FDG-PET/MR Imaging for Prediction of Lung Metastases in Soft-Tissue Sarcomas of the Extremities by Texture Analysis and Wavelet Image Fusion

Martin Carrier-Vallières

Master of Science

Medical Physics Unit

McGill University

Montreal, Québec, Canada

December 2012

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Medical Radiation Physics

ABSTRACT

Soft-tissue sarcoma (STS) of the extremities forms a relatively uncommon yet aggressive group of neoplasms with high metastatic risk of the disease. The vast majority of STS metastases occur in the lungs. Due to the general poor prognosis of patients diagnosed with STS lung metastases, there is a clinical need to identify relevant prognostic factors as early as possible in the course of staging and treatment management. Recent evidence suggests that positron emission tomography (PET) using fluorodeoxyglucose (FDG) and magnetic resonance (MR) imaging texture features have the potential to predict the outcome of tumours through the assessment of their microenvironment heterogeneity characteristics. The goal of this work is therefore to investigate FDG-PET and MR texture features as potential early predictors of lung metastasis risk in STS cancer of the extremities.

In this study, a dataset of 35 patients with histologically proven STS of the extremities was retrospectively analyzed. All patients received pre-treatment FDG-PET and MR scans. MR imaging data comprised of T_1-weighted, T_2 fat-saturation (T2FS) and short tau inversion recovery (STIR) sequences. The median follow-up period was 29 months (range: 4 to 85 months). Thirteen patients from the dataset developed lung metastases. Six texture features from the gray-level co-occurrence matrix (GLCM) were extracted from the FDG-PET, MR and fused FDG-PET/MR scans. In addition, the maximum standard uptake value ($SUV_{\max}$) of the tumours was included in the feature set. The fusion of FDG-PET and MR scans was carried out using the discrete wavelet transform (DWT) and a band-pass frequencies enhancement technique. Statistical analysis was performed using Spearman's correlation (*rho*), and multivariable modeling using logistic regression. The prediction performance of the different multivariable models was assessed using bootstrap resampling by calculating the area under the receiver-operating characteristics curve (AUC) and Matthews' correlation coefficient (MCC).

The highest univariate prediction of lung metastases was attributed to the SUV_{max} metric ($rho = 0.6382,\ p < 0.0001$). Most texture features extracted from fused scans had higher Spearman's correlation with lung metastases than those extracted from separate scans. On separate scans, FDG-PET texture features were generally dominant over MR texture features. The highest multivariable prediction of lung metastases was found using fused scans and the following 4-parameters model: $0.94*SUV_{max} - 0.401*$*PET-T2FS/STIR--Variance* $- 6.7*$*PET-T1--Contrast* $- 165*$*PET-T1--Homogeneity* $+ 140$. This model reached $rho = 0.8255,\ p < 0.0001$ on the entire dataset and $\text{AUC} = 0.956 \pm 0.002$, $\text{MCC} = 0.829 \pm 0.002$ in bootstrap testing sets.

Overall, this work indicates the strong potential of FDG-PET and MR texture features for the prediction of lung metastases in STS cancer of the extremities. Substantial prediction improvements were found using texture features from fused scans and multivariable modeling strategies compared to texture features extracted from separate scans and univariate analysis. Potentially, this could improve patient outcomes by allowing better personalization of treatments and the application of pre-emptive strategies to mitigate disease spread.

RÉSUMÉ

Les sarcomes des tissus mous (STM) provenant des extrémités forment un groupe relativement rare de néoplasme avec un risque métastatique élevé. La grande majorité des métastases provenant des STM ont lieu dans les poumons, et le pronostique résultant est généralement faible. En ce sens, il est important d'identifier autant de facteurs pronostiques pertinents que possible au moment du diagnostique et de la gestion du traitement. Certains travaux récents ont permis de démontrer que les caractéristiques texturales d'images provenant de la tomographie par émission de positrons (TEP) utilisant le fluorodéoxyglucose (FDG) et l'imagerie par résonance magnétique (IRM) ont le potentiel de prédire l'évolution tumorale grâce à l'évaluation des propriétés d'hétérogénéité biologique des tumeurs. Donc, le but de ce travail est d'évaluer le potentiel des caractéristiques texturales d'images FDG-TEP et IRM en tant que prédicteur du risque de métastases aux poumons pour le cancer des STM provenant des extrémités.

Dans cette étude, une cohorte de 35 patients diagnostiqués avec des STM aux extrémités a été rétrospectivement analysée. Tous les patients ont reçu un scan FDG-TEP et un scan IRM avant leur traitement. Les séquences IRM qui ont été utilisés dans l'analyse sont: T_1, T_2 par saturation des gras (T2FS) et STIR. Les patients ont été suivis sur une période médiane de 29 mois (intervalle: 4 à 85 mois). Treize patients de la cohorte ont développé des métastases aux poumons. Six caractéristiques texturales d'images provenant de la matrice de co-occurrence des niveaux de gris (GLCM) ont été extraites des scans FDG-PET, IRM et FDG-PET/IRM fusionnés. De plus, la valeur maximale de consommation standard des tumeurs ($SUV_{\max}$) a été incluse dans l'analyse. La fusion des scans a été effectuée grâce à la transformée d'ondelettes discrètes et grâce à une technique de renforcement des fréquences passe-bandes. L'analyse statistique a été effectuée en utilisant la corrélation de Spearman (*rho*), et l'analyse multivariable en utilisant la régression logistique. Les performances de prédiction des différents modèles multivariables ont été évaluées en calculant 2 métriques à partir de la technique de ré-échantillonnage « bootstrap »: L'aire sous la courbe de fonctionnement (AUC) et le coefficient de corrélation de Matthews (MCC).

La plus haute prédiction univariée est attribuée à $SUV_{\max}$ ($rho = 0.6382$, $p < 0.0001$). La plupart des caractéristiques texturales extraites des scans fusionnés possèdent des coefficients de corrélation Spearman plus haut que celles extraites des scans séparés. Dans le cas des scans séparés, les caractéristiques texturales provenant de FDG-TEP sont généralement dominantes par rapport à celles provenant des scans IRM. La plus haute prédiction multivariable est provenue des scans fusionnés avec le model suivant: $0.94*SUV_{\max} - 0.401*$*PET-T2FS/STIR--Variance* $- 6.7*$*PET-T1--Contrast* $- 165*$*PET-T1--Homogeneity* $+ 140$. Ce model a atteint des résultats de $rho = 0.8255$, $p < 0.0001$ sur l'ensemble des patients et $\text{AUC} = 0.956 \pm 0.002$, $\text{MCC} = 0.829 \pm 0.002$ sur les ensembles de tests « bootstrap ».

De façon générale, cette étude indique le fort potentiel des caractéristiques texturales provenant des images FDG-TEP et IRM pour prédire les métastases aux poumons dans le cas des patients atteints des STM aux extrémités. Une amélioration substantielle des prédictions a pu être obtenue en utilisant les caractéristiques texturales des scans fusionnés et des stratégies d'analyse multivariable comparativement aux caractéristiques texturales des scans séparés et à l'analyse univariée. Potentiellement, cela pourrait mener à l'application de stratégies préventives pour atténuer la propagation du cancer des STM et à l'application de traitements mieux adaptés aux besoins des patients.

ACKNOWLEDGEMENTS

First, I would like to sincerely thank my supervisor Dr. Issam El Naqa for his great help and guidance throughout my master research. The dedication he has for the success of his students is simply enormous. In the future, I expect that great things will happen to me under his supervision.

My work would not have been possible without the help and useful advice of Dr. Sonia Skamene, a medical resident at McGill University Health Centre (MUHC). I thank her not only for initiating the project by gathering patient information from medical records, but also for updating patient information two times during my research.

My sincere thanks go also to Dr. Carolyn Freeman, an expert radiation oncologist that fed me with plenty of medical advice since the start of my research. I am grateful for all the time she took with me for patient contouring. Her dedication to research in general is something that will keep inspiring me.

Also, let me appreciate the excellent work of Dr. Jan Seuntjens as the Director of the Medical Physics Unit (MPU) of McGill University. He is definitely the leader of MPU and truly takes great care of his students.

Furthermore, I want to acknowledge the help of Mr. Sangkyu Lee during my research. He has always been there to answer my questions and explain concepts. I really appreciate it. Thanks also to Dr. Murray F. Brennan from the Memorial Sloan-Kettering Cancer Center (MSKCC) for his data on primary extremity soft-tissue sarcoma histopathology.

Mes plus sincères remerciements reviennent à Suzanne, Joannie, Sophie et Simon. Absolument rien de mes recherches n'aurait été possible sans leur support inconditionnel et infini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 FDG-PET Imaging

The development of hybrid PET/CT imaging in the past decade has created a revolution in oncology, a branch of medicine concerned with the diagnosis and treatment of cancer [1] [2]. Indeed, PET/CT was selected as the invention of the year by *TIMES* magazine in 2000. The combination of functional imaging in positron emission Tomography (PET) with the anatomical information in computed tomography (CT) scans provides an efficient tool to accurately localize metabolic abnormalities in the human body. Thanks to the significant improvements made in the performance of this imaging technology in the last decade, PET now plays an increasing role in the diagnosis, staging and monitoring of therapy response of cancer [3]. For example, PET provides great assistance in the detection of distant metastases that may not be apparent by routine staging procedures and thus has profound impact on clinical management and therapy decision-making; just to list a few, examples in different cancer types such as non-small cell lung cancer (NSCLC), lymphoma, colorectal cancer and malignant melanoma can be found in the literature [4] [5] [6] [7].

Fundamentally, PET imaging starts with the injection of a radiopharmaceutical tracer in the body. The radiopharmaceutical is comprised of a radionuclide that is attached to a chemical compound, which acts like a physiological analog known as the tracer. The tracer is fabricated in order to target the metabolic function of interest of tumours undergoing a certain biological process such as, for instance, glucose uptake. On the other hand, the radionuclide is used in the imaging acquisition process and acts as a source of radiation emission captured by the imaging scanner. Fundamentally, radionuclides are unstable isotopes undergoing transient radioactive decay. Proton-rich isotopes with low atomic mass number are used in PET imaging as they undergo the following positron decay process:

$$p \rightarrow n + e^{+} + \nu \tag{1.1}$$

In the decay process of Equation (1.1), one proton (p) of the unstable nucleus of the radionuclide gets converted into a neutron (n). The energy liberated in the conversion process is transferred to a positron (e^+) and a neutrino (ν), which are ejected from the nucleus with a continuous kinetic energy spectrum. Figure 1.1 depicts the physical principles of PET starting from the emission of the positron.

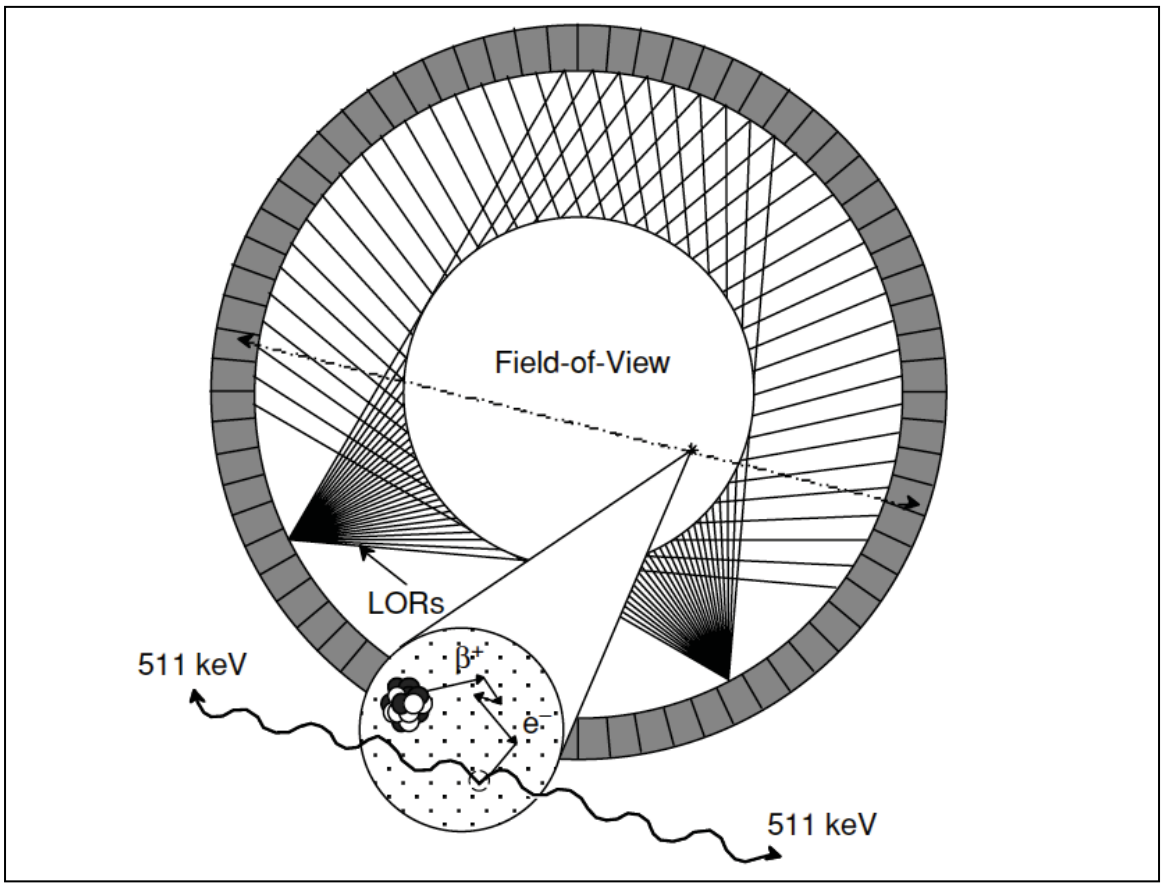

Figure 1.1. Physical principles of PET (taken from [8]).

Once the positron is emitted at a specific location in the body, it travels a few millimeters in tissues depending on its energy and undergoes several scattering events. At the end of its track, the positron annihilates with an electron (e^-) and the rest mass energy of the two particles is converted into two photons each of energy of 511 keV and nearly anti-parallel to each other. The detection of these two coincident photons along this line of response (LOR) allows inference about the location of the radiopharmaceutical in the body. The detection of annihilating photons is recorded in coincidence by several rings of

radiation detectors that are placed around the outside of the patient in the PET scanner (and therefore detects photons escaping from the inside of patients). A single detector on this ring is made of a scintillating crystal that converts the high-energy photons into brief pulses of visible light every time it is struck by an annihilation photon. The crystal is optically coupled to a photomultiplier tube (PMT) that converts and amplifies the scintillation light into an electrical signal. By capturing coincident annihilating photons along many LORs, it is possible to reconstruct a spatial map of the radioactivity concentration [MBq/kg] of the radiopharmaceutical in the body using basic principles of the Radon projection-slice theorem [9]. For a more detailed description of the underlying physics of PET imaging and image reconstruction, the reader is referred to some excellent reviews in references [8] [10] [11].

Nowadays, the most widely used radiopharmaceutical in the clinic for cancer detection and staging is fluorodeoxyglucose (FDG). As shown in Figure 1.2, the FDG tracer is a glucose analog in which the positron-emitting radionuclide fluorine-18 (^{18}F) with half-life of 110 minutes substitutes a normal hydroxyl group in the glucose molecule.

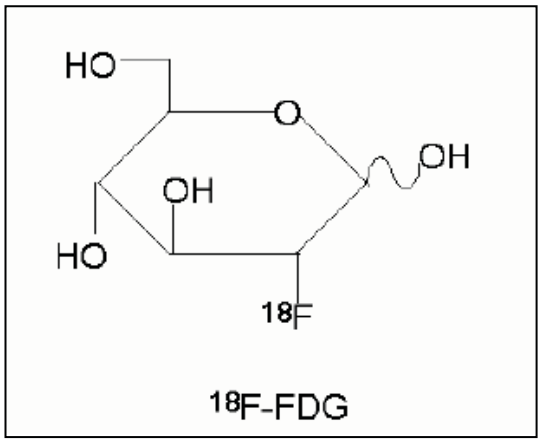

Figure 1.2. Fluorodeoxyglucose (FDG) compound (taken from [12]).

Glucose plays a central role in living cells as the fuel to cellular energy metabolism. Warburg [13] was among the first to demonstrate the altered glucose metabolism of malignant tumour cells. Essentially, one common property of cancer cells

is an enhanced rate of glucose uptake in the presence of oxygen. The observation of this effect in cancer cells has been repeatedly verified [14]. FDG-PET scans reveal regions of significantly increased glucose uptake in the human body, a dominant characteristic of tumour cells over normal tissues due to their high metabolic activity in support for rapid growth. FDG-PET imaging thus facilitates the quantification of glucose metabolic rates and subsequently the detection of most primary and metastatic cancers [15].

In most institutions, the injection of the radiopharmaceutical is done approximately 1 hour prior to the FDG-PET scan [1]. This uptake phase allows for the tracer to circulate through patient's blood and reach tissue targets. Once the scan is performed, the radioactivity concentration map of the human body is converted into a semi-quantitative measure known as the standard uptake value (SUV) in order to account for injection and body weight variability as defined in Equation (1.2). This metric provides a standard in the medical community for reporting the uptake measurements of FDG-PET scans in the form of SUV maps.

$$SUV(t) = \frac{radioactivity\ concentration\ at\ time\ t\ [MBq/kg]}{injected\ dose\ at\ time\ t_0\ [MBq]} * body\ weight\ [kg] \quad (1.2)$$

An important metric in oncology studies is the value of the voxel that yields the maximal SUV within the tumour region. This metric is denoted as $SUV_{\max}$. Nevertheless, there are several factors that can contribute to SUV variability and may impact its usefulness [16] [17].

1.2 MR Imaging

The importance of magnetic resonance (MR) in clinical imaging has exceeded most hopes of researchers from the 1980's due to its ability to manipulate and adjust tissue contrast with increasingly complex pulse sequences [18]. Indeed, MR imaging is without a doubt one of the wonders of modern medicine as one can generate contrast images that report a very large number of physical (e.g., proton density, T_1 or T_2 based contrast, etc.) and physiologic phenomena (e.g., water diffusion, tissue perfusion, oxygen levels, susceptibility variations, etc.) based on the rich physics of nuclear magnetic resonance (NMR) [19]. From an anatomical point of view, MR imaging provides much superior

soft-tissue contrast than CT images without loss of spatial resolution. More importantly, MR scans are non-invasive, as they do not expose patients to the undesirable x-ray radiation as in the case of CT scanning. These advantages of MR over CT scans has led, in the last few years, to the developments of whole-body imaging systems which integrate PET and MR imaging into one device [20]. Many experts consider the recent development of the hybrid PET/MR imaging technology as a major breakthrough that could potentially revolutionize clinical imaging practice [21] [22].

The NMR phenomenon occurs in atoms possessing a non-zero nuclear spin angular momentum. Due to its abundance in the human body, the hydrogen atom (^{1}H) is most often used in MR imaging. In the presence of a large external and constant magnetic field B_0, a net ensemble of proton spins (^{1}H) will align in the B_0 direction such that a net magnetization vector M_0 will be created in tissues. This effect is depicted in Figure 1.3.

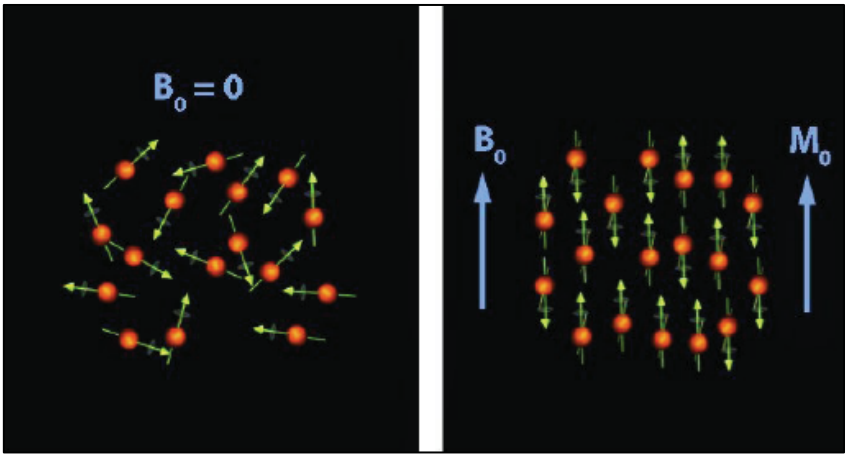

Figure 1.3. Alignment of nuclear spins in the presence of an external field (taken from [23]).

The coupling of the magnetic moment of nuclear spins with the angular momentum of nucleons causes the magnetization vector M_0 to precess around B_0 with an angular frequency known as the Larmor frequency (ω_0):

$$\omega_0 = \gamma B_0 \tag{1.3}$$

where γ is defined as the gyromagnetic ratio and is nuclei-specific. In the case of ^{1}H, γ is roughly equal to 42.58 MHz/Tesla. Let B_0 lie in the z-direction with M_0 precessing

around it in its equilibrium position. By using a radiofrequency coil, a radiofrequency pulse (RF) with time-varying (general case) magnetic field $B_1(t)$ tuned to Larmor frequency ω_0 can be applied in the transverse plane (*xy*-plane). As a result, M_0 is excited into the transverse plane as depicted in Figure 1.4. Then, M_0 will precess towards the transverse plane for a duration τ by which $B_1(t)$ is applied. The resulting angular displacement θ by which M_0 is rotated away from the longitudinal axis (*z*-axis) is given by:

$$\theta = \int_0^{\tau} \gamma B_1(t)\, dt \tag{1.4}$$

After the RF pulse is completed, the transverse magnetization will decay with characteristic relaxation time T_2 (spin-spin relaxation time constant), and the longitudinal magnetization will recover with characteristic time T_1 (spin-lattice relaxation time constant) to its previous equilibrium state governed by B_0. The rotating magnetization in the transverse plane then induces an oscillating electrical signal that can be captured and demodulated by two amplified radiofrequency coils placed at right angles in the transverse plane.

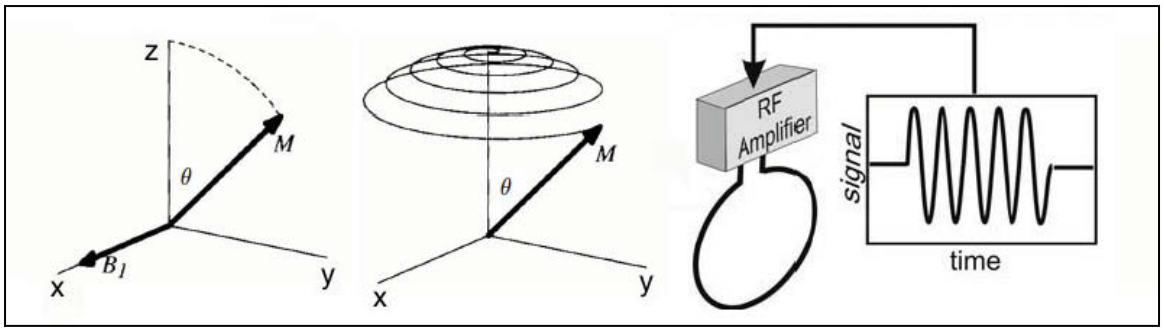

Figure 1.4. Principles of magnetization excitation and signal acquisition (adapted from [19] and [24]).

A general formalism known as the Bloch equation describes both the precession of the magnetization vector in 3D space due to arbitrary applied magnetic fields as well as the transverse and longitudinal relaxations:

$$\frac{d\mathbf{M}}{dt} = \mathbf{M} \times \gamma \mathbf{B} - \frac{M_x \mathbf{i} + M_y \mathbf{j}}{T_2} - \frac{(M_z - M_0)\mathbf{k}}{T_1} \tag{1.5}$$

Now, in order to generate an image, spatial localization is necessary. This is achieved by applying different gradients of magnetic fields in addition to the main field B_0 such that the total field strength varies in space. In this manner, the frequency of precession of spins varies with location since it is proportional to the magnetic field strength as shown in Equation (1.3). Fundamentally, the signal acquired at a given time from the receiving coils contains the contributions of all excited spins spatially oscillating with different frequencies as governed by the gradient fields applied at that particular time. In other words, the intensity of the signal at a given time represents one point of the Fourier space of the MR image, known as the *k-space*. Taking the inverse Fourier transform of the time signal thus linearly maps the contribution of each frequency component to its corresponding spatial location. This is the central concept allowing the formation of MR images. The vast variety of contrasts offered in MR imaging depends on the timing and strength of the different gradients used in the MR sequences. For example, a spin-echo sequence can be used to form an image where the contrast is based on the differences in T_1 or T_2 relaxation times of the different tissues of the human body, which is due to their different molecular environments. Figure 1.5 illustrates the formation of a typical spin-echo sequence. First, a 90° RF pulse is employed at the beginning of the sequence. Simultaneously, a slice selection gradient (Gss) is applied such that only spins within a slice of interest are excited. At time TE/2, spins have started to dephase by a certain amount and a 180° RF pulse is applied to invert the phase of spins. The spins then start to rephase such that at the echo time (TE), they are refocused and a high intensity spin-echo (SE) signal is created from the constructive interference of the spins. At time TE, the readout is performed by the frequency-encoding gradient (Gfe) in order to fill one line of the *k-space*. The process is reproduced at the repetition time (TR) for another line of the *k-space* by using a different phase-encoding gradient (Gpe) strength. The process goes on for many TRs until the whole *k-space* is sampled. TR governs the time by which the longitudinal magnetization can recover, whereas TE governs the time by which the transverse magnetization can decay. Hence, a T_1-weighted image will be formed for

short TR ~ T_1 and for short TE. A T_2 - weighted image will be formed for long TR and long TE ~ T_2.

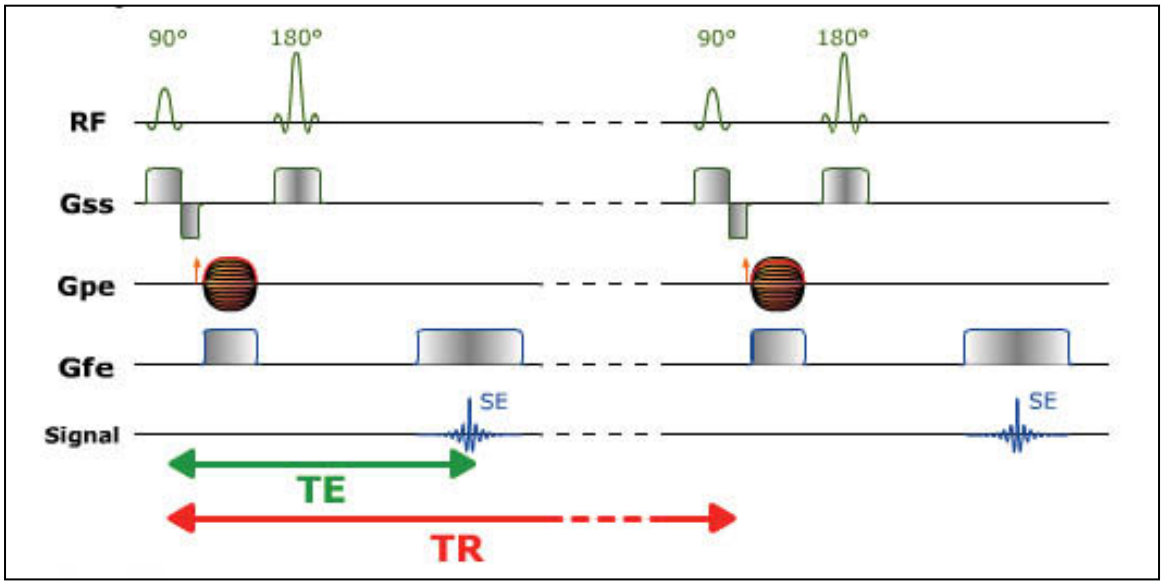

Figure 1.5. Typical spin-echo sequence (taken from [25]). RF: Radiofrequency pulses, Gss: slice selecting gradient, Gpe: phase-encoding gradient, Gfe: frequency encoding gradient, SE: spin-echo, TE: echo time, TR: repetition time.

