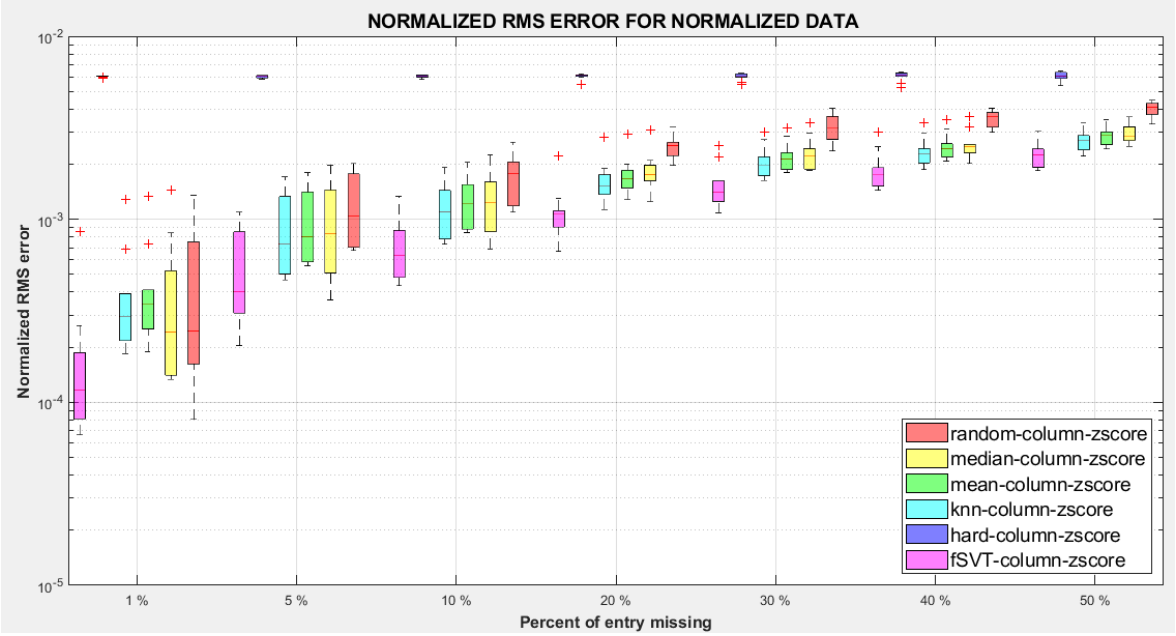

Figure 3.11 *Boxplots representing the normalized root mean square (RMS) error for normalized data for different combinations of missingness and imputation method.*