In order to form images with other types of contrast, many other kinds of MR sequences exist for different purposes. One class is defined as *fat-suppression sequences* and its main purpose is to enhance tumour visualization from its surrounding. T_2 fat-saturation (T2FS) and short tau inversion recovery (STIR) sequences are part of this class. T2FS ([26] [27]) and STIR ([28] [29]) sequences are both fat-suppression techniques that have been clinically employed since the 1980's to emphasize the soft-tissue components in human body by supressing the signal coming from fat. To conclude this section, a brief description of these two sequences is presented.

STIR

Inversion-recovery methods exploit the fact that the characteristic time T_1 of fat is shorter than that of water. The sequence first starts with a 180° RF pulse such that the spins become anti-parallel to the main magnetic field. Subsequently to the pulse, the longitudinal magnetization of fat will return to equilibrium faster than the longitudinal

magnetization of water. At one point in time, the longitudinal magnetization of fat will be null as it crosses the *xy*-plane. If a 90° pulse is applied at that time, only the magnetization of water will be transferred to the *xy*-plane to produce the signal of interest. Hence, for the rest of any subsequent standard MR sequence, the fat spins will not contribute to the signal. The sequence has to be repeated with a long enough TR such that all spins have time to recover.

T2FS

This form of fat suppression technique exploits the small difference in resonant frequency between fat and water protons, which is related to their different electronic environments (chemical-shift effect). The sequence starts with a spectrally selective 90° pulse that ideally tips only the fat spins into the transverse plane. Only fat spins would contribute to the signal at this point. However, a spoiling gradient is applied immediately after the 90° pulse in order to dephase the fat spins in the transverse plane. As a result, fat signal decays to zero without affecting the water spins in their equilibrium state. The fat signal is then said to be "saturated" such that its contribution is suppressed in the subsequent standard MR sequence. The fat saturation step must then be repeated for every repetition of the MR sequence.

For the interested reader, several excellent reviews can be found in the literature and would give more detailed information about MR physics and pulse sequences variety, such as the ones from Nishimura [30] and Bernstein et al. [31].

1.3 Soft-tissue Sarcomas

Sarcomas are divided into two main groups of neoplasms: bone and soft-tissue sarcomas. Soft-tissue sarcomas (STS) constitute a heterogeneous group of malignant neoplasms of mesenchymal cell origin (connective tissue derived from mesoderm). More than 50 sub-types are recognized by the World Health Organization (WHO) [32] [33]. STS form a relatively uncommon type of cancer representing approximately 0.7% of new adult malignancies in the United States [34], but the majority (approximately two-thirds) of new cases are either intermediate or high-grade tumours [35]. STS tumours may arise in virtually all sites (e.g., head and neck, chest wall, retro-abdominal, etc.), but the

extremities is the most common site of origin (> 50%) with about twice as much primary sites being in the lower extremities as compared to the upper extremities [35] [36]. STS are relatively large compared to other types of tumours. In general, the different forms of therapy lead to excellent local control of STS of the extremities, but approximately 25% of these patients develop distant metastases [37]. In the case of high-grade tumours specifically, the metastatic rate goes up to approximately 50% [35]. The lungs are the site that accounts for approximately 80% of the metastatic cases in STS cancer of the extremities, more than in any other STS primary site [38]. In 1999, Billingsley et al. [39] retrospectively analyzed the outcomes of 3149 STS patients admitted at their institution from 1982 to 1997. A total of 719 patients either developed or presented with lung metastases. The study established that the 3-year actuarial rate of patients with lungs metastases is approximately 25%. Furthermore, Rehders et al. [40] claimed that STS patients with lung metastases do benefit from surgical treatments, but the authors also established that after the complete resection of the lung metastasis, the 5-year survival rate is still at the low rate of 25%. Due to this general poor prognosis, it is thought that better systemic therapies at earlier stages are needed for the management of lung metastases in STS of the extremities [35]. Standard treatments of primary STS of the extremities currently involve a combination of radiotherapy and surgery, but other adjuvant therapies could be considered for patients with high risk of lung metastases. For example, aggressive chemotherapy regimens or targeted cancer therapy adapted to the histopathology of the tumour could be considered once a patient is recognized to be more at risk of developing distant metastases [41]. In general, the development of highly personalized treatments and mechanisms for early metastasis detection should give a better chance to patients to overcome cancer.

The prognosis of STS is strongly related to several factors [42]. In 2002, Kattan et al. [43] published the results on the development of a nomogram that combines clinical and pathological predictor variables for the endpoint of postoperative sarcoma-specific death. This nomogram is nowadays used in many clinics to assess the aggressiveness of STS at diagnosis. The authors used a database of 2136 STS tumours from prospectively followed adult patients at the Memorial Sloan-Kettering Cancer Center (MSKCC) and

used bootstrapping techniques to validate the nomogram. The nomogram accepts *age at diagnosis*, *tumour size* (< 5, 5 to 10 cm or > 10 cm), *histologic grade* (high or low), *histologic sub-type* (fibrosarcoma, leiomyosarcoma, liposarcoma, malignant fibrous histiocytoma, malignant peripheral nerve tumour, synovial or other), *depth* (superficial or deep) and *site* (upper extremity, lower extremity, visceral, thoracic, trunk, retrointraabdominal or head and neck) as input variables to calculate the postoperative probability of 4, 8 or 12-year sarcoma-specific death. Recently, another model has been suggested as a complement to the one from the MSKCC. Carneiro et al. [44] proposed to use the combination of *size*, *vascular invasion*, *necrosis* and *peripheral tumour growth patterns* as a new prognostic model in STS of the extremities. The investigators used a dataset of 239 patients and achieved a prediction of high risk of metastasis with 74% sensitivity and 85% specificity.

The diagnosis and staging of STS is performed primarily with the use of MR and CT scans as well as with targeted biopsies. Due to the resulting excellent soft-tissue contrast, MR imaging is the procedure of choice for diagnosing STS [45]. It has also been shown that dynamic contrast-enhanced MR imaging (DCE-MRI) can be used to assess treatment response of STS [46] [47] [48]. However, staging studies must also include a CT scan of the chest in order to investigate for possible metastatic disease [42] [49]. Moreover, several studies have demonstrated that PET imaging could be successfully used in STS cancer for predicting the prognosis, for staging the disease and for assessing the response to therapy [50]. For example, from a dataset of 50 STS patients, Schwarzbach et al. [51] identified a trend where different SUV metrics (mean, median, maximal, etc.) increase with tumour grade, although the authors did not report any specific correlation coefficients. Although this is still a controversial debate, SUV_{max} is perceived to be a more useful value than the average SUV of the tumour region in high-grade STS since it is not diminished by large necrotic areas or haemorrhage [32]. In fact, Eary et al. [52] demonstrated the correlation of SUV_{max} with typical cancerogenesis processes such as tumour cellularity, mitosis Ki-67 level (proliferation marker) and p53 level (cell growth regulation product). Also, Eary et al. [53] used a database of 238 patients and showed that SUV_{max} is an independent predictor of survival and disease

progression using Kaplan-Meier curve analysis. All these findings demonstrate that SUV_{max} could be used as a marker of tumour biological aggressiveness in STS of the extremities. More importantly, this metric can be acquired with minimal invasiveness and without the sampling bias inherent to biopsies. Indeed, the biopsy from a small portion of the tumour may miss clinically significant high-grade areas [54]. However, we should keep in mind that a single voxel of FDG uptake (SUV_{max}) may be subject to variability and may not properly explain heterogeneous tumour behaviour. In order to obtain a more complete description of STS characteristics, O'Sullivan and co-workers [55] [56] analyzed the spatial heterogeneity of the FDG uptake of these tumours. From a dataset of 179 STS patients, the investigators developed a metric based on the deviation of the FDG distribution within the tumour region from a unimodal elliptically spatial pattern incorporating tumour boundary information. They established that the degree of heterogeneity of STS is a major risk factor that is associated with patient death.

To summarize this section, the management of STS represents a significant diagnostic and therapeutic challenge. Sarcomas are relatively uncommon but yet comprise a wide variety of entities. The prognostic evaluation of STS by an expert team is thus desirable [57]. A better distinction between low-grade and high-grade tumours is clinically needed in order to individualize treatment and consequently improve survival. This is especially true in the case of patients with lung metastases due to the resulting general poor prognosis. It is important to compile as many complementary prognostic factors as possible allowing for better prediction of lung metastases in STS patients, since their presence or absence can significantly influence the choice of treatment [58]. The development of new image analysis techniques that could provide better understanding and assessment of tumour aggressiveness than what is currently provided by traditional means is therefore desirable. Patients with STS of the extremities would certainly benefit from such techniques since the extremities is the most common site of origin of primary STS and the one from which lung metastases develop the most frequently.

1.4 Assessment of tumour biological heterogeneity

1.4.1 Overview

It is now recognized that tumours do not represent a homogeneous entity but rather can be composed of multiple clonal sub-populations of cancer cells. Differing properties can be attributed to the different sub-populations in terms of growth rate, expression of biomarkers, ability to metastasize, and immunological characteristics [59]. Such intratumoral differences are related to the concept of tumour heterogeneity, a characteristic that can be observed with significantly different extents even amongst tumours of the same histopathological type. As shown in the image of Figure 1.6, intratumoral variations in STS can be observed from diagnostic images such as MR T2FS scans. Different sub-regions representing different cell types can clearly be seen in that image. Just to name a few, differences in metabolic activity, cell proliferation, oxygenation levels, pH, drug delivery, blood vasculature and necrotic areas characterize the different cell sub-populations within a tumour. These variations greatly influence the sensitivity to therapeutic response [60], as observed in chemotherapy ([61] [62]) and radiation therapy ([63] [64]) for different cancer types.

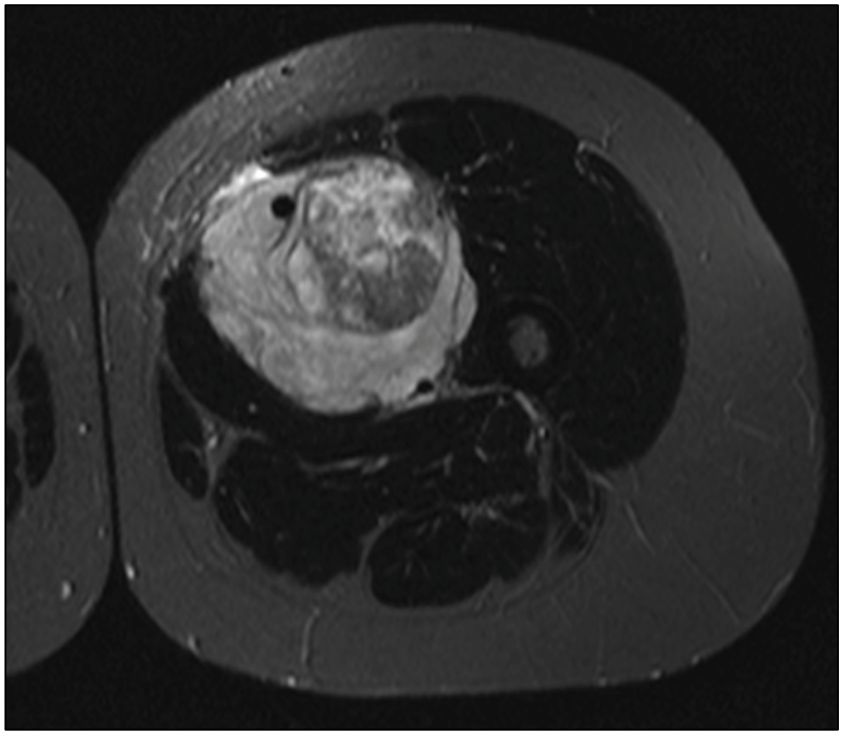

Figure 1.6. Example of a soft-tissue sarcoma of the leg from a MR T2FS scan.

Ideally, the study of tumour heterogeneity should provide molecular signatures specific to the patient to be treated such that tumour aggressiveness and sensitivity to therapeutic response can be assessed prior to treatments. Thereafter, this would allow the individualization of treatments and eventually, improved outcomes. Different techniques can be employed in order to identify prognostic factors from histopathological biopsies. A technique that has received much attention in the last decade performs digital analysis of nuclei images of cancer cells. This technique attempts to find nuclear signatures that are statistically significant prognostic factors of tumours through the spatial arrangement of gray levels in histopathological images. An excellent review of this method known as *statistical nuclear texture analysis* can be found in reference [65]. However, one major drawback of such approach is that the image analysis can only be performed *in vitro* and on small sections of the tumour. Studying tumour heterogeneity from histopathological biopsies is extremely difficult since the "answer" significantly varies depending on which part of the tumour was sampled [66]. An additional difficulty is that the knowledge of the characteristics of individual components of a tumour is not sufficient to predict the behaviour of the whole [59]. As a result, texture analysis of the whole tumour volume and from *in vivo* images can be further appreciated as a promising approach for analyzing tumour heterogeneity. This will be the subject of the next sub-section as well as the main subject of the work presented in this thesis.

1.4.2 Texture analysis in FDG-PET and MR imaging

The quantification of tumour heterogeneity using FDG-PET and MR imaging is an active area of research incorporating many different techniques such as selective regions of interest, cluster analysis, selective classification, histogram analysis, texture mapping and spatial geometric approaches [67]. In the case of FDG-PET, it has been demonstrated that FDG uptake is dependent on the tumour microenvironment such that different regions of low oxygenation levels (hypoxia), cellular proliferation, blood flow and necrosis correlates either positively or negatively with FDG uptake [68] [69]. In the case of MR imaging, the vast variety of contrasts allowed by the many different types of sequences can definitely play a role in the assessment of tumour physiology and the identification of cell sub-populations [70]. In addition to anatomical imaging, MR allows functional imaging of biological processes in the human body. For example, diffusion-weighted MR

imaging (DW-MRI) quantifies the degree of isotropic water diffusion in extracellular space as affected by the size and the distribution of cellular populations. It has been shown that DW-MRI can be used to assess regional cellularity and the aggressiveness of tumours [71] [72]. FDG-PET and MR scans can thus reflect intratumoral heterogeneity and can provide methods that would potentially lead to a better understanding of tumour biological complexity. However, the study of tumour heterogeneity should also be directed towards the identification of useful prognostic factors such that physicians can choose the best therapeutic approach for patients affected by cancer. It is accepted in the literature that tumours with highly heterogeneous metabolism are likely associated with high metastatic potential that thereafter can lead to poor outcomes [73] [74]. Hence, the development of new prognostic factors that can robustly identify aggressive tumours with high metastatic risk as early as possible in the course of diagnosis and treatment management is profoundly desirable. With this is in mind, the application of texture analysis for tumour outcome prediction has very recently gained more attention in the scientific community.

Texture analysis is concerned with the spatial distribution of gray level variations within an image. Textures are useful to quantify the complexity of an image as it pertains to the extent, frequency and spatial arrangement of these variations [75]. As an example, Figure 1.7 shows an image with five different regions, each represented by different textural properties. It can clearly be seen that the five different regions of that image have different spatial arrangements of gray levels.

Figure 1.7. Image texture example (taken from [76]). The image contains 5 regions with different textural properties.

The statistical distribution of gray levels defining the texture of an image can be quantified in many ways depending on what type of information is analyzed. Just to name a few, typical texture features can be extracted from co-occurrence matrices, run-length matrices, autocorrelation features, fractals or from the wavelet transform. Chapter 2 will describe in details one texture analysis method that have been used for more than 30 years in the pattern recognition community. The reader is referred to some excellent reviews in the literature ([76] [77] [78]) for more information about the variety of textures that can be extracted from an image.

In medical imaging, textures are primarily used for the classification of benign versus malignant lesions as well as for the segmentation of tumours. This has been demonstrated in the literature for different cancer types for both FDG-PET ([79] [80] [81] [82]) and MR imaging ([83] [84] [85] [86] [87] [88]). Some groups have also studied the variability and the reproducibility of FGD-PET textural measurements in terms of differing acquisition protocols, reconstructions parameters and time elapsed between acquisitions [89] [90]. Likewise, the influence of MR acquisition protocols and parameter variations on textural measurements has been investigated [91] [92] [93]. Moreover, some studies have investigated the potential of textural features to assess cancer treatment response [94] [95] [96]. In this case, the change in texture features from before to after the treatments is reported. Yet, the application of interest in this study is texture analysis performed on FGD-PET and MR whole tumour volumes for improved **outcome prediction** such as local recurrence, distant metastases and survival. The proof of concept implies that textural measurements are to be done retrospectively on diagnostic images obtained prior to any treatment and analyzed after a certain period of patient follow-up. Thereafter, the association of features with the different outcomes is investigated. The goal of this general methodology is to retrospectively identify prognostic factors to be thereafter used prospectively in order to assess the potential risk of a given outcome at the diagnosis of cancer. To date, very few publications have reported this type of work. In 2009, El Naqa et al. [97] were among the first to present a robust methodology dedicated to the prediction of tumour outcomes using texture features. The study retrospectively evaluated intensity-volume histogram metrics, shape features and texture features from the gray-level co-occurrence matrix (GLCM) of pre-chemoradiotherapy diagnostic FGD-

PET scans of 14 cervix cancer patients and 9 head and neck cancer patients. Using logistic regression, the study respectively combined an intensity-volume histogram metric with *Energy* texture and an intensity-volume histogram metric with a shape metric in order to respectively separate disease persistence patient classes in cervix cancer and overall survival patient classes in head and neck cancer with good statistical power, as demonstrated by the respective Spearman's correlation coefficient obtained on the entire patient cohorts ($rho = 0.49,\ p = 0.04$ and $rho = 0.87,\ p = 0.0012$). In 2011, Maday et al. [98] retrospectively analyzed DCE-MRI images of 17 breast cancer patients in order to predict response to neoadjuvant chemotherapy. The authors extracted Gabor texture features from the baseline scans before treatment and were able to classify responders and non-responders with an accuracy of 69%. Also in 2011, Tixier et al. [99] evaluated 6 texture features from the GLCM of pre-treatment FDG-PET images for the prediction of response to chemoradiotherapy of 41 patients with esophageal cancer. Response was assessed one month following therapy and patients were classified as non-responders (NRs), partial responders (PRs) and complete responders (CRs). The authors showed that *Entropy* and *Homogeneity* GLCM texture features could differentiate between the 3 classes of response with a higher sensitivity (76% to 92%) than any SUV measurement, where the best discrimination was obtained between NRs and (PRs + CRs). Finally, in 2012, Vaidya et al. [100] retrospectively analyzed pre-treatment FDG-PET and CT images of 27 patients with non-small cell lung cancer. Local failure and loco-regional failure to radiotherapy was assessed at least 6 months after the completion of treatment for all patients. It was shown that the combination of one intensity-volume histogram metric from FDG-PET and one from CT provided better separation of local failure ($rho = 0.5908,\ p = 0.0013$) and loco-regional failure ($rho = 0.4853,\ p = 0.0067$) patient classes than texture features.

To summarize this section, it has been shown in the literature that tumour heterogeneity is associated to aggressive tumours with high metastatic potential. In order to improve patient outcomes, the importance of identifying relevant prognostic factors that can better assess the metastatic risk of tumours at the moment of diagnosis is crucial. Texture analysis of medical images has recently been shown to be a promising technique

for tumour outcome prediction. This progress could pave the way to better individualization of treatments, but much research efforts are still required in order to develop and validate new image processing techniques that enable to robustly assess and predict tumour aggressiveness.

1.5 Thesis workflow, objectives and organization

Considering the high risk of lung metastases in STS cancer of the extremities and the resulting poor survival rate, the general objective of this work is to develop new and robust prognostic factors for the prediction of lung metastasis risk at the time of diagnosis of the primary tumour. This information could eventually assist physicians in their choice of treatment and potentially improve survival. Figure 1.8 summarizes the major concepts and the general methodology used in this work.

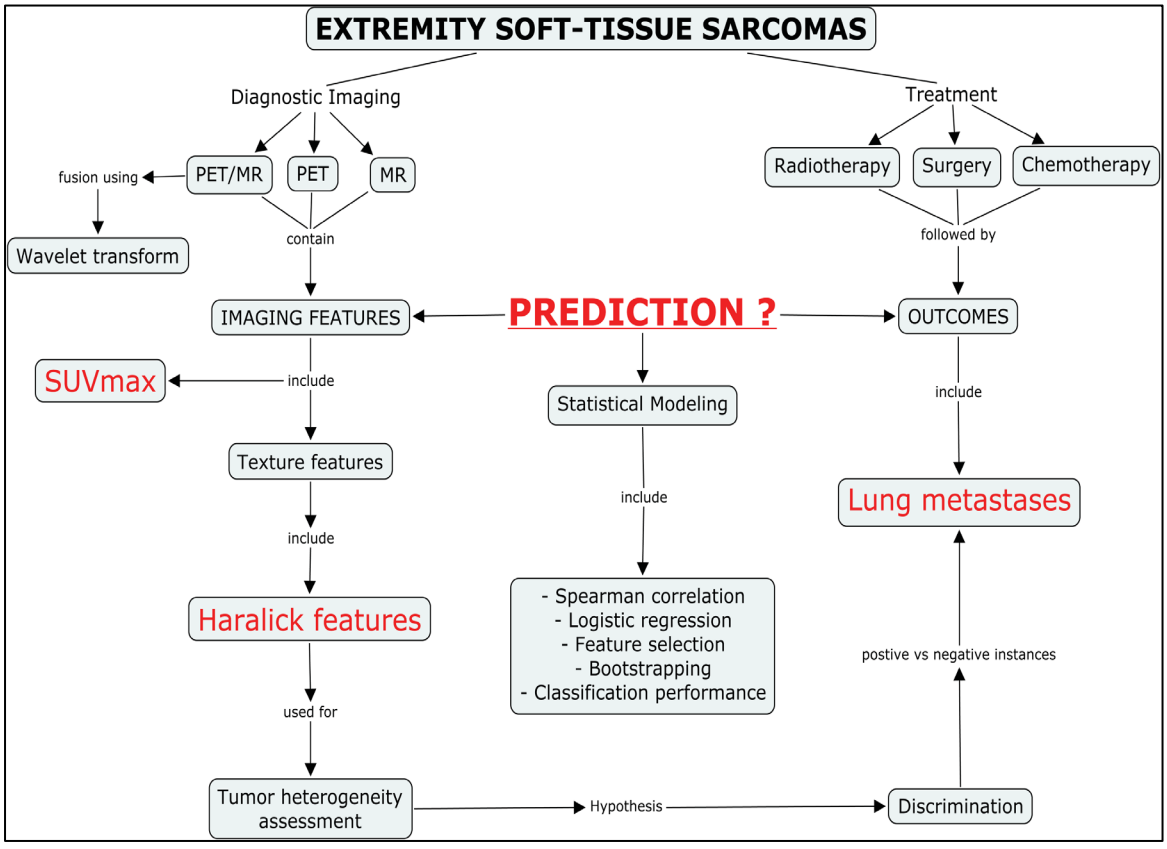

Figure 1.8. Thesis workflow.

Overall, several *Haralick texture features* extracted from the gray-level co-occurrence matrix (GLCM) of FDG-PET, MR and fused FDG-PET/MR images are used to assess the tumour heterogeneity of STS of the extremities. Our working hypothesis is that these features (via the differences in tumour heterogeneity and cell composition) should allow to discriminate between two categories of STS of the extremities: those that metastasize to the lungs (MetsLungs) and those that do not metastasize to the lungs (No MetsLungs). Thereafter, different statistical modeling techniques are employed in order to identify a general model formed from a subset of features and optimized for the prediction of lung metastases. More specifically, the major objectives of this study are thus to:

1. Investigate the potential of texture features from pre-treatment FDG-PET, MR and fused FGD-PET/MR scans for the prediction of lung metastases in STS cancer of the extremities.
2. Investigate linear combinations of texture features for the prediction of lung metastases in STS cancer of the extremities.
3. Identify a model of linear combination of texture features that would offer best prediction performance on an independent patient dataset.

To the best of our knowledge, this is the first study that explores the potential of texture features for the prediction of lung metastases in STS cancer. More importantly and still to the best of our knowledge, this is the first study exploring the potential of texture features from fused FDG-PET/MR scans for the assessment of biological properties of any type of cancer. Our hypothesis is that the combination of FDG-PET and MR spatial information does result in the creation of new composite textures that could help to better identify STS tumours with high metastatic risk.

The organization of the thesis goes as follows: Chapter 2 briefly describes the theory behind some of the techniques employed in this work. The theory of texture analysis using the gray-level co-occurrence matrix (GLCM) is first elaborated. Then, the theory behind the discrete wavelet transform is detailed, as it provides the mathematical tool of choice in this work to perform the fusion of FDG-PET and MR scans. Finally, the

statistical methods used to assess the prediction of the clinical endpoint of interest are described. Chapter 3 thereafter describes in details the methodology of this work. Chapter 4 is dedicated to the presentation of results. A discussion about the implication of the results is offered in Chapter 5. In this chapter, an uncertainty analysis of the best linear model found in this work as well as a discussion about the validity of the model evaluated from permutation tests is also provided. Finally, Chapter 6 concludes the thesis and presents future work to be done subsequently to this study. An appendix providing further justification of the methods employed in this work is also provided at the end of the thesis.

CHAPTER 2: BACKGROUND

2.1 GLCM-based texture features

In 1973, Haralick et al. [101] proposed the concept of texture analysis from the gray level co-occurrence matrix (GLCM). In their original pioneering work, the investigators took into account the statistical nature of textures, which is based on the assumption that texture information is contained in the overall spatial relationship that the gray levels have to one another. More specifically, the authors made the hypothesis that the texture information in an image could be adequately specified by the matrix of frequencies of occurrence $P_{d,\theta}(i,j)$ with which gray level i and gray level j are neighbours by a distance d and angle θ (i.e. the GLCM). This concept is shown in Figure 2.1 for $d=1$ and 4 different angles $\theta = 0°, 45°, 90°, 135°$.

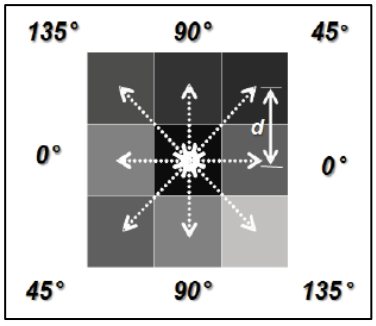

Figure 2.1. GLCM concept (taken from [102]). The GLCM is computed by considering adjacent neighbours of every pixel of an image that are separated by a distance d and angle θ.

The size of the GLCM is dependent on the number of gray levels (N_g) in a given image. The entries of row i and column j of the GLCM specify the number of times gray levels i and j are neighbours by a given distance and angle. Hence, the GLCM is of size ($N_g \times N_g$). In order to compute the GLCM, the image is scanned such that the neighbouring properties of all pixels are verified. Figure 2.2 provides an example of the GLCM computation of a test image with $d=1$ and $\theta = 0°, 45°, 90°, 135°$.