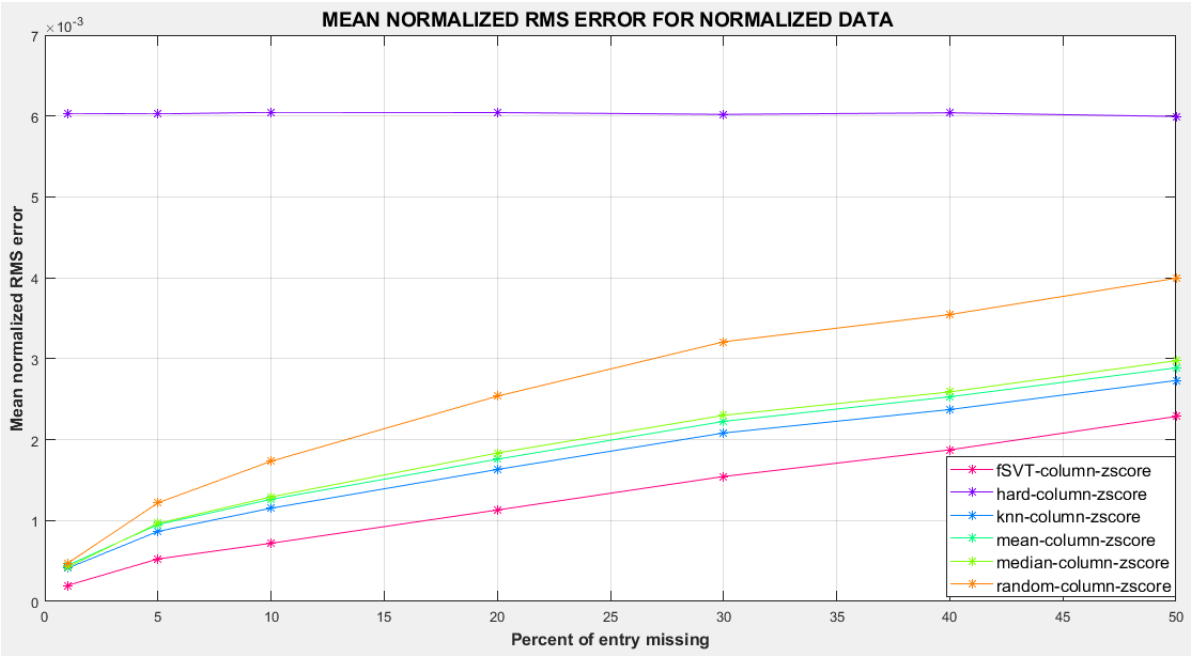

Figure 3.12 *Plots representing normalized root mean squared (RMS) error as a function of level of missingness for different levels of missingness and imputation method for normalized data.*

Imputation method	Percentage of missing data						
	1%	5%	10%	20%	30%	40%	50%
fSVT-column-zscore	0.0002	0.0005	0.0007	0.0011	0.0015	0.0019	0.0023
hard-column-zscore	0.006	0.006	0.006	0.006	0.006	0.006	0.006
knn-column-zscore	0.0004	0.0009	0.0012	0.0016	0.0021	0.0024	0.0027
mean-column-zscore	0.0004	0.0009	0.0013	0.0018	0.0022	0.0025	0.0029
median-column-zscore	0.0004	0.001	0.0013	0.0018	0.0023	0.0026	0.003
random-column-zscore	0.0005	0.0012	0.0017	0.0025	0.0032	0.0035	0.004

Table 3.4 *Level of normalized root mean squared error (NRMSE) as a function of the level of missingness and the imputation method, for the normalized dataset.*

3.3.2 Statistical comparison of imputation error

The test post-hoc comparisons were performed to each percentage associated with missing data. Figure 3.13 shows the results for the multiple comparison in case of 1% of missing data. The method with the lowest error was 'fSVT', which had significantly lower error compared to 'mean' and 'hard' methods.

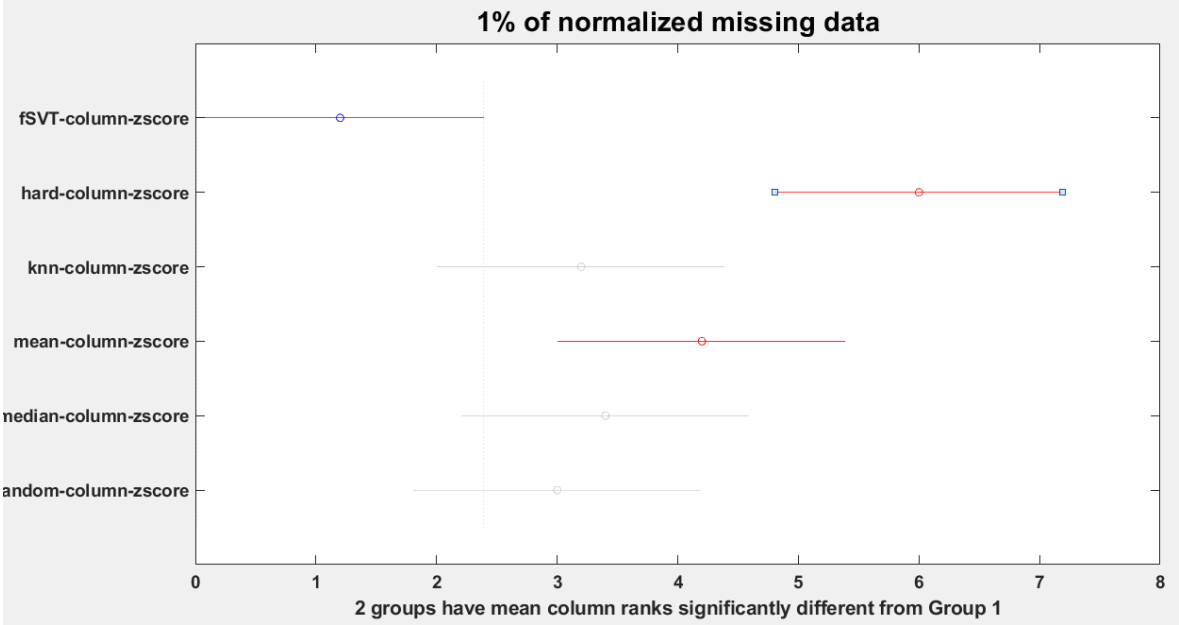

Figure 3.13 *Multicomparison analysis for 1% of normalized missing data.*

When the percentage of missing data was 5% (figure 3.14) the 'fSVT' method was still the best, with significantly lower NMRSE compared to all the other methods except 'knn'. This trend was observed also for 10% and 40% level of missingness (Figure 3.15 and Figure

3.18). For missingness level 20%, 30% and 50%, the NMRSE of fSVT was significantly lower than all the methods except ‘mean’ and ‘knn’ (Figure 3.16-3.17 and Figure 3.19).