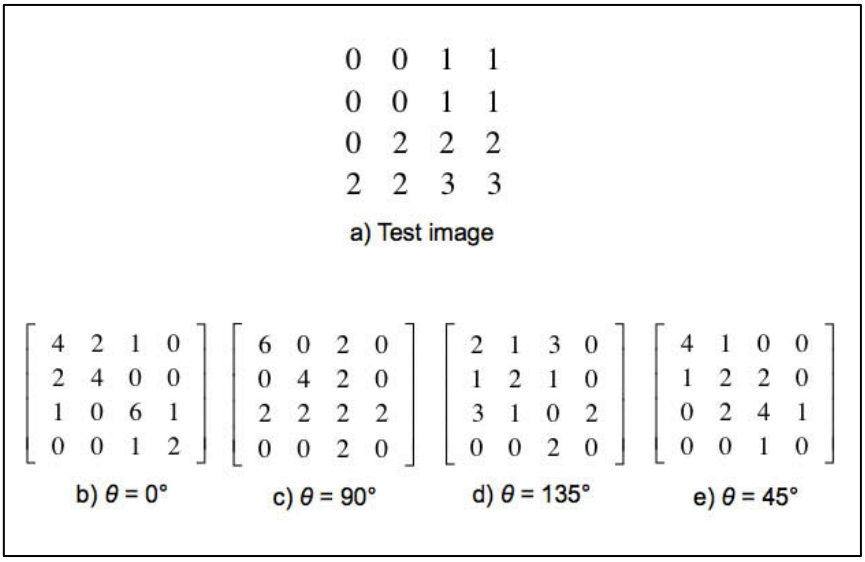

Figure 2.2. GLCM computation example (adapted from [101]). a) Test image, b) GLCM with $d=1$ and $\theta=0°$, c) GLCM with $d=1$ and $\theta=90°$, d) GLCM with $d=1$ and $\theta=135°$, e) GLCM with $d=1$ and $\theta=45°$. Every GLCM is of size 4X4 since the test image contains 4 different gray levels.

In applications for which the directional dependence of textures is not studied (such as the one presented in this work), a common practice is to average the corresponding entries of the 4-directional GLCMs such that only one matrix describes the isotropic textural properties of the image. The resulting entries of the GLCM then become:

$$P_d(i,j)=\frac{P_{d,0°}(i,j)+P_{d,90°}(i,j)+P_{d,135°}(i,j)+P_{d,45°}(i,j)}{4},\quad \text{for } i,j=1,...,N_g \qquad (2.1)$$

Subsequently to this operation, a normalization factor representing the sum of all occurrences in the image is applied to the GLCM such that:

$$p_d(i,j)=\frac{P_d(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P_d(i,j)},\quad \text{for } i,j=1,...,N_g \qquad (2.2)$$

Thereafter, features containing different textural characteristics can be extracted from the GLCM described by $P_{d,\theta}(i,j)$. In their original work, Haralick and co-workers proposed a set of 14 features (now known as the *Haralick features*) to be extracted from the GLCM through different mathematical operations. In the current study, 6 *Haralick features* were

tested. The name commonly used in the literature, the mathematical description ([101]) and a qualitative description of each of these features are presented in Table 2.1.

Table 2.1. Haralick Features used in this study

Texture feature	Formula	Description
Energy	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_d^2(i,j)$	Measure of repeatability of gray levels
Entropy	$-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_d(i,j)\log(p_d(i,j))$	Measure of randomness in an image
Contrast	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2\, p_d(i,j)$	Measure of the amount of local variations in an image
Homogeneity	$\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{p_d(i,j)}{1+\lvert i-j\rvert}$	Measure of image uniformity
Sum-Mean	$\frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i\cdot p_d(i,j)+j\cdot p_d(i,j))$	Average of the distribution of occurrence of gray levels
Variance	$\frac{1}{2}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}((i-\mu)^2\cdot p_d(i,j)+(j-\mu)^2\cdot p_d(i,j))$	Deviation of the distribution of occurrence of gray levels from the mean

Each texture feature extracted from the GLCM characterizes a different property of the image. The first four features of Table 2.1 are commonly used in the literature. The last two features of Table 2.1 were shown to be important discriminants of benign versus malignant lesions in T_1 and T_2 MR images of glioblastoma multiforme [88]. Hence, the choice was made to investigate their discriminant power in this work. Note that in the formula of *Variance* shown in Table 2.1, the symbol μ represents the average of the GLCM entries.

An important concept related to texture extraction that needs to be described is the quantization of the image prior to the computation of the GLCM. In this procedure, the full range of gray levels is reduced to a smaller subset number N_g. Every pixel is assigned

a new value from a set of values $\mathbf{r} = \{r_k \in \mathbb{R} : k = 1,2,...,N_g\}$. This procedure is referred to as intensity quantization. Smaller number of gray levels accelerates the computation of the GLCM and reduces the influence of noise on calculated texture features, but this is offset by a potential loss of information. With higher N_g, more information is provided and it would be expected, from a pattern recognition point of view, that classification accuracy would be improved assuming high signal-to-noise ratio. However, Clausi [103] showed that this is not always the case for *Haralick features* and that some of these features exhibit the opposite behaviour. At full dynamic range, very few gray level pairs are repeated and the GLCM is rather sparse. As a consequence, the classification power of most features is diminished. Therefore, quantization is a necessity and an optimal N_g specific to a given texture analysis application must be found such that the trade-off between the amount of stochastic noise and the amount of information in a set of images allows best classification results. Typical numbers of gray levels used in the literature are 8, 16, 32 or 64.

2.2 Discrete wavelet transform

The wavelet transform has found applications in many areas since its introduction in the 1980's. Just to name a few, wavelets are nowadays used in image compression, image de-noising, image fusion and pattern recognition. The goal of wavelet analysis is to decompose a given signal over a family of wavelet basis functions generated from a mother wavelet by dilatation and translation [104]. Unlike the Fourier transform which decomposes a given signal according to its frequency content only, the wavelet transform allows efficient localization of a signal in both space (or time) and frequency domains. In this section, a review of the discrete wavelet transform (DWT) theory is provided in the first sub-section. Then, the last sub-section will show how the wavelet transform can be used to fuse FDG-PET and MR volumes. Throughout the whole section, only a brief overview of the theory could be provided due to space constraints. The reader is referred to excellent comprehensive reviews by Strang [105] and Burrus [106] for further details about the wavelet theory.

2.2.1 General theory

As previously mentioned, the goal of wavelet analysis is to decompose a signal over a family of wavelets generated from a mother wavelet. Let the signal of interest be in space x and the mother wavelet of interest be $\psi(x)$. The mother wavelet is a squared-integrable function over all space with zero average such that:

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 \, dx \quad \textit{and} \quad \int_{-\infty}^{+\infty} \psi(x) dx = 0 \tag{2.3}$$

In wavelet theory, the class of expansion functions generated from the mother wavelet are written in their most general form as:

$$\psi_k^j(x) = c^{-j/2} \psi(c^{-j}(x - kbc^j)) \tag{2.4}$$

For most wavelet families, $c = 1/2$ and $b = 1$ such that:

$$\psi_k^j(x) = 2^{j/2} \psi(2^j x - k) \tag{2.5}$$

This implies that all wavelets $\psi_k^j(x)$ are dilated (or scaled) and translated versions of $\psi(x)$ as defined by the integers j and k respectively. The goal of wavelet expansion is to generate a set of functions $\psi_k^j(x)$ such that any signal in the space of squared-integrable functions $L^2(\mathbb{R})$ can be represented by the series:

$$f(x) = \sum_k \sum_j w_k^j \, 2^{j/2} \psi(2^j x - k) \tag{2.6}$$

In Equation (2.6), the set of expansion coefficients w_k^j is called the discrete wavelet transform (DWT) of $f(x)$. If the expansion is unique, the set of functions $\psi_k^j(x)$ is called a *basis* for the class of functions that can be so described. The power of such a basis is that it can simultaneously express a signal at different scales and spatial locations. However, in wavelet theory, the formulation of such multiresolution analysis is made in terms of two closely related basis functions. In addition to the mother wavelet, we introduce the scaling function $\phi(x)$ that can be expressed in terms of a weighted sum of translated versions of $\phi(2x)$ such that:

$$\phi(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} l(n) \phi(2x - n) \tag{2.7}$$

The mother wavelet $\psi(x)$ is also expressed in the same manner:

$$\psi(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} h(n) \phi(2x - n) \tag{2.8}$$

Equation (2.7) is governed by "low-pass" coefficients $l(n)$ of the wavelet expansion and Equation (2.8) is governed by "high-pass" coefficients $h(n)$. The relation between these coefficients is:

$$h(n) = (-1)^n l(1 - n) \tag{2.9}$$

Figure 2.3 displays an example of the scaling and wavelet functions of a specific class of wavelets known as *sym8*.

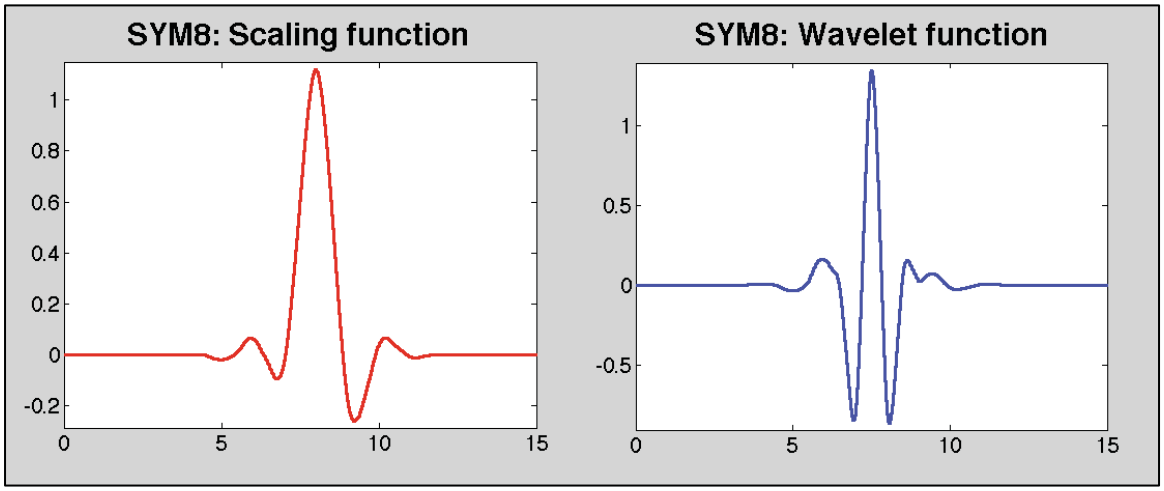

Figure 2.3. Scaling and Wavelet functions *sym8*.

With such double-basis representation, the decomposition of a signal into a finite number of levels J becomes:

$$f(x) = \sum_k a_k^J 2^{J/2} \phi(2^J x - k) + \sum_k \sum_{j=1}^{J} d_k^j 2^{j/2} \psi(2^j x - k) \tag{2.10}$$

Equation (2.10) implies that the a_k^J coefficients are used to represent the approximation of signal at the lowest level (or scale) J with the scaling function $\phi(x)$. As such, $\phi(x)$ is used to represent the coarse details of the signal, or its low-frequency components. The rest of the decomposition coefficients (d_k^j) are used to represent the fine details of the signal, or its high-frequency components. These coefficients are obtained at all scales using the family of functions $\psi_k^j(x)$. Finally, all coefficients at scale j can be expressed in terms of the coefficients of the previous scale using the following recursive equations:

$$\begin{aligned} a_k^j &= \sum_{n\in\mathbb{Z}} a_k^{j-1} l(n-2k) \\ d_k^j &= \sum_{n\in\mathbb{Z}} a_k^{j-1} h(n-2k), \quad \text{for } j=1,2,...,J \end{aligned} \tag{2.11}$$

To summarize, the discrete wavelet decomposition of a signal starts with an educated choice of the scaling function $\phi(x)$. The mother wavelet $\psi(x)$ is then extracted from the scaling function and made orthogonal to the scaling function through Equations (2.7) to (2.9). A family of dilated and translated wavelets $\psi_k^j(x)$ is then extracted from the mother wavelet through Equations (2.4) and (2.5). As shown in Figure 2.4, the discrete wavelet decomposition up to level (or scale) J is thereafter performed through a cascade-tree of low-pass and high-pass filters followed by downsampling by a factor of 2. The wavelet coefficients a_k^j and d_k^j are obtained by the convolution over space of the proper scaling and wavelet functions defined at each level j.

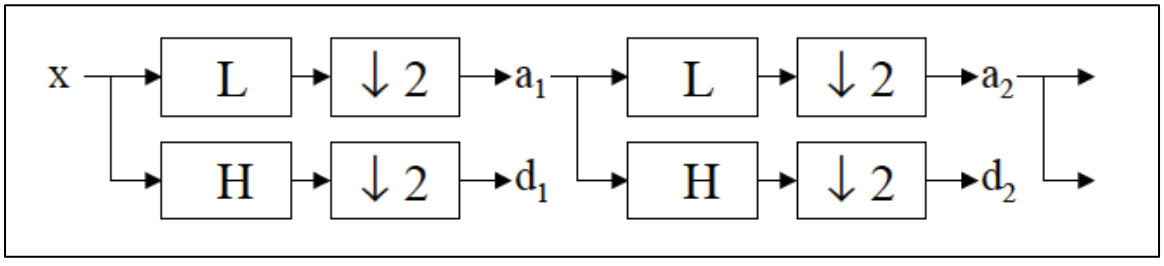

Figure 2.4. DWT tree (taken from [107]). A cascade of low-pass filter (L) and high pass filter (H) decomposes the signal (x). At every level j, new approximation coefficients (a_j) and detail coefficients (d_j) are obtained. ↓ 2 denotes the process of downsampling by a factor of 2.

Up to now, the discussion about wavelet theory has involved 1-dimensional (1D) signals. The 1D multiresolution wavelet decomposition can be extended to two dimensions by introducing 2D scaling and wavelet functions as the tensor product of their 1D complements ([108]) such that:

$$\begin{aligned} \phi_{LL}(x,y) &= \phi(x)\phi(y), \quad \psi_{LH}(x,y) = \phi(x)\psi(y) \\ \psi_{HL}(x,y) &= \psi(x)\phi(y), \quad \psi_{HH}(x,y) = \psi(x)\psi(y) \end{aligned} \tag{2.12}$$

In reality, performing one level of a 2D discrete wavelet decomposition consists of filtering and down-sampling an image $I(x,y)$ both horizontally and vertically with the 1-D low-pass filter (L) ϕ and the 1D high-pass filter (H) ψ. As a result, the wavelet coefficients of four different sub-bands are produced: *LL, LH, HL* and *HH* (Figure 2.5*a*). Every sub-band has now half the initial size of $I(x,y)$ in both x and y directions. In order to obtain the 2D discrete wavelet decomposition at higher levels (level 2 shown in Figure 2.5*c*), the same process is performed on the *LL* sub-band generated from the previous decomposition level (level 1 shown in Figure 2.5*b*) and is repeated up to the desired level of decomposition.

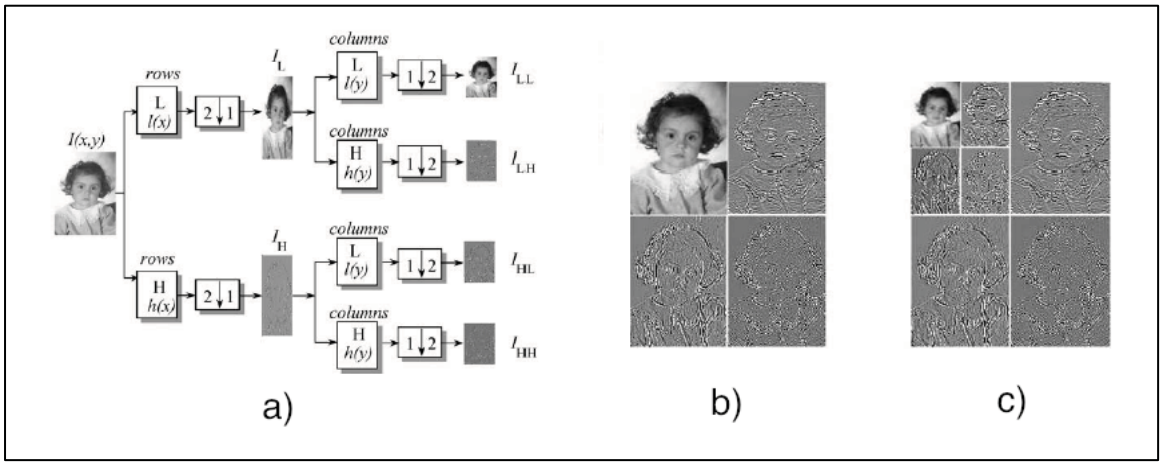

Figure 2.5. 2D DWT process (adapted from [108]). a) One stage of a 2D DWT image decomposition, b) Representation of a one-level decomposition, c) Representation of a two-level decomposition.

Finally, the procedure used to reconstruct the original signal from the wavelet coefficients is known as the inverse discrete wavelet transform (IDWT), which is simply

the reverse process of the DWT. In practice, each sub-band is first up-sampled by a factor of 2 by inserting zeros in-between the wavelet coefficients. Next, each sub-band is convolved with the appropriate reconstruction filters. For example, the *HL* sub-band is first upsampled and convolved horizontally with the 1D high-pass (wavelet) reconstruction filter. Then, it is upsampled and convolved vertically with the 1D low-pass (scaling) reconstruction filter. The reconstruction filters are the original scaling and wavelet filters flipped from left to right about their central position. Once this process has been applied to the four sub-bands, the results are added together to obtain the original image. The reader is referred to reference [108] for pedagogical examples of wavelet decomposition and reconstruction.

2.2.2 Fusion of FDG-PET and MR volumes

For the purpose of this study, image fusion can be described as the process of combining information from two different images into a single composite image that is more informative for texture analysis. During the past two decades, many different methods for performing image fusion were developed and tested on different types of images [109]. Some of these algorithms use spatial domain fusion features such as gradients, spatial frequencies or local standard deviations. Another category of fusion methods exploits transform domains, in which the source images are projected onto localized basis usually designed to detect meaningful salient features. The latter category includes the DWT fusion method, a concept proposed by Li et al. [110] in 1995. In their original work, the investigators successfully tested their methods on multi-focus images, synthetic-aperture radar (SAR) images as well as PET and MR images. The general framework of the DWT fusion method is described in this sub-section and depicted in Figure 2.6.

First, let us assume that the two images to be fused are co-registered (if misregistration occurs, artefacts will be present in the fused image) and have the same resolution. For the fusion scheme presented in Figure 2.6, it means that resampling and registration strategies have to be applied prior to fusion. Then, the fusion process starts with the application of the DWT to both images with a given scaling function of choice. The decomposition can go up to an arbitrary number of levels (2 decomposition levels are shown in Figure 2.6). Afterwards, the respective wavelet coefficients of the two images

are merged together. In other words, *LL* coefficients of image 1 are merged with *LL* coefficients of image 2, *HL* coefficients of image 1 are merged with *HL* coefficients image 2, and this process is repeated for all sub-bands. Subsequently to the latter step, only one set of fused wavelet coefficients exist. Finally, the IDWT is applied to the fused wavelet coefficients in order to reconstruct the fused image. Fundamentally, the key step in DWT image fusion is based on how the wavelet coefficients of the two different images are combined. The goal is to merge the wavelet coefficients in an appropriate way in order to obtain the image characteristics sought. For example, a *maximum selection* rule could be employed in which the maximal wavelet coefficient of the two images is chosen at every position and the other is discarded. A more advanced technique would involve *adaptive weighted averaging* in which the weight given to each coefficient of the two images is chosen based on the activity level around each coefficient as defined by a window of a small area. In general, the choice of the fusion rule is specific to the application of interest and does not need to be the same for all sub-bands. Finally, it is worth mentioning that the DWT is also very well suited for the fusion of images with different resolutions due to its multiscale representation, as long as each step of the fusion is carried out with sub-bands at the same resolution level.

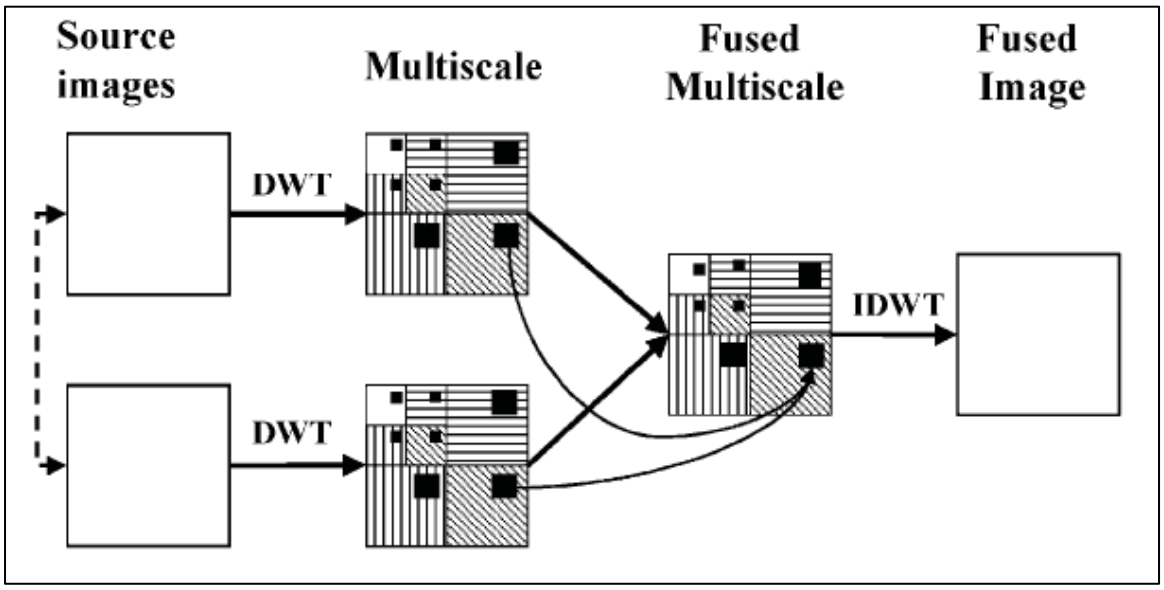

Figure 2.6. DWT fusion scheme (adapted from [108]).

In this study, the DWT is used to perform 3D fusion of FDG-PET and MR volumes. For this purpose, a 3D DWT is applied to both volumes. The generalization of the wavelet theory to 3 dimensions is straightforward: 3D wavelet decomposition consists of filtering and downsampling a volume $V(x,y)$ in the x, y and z directions with the 1D low-pass filter (L) ϕ and the 1D high-pass filter (H) ψ. As shown in Figure 2.7, the wavelet coefficients of eight different sub-bands are produced: *LLL, LHL, LHH, HLL, HHL, HLH* and *HH*.

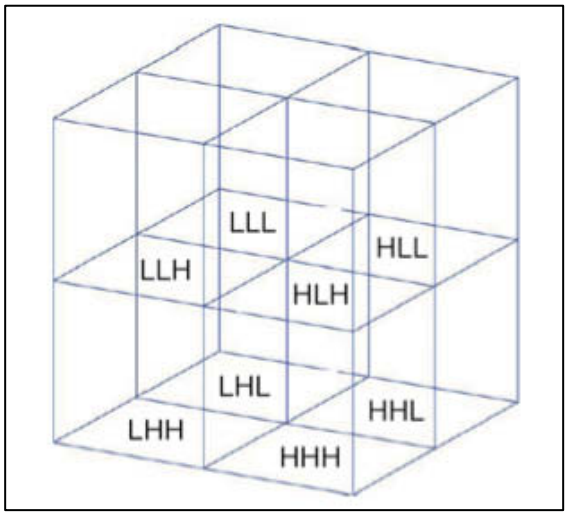

Figure 2.7. 3D wavelet decomposition sub-bands (adapted from [111]).

Thereafter, wavelet coefficients of the eight respective FDG-PET and MR sub-bands are grouped together, and the 3D IDWT is applied in order to reconstruct the 3D fused FDG-PET/MR volume.

2.3 Statistical Modeling

This section details the statistical methods employed in this work, from the investigation of the univariate association of textures features with the clinical endpoint of interest to the statistical validation of multivariable models.

2.3.1 Spearman's rank correlation

In this study, the correlation of texture features (or a combination of them) with lung metastases in STS of the extremities is assessed using Spearman's rank correlation. Let **x**

and $\mathbf{y}$ be two datasets of equal size N such that $\mathbf{x},\mathbf{y} = \{x_i, y_i \in \mathbb{R} : i = 1,2,...,N\}$. First, the individual values x_i and y_i are converted to the ranks X_i and Y_i that they take in their respective dataset such that $X_i, Y_i \in \{1,2,...,N\}$ with the exception of ties which are assigned a rank equal to the average of their positions in the ascending order of the values. Then, Spearman's rank correlation coefficient *rho* between $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$rho = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{2.13}$$

In Equation (2.13), $\bar{X}$ and $\bar{Y}$ represent the average of the ranks X_i and Y_i. Spearman's rank correlation describes how well two variables are monotonically related, independently of their linear association as it is the case with Pearson's coefficient. A result of 1 implies perfect positive correlation, a result of -1 implies perfect negative correlation and a result of 0 implies no correlation between the variables.

2.3.2 Logistic regression

One objective of this study is to build a multivariable model of features allowing for maximal prediction of lung metastases in STS of the extremities. Suppose that we have a sample set of n independent observations (number of patients) of the pair $(\mathbf{x}_i, y_i)$, where y_i denotes the value of a dichotomous outcome variable and $\mathbf{x}_i$ is the vector of input variables (e.g., imaging data) of the i^{th} patient for $i = 1,2,...,n$. The value of the dichotomous variable takes the form $y_i \in \{0 : \text{No MetsLungs}, 1 : \text{MetsLungs}\}$, and the vector of input variables takes the form $\mathbf{x}_i = \{x_{ij} \in \mathbb{R} : j = 1,2,...,p\}$, where p is the number of variables in the model, or the model order. This notation will be used throughout the text. We are interested in finding a linear combination of the p variables such that the multivariable model of interest takes the form:

$$g(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } i = 1,...,n \tag{2.14}$$

The set $\boldsymbol{\beta} = \{\beta_j \in \mathbb{R} : j = 1,2,...,p\}$ is the set of regression coefficients of the model to be determined such that the conditional probability of the set of outcome states {0,1} given the input data $\mathbf{x}_i$ is maximized for $i = 1,2,...,n$. This operation is carried out using a logistic regression model (logit transformation) of the form:

$$\pi(\mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i) = \frac{e^{g(\mathbf{x}_i)}}{1 + e^{g(\mathbf{x}_i)}}, \quad \text{for } i = 1,...,n \tag{2.15}$$

Excellent reviews of the logistic regression method can be found in references [112] and [113]. The form of the logistic regression model shown in Equation (2.15) is commonly used as it models a sigmoidal relationship between the input variables and the response endpoint within the range [0,1], lending itself to a clinically meaningful interpretation of observed responses. To be more specific, $\pi(\mathbf{x}_i)$ express the conditional probability that outcome y_i equals 1 (MetsLungs) given the input $\mathbf{x}_i$. Consequently, the conditional probability that outcome y_i equals 0 (No MetsLungs) given the input $\mathbf{x}_i$ is $P(y_i = 0 \mid \mathbf{x}_i) = 1 - \pi(\mathbf{x}_i)$. If we assume the n observations to be independent, it follows that a convenient way to express the conditional probability of a set of dichotomous outcome states given the set of input data is:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i) = \prod_{i=1}^{n} \begin{cases} \pi(\mathbf{x}_i) & \text{if } y_i = 1 \\ 1 - \pi(\mathbf{x}_i) & \text{if } y_i = 0 \end{cases}$$

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} \tag{2.16}$$

where $l(\boldsymbol{\beta})$ is known as the *likelihood function*. It is defined as in Equation 2.16 ([112]) for logistic regression. For greater mathematical simplicity, the logarithm of $l(\boldsymbol{\beta})$ is used:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]\} \tag{2.17}$$

The set of regression coefficients that maximize the likelihood function is found by differentiating $L(\boldsymbol{\beta})$ with respect to all β_j coefficients and then equating to zero. This yields a set of $p+1$ non-linear likelihood equations to be solved simultaneously:

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

and (2.18)

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0, \quad \text{for } j = 1,2,...,p$$

Several available logistic regression software such as SAS and SPSS have implemented methods to solve this set of non-linear equations for the set of β_j values. Numerically, the set of p+1 non-linear likelihood equations are solved using an iterative weighted least-square method. The presentation of this methodology goes beyond the scope of this text, but the interested reader is referred to reference [114] for a general description of the methods used by most programs. Once the solution to the set of β_j values is found, the multivariable model of Equation (2.14) can be constructed and the association between **g** and **y** can be tested using Spearman's rank correlation in the same manner as in the univariate case, where $\mathbf{g} = \{g(\mathbf{x}_i) : i = 1,2,...,n\}$ and $\mathbf{y} = \{y_i : i = 1,2,...,n\}$.

2.3.3 Feature selection

Generally, the construction of a multivariable model of order p involves the selection of a subset of p variables from a larger set of m variables. The search for the most parsimonious model (or the simplest plausible model with fewest number of variables) is an important step in any multivariable approach. Such a model needs to be protected against overfitting such that it is not too dependent on the observed data and thus could subsequently be generalized to unseen data. However, enough variables need to be selected in order to reach maximum predictive power. Once an optimal model order p is found, the process of selecting p variables can be done in several ways and the most common methodologies are known as *forward selection* and *backward selection*. Stepwise forward selection involves starting with no variables in the model and then testing the contribution of the stepwise addition of all variables to the model by taking into account the previously chosen variables until model order p is reached. Stepwise backward selection involves starting with all variables in the model and then testing the contribution of the stepwise deletion of all variables to the model until model order p is reached.

In this work, a forward selection scheme was chosen and the statistical significance of stepwise addition of regression coefficients β_j was determined using the Wald's test (*W*). Let a subset of *p* variables to be selected from a larger set of *m* features such that a model of order *p* of the form of Equation (2.14) is constructed with coefficients $\boldsymbol{\beta} = \{\beta_j \in \mathbb{R} : j = 1,2,...,p\}$. The forward selection scheme go as follows:

1. Begin with a fit of the offset β_0 found using logistic regression in order to build the model M_0.
2. Investigate the fit of all *m* variables with M_0. Choose the variable with maximum *W* by performing logistic regression on the *m* variables separately with M_0. Add the chosen variable to M_0 in order to build the model M_1 using logistic regression.
3. Investigate the fit of the $m-1$ variables left from the complete set with M_1. Choose the variable with maximum *W* by performing logistic regression on the $m-1$ variables separately with M_1. Add the chosen variable to M_1 in order to build the model M_2.
4. Repeat step 3 until a model of order *p* has been constructed.

The solution that we obtain from logistic regression is an estimate of the regression coefficients. Inherently, these *estimated coefficients* $\hat{\beta}_j$ have also estimated standard errors $\widehat{\mathrm{SE}}(\hat{\beta}_j)$. The Wald test is then defined as ([112]):

$$W_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \tag{2.19}$$

It has already been shown how to compute the solution to logistic regression in order to obtain the coefficient estimates $\hat{\beta}_j$. The estimated standard errors of regression coefficient estimates can be obtained from the covariance matrix of all regression coefficients. The methodology for estimating the variances and covariances of the regression coefficients follows the well-established theory of maximum likelihood

estimation that can be found, for example, in reference [115]. This theory states that the variance and covariance estimates are obtained from the second-order partial derivatives of the log-likelihood function $L(\boldsymbol{\beta})$ such that:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}\,\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)), \quad for\ j = 0,1,...,p$$

and (2.20)

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j\,\partial \beta_l} = -\sum_{i=1}^{n} x_{ij}\,x_{il}\,\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i)), \quad for\ j,l = 0,1,...,p$$

Then, the observed information matrix $\mathbf{I}(\boldsymbol{\beta})$, known as the Fisher information matrix, is obtained by taking the negative of the terms in Equation (2.20):

$$\mathbf{I}(\boldsymbol{\beta}) = -\begin{pmatrix} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1^2} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{pmatrix} \tag{2.21}$$

The variance and covariance estimates of the regression coefficients are thereafter obtained by taking the inverse of $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})$ such that:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})^{-1} \tag{2.22}$$

The diagonal terms then represent the variances and the off-diagonal terms the covariances. Finally, the standard error estimate of the individual regression coefficients $\widehat{SE}(\hat{\beta}_j)$ is obtained by taking the square root of the diagonal terms of the inverse of $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})$ such that:

$$\widehat{SE}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} \tag{2.23}$$

2.3.4 Bootstrapping

Bootstrapping is a statistical resampling method introduced by Efron [116] in 1979. The motivation of his pioneering work was to develop a more general yet simple alternative to cross-validation techniques for the estimation of unknown probability distribution of some random variable based on the observed data. In the current study, bootstrapping is used as the resampling method of choice to assess the prediction performance of texture models. Two important reasons have motivated this choice. First, bootstrap techniques are known to reduce the variance of the error estimates compared to cross-validation techniques (leave-one-out or *n*-fold cross-validation), although it introduces a bias due to the resampling of data. More importantly, bootstrapping provides more realistic scenarios, as it is less prone to overestimation of statistical significance than cross-validation techniques, one of the major pitfalls of data mining. Bootstrap tutorials ([117]), reviews ([118]) and considerable number of applications in medicine ([119] [120]) are readily found in the literature. A key issue in our analysis is to determine which models learned from the current patient cohort would best predict lung metastases on unseen data (prospective patient cohort). The general methodology used to simulate this situation is to divide the original dataset into training and testing sets. Then, the complete regression model is learned from the training set and its predictive power is assessed in the testing set using different classification performance metrics. In this work, the resampling method used to divide the patient population into training and testing sets is based on the *bootstrap technique*. An example of the application of this method is presented here:

1. Let a set of 10 patients denoted as $k = \{n_1, n_2, ..., n_{10}\}$.
2. Randomly sample a patient 10 times with replacement from k. For example, let the selected patients be $\{n_3, n_4, n_1, n_2, n_2, n_1, n_5, n_{10}, n_3, n_5\}$.
3. Patients that were selected form the training set, no matter how many times they were selected. Patients that were not selected form the testing set.
4. Following the example given in step 2, the training set is $\{n_1, n_2, n_3, n_4, n_5, n_{10}\}$. The testing set is $\{n_6, n_7, n_8, n_9\}$.

Afterwards, a multivariable model can be built from the training set and its prediction performance can be assessed in the testing set. This process is performed over many bootstrap samples in order to achieve convergence of the statistical quantity of interest. In the limit of an infinite number of bootstrap samples, the average size of the training sets is approximately 63% that of the original data set while approximately 37% is withheld for testing [118]. The inherent drawback of the method is increased computing time.

2.3.5 Classification performance metrics

In this sub-section, the performance metrics used to assess the predictive power of multivariable models in bootstrap testing sets are described. Let a multivariable model $g(\mathbf{x}_i)$ of the form of Equation (2.14) be found from a given training set. As usual, y_i denotes the value of a dichotomous outcome variable of the form {0: No MetsLungs, 1: MetsLungs} and $\mathbf{x}_i$ is the vector of input variables for the i^{th} patient. The testing set is first used to assess the classification of positive instances $y_i = 1$ and negative instances $y_i = 0$. In binary classification theory, 4 quantities of interest need first to be calculated: 1) TP: number of true positive instances; 2) FP: number of false positives instances; 3) TN: number of true negative instances; and 4) FN: number of false negative instances. Given the offset β_0 of the multivariable model, a threshold of $g(\mathbf{x}_i) = 0$ is used to determine if a patient is to be classified as a TP, FP, TN or FN instance. Table 2.2 resumes this classification scheme.

Table 2.2. Classification of TP, FP, TN and FN

Model value	Outcome value	Classification
$g(\mathbf{x}_i) > 0$	$y_i = 1$	TP
$g(\mathbf{x}_i) > 0$	$y_i = 0$	FP
$g(\mathbf{x}_i) < 0$	$y_i = 0$	TN
$g(\mathbf{x}_i) < 0$	$y_i = 1$	FN

Given TP, FP, TN and FN, the *sensitivity*, *specificity* and *accuracy* of classification is calculated as:

$$\begin{aligned} sensitivity &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ specificity &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ accuracy &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \end{aligned} \tag{2.24}$$

Another measure of the quality of classification known as Matthews' correlation coefficient (MCC) is also used in this work. MCC is regarded as a balance measure that can be used even if the number of positive and negative instances is very different. Essentially, MCC provides a measure of the correlation between the observed and predicted binary classification. A coefficient of +1 corresponds to a perfect prediction, a coefficient of 0 corresponds to a random prediction and a coefficient of -1 corresponds to a total disagreement between the observations and predictions. This metric is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{2.25}$$

Finally, another metric assessing model performance can be extracted from the receiver-operating characteristics (ROC) curve of a binary classifier. The ROC curve is a plot of *sensitivity* against (1-*specificity*) as new subsets of {TP, FP, TN, FN} are obtained for varying decision thresholds. The metric of interest enabling the assessment of the quality of the classifier is:

$$\mathrm{AUC} = \textit{Area under the ROC curve} \tag{2.26}$$

An AUC of 0.5 corresponds to a random classifier whereas an AUC of 1 corresponds to a perfect classifier. One way of interpreting this metric is that the greater is the AUC, the better is the separation between the positive and negative instances.

CHAPTER 3: MATERIALS AND METHODS

3.1 Dataset

3.1.1 Patient cohort

Subsequently to research ethics board (REB) approval, a database of 35 patients (18 males, 17 females) with a histologically proven primary soft-tissue sarcoma of the extremities (upper or lower extremity STS; denoted as "STS" only for the rest of the text, unless otherwise specified) and having received a pre-treatment FDG-PET/CT and MR scans at the Montreal University Health Center (MUHC) was retrospectively queried. Patients with recurrent STS at presentation or with STS of the head and neck, chest wall, retro-abdominal or other non-extremity sites were excluded from the study. The age of the group of patients was 17-82 (mean: 54 ± 18, 1 standard deviation). Four patients from the dataset presented with metastases at the time of diagnosis of the primary tumour (2 bone, 2 lungs) and 11 patients developed lung metastases during the follow-up period. Lung metastases were either proven by biopsy or diagnosed by an expert physician from the appearance of typical pulmonary lesions on CT and/or FDG-PET images. In this study, the dataset was hence divided into two lung metastases classes: 13 patients with lung metastases (MetsLungs) and 22 patients without lung metastases (No MetsLungs). Table 3.1 provides different characteristics of patients including age, sex, histological type of primary STS, location of primary STS, grade of primary STS, spread at diagnosis and recurrence information. Also, Table 3.2 presents the follow-up time (time elapsed between the date of diagnosis of the primary STS to the last date of follow-up) for all patients and for the two sets of patients with different lung metastasis status.

Table 3.1. Characteristics of patients

Patient	Age	Sex (M/F)	Histology	Location (Upper/Lower extremity)	Grade	Spread @ diagnosis	Recurrence
1	59	M	Liposarcoma	Lower	High	Local	Distant (foremarm)
2	22	F	MFH	Lower	Low	Local	None
3	64	M	Extraskeletal bone sarcoma	Lower	High	Local	Distant (lungs)
4	28	F	MFH	Upper	Low	Local	None
5	82	M	Liposarcoma	Lower	High	Local	None
6	76	M	Liposarcoma	Lower	Intermediate	Local	None
7	54	M	Extraskeletal bone sarcoma	Lower	High	Local	Distant (lungs)
8	48	F	MFH	Upper	High	Local	None
9	76	M	MFH	Lower	High	Local	None
10	51	F	Liposarcoma	Lower	High	Local	Distant (spine)
11	62	F	MFH	Lower	High	Distant (lungs)	Distant (lungs)
12	69	M	Liposarcoma	Lower	High	Local	Distant (lungs)
13	65	F	MFH	Lower	High	Local	None
14	24	F	Leiomyosarcoma	Lower	High	Local	None
15	63	M	Liposarcoma	Lower	High	Local	Distant (lungs)
16	57	F	Epithelioid sarcoma	Lower	Ungraded	Local	Distant (lungs)
17	29	M	Liposarcoma	Lower	Low	Local	None
18	46	M	Liposarcoma	Lower	Low	Local	None
19	27	M	Extraskeletal bone sarcoma	Lower	High	Local	None
20	57	F	MFH	Lower	High	Distant (lungs)	Distant (lungs) + Local
21	62	F	Liposarcoma	Lower	Ungraded	Local	None
22	34	M	Liposarcoma	Lower	Intermediate	Local	Distant (abdominal)
23	54	M	Synovial sarcoma	Lower	Intermediate	Local	Local (lymph nodes)
24	61	F	MFH	Lower	Intermediate	Distant (bone)	Distant (bone)
25	61	M	MFH	Lower	High	Local	Distant (lungs)
26	62	F	MFH	Lower	High	Local	None
27	57	M	Leiomyosarcoma	Lower	Intermediate	Local	Distant (lungs)
28	76	F	Liposarcoma	Lower	Intermediate	Local	None
29	69	M	Leiomyosarcoma	Lower	Intermediate	Local	Distant (lungs)
30	19	M	Rhabdomyosarcoma	Upper	High	Distant (bone)	Distant (bone)
31	71	F	Extraskeletal bone sarcoma	Lower	High	Local	Distant (lungs)
32	53	M	Liposarcoma	Lower	Low	Local	None
33	75	F	Leiomyosarcoma	Lower	High	Local	Distant (lungs)
34	17	F	Synovial sarcoma	Lower	Intermediate	Local	None
35	72	F	Leiomyosarcoma	Lower	High	Local	Distant (lungs)

Extraskeletal bone sarcoma: *Includes Ewing and osteogenic sarcoma developing in soft-tissues*
MFH (Malignant Fibrous Histiocytoma): *Includes myxofibrosarcoma and undifferentiated sarcoma*

Table 3.2. Follow-up time of patients

Patients	Median follow-up (months)	Range (months)
All patients	29	4-85
No MetsLungs	47	4-85
MetsLungs	20	5-34

Some patients of the dataset were lost to follow-up early after the time of diagnosis of the primary STS. In principle, short follow-up times lead to a significant uncertainty on the lung metastasis status only for the class of patients that did not develop lung metastases during the follow-up period. In fact, all patients from this class have a non-zero probability of having the wrong lung metastasis status since they could develop the distant disease after the follow-up period. This probability increases with shorter follow-up time. It is thus important at this point to establish the standard time duration between the diagnosis of a primary STS of the extremities and the spread to the lungs in order to estimate the significance of that probability in our patient cohort. Due to the specificity of the question, only one report that directly estimate this timeframe was found in the literature, whereas other reports included other variants of sarcomas and/or included metastatic events to sites other than the lungs. Gadd et al. [121] followed 716 extremity STS patients admitted at their institution between 1983 and 1990. A total of 77 patients developed metachronous lung metastases after the diagnostic of a primary tumour. The median time between the treatment of the primary tumour and the development of metachronous lung metastases for these 77 patients was 14 months (range: 1 to 152 months), with 72% of the cases occurring within 2 years after treatment. On the other hand, it occurred that the median time between the diagnosis of the primary STS and the development of lung metastases for the 11 patients of our patient cohort (excluding patients 11 and 20) was found to be 5.5 months (range: 2.5 to 17 months). For the other class of patients of our cohort (22 patients; No MetsLungs), it occurred that 1 patient had a follow-up time smaller than one year (patient 6: 4 months) and 2 patients had a follow-up time ranging between 1 and 2 years (patient 24: 22.5 months, patient 34: 23 months). According to these numbers, we believe that only patient 6 suffers from a significant probability of having the wrong lung metastasis status.

Finally, Table 3.3 presents the occurrence of histologic sub-types in our patient cohort. It is compared to the database of 3073 patients with primary STS of the extremities admitted at the Memorial Sloan-Kettering Cancer Center (MSKCC), NY between July 1982 and December 2007. Appendix A also displays the data obtained from the MSKCC.

Table 3.3. Histologic sub-types of STS of the extremities

Histology	Number of patients (%)	
	This study	MSKCC*
Liposarcoma	12 (34)	718 (23)
MFH	10 (29)	847 (28)
Leiomyosarcoma	5 (14)	241 (8)
Synovial sarcoma	2 (6)	319 (10)
Fibrosarcoma	0 (0)	113 (4)
Extraskeletal bone sarcoma	4 (11)	835 (27)
Other	2 (6)	
Total	35 (100)	3073 (100)

* Reproduced with the permission of Murray F. Brennan from the MSKCC

From Table 3.3, it is reasonable to think that the variety of histologic sub-types in our patient cohort adequately represents the variety of histologic sub-types present in the population of STS of the extremities. As a consequence, the methodology developed in this work should be generalizable to other patient cohorts of STS of the extemities.

3.1.2 Imaging data

FDG-PET and MR scans were retrieved from the PACS server at MUHC. Scans were performed at our institution between May 2004 and May 2010 on a hybrid PET/CT scanner (Discovery ST, General Electric Medical Systems, Waukesha) and a 1.5T MR scanner (SIGNA EXCITE, General Electric Medical Systems) respectively.

In the case of FDG-PET/CT scans, CT and PET images were consecutively acquired from the base of the skull to the upper thighs and from the base of the feet to the upper thighs, with additional images of the extremities acquired according to the sarcoma location if necessary. The CT slice thickness resolution was 3.75 mm and CT in-plane resolution was 0.98 mm for all patients. For the PET portion of the scans, between 370

and 500MBq of FDG were injected intravenously. Approximately 60 minutes following the injection, a 2D acquisition was performed for 5 to 6 bed positions (height dependent) with 4-5 min per bed position (body weight dependent). PET attenuation corrected images were reconstructed in the axial plane using an ordered subset expectation maximization (OSEM) iterative algorithm. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 4.69X4.69 mm^2 (range: 3.91-5.47 mm).

In the case of MR scans, several different sequence types acquired according to the sarcoma location were performed in the clinic for each patient. Three types of MR sequences were selected for this study, namely T_1 - weighted, T2FS and STIR sequences. The MR scans resulted from clinical acquisitions with heterogeneous protocols across patients. The median slice thickness resolution was 5 mm (range: 3-10 mm) and the median in-plane resolution was 0.78X0.78 mm^2 (range: 0.20-1.72 mm). T_1 sequences were common to all patients. On the other hand, T2FS and STIR were not common to all patients, but all patients were scanned with at least one of these two sequences. From a physics point of view, T2FS and STIR sequences are not the same, but macroscopically (fat signal suppression) and thus from a texture point of view, the images appear similar. In order to have the largest dataset possible, the choice was made to use both sequences in a similar manner for texture analysis. As an example, Figure 3.1 shows the four types of scans used in this study (FGD-PET, T_1, T2FS, STIR) taken from patient 3. Table 3.4 shows the availability of the scans for all patients. T2FS scans were selected by default due to their overall better visual quality over STIR in our dataset and also due to their axial scan availability. When T2FS scans were not available, STIR scans were used. Appendix B further explores the consequences of that choice of priority of fat-suppression scan type in our analysis. For the rest of the text, we will refer to the three separate scans that were used as PET, T1 and T2FS/STIR (i.e. T2FS or STIR).

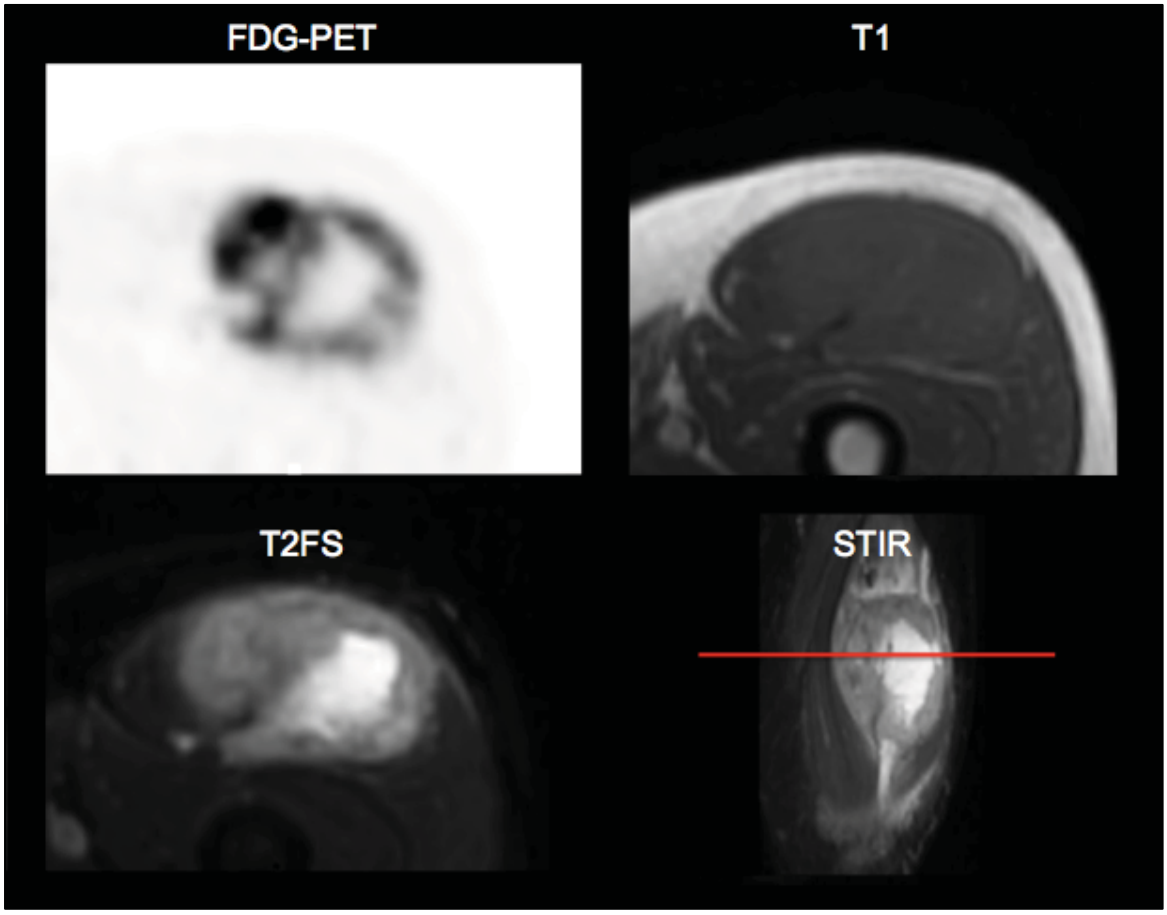

Figure 3.1. Example of STS imaging from patient 3. Top-left: FDG-PET image, Top-right: T1 image, Bottom-left: T2FS image, Bottom-right: STIR image. FDG-PET, T2FS and T1 images were acquired in the axial plane and STIR images were acquired in the coronal plane. The red line in the STIR image corresponds to the plane shown in the three other images.

Table 3.4. Scan availability

Patient number	PET	T1	T2FS	STIR
1	✔	✔	✔	-
2	✔	✔	-	✔
3	✔	✔	✔	✔
4	✔	✔	-	✔
5	✔	✔	-	✔
6	✔	✔	✔	✔
7	✔	✔	-	✔
8	✔	✔	✔	✔
9	✔	✔	-	✔
10	✔	✔	-	✔
11	✔	✔	✔	✔
12	✔	✔	✔	-

13	✔	✔	✔	✔
14	✔	✔	✔	✔
15	✔	✔	✔	✔
16	✔	✔	-	✔
17	✔	✔	✔	✔
18	✔	✔	✔	✔
19	✔	✔	✔	-
20	✔	✔	✔	✔
21	✔	✔	✔	-
22	✔	✔	✔	✔
23	✔	✔	✔	✔
24	✔	✔	✔	✔
25	✔	✔	✔	✔
26	✔	✔	✔	✔
27	✔	✔	✔	✔
28	✔	✔	-	✔
29	✔	✔	✔	✔
30	✔	✔	-	✔
31	✔	✔	✔	✔
32	✔	✔	✔	✔
33	✔	✔	✔	✔
34	✔	✔	-	✔
35	✔	✔	-	✔

3.2 Tumour volume definition

In order to perform texture analysis on PET, T1 and T2FS/STIR scans and allow a direct comparison amongst them, tumour volume definition must be the same on all patient's respective scans. To achieve this procedure, contouring of tumours was first done manually slice by slice. Then, image registration was carried out such that the different volumes were spatially related in order for the same contours to be propagated on all scans. In this section, the manual contouring, contour propagation workflow and image registration steps are detailed.

3.2.1 Tumour contouring

Tumour contours were drawn on T2FS/STIR scans and were verified by an expert radiation oncologist. The T2FS/STIR scans were first imported into a clinical workstation at MUHC and contours were manually drawn on a slice-by-slice basis using the contour tools from the clinical software Eclipse™ (Varian Medical Systems, Inc., Palo Alto, CA;

http://www.varian.com/us/oncology/radiation_oncology/eclipse/). Visible edema or inflammation could be clearly identified in the vicinity of the tumours for 16 patients of the cohort. For these patients, two types of contours were drawn; one incorporating the visible edema and one excluding it. For the rest of patients, only one contour was drawn and was considered as containing no edema. As an example, Figure 3.2 shows two contours drawn on the T2FS scan of patient 12.

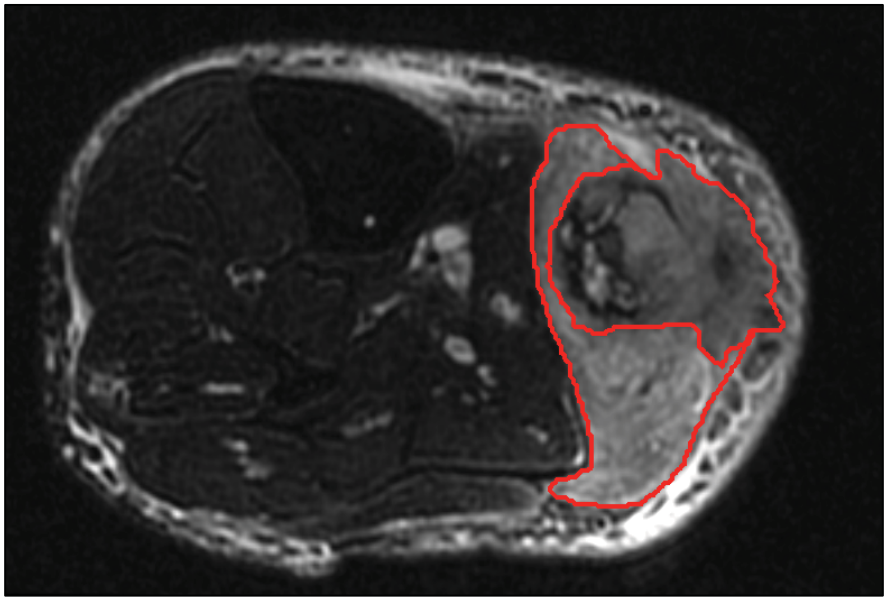

Figure 3.2. Contouring example on T2FS scan of patient 12.