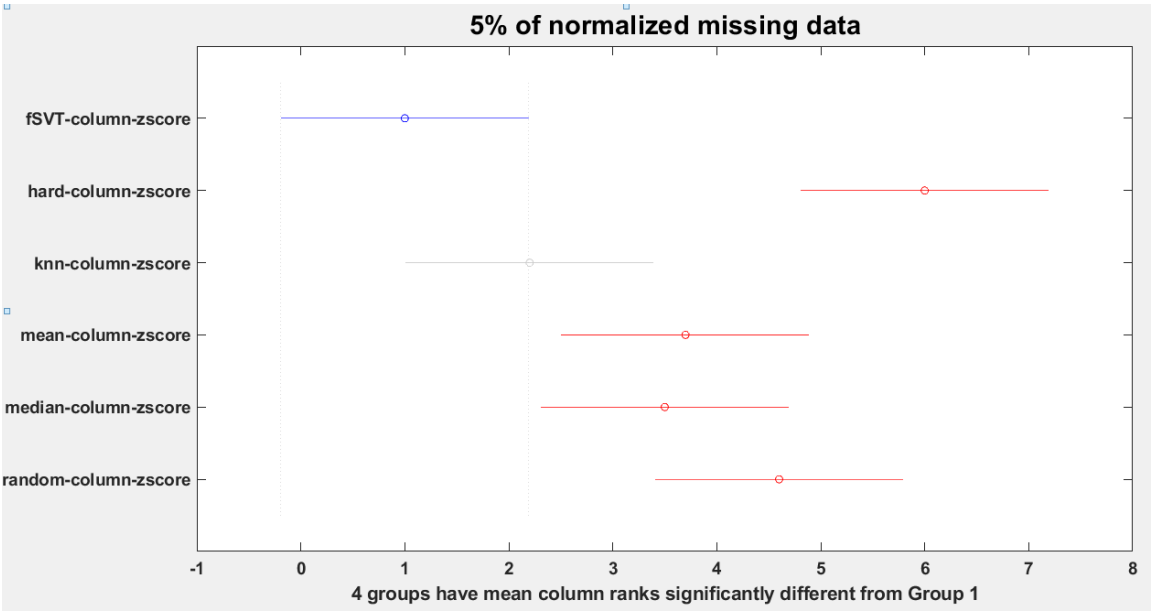

Figure 3.14 *Multicomparison analysis for 5% of normalized missing data.*

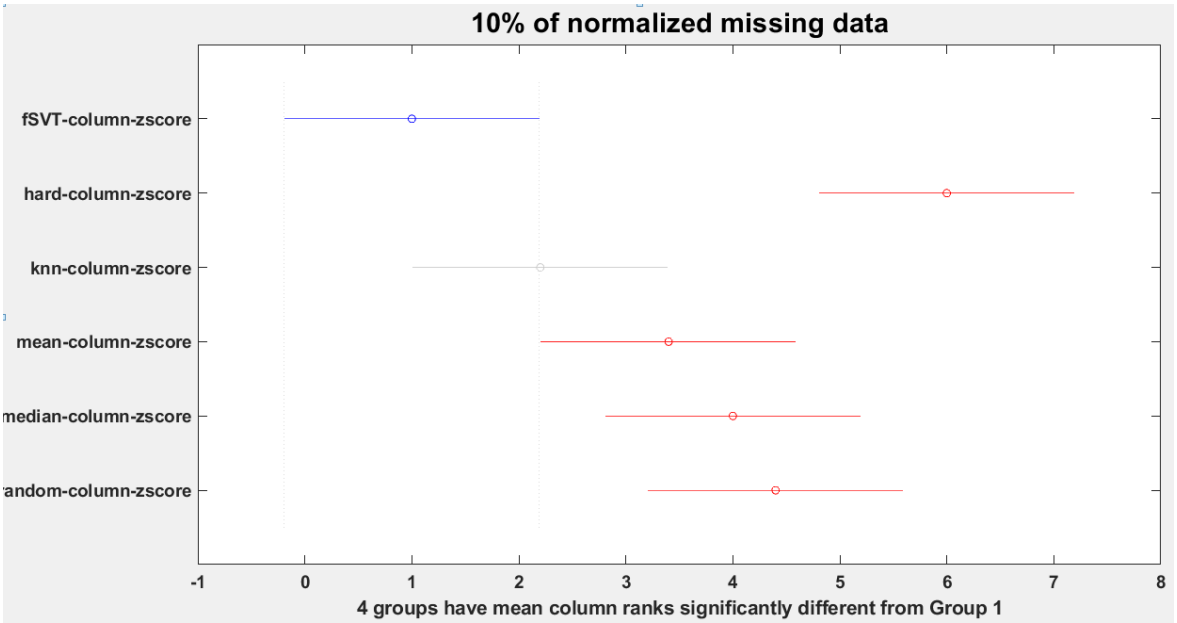

Figure 3.15 *Multicomparison analysis for 10% of normalized missing data.*

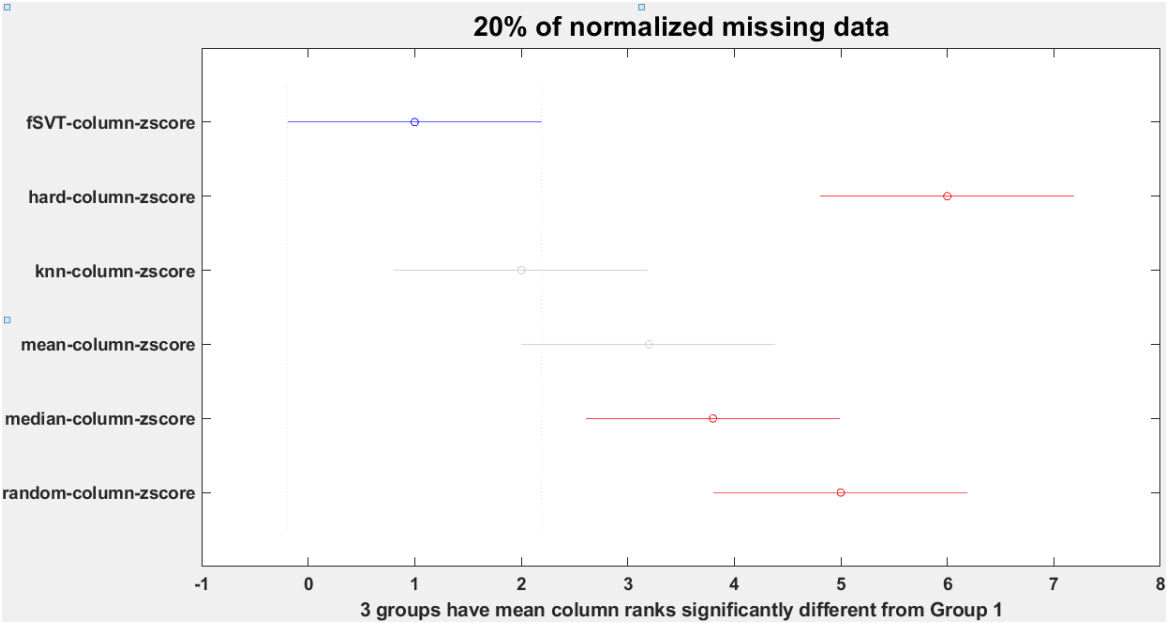

Figure 3.16 *Multicomparison analysis for 20% of normalized missing data.*

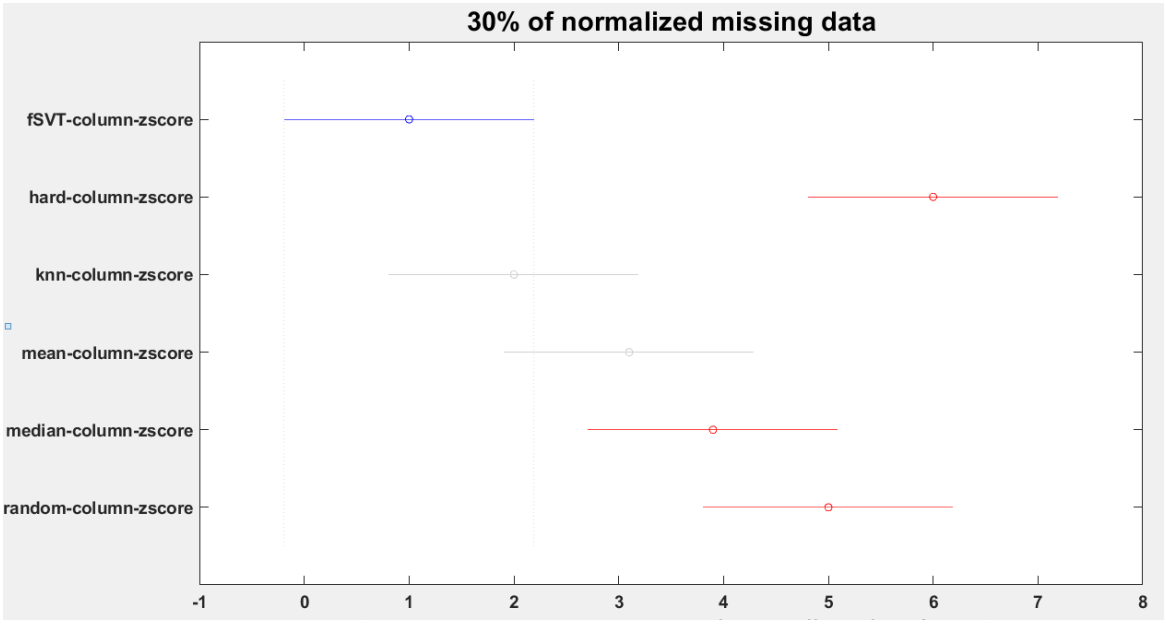

Figure 3.17 *Multicomparison analysis for 30% of normalized missing data.*

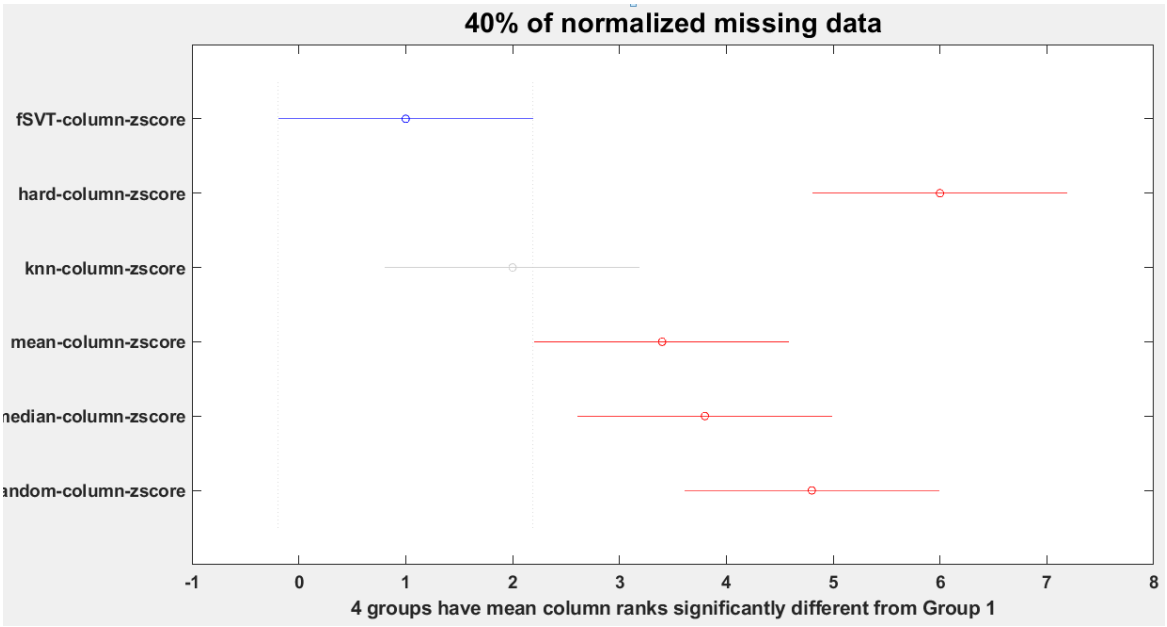

Figure 3.18 *Multicomparison analysis for 40% of normalized missing data.*

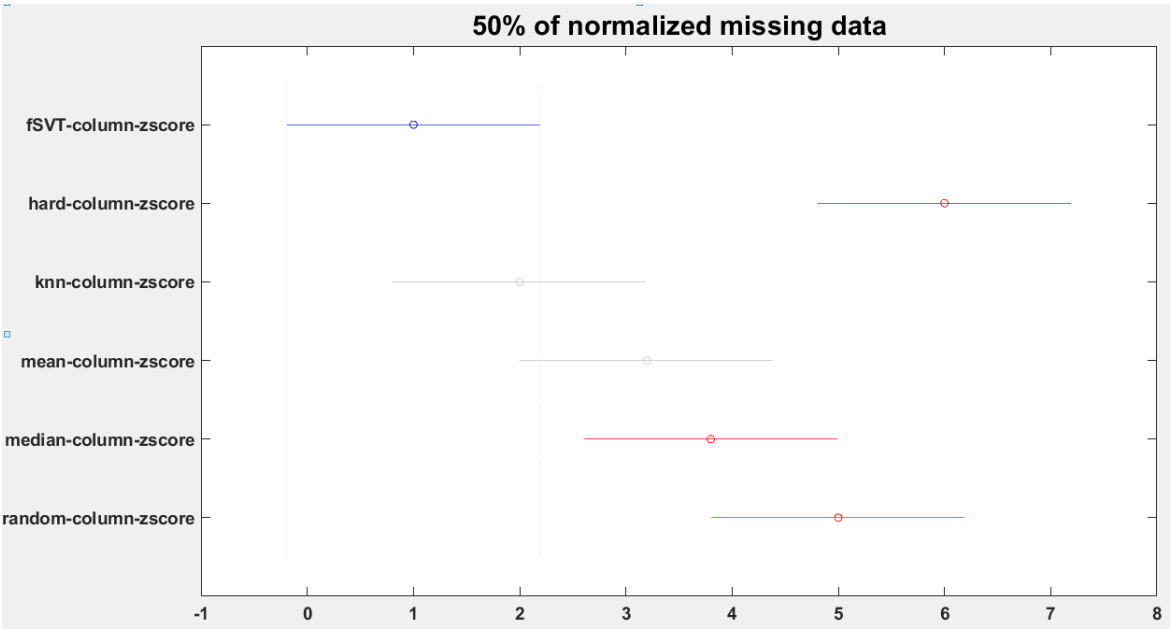

Figure 3.19 *Multicomparison analysis for 50% of normalized missing data.*

3.3.3 Time of imputation

The normalization does not impact the time of imputation, which depends only on the percentage of missing data. Therefore, the results in terms of time of imputation for the normalized dataset are the same one displayed in figure 3.10 and table 3.3.