The results presented in this work were obtained from texture analysis performed only on the inner portion of the tumours, that is to say, using the contours containing no edema. The contours incorporating edema were used only for the uncertainty analysis presented in section 5.2.

3.2.2 Contour propagation workflow

Subsequently to tumour contouring, PET, T1, T2FS/STIR scans as well as the CT images from the PET/CT scans were transferred to another clinical workstation to propagate the contours from T2FS/STIR to PET and T1 scans using the commercial software MIMvista™ (MIM software, Inc., Cleveland, OH; http://www.mimsoftware.com/). This operation involved rigid registration of scans which will be detailed in the next sub-section. For now, the contour propagation steps are resumed in Figure 3.3.

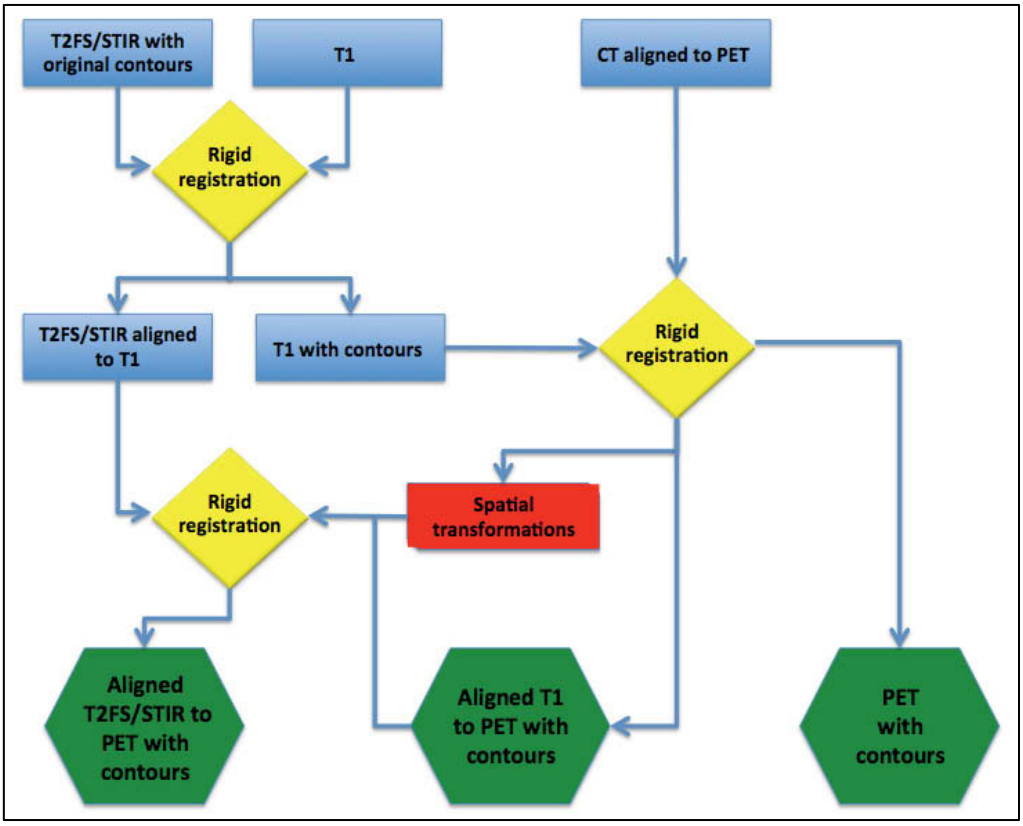

Figure 3.3. Contour propagation workflow.

1. Rigid registration of T2FS/STIR with original contours to T1.
2. Propagation of T2FS/STIR contours to T1.
3. Rigid registration of T1 with contours to CT ➔ *Aligned T1* to PET.
4. Propagation of *Aligned T1* contours to PET.
5. Rigid registration of T2FS/STIR (resulting from step 1) to *Aligned T1* using spatial transformations of step 3 ➔ *Aligned T2FS/STIR* to PET.
6. Propagation of *Aligned T1* contours to *Aligned T2FS/STIR*.

All six steps were performed using MIMvista™. In this workflow, step 3 is the most important one, as it allows to spatially relate MR scans to PET scans. This step is also the one that brings most registration errors into the process. In order to minimize registration errors between PET and MR scans, it is important to attempt the registration of two high-resolution modalities, namely CT and MR in our case. This was made possible from the fact that PET and CT are already co-registered since both scans were

acquired with patients in the same scanning position. The choice was made to perform MR and CT rigid registration using the T1 scan since the latter provides better overall anatomical contrast than T2FS/STIR. It was verified that this method brought less registration errors than the registration of T2FS/STIR to CT. T2FS/STIR scans had first to be registered to T1, but this step required minor adjustments around the tumour region only, since the different MR scans were acquired with patients in the same scanning position.

Subsequently to step 6, all scans (PET, T1 and T2FS/STIR) were spatially co-registered and the initial T2FS/STIR contours were propagated to all scans.

3.2.3 Image registration

The registration of two volumes refers to the process in which an explicit coordinate transformation is applied to one volume to spatially map all points to its corresponding points in the other volume. Rigid registration is subject to rigid body constraints only and does not involve affine transformations of volumes. It models the transformation of space as a combination of translations and rotations. In general, the transformation is performed in an iterative manner in which a measure of similarity is calculated at every iteration such that the transformation can be recursively updated until convergence of similarity scoring. MIMvista™ provides an assisted alignment tool that uses normalized mutual information (NMI) as the similarity measure. It implements an algorithm that attempts to maximize the NMI between the two volumes as defined by Pluim et al. [122].

The MIMvista™ software provides several tools to assist in the rigid alignment of two volumes. For example, one can draw a box around tumours or bony structures and maximize NMI between the two volumes only within that box. Figure 3.4 shows an example of how that tool can be used in 3D. In that figure, the patient was scanned head first in a supine position. The tumour is located in the left leg (right of the image). A box enclosing the tumour and most of the left femur was created, and NMI between the CT and T1 volumes was maximized only within that box. The registration of CT and T1 scans was performed in a similar manner for all patients. Great care was taken throughout this procedure to minimize registration errors and hence maximize the similarity between

PET and MR scans later on. Once convergence was achieved, NMI was computed between CT and T1 volumes for the tumour region only. A threshold of NMI = 0.1 was arbitrarily set up as an inclusion criterion. Three patients did not satisfy this criterion and were rejected from the study. Figure 3.5 details the final NMI between CT and T1 volumes for the 35 patients used in this study.

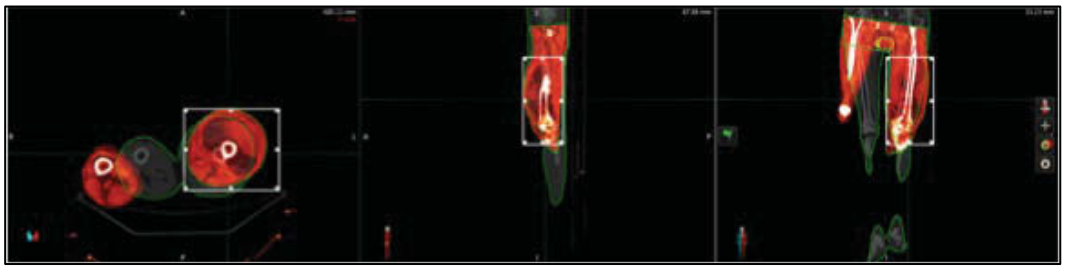

Figure 3.4. Example of registration tool.

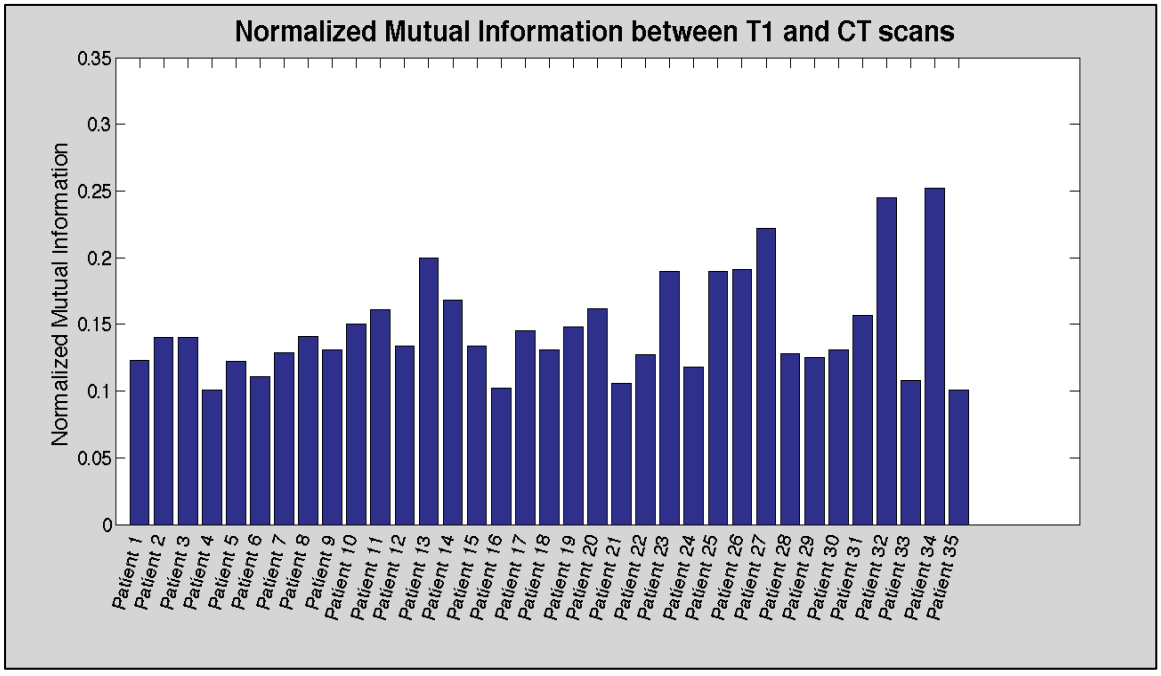

Figure 3.5. Normalized Mutual Information between T1 and CT scans.

3.3 Pre-processing of data

Prior to texture analysis, FDG-PET and MR scans were transferred into MATLAB® (MathWorks, Inc., Natick, MA; http://www.mathworks.com) format using the DICOM protocol and the research software CERR ([123]). FDG-PET scans were converted to

SUV maps using CERR and MR scans were kept in raw data form. Then, pre-processing of all scans was performed. This section details the operations performed on FDG-PET and MR scans separately.

3.3.1 FDG-PET pre-processing

As described in section 1.1, PET measurements are based on counting the number of coincident annihilation photons that were created from the disintegration of a positron and an electron. In the case of FDG-PET, the positron is emitted from the unstable ^{18}F nucleus. Likewise any radioactive decay, the disintegration rate of the ^{18}F nucleus undergoes random variations over time. Hence, the number of counts in FDG-PET measurements is also subject to random variations. These random fluctuations in photon counts constitute the main source of noise in FDG-PET image intensities and can be modeled by a Poisson distribution ([9] [10]) such that:

$$x(\mathbf{r}) \sim Poisson[\lambda(\mathbf{r})] \tag{3.1}$$

where $x(\mathbf{r})$ is the observed image, $\lambda(\mathbf{r})$ is the true image and $\mathbf{r}$ is the index vector of voxel positions $\mathbf{i},\mathbf{j},\mathbf{k}$. It can be directly seen that the Poisson noise parameter $\lambda(\mathbf{r})$ depends on the spatial location $\mathbf{r}$. Since texture analysis is concerned with the spatial distribution of gray levels across an image, noise removal algorithms should be implemented in order to remove the spatial dependence of noise across the volume and hence minimize its effect on textures. In 1948, Anscombe [124] established that a Poisson variable r with mean m could be transformed to a variable y nearly normally distributed with variance ¼ for large m if and only if:

$$y = \sqrt{r} \tag{3.2}$$

Therefore, a square root transform was applied to FDG-PET intensities in order to stabilize the variance of Poisson noise to Gaussian noise such that the spatial dependence of PET noise was minimized prior to texture analysis.

3.3.2 MR pre-processing

It is well established that textures extracted from MR images are sensitive to acquisition conditions such as variations in MR protocols, scanners and adjustments [92] [125].

These intra- and inter-acquisition variations alter the reproducibility of texture measurements on MR images [91]. In 2004, Collewet et al. [75] proposed that gray level normalization could be one way of making texture measurements on MR images more reliable. The study reports the influence of four different gray level normalization schemes on the discrimination power of texture analysis (including GLCM-based texture features) of two classes of food samples. The authors applied gray level normalization methods prior to computing texture features on proton density and T_2 - weighted MR images. They concluded that the differences between the two classes of food samples were best enhanced when using the limitation of dynamics to $\mu + 3\sigma$, where μ is the mean value and σ the standard deviation of gray levels inside the region of interest. In the current study, the latter gray level normalization scheme was applied to the tumour volume of all MR scans. Tumour voxels with intensities outside the range $\mu + 3\sigma$ were rejected and not considered in subsequent GLCM computations.

3.4 FDG-PET/MR Fusion

In this work, texture analysis was performed on 5 types of scans: PET, T1, T2FS/STIR, PET-T1 and PET-T2FS/STIR scans. PET-T1 and PET-T2FS/STIR scans come from the fusion of PET with T1 or T2FS/STIR. Textures were extracted from fused scans in the same manner as separate scans in order to investigate the capacity of these new textures to predict lung metastases in STS cancer. Two fusion methods were implemented: a simple weighted averaging technique and a wavelet-based technique. The general scheme for the fusion of PET and MR co-registered volumes is shown in Figure 3.6.

The fusion of PET and MR volumes was performed at the respective resolution of patient's PET scans (detailed in section 3.1.2). The MR volumes were downsampled to PET resolution using cubic interpolation and an antialiasing kernel from MATLAB® version R2011b. The rationale behind the fusion at the resolution of PET scans is further emphasized in results section 4.2.2, as it is shown that texture features from FDG-PET scans offer better prediction of lung metastases than MR scans. In this respect, the choice was made to leave FDG-PET image characteristics unchanged prior to fusion. The next sub-sections describe the two fusion schemes tested in this work.

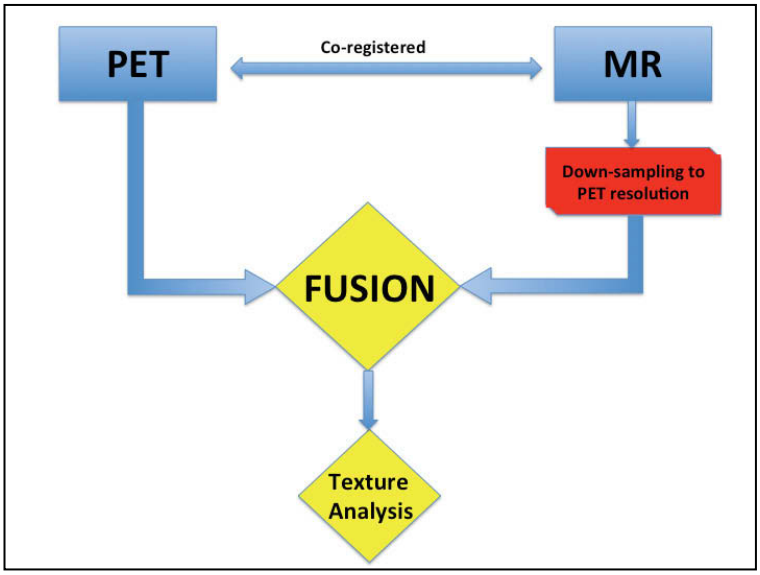

Figure 3.6. General fusion scheme.

3.4.1 Weighted Averaging

The simplest fusion scheme was implemented as follows:

1. Respectively normalize PET and MR tumour volumes to 1.
2. Combine the i^{th} voxel of PET and MR volumes with intensity j such that:

$$Fused\ Volume(i) = w * PET_j(i) + (1-w) * MR_j(i), \quad w = 0.1, 0.2, ..., 0.9. \qquad (3.3)$$

As described in Equation (3.3), nine different weights w were tested.

3.4.2 Wavelet-based fusion

The fusion scheme using the DWT was implemented as follows:

1. Respectively normalize PET and MR volumes to 1.
2. Apply the 3D DWT to PET and MR volumes up to 1 decomposition level by using the wavelet basis function *sym8*.
3. Multiply all PET and MR sub-bands of the first decomposition level by a weighting factor w_{BP} in the case of band-pass sub-bands (*LLL*, *LHL*, *LHH*, *HLL*, *HHL*, *HLH*), and a factor w_{HL} for the rest of the sub-bands (*LLL*, *HHH*). The ratio *R* is then defined as:

$$R = \frac{w_{BP}}{w_{HL}}$$
$$with \;\; 6w_{BP} + 2w_{HL} = 1 \tag{3.4}$$
$$such\ that \;\; w_{BP} = \frac{R}{2(3R+1)}, w_{HL} = \frac{1}{2(3R+1)}.$$

4. Average the wavelet coefficients PET and MR eight respective sub-bands to obtain combined wavelet coefficients at each sub-band.
5. Apply 3D IDWT to combined coefficients using the reconstruction wavelet basis function *sym8*.
6. Normalize the fused volume to 1.

The wavelet basis function *sym8* from the MATLAB® Wavelet ToolBox™ was chosen based on its compact support, biorthogonality and near-symmetric properties. Appendix C further justifies that choice by quantitatively comparing the discriminative power of the resulting fused scans constructed using other wavelet basis functions. Figure 3.7 shows an example of the fusion of PET and T2FS scans of patient 6 implemented with the DWT technique ($R = 1.5$).

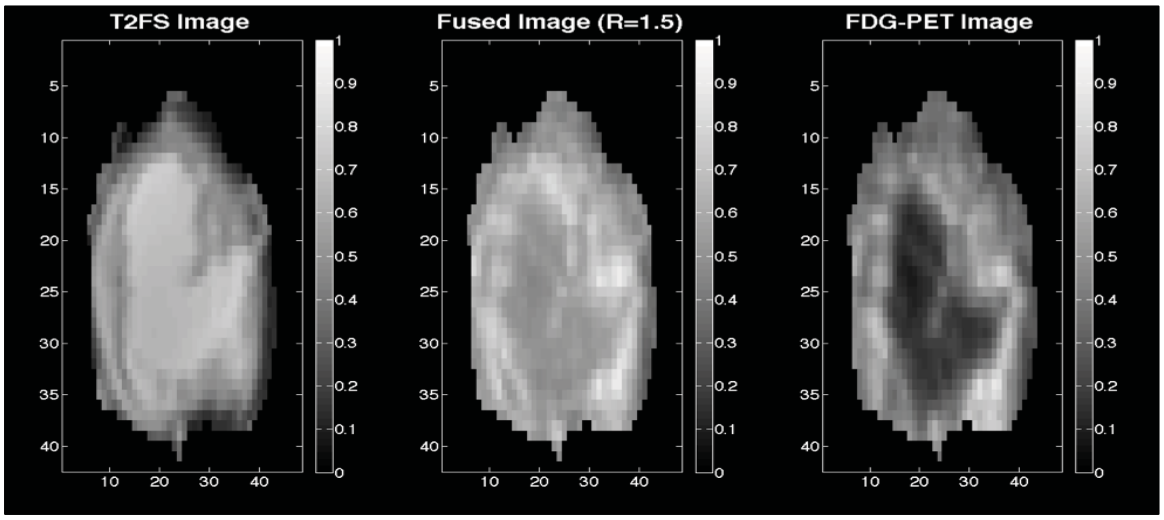

Figure 3.7. Fusion example using the DWT on PET and T2FS scans of patient 6 (Scale 3.27 mm).

R ratios $\in \{1/3,\ 1/2,\ 2/3,\ 1,\ 3/2,\ 2,\ 3\}$ were tested in this work. The rationale behind applying a different weight to band-pass sub-bands relies on the premise that the textures of an image are made up of spatially localized objects forming different patterns. Also, the spatial definition of an object can be enhanced by band-pass filtering the image with a proper filter bandwidth. The proposed scheme therefore attempts to enhance the spatial definition of fused FDG-PET/MR textures by modifying their corresponding frequency properties in the wavelet domain. Figure 3.8 shows the wavelet band-pass frequencies enhancement effect for PET and T1 fused volumes of patient 14 with $R=1$ and $R=1.5$.

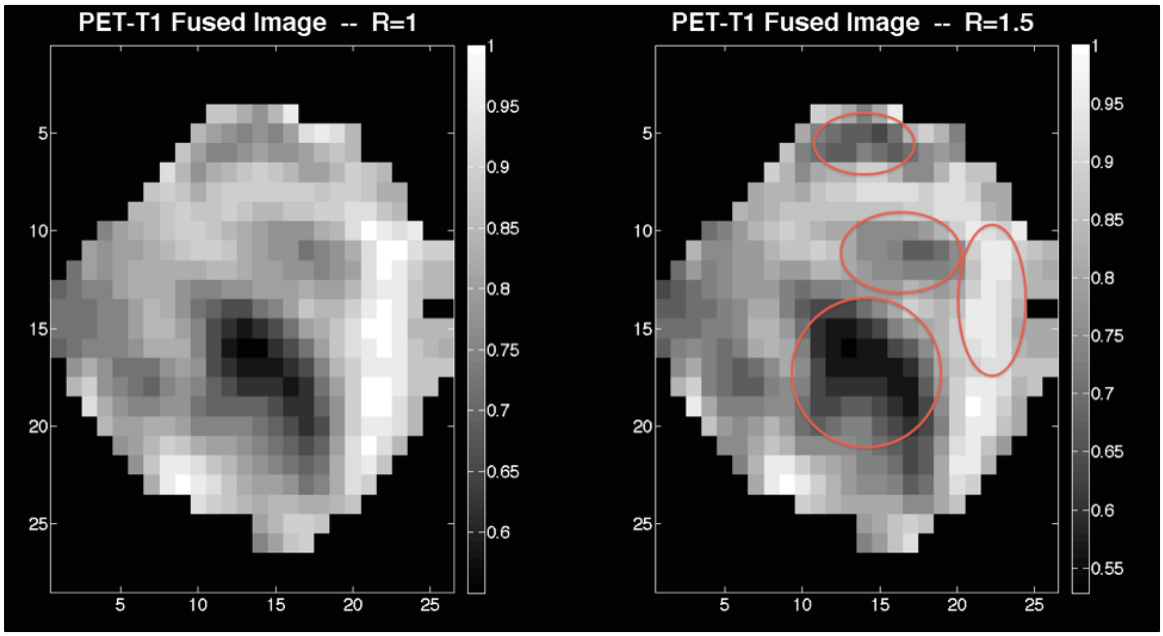

Figure 3.8. Wavelet band-pass filtering effect on PET-T1 fused scan of patient 14 (Scale 3.27 mm).

3.5 Quantization of gray levels

Prior to the computation of the GLCM of the 5 types of volumes, volumes intensities were normalized to 1 and the full range of gray levels $\in$ [0,1] were quantized to a smaller number of gray levels N_g. The quantization process maps the voxel values of a volume to a finite set $\mathbf{r} = \{r_k \in \mathbb{R} : k = 1,2,...,N_g\}$ of *reconstruction levels* by defining a set

$\mathbf{t} = \{t_k \in \mathbb{R} : k = 1,2,...,N_g + 1\}$ of *decision levels*, with $t_1 = 0$ and $t_{N_g+1} = 1$ after normalization. If a voxel value lies in the interval $[t_k, t_{k+1})$, it is mapped to r_k.

All quantization algorithms attempt to resolve, for a given number of gray levels, the reconstruction and decision levels of an input volume. The simplest quantization scheme is called the *uniform quantizer*. Let a given volume have voxel values in the interval $[a,b]$. If the volume is uniformly quantized to N_g levels, the transition and reconstruction levels become:

$$t_k = a + \frac{(b-a)(k-1)}{N_g}, \quad \text{for } k = 1,2,...,N_g + 1$$
$$r_k = \frac{t_k + t_{k+1}}{2}, \quad \text{for } k = 1,2,...,N_g \tag{3.5}$$

Figure 3.9 shows the histogram of the tumour volume from the T2FS scan of patient 3 built using uniform quantization with 256 gray levels.

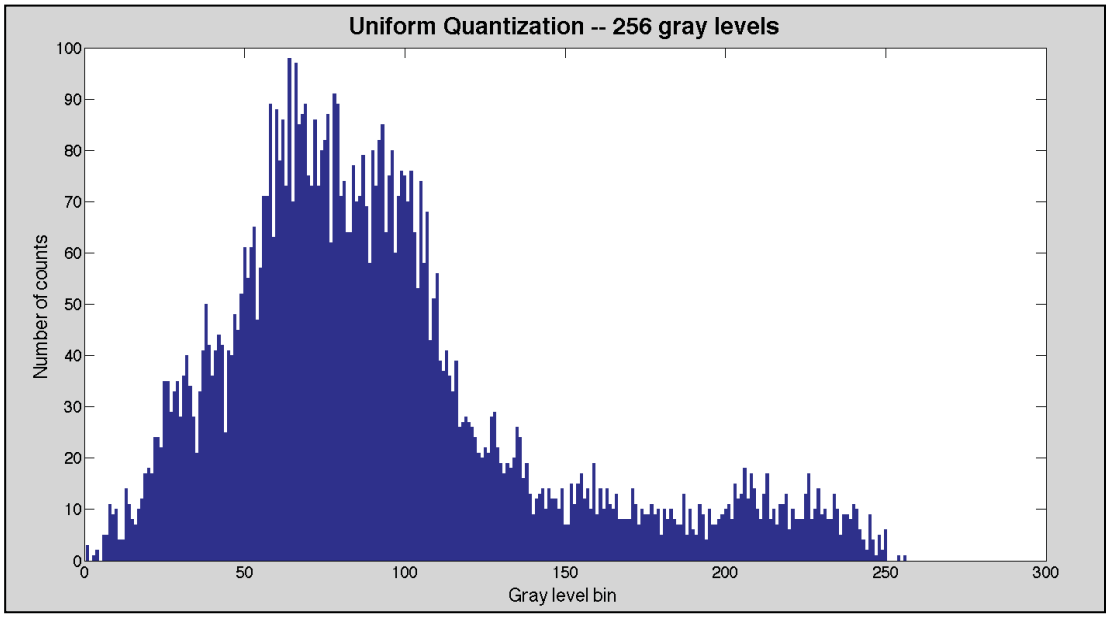

Figure 3.9. Uniform quantization of the tumour volume from the T2FS scan of patient 3 (Scale 3.27 mm).

In the last figure, the number of gray levels was intentionally put to a high number (256) to reflect the real histogram of patient 3. The next two sub-sections detail two quantization schemes investigated in this work.

3.5.1 Equal-probability quantization

In their original work on GLCM-based textures features, Haralick et al. [101] proposed that the quantization of images prior to computation of the GLCM should be done using an equal-probability quantization scheme in order for the textures to be invariant under monotonic gray-tone transformations. This quantization scheme defines decision thresholds in an image such that the number of pixels with reconstruction level r_k is the same in the quantized image for all k (i.e. for all N_g, or bin). Similarly to the *uniform quantizer*, the reconstructions levels are taken as the average of two consecutive decision levels. In this work, an *equal-probability quantization* algorithm similar to the one described by Haralick et al. [101] was implemented. Figure 3.10 shows the histogram of the tumour volume from the T2FS scan of patient 3 built using equal-probability quantization with 64 gray levels. The equal-probability quantization algorithm implemented in this work attempts to provide the same number of counts in every bin, but some discretization errors are present and can be seen in Figure 3.10.

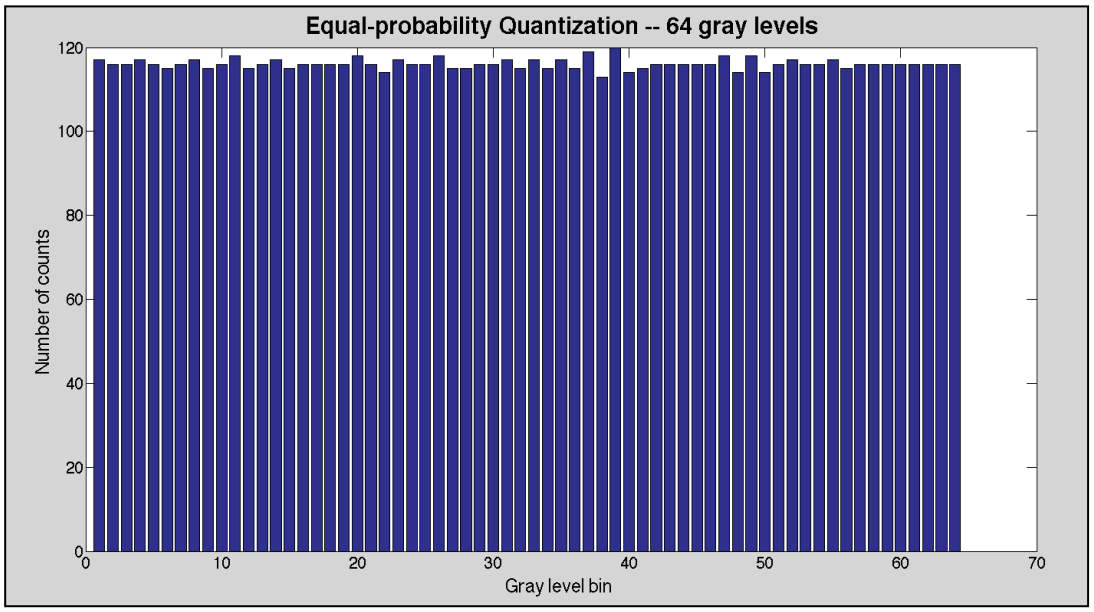

Figure 3.10. Equal-probability quantization of the tumour volume from the T2FS scan of patient 3 (Scale 3.27 mm).

3.5.2 Lloyd-Max quantization

In 1982, Lloyd [126] formulated the concept of Max [127] (of quantizing an input signal to achieve minimal distortion) into a coherent quantization theory now known as the *Lloyd-Max quantization* algorithm. Lloyd enounced his optimization criterion as the minimization of "average quantization noise power". Essentially, this scheme optimally minimizes the mean square quantization error of the output [128]. Let X be the input volume and $Q(X)$ the output of quantization. For N_g gray levels, the mean-squared error ε is:

$$\varepsilon = E\left[(X - Q(X))^2\right] = \int_{t_1}^{t_{N_g+1}} (X - Q(X))^2 \, p_X(X) \, dX \tag{3.6}$$

where $p_X(X)$ is the amplitude probability density of the input volume X. The necessary conditions for minimization of ε are obtained by differentiating Equation (3.6) with respect to the decision levels t_k and the reconstruction levels r_k. By equating to 0 and from the fact that $t_{k-1} \leq t_k$, we obtain:

$$t_k = \frac{r_k + r_{k-1}}{2} \tag{3.7}$$

$$r_k = E\left[X \middle| X \in [t_k, t_{k+1})\right] = \frac{\int_{t_k}^{t_{k+1}} X \cdot p_X(X) dX}{\int_{t_k}^{t_{k+1}} p_X(X) dX} \tag{3.8}$$

Practically speaking, Equations (3.7) and (3.8) have to be solved simultaneously (given boundary conditions t_1 and t_{N_g+1}) using an iterative scheme. In this work, this procedure was performed using the function *lloyd.m* from MATLAB®. Figure 3.11 shows the histogram of the tumour volume from the T2FS scan of patient 3 built using the Lloyd-Max quantization algorithm with 64 gray levels.

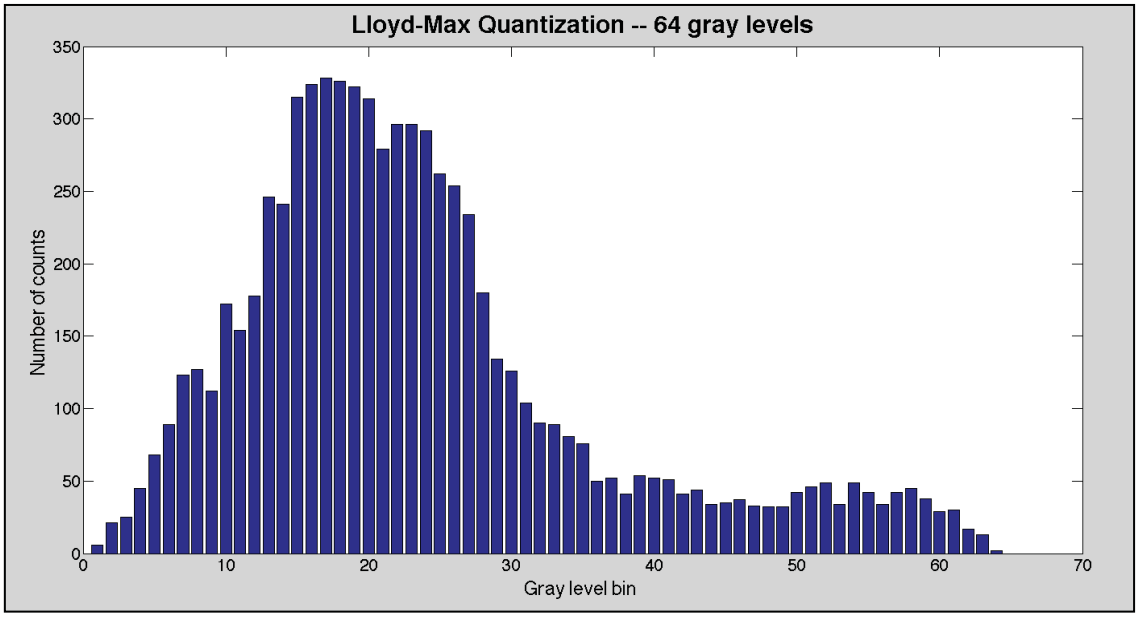

Figure 3.11. Lloyd-Max quantization of the tumour volume from the T2FS scan of patient 3 (Scale 3.27 mm).

3.6 3D texture analysis

Subsequently to the quantization procedure, textures were extracted from the 5 types of volumes used in this study. This section details the GLCM computation and gives an overview of the texture extraction workflow.

3.6.1 3D gray-level co-occurrence matrix computation

The gray-level co-occurrence matrix (GLCM) describes the frequencies of occurrence of the different pairs of gray levels i and j separated by a distance d and an angle θ in an image. In 3D discrete space, the voxel connectivity description can take one of the three forms shown in Figure 3.12.

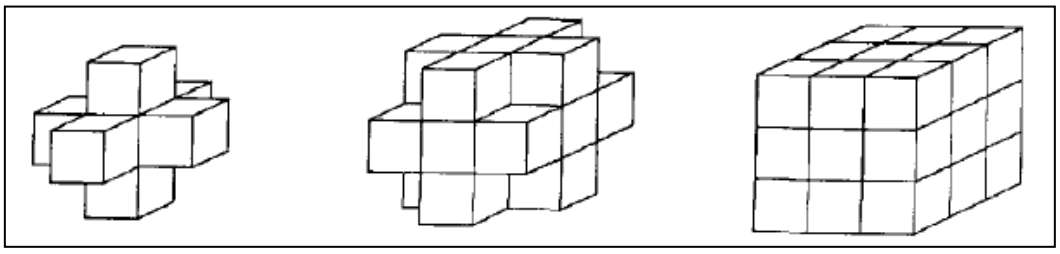

Figure 3.12. Voxel connectivity in 3D space (taken from[129]). Left: 6-voxel connectivity around the center voxel, Middle: 18-voxel connectivity around the center voxel, Right: 26-voxel connectivity around the center voxel.

In this work, the GLCMs were computed using 26-voxel connectivity by applying a 3D generalization of the theory presented in section 2.1. The directional dependence θ was not taken into account. Therefore, the GLCMs computed in this work represent the sum of occurrences of pairs of gray levels for the 26 directions around a center voxel such that the computation mapped all neighbours of all voxels in the tumour. Furthermore, the distance parameter d was set to 1, similarly to other related works ([97] [99] [100]). For this parameter to be meaningful in 3D space, isotropic voxel size is needed. Hence, prior to quantization and GLCM computation, the 5 volumes were resampled using cubic interpolation to an isotropic voxel size set to the desired scale. As a final step prior to texture features extraction, the GLCMs were normalized by the sum of the values at all positions (i, j). Figure 3.13 shows a 3D representation of two GLCMs with 32 gray levels computed using Lloyd-Max quantization algorithm from the PET-T1 fused scans of patients 11 (MetsLungs) and 17 (no MetsLungs).

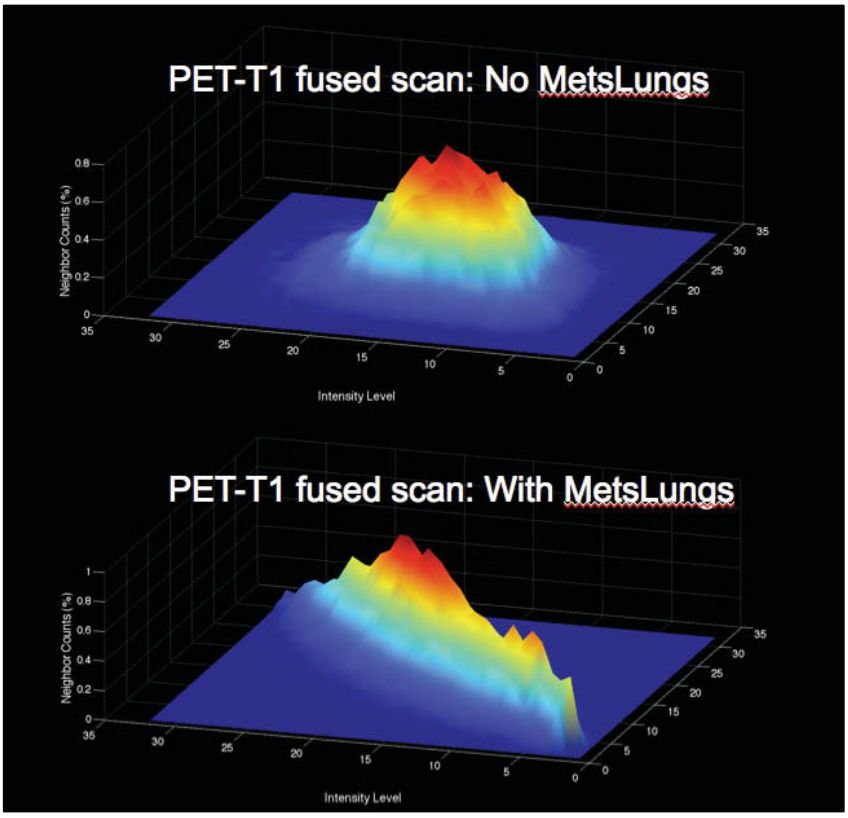

Figure 3.13. 3D representation of the GLCM from PET-T1 fused scan (R = 1.5, Scale 3.27 mm) of patient 11 (MetsLungs) and patient 17 (no MetsLungs). Percentages of occurrence of the different pairs of gray levels are reported for illustrative purposes.

3.6.2 Texture extraction workflow

From the normalized GLCMs with N_g gray levels computed after quantization of the 5 types of volumes with isotropic voxel size resampled at a desired scale, 6 texture features were extracted according to the mathematical operations shown in Table 2.1. The workflow from the resampling of scans to texture feature extraction is resumed in Figure 3.14.

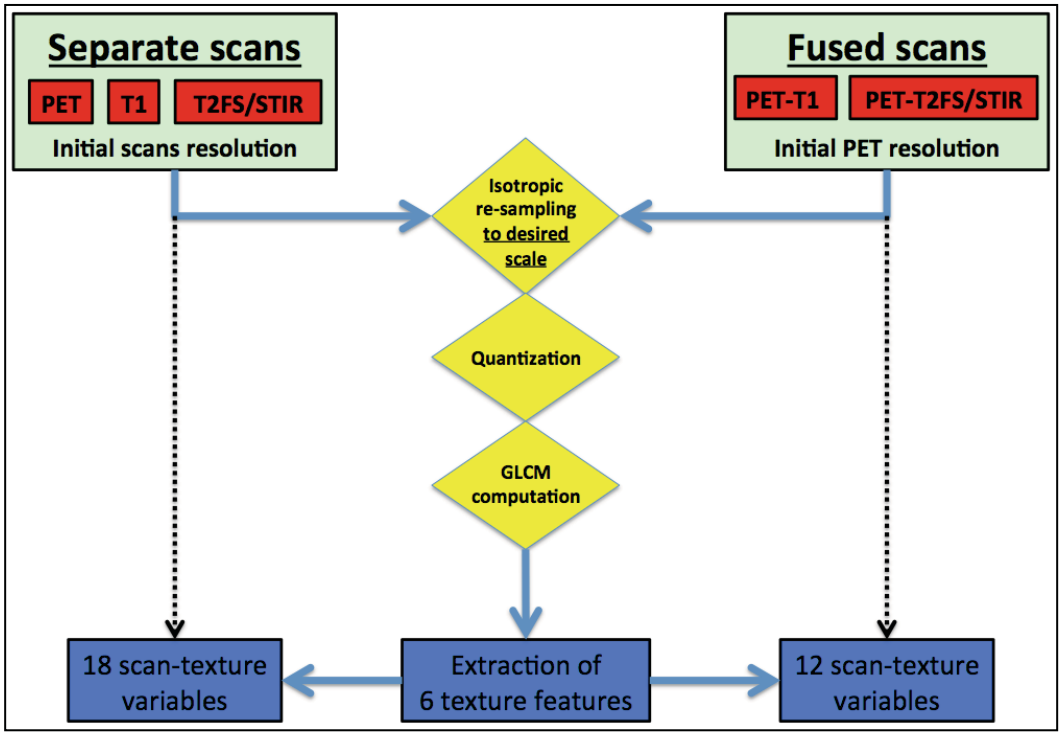

Figure 3.14. Workflow of texture feature extraction.

In this work, the effect of 7 different scales on the prediction performance was investigated: 1.64 mm, 2.18 mm, 2.45 mm, 3.27 mm, 4.36 mm, 4.91 mm and 6.54 mm.

3.7 Univariate analysis

Up to this point, 18 scan-texture combinations were extracted from separate scans and 12 scan-texture combinations from fused scans. In addition, SUV_{max} was extracted from FDG-PET tumour volumes and was added to the feature set. Then, the association between the different features and lung metastases was investigated. The association

between the vector $\mathbf{f}_j = \{x_{ij} \in \mathbb{R} : i = 1,2,...,n\}$ and the outcome vector $\mathbf{y} = \{y_i : i = 1,2,...,n\}$, where x_{ij} is the j^{th} input variable for the i^{th} patient with $n = 35$ and $y_i \in \{0 : \text{No MetsLungs}, 1 : \text{MetsLungs}\}$, was calculated with Spearman's rank correlation by using the function *corr.m* of MATLAB®. The choice was made to include all patients from Table 3.1 in the analysis even though the ultimate goal is to find relevant image characteristics for the prediction of lung metastases at diagnosis of STS of the extremities. In other words, patients 11, 20, 24 and 30 were included in the univariate and multivariable analysis. In addition to the primary sarcoma, patients 11 and 20 were diagnosed with lung metastases at presentation, and patients 24 and 30 were diagnosed with bone metastases at presentation. It was verified that the exclusion of patients with lung metastases at presentation generally lowered significantly the Spearman's coefficients of textures extracted from separate scans. On the other hand, the exclusion of patients with bone metastases at presentation generally did not significantly affect the Spearman's coefficients of textures extracted from separate scans. Appendix D presents the details of the latter analysis. These results tend to prove that patients diagnosed with lung metastases at presentation possess similar image characteristics as those that developed lung metastases during the follow-up period. As a consequence, all patients were included in the study.

3.8 Multivariable analysis

In addition to univariate analysis, linear combinations of features were investigated for the prediction of lung metastases. Multivariable models were built using forward selection and logistic regression with 19 variables in the case of separate scans (18 scan-textures + SUV_{max}) and 13 variables in the case of fused scans (12 scan-textures + SUV_{max}). In this work, many parameters had to be optimized in order to identify a general model with best prediction performance. In this respect, this section first details the tests performed to identify the optimal image analysis parameters. Finally, the methodology used to identify a general multivariable model and to evaluate its prediction performance is described.

3.8.1 Optimization of parameters

This sub-section describes the methodology used to identify the optimal quantization algorithm and optimal number of gray levels to be used in subsequent analysis. Two quantization algorithms were tested: equal-probability and Lloyd-Max quantization. In addition, four numbers of gray levels $N_g \in \{8, 16, 32, 64\}$ were tested. In order to identify the optimal quantization scheme, three metrics were used.

METRIC 1: *rho5Best*

$$rho5Best = \frac{1}{5}\sum_{i=1}^{5} \sup_5 \left| Spearman's\ Coefficient(i) \right| \tag{3.9}$$

For a given model order, all possible feature combinations were tested. For every combination, logistic regression was applied on the entire dataset (35 patients) in order to find the regression coefficients of $g(\mathbf{x}_i)$. Spearman's coefficient was then computed between the vector of the linear combination values $\mathbf{g} = \{g(\mathbf{x}_i) : i = 1,2,...,n\}$ and the outcome vector $\mathbf{y}$. The value *rho5Best* was then calculated as the average of the 5 highest absolute Spearman's coefficients. This metric can be used only to evaluate the performance of models for different quantization schemes on training data (here, the entire dataset).

METRIC 2: *rhoBoot*

$$rhoBoot = \sum_{i=1}^{Number\ of\ Models} Percentage\ Frequency(i) * Spearman's\ Coefficient(i) \tag{3.10}$$

For every quantization scheme tested and a given model order, one *rhoBoot* was computed. First, the entire dataset was separated into a training set and a testing set using bootstrapping. The process was repeated for 1000 bootstrap samples. For every bootstrap sample, variables were selected from the training set by using the open-source software DREES [130]. For a given model order, the software implements an efficient forward selection scheme, in which the stepwise statistical significance of added regression coefficients is determined via the maximization of the Wald's test. Subsequently to all bootstrap runs, the frequency by which each group of variables was selected was again

calculated using DREES. For this purpose, the software implements a *model coalescing* step, in which "models differing in terms of variables that are correlated by greater than a threshold Spearman's coefficient are reduced to the same model" [130]. Then, for all *coalesced models* (or group of variables) and up to 5 models, logistic regression was applied on the entire dataset using DREES in order to find the regression coefficients and Spearman's coefficient between **g** and **y** was computed. The metric *rhoBoot* was then calculated using Equation (3.10). For different quantization schemes, this metric allows to weight the goodness-of-fit on the entire dataset based on the dominance of group of variables found in bootstrap training sets.

METRIC 3: *Stability*

$$Stability = \sum_{i=1}^{Number\ of\ Models} (Percentage\ Frequency(i))^2 \tag{3.11}$$

For every quantization scheme tested and a given model order, one *stability* metric was computed. First, the dataset was separated into training and testing sets using bootstrapping for 1000 bootstrap samples. For every bootstrap sample, variables were selected in the training set using the forward selection scheme described earlier. Subsequently to all bootstrap runs, the frequency by which each group of variables was selected was again calculated using the same method described earlier. The metric *stability* was then calculated using Equation (3.11). A given quantization scheme is considered stable if the same group of variables are often picked up in bootstrap training sets. The premise is that for a given model order, a quantization scheme with higher stability will perform better on unseen data.

3.8.2 Evaluation of prediction performance of multivariable models

This sub-section presents the methodology used to evaluate the prediction performance of imaging feature models. Essentially, performance metrics were calculated in different testing sets from models built in different training sets for a high number of bootstrap samples. First, the methodology used to estimate the optimal order of a prediction model is described here:

Model Order Estimation

1. For model orders 1 to 10, apply bootstrap technique to divide the dataset into training and testing sets for 1000 bootstrap samples.
2. For every bootstrap sample of a given model order, select variables and their corresponding regression coefficients in the training set using forward selection and logistic regression algorithms from DREES. Calculate Spearman's coefficient between the vector of the linear combination values $\mathbf{g}_{train} = \{g(\mathbf{x}_i) : i \in \text{training set}\}$ obtained from the regression coefficients found in the training set and the outcome vector $\mathbf{y}_{train} = \{y_i : i \in \text{training set}\}$.
3. For a given model order, average Spearman's coefficients of the 350 bootstrap runs.

Finally, the methodology used to evaluate the performance of prediction models for situations involving different fusion schemes, *R* ratios, scales and model orders is described here:

Prediction Performance Evaluation

1. Apply bootstrap technique to divide the dataset into training and testing sets for 10000 bootstrap samples. For every bootstrap run and for a given model order, select variables and their corresponding regression coefficients in the training set using the forward selection and logistic regression algorithms from DREES
2. Calculate the frequency of occurrence of group of variables. Pick up the most frequent group of variables.
3. Apply bootstrap technique to divide the dataset into training and testing sets for another 10000 bootstrap samples. For every bootstrap run, use the most frequent group of variables found in step 2 and calculate its corresponding regression coefficients in the training set. Use the regression coefficients found in the training set to calculate a vector of linear combination values $\mathbf{g}_{test} = \{g(\mathbf{x}_i) : i \in \text{testing set}\}$. Use $\mathbf{g}_{test}$ and the outcome vector $\mathbf{y}_{test} = \{y_i : i \in \text{testing set}\}$ in order to calculate

performance metrics in the testing set, namely AUC and MCC. Average the performance metrics of the 10000 bootstrap runs and use their means in order to rank the different situations.

The only constraint set for the choice of training and testing sets was that their size had to be greater than 1. Without other constraints (e.g., keeping the proportion of patients with lung metastases the same in the two sets), the bias inherent to resampling was minimized. MCC was calculated using Equation (2.25) and a threshold of 0 as defined by the offset in Equation (2.14). AUC was calculated using the function *perfcurve.m* of MATLAB®. Essentially, the whole *prediction performance evaluation* methodology allows to identify an optimal model for a given situation and to thereafter simulate its prediction performance on unseen data using bootstrap testing sets.

CHAPTER 4: RESULTS

4.1 Optimization of quantization scheme

The presentation of results starts with the optimization of the quantization algorithm and the number of gray levels. The experiments performed in this section had for goal to identify a quantization scheme of choice before proceeding to other experiments presented in sections 4.2 and 4.3.

4.1.1 Quantization algorithm

Table 4.1 presents correlation results obtained with equal-probability quantization, and Table 4.2 presents those obtained with Lloyd-Max quantization. The tests were performed using the 19 variables extracted from separate scans at scale 3.27 mm for both MR and PET volumes. All combinations of the number of gray levels $N_g \in \{8, 16, 32, 64\}$ were investigated for the MR and PET volumes. For every situation, the three metrics described in section 3.8.1 were calculated for a linear combination of 5 variables.

Table 4.1. Performance of Equal-Probability quantization (model order 5)

N_g		Metric		
MR	**PET**	***rho5Best***	***rhoBoot***	***Stability***
8	8	0.8267	0.7886	0.20
8	16	0.8267	0.7967	0.21
8	32	0.8267	0.8021	0.27
8	64	0.8267	0.7916	0.21
16	8	0.8173	0.7566	0.22
16	16	0.8197	0.7668	0.20
16	32	0.8197	0.7822	0.22
16	64	0.8197	0.7716	0.23
32	8	0.8091	0.7661	0.29
32	16	0.8103	0.7692	0.30
32	32	0.8091	0.7747	0.33
32	64	0.8103	0.7506	0.30
64	8	0.7822	0.7695	0.30
64	16	0.7843	0.7711	0.28

64	32	0.7857	0.7774	0.27
64	64	0.7857	0.7693	0.20

Table 4.2. Performance of Lloyd-Max quantization (model order 5)

N_g		Metric		
MR	**PET**	***rho5Best***	***rhoBoot***	***Stability***
8	8	0.7682	0.7651	0.70
8	16	0.7658	0.7521	0.65
8	32	0.767	0.7668	0.74
8	64	0.7611	0.7560	0.56
16	8	0.8009	0.7907	0.59
16	16	0.7974	0.7856	0.54
16	32	0.7986	0.7864	0.59
16	64	0.7869	0.7725	0.59
32	8	0.7693	0.7413	0.29
32	16	0.7857	0.7670	1
32	32	0.7857	0.7809	0.70
32	64	0.7623	0.7247	0.32
64	8	0.7682	0.7571	0.57
64	16	0.7693	0.7611	0.43
64	32	0.7728	0.7538	0.50
64	64	0.7529	0.7329	0.35

A direct comparison between Table 4.1 and Table 4.2 shows that metrics *rho5Best* and *rhoBoot* appear similar in both quantization algorithms, although *rho5Best* is slightly higher for equal-probability quantization. However, prediction models from Lloyd-Max quantization appear much more stable than those from equal-probability quantization. We believe that the stability of prediction models also plays a very important role in the prediction performance on unseen data. Therefore, the choice was made to use Lloyd-Max quantization algorithm for all subsequent analysis.

4.1.2 Number of gray levels

Next, other tests were performed in order to identify the optimal number of gray levels to be used in the Lloyd-Max quantization algorithm. From Table 4.2, it can be seen that 16

gray levels in MR volumes allows for the optimization of *rho5Best* and *rhoBoot*. The number of gray levels for MR volumes was thus chosen to be 16. Table 4.3 presents the tests performed to thereafter optimize the number of gray levels in PET volumes. To perform such operation, the procedure described in section 4.1.1 was repeated for lower model orders. The tests were performed using the 19 variables extracted from separate scans at scale 3.27 mm for both MR and PET volumes. For every situation, the three metrics described in section 3.8.1 were calculated for linear combinations of 4 and 3 variables.

Table 4.3. Performance of Lloyd-Max quantization (model orders 4 and 3)

N_g		Metrics					
		Order 4			Order 3		
MR	**PET**	***rho5Best***	***rhoBoot***	***Stability***	***rho5Best***	***rhoBoot***	***Stability***
16	8	0.7295	0.7205	0.3397	0.6909	0.6855	0.2274
16	16	0.7389	0.7194	0.3939	0.6979	0.6845	0.3285
16	32	0.7377	0.7288	0.3857	0.6991	0.6865	0.3359
16	64	0.7330	0.7220	0.3678	0.6956	0.6844	0.2932

From Table 4.3, it can be seen that 16 and 32 gray levels in PET appear slightly better than 8 and 64 gray levels. However, results are about the same between 16 and 32 gray levels. At this point, the choice was made to consistently use 16 gray levels for MR, PET and fused PET/MR scans, such that a direct comparison could be established in subsequent analysis.

To recapitulate the results obtained from section 4.1, the tests that were performed allowed to identify Lloyd-Max algorithm with 16 gray levels as the quantization method of choice for the application presented in this study. This quantization scheme was used for all experiments presented in the next sections.