3.4 Discussion

It could be seen that using radiomics is possible to extract features from the MRI, but in case of missing data it is necessary to use appropriate algorithms to solve that issue. The lack of features from T1wCont and ADC images was simulated by random imputation of missing data. Then, different statistical imputation methods were used to fill those missing data. There were six methods used to complete data, and they were all analyzed for different level of missingness. As results, it was observed that the non-normalized information has a wider range of error in comparison to normalized. This once more highlights the importance of data imputation as an immediate postprocessing step for the radiomic features. As for the performance of each method, it is seen that the 'hard' approach has the same level of error in all the levels of missingness, however in its normalized version, it has a slight lower performance; the same occurs with 'knn' in terms of performance. The 'mean' and 'median' have quantitative the same efficiency regardless the condition of normalization. The random approach has slightly higher level of error than 'mean' and 'median', however it is numerically observed the reduction of it when applying a normalized process.

In terms of accuracy, the best methods for reducing the imputation error are the 'fSVT' method for the normalized dataset and the 'knn' method for the non-normalized dataset. The difference in the ranking of the methods may be explained by the fact that the normalization changes the way that distances are computed in the knn and consequently the nearest neighbour that are used for the imputation, leading to worse results.

The comparison of methods must be assessed not only from a quantitative point of view, but also by looking at the time of imputation. It can be noted that the simple methods (random, mean and media) are the fasters, and that they are also independent from the level of missingness, followed by 'hard' and 'fSVT' approach. It was also noted that time of imputation for 'knn' imputation increased as a squared root of the percentage of missing data.

In conclusion, when dealing with missing data imputation for radiomic application, one has to think about the priority (accuracy over speed or vice versa), the normalization of the data and the level of missingness. If the focus is speed, the simple methods are the best and they are also the best accuracy/speed trade-off. However, if the focus is pure accuracy, more complex methods like 'fSVT' or 'knn' should be considered (the latter especially if the percentage of missing data is low).