4.2 Univariate results

This section details the univariate results of all features extracted from separate and fused scans. First, the correlation of SUV_{max} with lung metastases is shown. Then, the correlation of univariate texture features with lung metastases is presented for both separate and fused scans. Two situations are shown for both separate and fused scans: 1) the situation yielding the best average univariate correlation results; and 2) the situation optimized for the best multivariable prediction performance. For the sake of clarity, all univariate results are presented together, although the justification for the optimization of prediction performance will be presented in section 4.3.

4.2.1 SUV_{max}

The correlation between the vector of SUV_{max} values and the outcome vector **y** yielded Spearman's coefficient $rho = 0.6382$ with $p < 0.0001$. Figure 4.1 presents a plot of the outcome vector **y** as a function of SUV_{max}. In that plot, each green square represents one patient of the cohort.

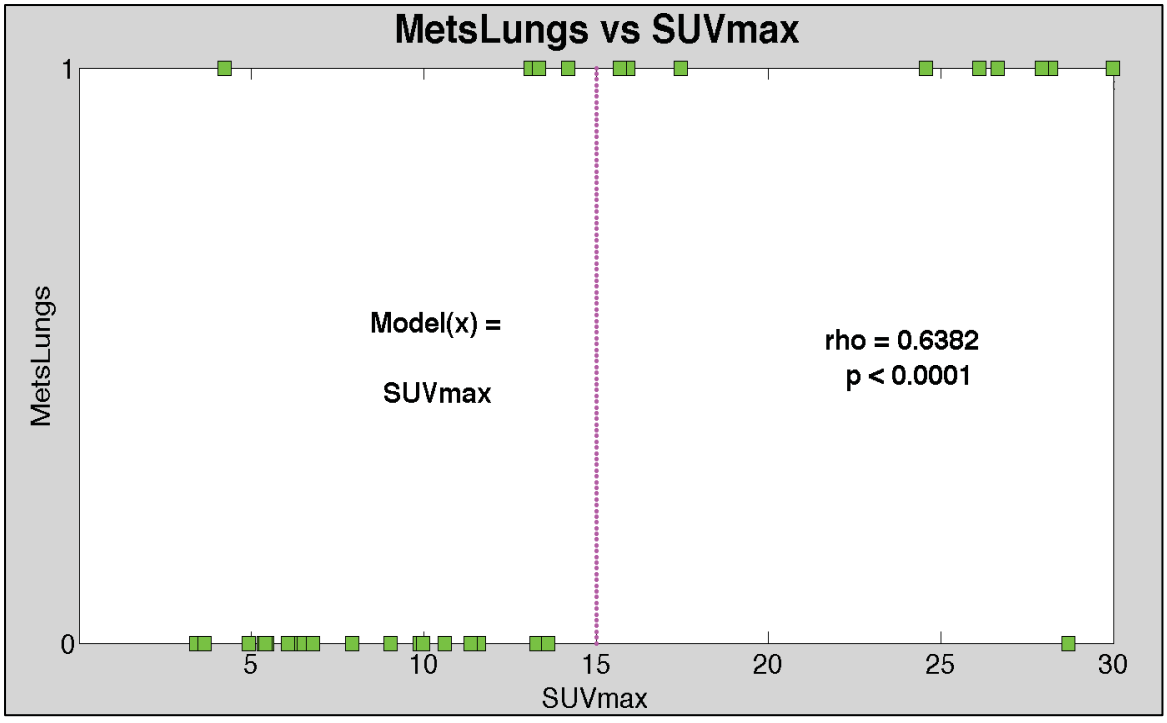

Figure 4.1. SUV_{max} correlation with lung metastases.

The results of Figure 4.1 confirm the usefulness of SUV_{max} as a predictor of lung metastases in STS of the extremities. Although not perfect, a clear separation can be seen between low (<15) and high (>15) SUV_{max} patients. One of the major goals of this work is to find complementary features to SUV_{max} in order to obtain a better separation of patients with and without lung metastases. This issue will be addressed in the multivariable analysis section 4.3.

4.2.2 Separate scans

Table 4.4 presents the univariate correlation of texture features extracted on separate scans with lung metastases. Two situations are shown: 1) scale 1.64 mm; and 2) scale 3.27 mm. For the seven scales tested in this work (1.64 mm, 2.18 mm, 2.45 mm, 3.27 mm, 4.36 mm, 4.91 mm, 6.54 mm), the first situation is the one yielding the best average of absolute Spearman's correlation coefficients, whereas the second situation is the one optimized for the best prediction performance (see section 4.3). Appendix E.1 gives more information about the average univariate correlation results at all scales.

Table 4.4. Univariate correlation of features from separate scans with lung metastases

Feature	Scale	
	1.64 mm	**3.27 mm**
SUV_{max}	$rho = 0.6382, p < 0.0001$	$rho = 0.6382, p < 0.0001$
PET--Contrast	$rho = -0.4567, p = 0.0058$	$rho = -0.3806, p = 0.0241$
PET--Entropy	$rho = -0.4567, p = 0.0058$	$rho = -0.3279, p = 0.0545$
PET--Homogeneity	$rho = 0.4450, p = 0.0074$	$rho = 0.4391, p = 0.0083$
PET--Energy	$rho = 0.3923, p = 0.0198$	$rho = 0.2986, p = 0.0814$
T1--Contrast	$rho = -0.3513, p = 0.0385$	$rho = -0.2283, p = 0.1871$
T1--Entropy	$rho = -0.3396, p = 0.0460$	$rho = 0.1874, p = 0.2812$
PET--SumMean	$rho = -0.3396, p = 0.0460$	$rho = -0.2869, p = 0.0947$
PET--Variance	$rho = -0.3103, p = 0.0696$	$rho = -0.2752, p = 0.1096$
T2FS/STIR--Contrast	$rho = -0.2693, p = 0.1177$	$rho = -0.1405, p = 0.4207$
T1--Homogeneity	$rho = 0.2693, p = 0.1177$	$rho = 0.0468, p = 0.7893$
T1--SumMean	$rho = 0.1991, p = 0.2516$	$rho = 0.1112, p = 0.5246$
T1--Variance	$rho = 0.1815, p = 0.2967$	$rho = 0.1405, p = 0.4207$
T2FS/STIR--SumMean	$rho = -0.1639, p = 0.3467$	$rho = -0.2283, p = 0.1871$
T2FS/STIR--Variance	$rho = -0.1639, p = 0.3467$	$rho = -0.1698, p = 0.3295$

T2FS/STIR--Homogeneity	$rho = 0.1347, p = 0.4406$	$rho = 0.1054, p = 0.5468$
T2FS/STIR--Entropy	$rho = -0.0761, p = 0.6639$	$rho = 0.0761, p = 0.6639$
T2FS/STIR--Energy	$rho = 0.0585, p = 0.7383$	$rho = -0.0410, p = 0.8152$
T1--Energy	$rho = 0.0527, p = 0.7637$	$rho = -0.1991, p = 0.2516$
Average absolute *rho*	0.2789	0.2274

It can be seen from Table 4.4 that the best average univariate correlation results of separate scans are not obtained at the same scale as the one optimized for the best multivariable prediction performance. Also, PET texture features generally appear superior to MR texture features.

4.2.3 Fused scans

Table 4.5 presents the univariate correlation of texture features extracted on fused scan with lung metastases. Two situations are presented: 1) scale 1.64 mm, $R = 1.5$ and 2) scale 4.91 mm, $R = 1.5$. For the seven scales (1.64 mm, 2.18 mm, 2.45 mm, 3.27 mm, 4.36 mm, 4.91 mm, 6.54 mm) and seven R ratios (1/3, 1/2, 2/3, 1, 3/2, 2, 3) tested in this work, the first situation is the one yielding the best average of absolute Spearman's correlation coefficients, whereas the second situation is the one optimized for the best prediction performance (see section 4.3). Appendix E.2 gives more information about the average univariate correlation results for all scales and R ratio combinations.

Table 4.5. Univariate correlation of features from fused scans with lung metastases

Feature	**Scale**	
	1.64 mm	**4.91 mm**
SUV_{max}	$rho = 0.6382, p < 0.0001$	$rho = 0.6382, p < 0.0001$
PET-T1--Contrast	$rho = -0.5504, p = 0.0006$	$rho = -0.4684, p = 0.0045$
PET-T1--Homogeneity	$rho = 0.5504, p = 0.0006$	$rho = 0.5035, p = 0.0020$
PET-T1--Energy	$rho = 0.4977, p = 0.0024$	$rho = 0.1874, p = 0.2812$
PET-T1--Entropy	$rho = -0.4684, p = 0.0045$	$rho = -0.1991, p = 0.2516$
PET-T2FS/STIR--Homogeneity	$rho = 0.4684, p = 0.0045$	$rho = 0.4508, p = 0.0066$
PET-T2FS/STIR--Entropy	$rho = -0.4508, p = 0.0066$	$rho = -0.2635, p = 0.1262$
PET-T2FS/STIR--Contrast	$rho = -0.4450, p = 0.0074$	$rho = -0.4333, p = 0.0093$
PET-T2FS/STIR--Energy	$rho = 0.4391, p = 0.0083$	$rho = 0.1639, p = 0.3467$
PET-T2FS/STIR--Variance	$rho = -0.4333, p = 0.0093$	$rho = -0.4040, p = 0.0161$

PET-T2FS/STIR--SumMean	*rho* = -0.4157, *p* = 0.0130	*rho* = -0.3864, *p* = 0.0219
PET-T1--SumMean	*rho* = -0.3454, *p* = 0.0421	*rho* = -0.2693, *p* = 0.1177
PET-T1--Variance	*rho* = -0.3045, *p* = 0.0754	*rho* = -0.2225, *p* = 0.1989
Average absolute *rho*	0.4621	0.3535

It can be seen from Table 4.5 that the best average univariate correlation results of fused scans are not obtained at the same scale as the one optimized for the best multivariable prediction performance. However, the best average univariate correlation results of fused scans are obtained at the same scale as those of separate scans.

4.2.4 Summary

This sub-section presents a summary of the last 3 sub-sections. A direct comparison between univariate correlation results of SUV_{max}, separate scans and fused scans is made. Figure 4.2 presents the results in the case of the best univariate conditions. Figure 4.3 presents the results in the case of the best prediction performance conditions.

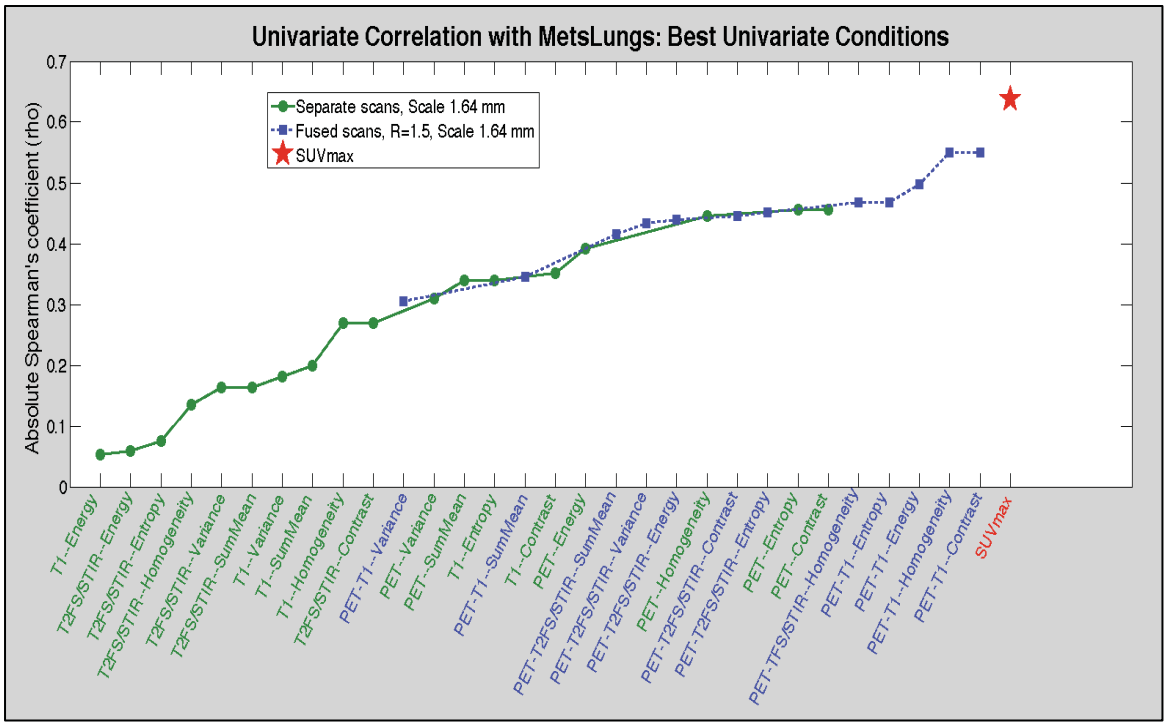

Figure 4.2. Overview of univariate correlation results, best univariate conditions.

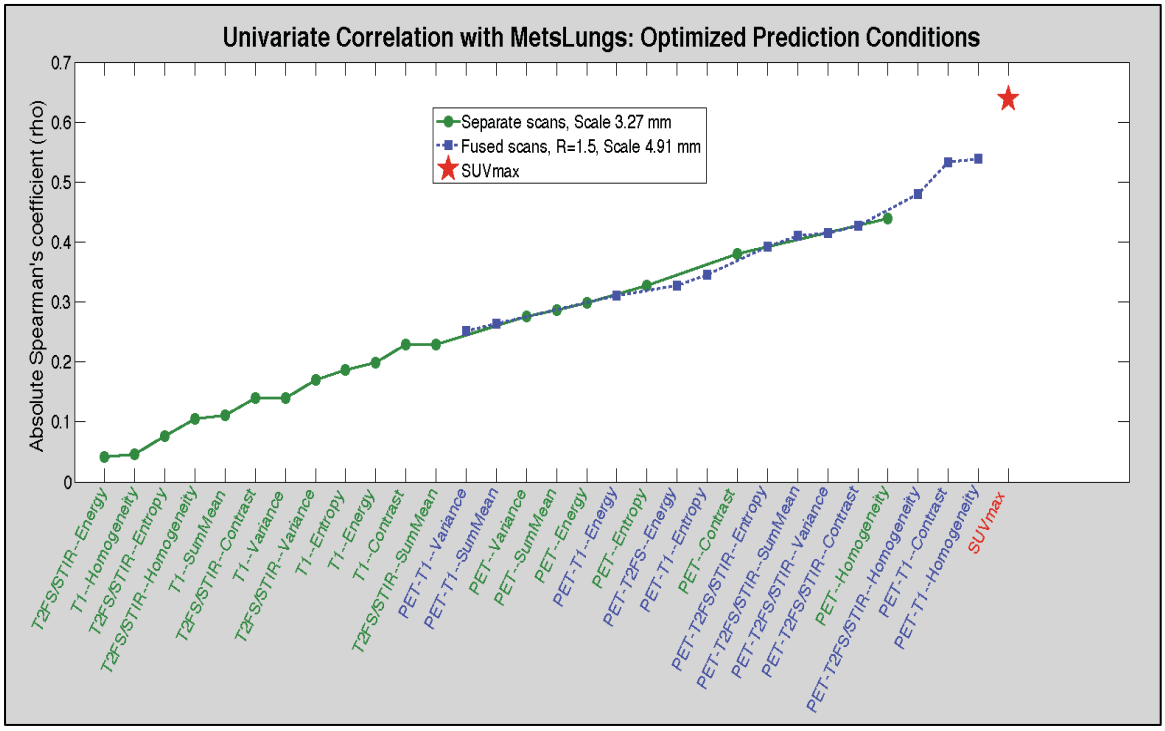

Figure 4.3. Overview of univariate correlation results, best prediction performance conditions.

Figures 4.2 and 4.3 show that on a univariate basis, texture features extracted from fused scans appear superior to those from separate scans for predicting lung metastases in STS of the extremities. However, all texture features from separate or fused scans have lower predicting power than SUV_{max}. Consequently to these results, one of the major goals of this work has been to build a linear combination of features including SUV_{max} in order to significantly improve the prediction of the clinical endpoint of interest.

4.3 Multivariable results

This section presents the majority of the work performed in this study. By combining logistic regression, forward selection and bootstrapping statistical techniques, the multivariable analysis had for goal to identify a general model that linearly combines imaging features to be tested on future patient cohorts for the prediction of lung metastases in STS of the extremities. The prediction performance was evaluated using the methodology described in section 3.8.2. First, the work done using separate and fused scans is separately presented. Then, an overview of results is presented to finally terminate with the proposal of a general multivariable model. Please note that the

confidence intervals on all tables and graphs of this section represent the standard error of the mean (*SEM*). This metric estimates the standard deviation of the sampling distribution of the mean. For a 95.5% confidence interval, $SEM = 2\sigma / \sqrt{N}$, where σ is the standard deviation of measurements and N is the number of bootstrap samples.

4.3.1 Separate scans

The prediction performance of the features from separate scans as a function of scale for model orders 2 to 5 is first presented in Figure 4.4.

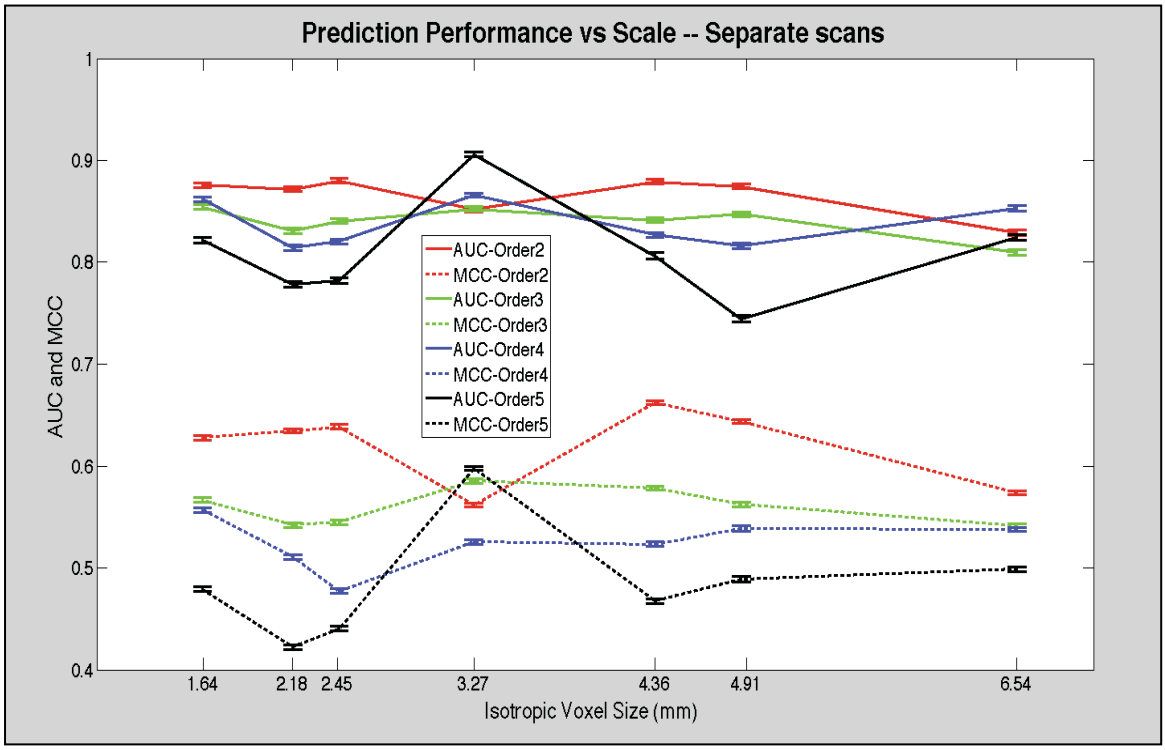

Figure 4.4. Prediction performance of separate scans as a function of scale.

The results of Figure 4.4 do not allow to clearly identify a scale at which separate scans have the best overall prediction performance. However, we believe that a higher weight should be given to the AUC metric over MCC. Higher AUC indicates a better separation between positive and negative instances of a classifier. On the other hand, MCC yields a balanced value between the sensitivity, specificity and accuracy properties of the classifier but does not indicates how well positive and negative instances are separated from each other. Hence, scale 3.27 mm was chosen as the optimal scale for

separate scans since it yields the highest AUC with acceptable MCC compared to other scales.

Next, model order estimation was performed to find the model order providing the best parsimonious model that prevents underfitting and overfitting. Results of the simulation are shown in Figure 4.5. The procedure was executed using features extracted from separate scans at scale 3.27 mm by following the methodology described in the *model order estimation* part of section 3.8.2.

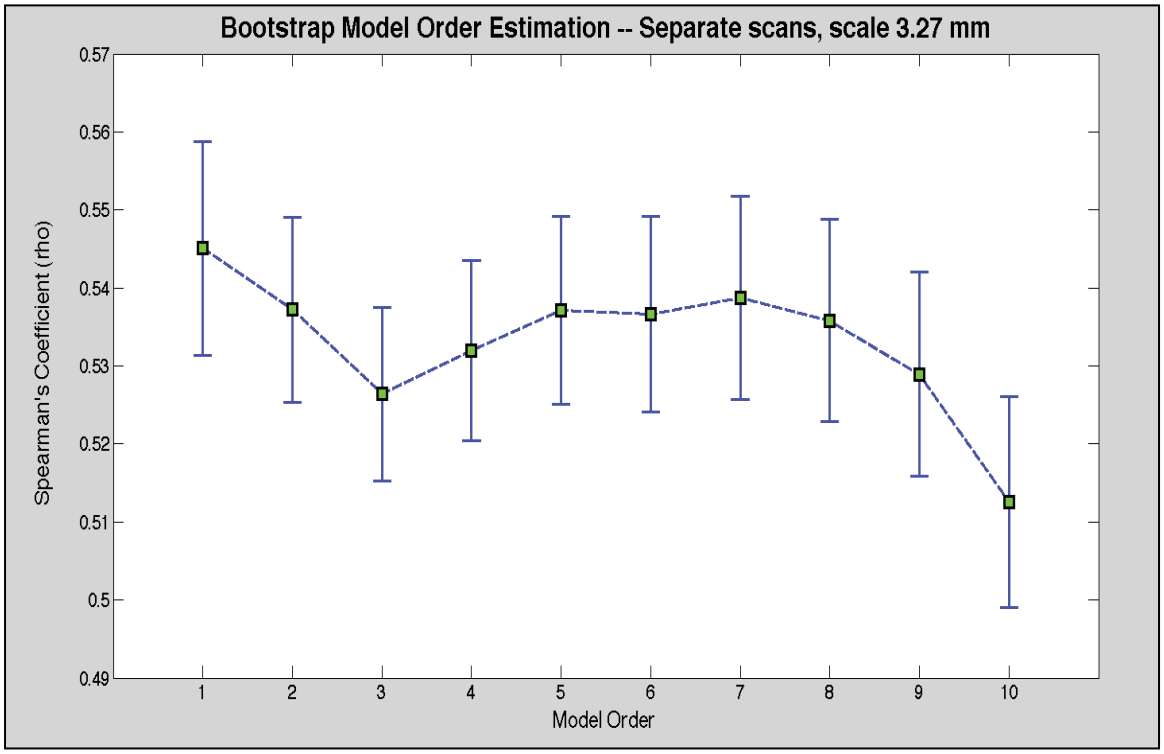

Figure 4.5. Model order estimation for separate scans (scale 3.27 mm).

The model order estimation method does not allow to find an optimal model order for separate scans. Ideally, a peak would have been found at the model order providing the optimal compromise between underfitting and overfitting. This abnormality is probably caused by the feature selection method, which does not seem to efficiently pick up the dominant features of separate scans.

Then, for the sake of completeness, a plot of the prediction performance of separate scans at scale 3.27 mm as a function of model order is presented in Figure 4.6. Table 4.6 then shows the most frequent group of variables used to perform the simulations of Figure 4.6 as defined in the *prediction performance evaluation* part of section 3.8.2.

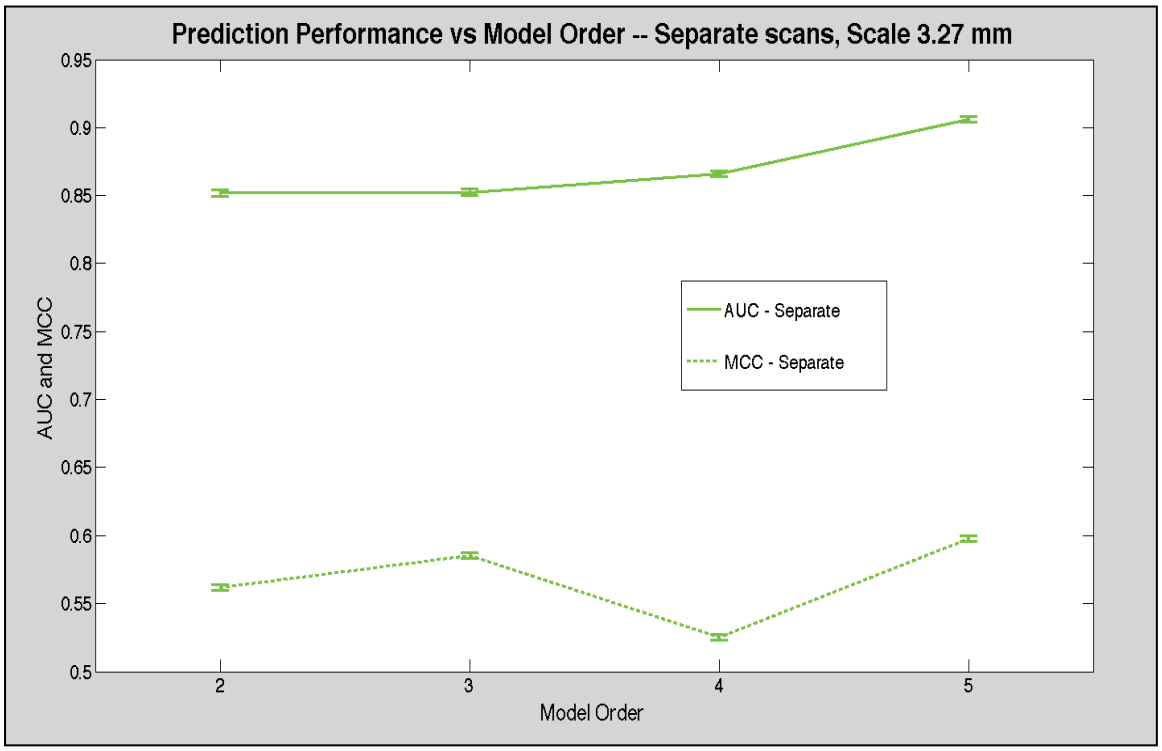

Figure 4.6. Prediction performance of separate scans (scale 3.27 mm) as a function of model order.

Table 4.6. Most frequent group of variables of separate scans (scale 3.27 mm)

Model Order	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	SUV_{max}	–	–	–	–
2	SUV_{max}	T1--SumMean	–	–	–
3	SUV_{max}	T1--SumMean	PET--Homogeneity	–	–
4	SUV_{max}	T1--SumMean	PET--Homogeneity	T1--Homogeneity	–
5	SUV_{max}	T1--SumMean	PET--Homogeneity	T1--Homogeneity	T1--Contrast

According to Figure 4.6, a model order of 5 has the best prediction performance for features extracted from separate scans at scale 3.27 mm. In order to complete the best linear model for separate scans, the regression coefficients for each of the variables at order 5 that are presented in Table 4.6 were computed as follows:

1. Apply bootstrap technique to divide the dataset into training sets and testing sets for 1000 bootstrap samples.
2. For every bootstrap run, apply logistic regression in the training set in order to find regression coefficients.
3. For each model variable, compute the mean of the corresponding regression coefficients over the 10000 bootstrap runs.