Chapter 4: Conclusion and future developments

This last part contains the main conclusion obtained from study. Limitations of the study and possible future developments to overcome them are also described.

In this study, 6 different methods for missing data imputation were compared on both original and normalized datasets of radiomic features. The main results of this study were the following:

- All the methods had a level of NRMS error below than 0.006.
- The process of normalization improves slightly the results making them be more accurate, except for 'fSVT' and 'knn' methods.
- For the non-normalized dataset, the best methods, ranked by accuracy, were the following: knn, mean, median, fSVT, random and hard imputation.
- For the normalized dataset, the best methods, ranked by accuracy, were the following: fSVT, knn, mean median, random and hard imputation
- The level of missingness information affects directly to the performance of each method: imputation error grew as a squared root of the level of missingness.
- In terms of time, for all the methods except for 'knn', the percentage of missing data is not relevant.
- The methods consider from the fastest to the slowest are (independently on the normalization): mean and mean, random, hard, fSVT.
- There is no best overall method: 'knn' and 'fSVT' are the best in terms of pure accuracy but the time of imputation is higher; simple methods, like mean or median imputation, are the best trade off between accuracy and time.

The limitation of the study is the fact that, although the error was quantified for each method and it is possible to say which method performs best, it is not possible to see what is the largest percentage of error acceptable to avoid problems in the following clinical analysis. In order to understand that, it is necessary to design ad-hoc experiments in which the radiomic features are used to train predictive and prognostic models and in which the performance is evaluated as a function of the initial level of missing data.

To sum up, different missing data imputation methods were used and analysed to perform a comparison in terms of imputation error and time of imputation. Hopefully this study will help to reduce problems due to the lack of some MRI sequences given that there are not standardized acquisition protocols at hospitals.

References

[1] American Cancer Society, “About Nasopharyngeal Cancer What Is Nasopharyngeal Cancer ?,” pp. 1–9, 2018.

[2] B. Brennan, “Nasopharyngeal carcinoma,” *Orphanet Journal of Rare Diseases*. 2006, doi: 10.1186/1750-1172-1-23.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA. Cancer J. Clin.*, 2018, doi: 10.3322/caac.21492.

[4] M. Pastor *et al.*, “SEOM clinical guideline in nasopharynx cancer (2017),” *Clin. Transl. Oncol.*, 2018, doi: 10.1007/s12094-017-1777-0.

[5] M. Bologna, “MRI- BASED RADIOMIC ANALYSIS OF RARE TUMORS : OPTIMIZATION OF A WORKFLOW FOR RETROSPECTIVE AND MULTICENTRIC STUDIES,” 2019.

[6] D. M. Gress *et al.*, “AJCC Cancer Staging Manual,” *AJCC Cancer Staging Man.*, 2017, doi: 10.1007/978-3-319-40618-3.

[7] H. B. Cobanoglu and S. Arslan, “Nasopharyngeal cancer: an update on diagnosis and treatment,” *Mucosa*, 2019, doi: 10.33204/mucosa.540845.

[8] A. Taylor and M. E. B. Powell, “Intensity-modulated radiotherapy - What is it?,” *Cancer Imaging*, vol. 4, no. 2, pp. 68–73, 2004, doi: 10.1102/1470-7330.2004.0003.

[9] B. Cho, “Intensity-modulated radiation therapy: A review with a physics perspective,” *Radiation Oncology Journal*. 2018, doi: 10.3857/roj.2018.00122.

[10] “Immunotherapy for Cancer - National Cancer Institute.” [Online]. Available: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy. [Accessed: 29-Mar-2020].

[11] “What is Immunotherapy - Cancer Research Institute (CRI).” [Online]. Available: https://www.cancerresearch.org/immunotherapy/what-is-immunotherapy. [Accessed: 11-Apr-2020].

[12] A. Berger, “Magnetic resonance imaging,” *BMJ*, 2002, doi: 10.1136/bmj.324.7328.35.

[13] D. Pollaco, “Magnetic resonance imaging.,” *Univ. Malta*, vol. 1, no. 1, pp. 93–116, 2016.

[14] R. A. Pooley, “AAPM/RSNA physics tutorial for residents: fundamental physics of MR imaging.,” *Radiographics*, 2005, doi: 10.1148/rg.254055027.

[15] K. Möllenhoff, A. M. Oros-Peusquens, and N. J. Shah, “Introduction to the basics of magnetic resonance imaging,” *Neuromethods*, 2012, doi: 10.1007/7657-2012-56.

[16] “RELAXATION.png (859×586).” [Online]. Available: https://mrimaster.com/images/RELAXATION.png. [Accessed: 12-Apr-2020].

[17] “Magnetic Resonance Imaging.” [Online]. Available: http://www.sprawls.org/mripmt/MRI06/index.html. [Accessed: 13-Apr-2020].

[18] “Magnetic Resonance Imaging.” [Online]. Available: http://www.sprawls.org/mripmt/MRI06/index.html. [Accessed: 12-Apr-2020].

[19] “MRI.” [Online]. Available: http://casemed.case.edu/clerkships/neurology/NeurLrngObjectives/MRI.htm. [Accessed: 06-Apr-2020].

[20] D. M. Koh and D. J. Collins, “Diffusion-weighted MRI in the body: Applications and challenges in oncology,” *Am. J. Roentgenol.*, 2007, doi: 10.2214/AJR.06.1403.

[21] W. R. Nitz and P. Reimer, “Contrast mechanisms in MR imaging,” *European Radiology*. 1999, doi: 10.1007/s003300050789.

[22] R. Bitar *et al.*, “MR pulse sequences: What every radiologist wants to know but is afraid to ask,” *Radiographics*. 2006, doi: 10.1148/rg.262055063.

[23] Z. Liu *et al.*, “The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges,” *Theranostics*. 2019, doi: 10.7150/thno.30309.

[24] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, “From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities,” *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 132–160, 2019, doi: 10.1109/MSP.2019.2900993.

[25] V. Parekh and M. A. Jacobs, “Radiomics: a new application from established techniques,” *Expert Review of Precision Medicine and Drug Development*. 2016, doi: 10.1080/23808993.2016.1164013.

[26] “State-of-the-art in radiomics of hepatocellular carcinoma: a review of basic principles, applications, and limitations. - PubMed - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31707435. [Accessed: 30-Mar-2020].

[27] L. Zhang *et al.*, “Radiomic nomogram: Pretreatment evaluation of local recurrence in nasopharyngeal carcinoma based on MR imaging,” *J. Cancer*, 2019, doi: 10.7150/jca.33345.

[28] B. Zhang *et al.*, “Advanced nasopharyngeal carcinoma: pre-treatment prediction of progression based on multi-parametric MRI radiomics,” *Oncotarget*, 2017, doi: 10.18632/oncotarget.19799.

[29] O. Troyanskaya *et al.*, “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, 2001, doi: 10.1093/bioinformatics/17.6.520.

[30] D. Adams and D. Beckett, “Summit ‘ 97 Normalization Is a Nice Theory,” 1997.

[31] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, 2009, doi: 10.1007/s10208-009-9045-5.

[32] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” *J. Mach. Learn. Res.*, 2010.

[33] J. F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, 2010, doi: 10.1137/080738970.

[34] S. Lee and D. K. Lee, “What is the proper way to apply the multiple comparison test?,” *Korean J. Anesthesiol.*, 2018, doi: 10.4097/kja.d.18.00242.

[35] “Radiomic Features — pyradiomics v3.0.post2+g896682d documentation.” [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/features.html#radiomics.firstorder.RadiomicsFirstOrder. [Accessed: 04-Mar-2020].

[36] “GitHub - Radiomics/pyradiomics: Open-source python package for the extraction of Radiomics features from 2D and 3D images and binary masks.” [Online]. Available: https://github.com/Radiomics/pyradiomics. [Accessed: 03-Mar-2020].

[37] “R2018a - MATLAB & Simulink - MathWorks Italia.” [Online]. Available: https://it.mathworks.com/help/parallel-computing/release-notes-R2018a.html;jsessionid=069d763d4756f8222d97cc8f45c5. [Accessed: 03-Mar-2020].

[38] “Standardized z-scores - MATLAB zscore - MathWorks Italia.” [Online]. Available: https://it.mathworks.com/help/stats/zscore.html. [Accessed: 05-Mar-2020].

[39] “Previous releases of R for Windows.” [Online]. Available: https://cran.r-project.org/bin/windows/base/old/. [Accessed: 13-Apr-2020].

[40] “CRAN - Package filling.” [Online]. Available: https://cran.r-project.org/web/packages/filling/index.html. [Accessed: 13-Apr-2020].