The logistic regression computation in DREES does not involve any type of regularization of the magnitude of regression coefficients. As a consequence, some coefficients were found to be very large over the 10000 bootstrap runs. To overcome this problem, coefficients with absolute magnitude greater than a value of 1000 were rejected (approximately 0.1% of the total number were rejected) in the computation of the mean. The whole procedure was also repeated for feature values normalized in the range [0,1] prior to bootstrapping runs. The mean regression coefficients that were obtained are summarized in Table 4.7.

Table 4.7. Regression coefficients of the best model from separate scans

Features	**Regression Coefficients (mean ± SEM)**	
	Non-normalized features	**Normalized features**
SUV_{max}	1.91 ± 0.02	56.5 ±0.5
T1--SumMean	-15.2 ± 0.1	-135 ± 1
PET--Homogeneity	232 ± 2	150 ± 1
T1--Homogeneity	-466 ± 3	-328 ± 2
T1--Contrast	-9.8 ± 0.1	-86 ± 1
Offset (β_0)	269 ± 2	269 ± 2

Finally, to summarize the results obtained with separate scans, we can calculate a new vector $\mathbf{g} = \{g(\mathbf{x}_i) : i = 1,2,...,n\}$ of combined features for all patients ($n = 35$) using the regression coefficients of the non-normalized features of Table 4.7. The plot of **g** with the outcome vector **y** is shown in Figure 4.7. The complete linear model is also shown on this figure. Spearman's correlation between **g** and **y** is found to be $rho = 0.8021$ with $p < 0.0001$.

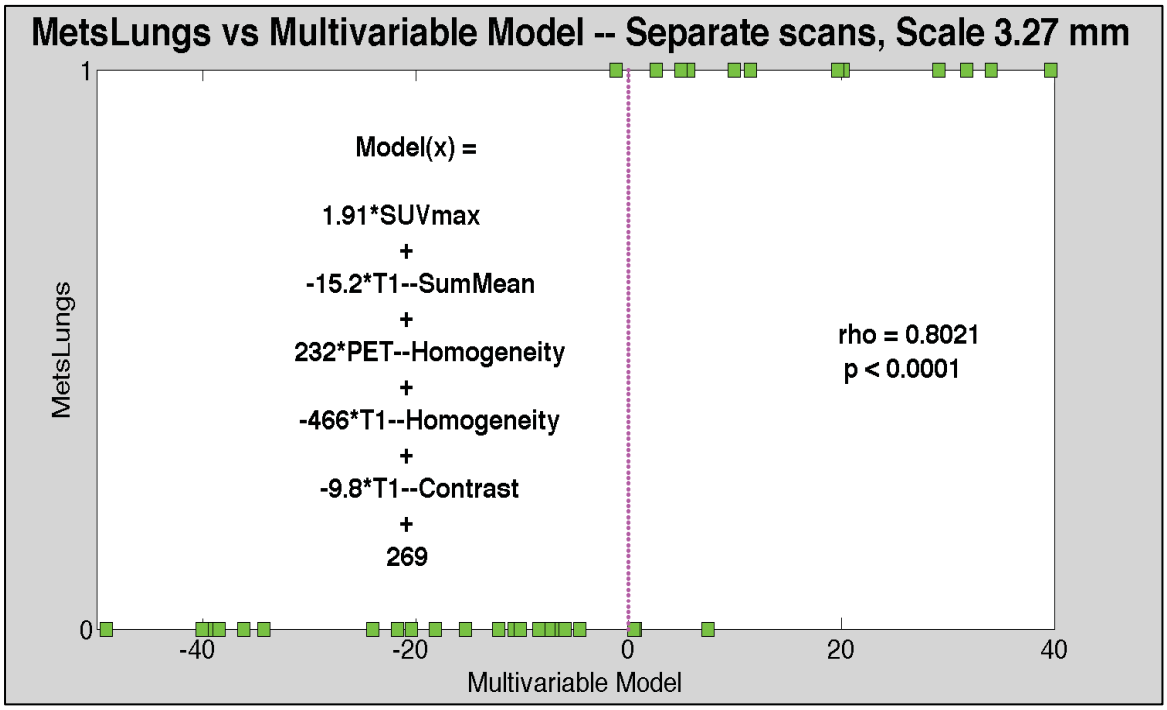

Figure 4.7. Plot of best multivariable model from separate scans.

Figure 4.7 shows that a much better separation of the two patient classes can be obtained using a linear combination of features from separate scans than with $SUV_{\max}$ alone (see Figure 4.1). The model variables shown in Figure 4.7 reached average results of $\mathrm{AUC} = 0.906 \pm 0.002$ and $\mathrm{MCC} = 0.598 \pm 0.002$ in bootstrap testing sets.

4.3.2 Fused scans

Similarly to sub-section 4.3.1, multivariable analysis using features extracted from fused scans is presented here. The optimization of fusion scheme, *R* ratio, scale and model order is described.

Figure 4.8 shows the prediction performance of features extracted from fused scans (scale 3.27 mm) computed with the weighted averaging technique described in section 3.4.1. The figure elaborates the prediction performance as a function of the weight of MR intensities in the fusion process for model orders 2 to 5.

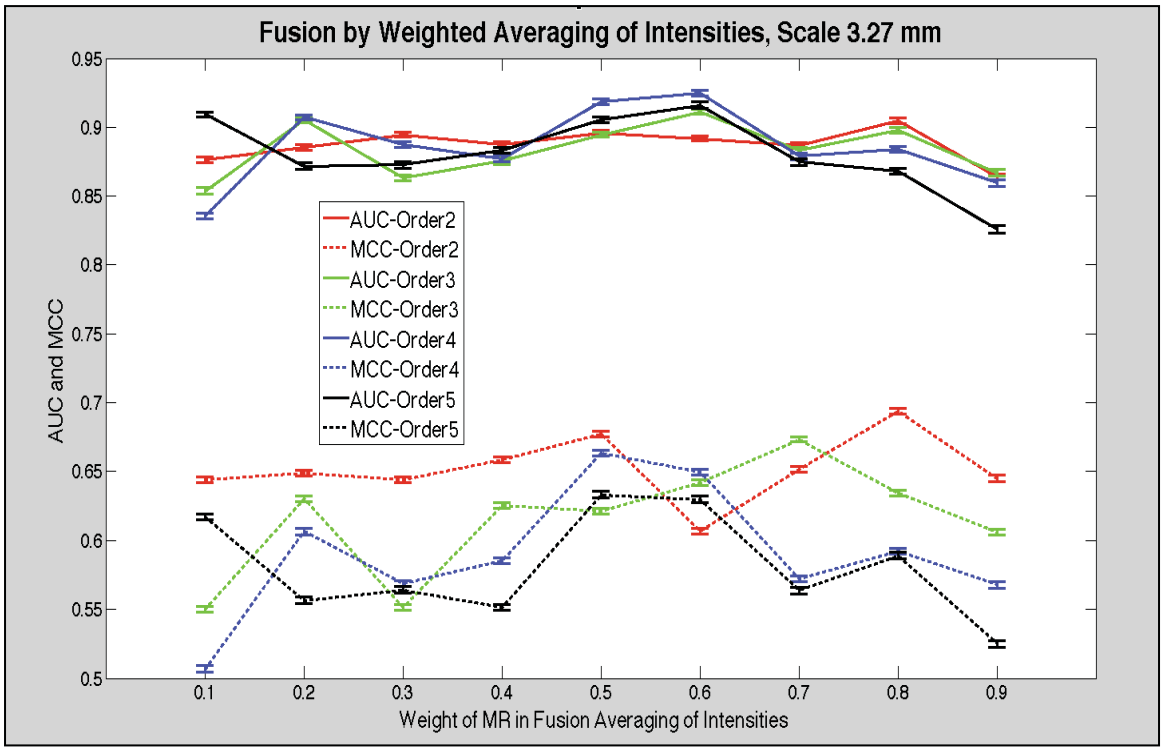

Figure 4.8. Prediction performance of fused scans (scale 3.27 mm), weighted averaging technique.

The results presented in Figure 4.8 suggest that PET and MR information may have equal importance on the prediction performance using multivariable modeling. As a matter of fact, the overall best prediction performance at all orders is obtained with a weight of about 0.5 MR in the fusion. This gives a basis for the implementation of the DWT fusion scheme presented next. Accordingly to these results, the choice was made to average the wavelet coefficients of both modalities as described in step 4 of the wavelet fusion scheme presented in section 3.4.2.

Then, in order to identify the prevalent fusion scheme, the same situation that was presented in Figure 4.8 (scale 3.27 mm) is now simulated with fused scans constructed

using the DWT technique described in section 3.4.2. Figure 4.9 elaborates the prediction performance as a function of the *R* ratio for model orders 2 to 5.

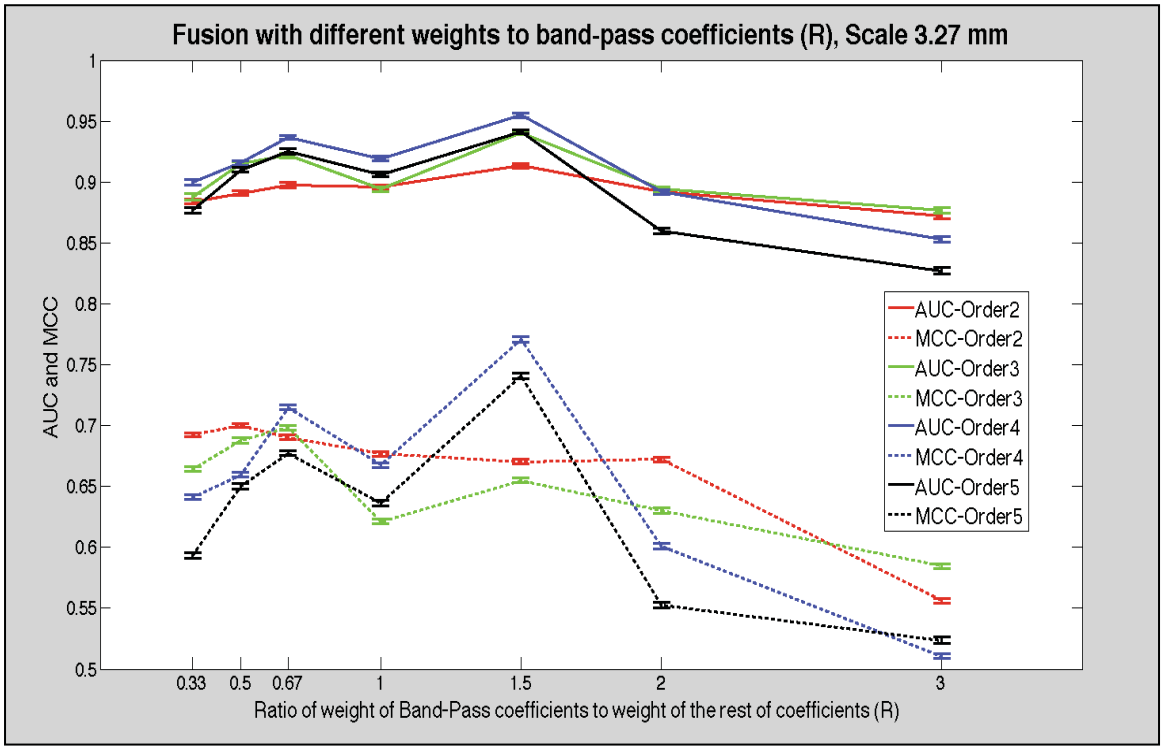

Figure 4.9. Prediction performance of fused scans (scale 3.27 mm), wavelet transform technique.

From Figure 4.8 and Figure 4.9, it can be seen that the fusion scheme using the DWT with $R = 1.5$ exhibit better prediction performance than the weighted averaging fusion scheme (especially at model orders 4 and 5) in terms of both AUC and MCC. Consequently to these results, the DWT was selected as the fusion scheme of choice for subsequent multivariable analysis.

Figure 4.9 shows that using $R = 1.5$ for fused scans allows for best prediction performance, but this result is valid for a scale of 3.27 mm. Figure 4.10 now presents the prediction performance of fused scans with $R = 1.5$ as a function of scale for model orders 2 to 5.

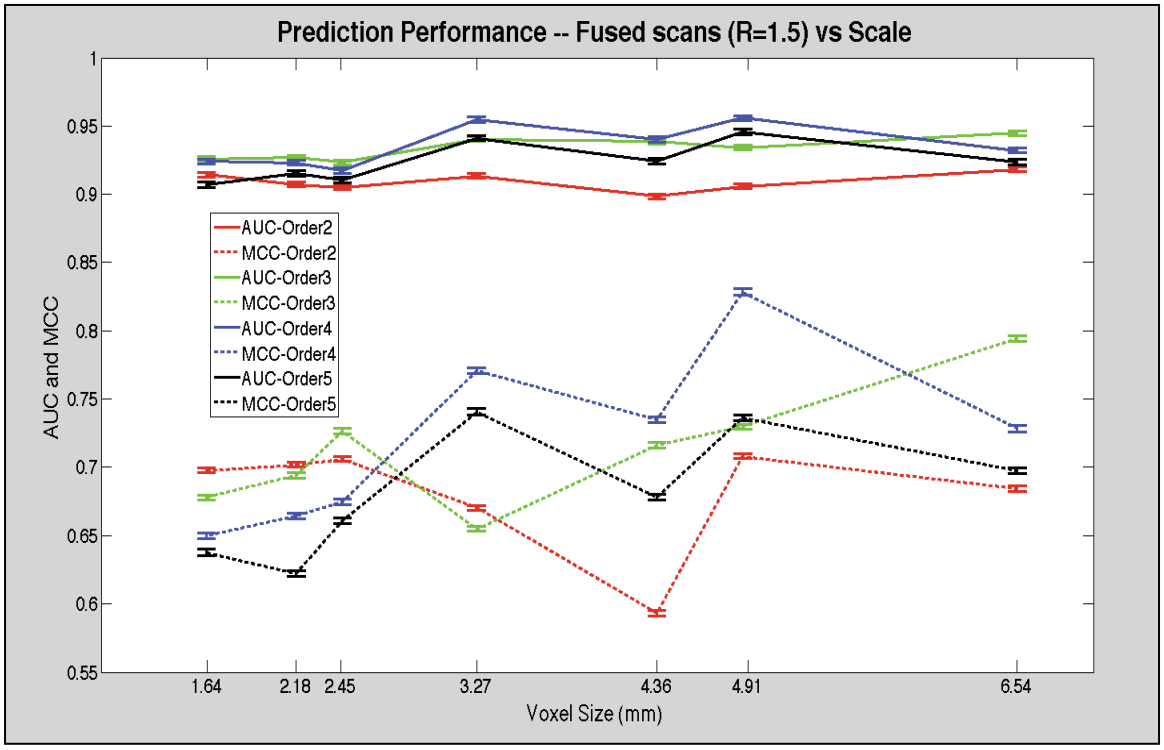

Figure 4.10. Prediction performance of fused scans (R = 1.5) as a function of scale.

According to Figure 4.10, scale 4.91 mm offers the best prediction performance for fused scans with $R = 1.5$. A significant increase in terms of MCC is observed from scale 3.27 mm to scale 4.91 mm at order 4, whereas AUC is not significantly affected. In order to validate model order 4 as the one providing the best parsimonious properties, model order estimation was performed using features extracted from fused scans with $R = 1.5$ at scale 4.91 mm by following the methodology described in the *model order estimation* part of section 3.8.2. Results of the simulation are shown in Figure 4.11.

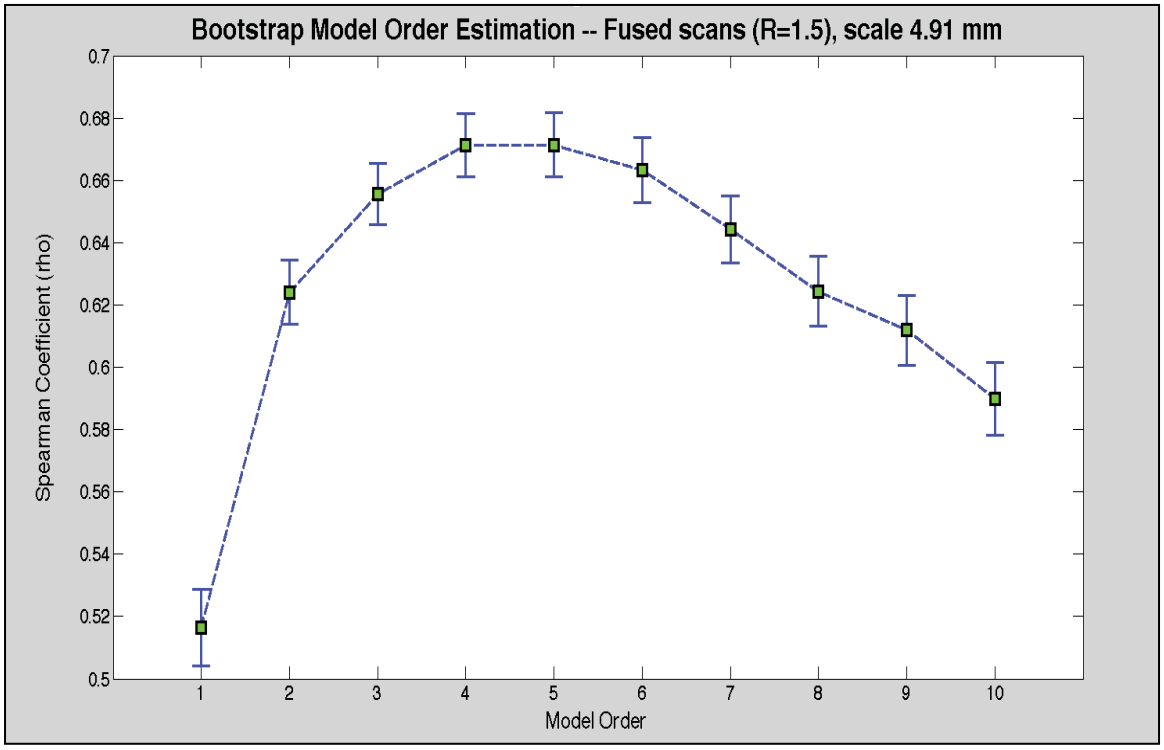

Figure 4.11. Model order estimation for fused scans (R = 1.5, scale 4.91 mm).

As expected, a peak at model order 4 is found in Figure 4.11, although a clear distinction between model orders 4 and 5 cannot be made. However, Figure 4.10 showed that the best prediction performance is obtained at order 4. This model order was thus chosen for the next simulation presented in Figure 4.12. The simulation of Figure 4.12 was performed in order to clearly identify the optimal combination of R ratio and scale. Figure 4.12 actually presents the prediction performance of fused scans computed with the whole range of R ratios as a function of scale for a specific model order of 4.

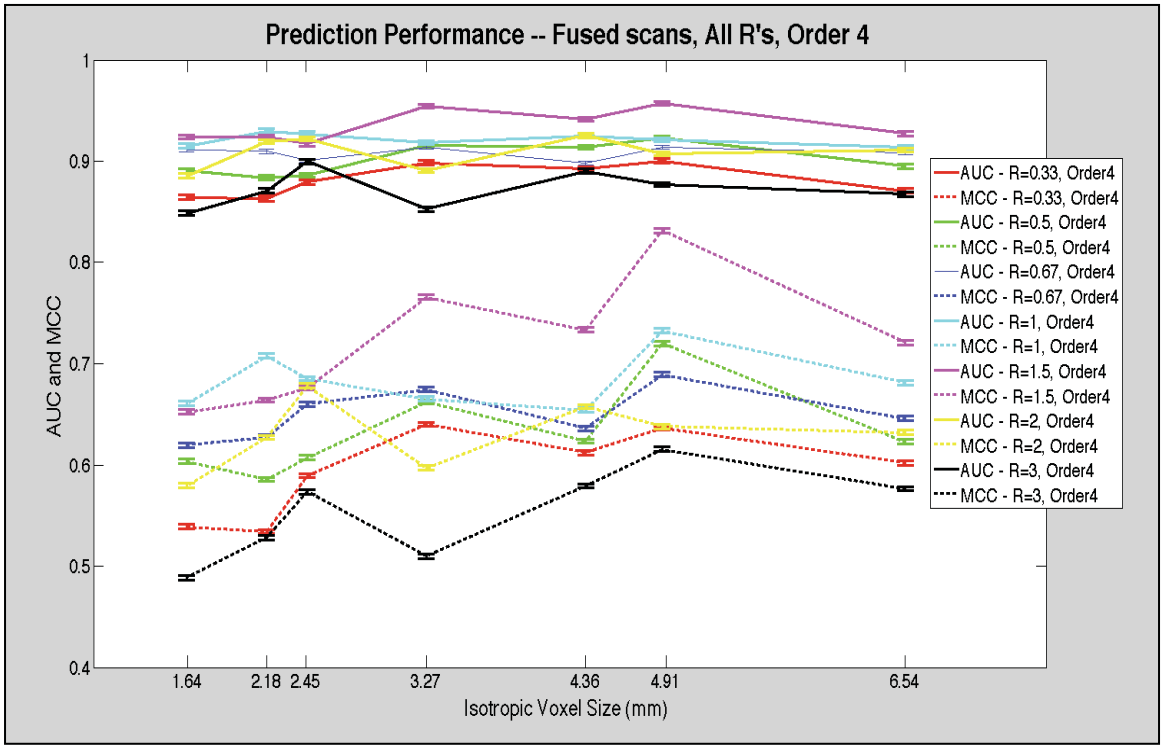

Figure 4.12. Prediction performance of fused scans (all *R* ratios, model order 4) as a function of scale.

Figure 4.12 confirms that the best prediction performance for fused scans is obtained with $R = 1.5$, scale 4.91 mm and model order 4. Then, for the sake of completeness, a plot of the prediction performance of fused scans with $R = 1.5$ and scale 4.91 mm as a function of model order is presented in Figure 4.13. Table 4.8 shows the most frequent group of variables used to perform the simulations of Figure 4.13 as defined in the *prediction performance evaluation* part of section 3.8.2.

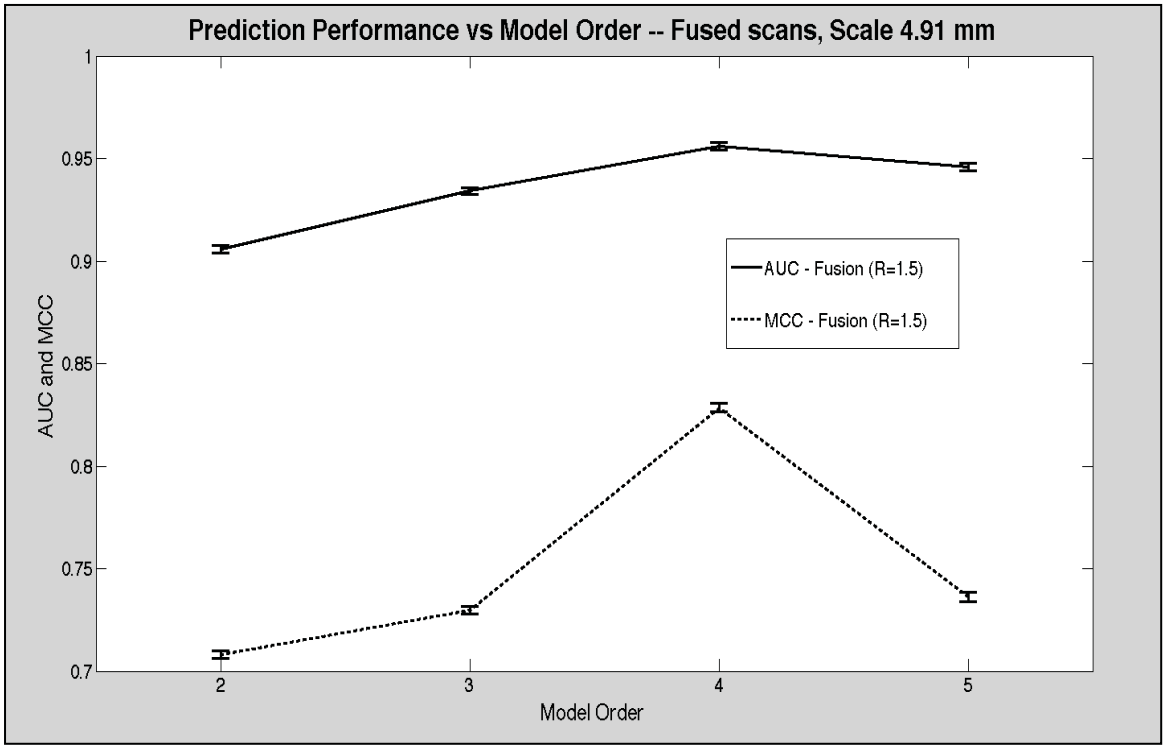

Figure 4.13. Prediction performance of fused scans (R = 1.5, scale 4.91 mm) as a function of model order.

Table 4.8. Most frequent group of variables of fused scans (R = 1.5, scale 4.91 mm)

Model Order	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	SUV_{max}	–	–	–	–
2	SUV_{max}	PET-T2FS/STIR--Variance	–	–	–
3	SUV_{max}	PET-T2FS/STIR--Variance	PET-T1--Contrast	–	–
4	SUV_{max}	PET-T2FS/STIR--Variance	PET-T1--Contrast	PET-T1--Homogeneity	–
5	SUV_{max}	PET-T2FS/STIR--Variance	PET-T1--Contrast	PET-T1--Homogeneity	PET-T2FS/STIR--SumMean

In order to complete the best linear model for fused scans, the regression coefficients for each of the variables at order 4 that are presented in Table 4.8 were computed using the same methodology as employed for separate scans (mean coefficients

over 10000 bootstrap training samples; see section 4.3.1). The whole procedure was also performed for feature values normalized in the range [0,1] prior to the bootstrapping runs. The mean regression coefficients that were obtained are summarized in Table 4.9.

Table 4.9. Regression coefficients of the best model from fused scans

Features	**Regression Coefficients (mean ± SEM)**	
	Non-normalized features	**Normalized features**
SUV_{max}	0.94 ± 0.02	28.4 ± 0.5
PET-T2FS/STIR--Variance	-0.401 ± 0.004	-52.1 ± 0.5
PET-T1--Contrast	-6.7 ± 0.1	-94 ± 1
PET-T1--Homogeneity	-165 ± 4	-95 ± 2
Offset (β_0)	140 ± 3	137 ± 3

Finally, to summarize the results obtained with fused scans, we can calculate a new vector $\mathbf{g} = \{g(\mathbf{x}_i) : i = 1,2,...,n\}$ of combined features for all patients ($n = 35$) using the regression coefficients of the non-normalized features of Table 4.9. The plot of **g** with the outcome vector **y** is shown in Figure 4.14. The complete linear model is also shown on that figure. Spearman's correlation between **g** and **y** is found to be $rho = 0.8255$ with $p < 0.0001$.

Figure 4.14 shows that a much better separation between patients with and without lung metastases can be obtained using a linear combination of features from fused scans than with SUV_{max} alone (see Figure 4.1). The separation is actually better with 4 variables from fused scans than with 5 variables from separate scans. This result suggests that texture features extracted from fused scans have better prediction power for lung metastases in STS cancer of the extremities than those from separate scans. The model variables shown in Figure 4.14 reached average results of $\mathrm{AUC} = 0.956 \pm 0.002$ and $\mathrm{MCC} = 0.829 \pm 0.002$ in bootstrap testing sets.

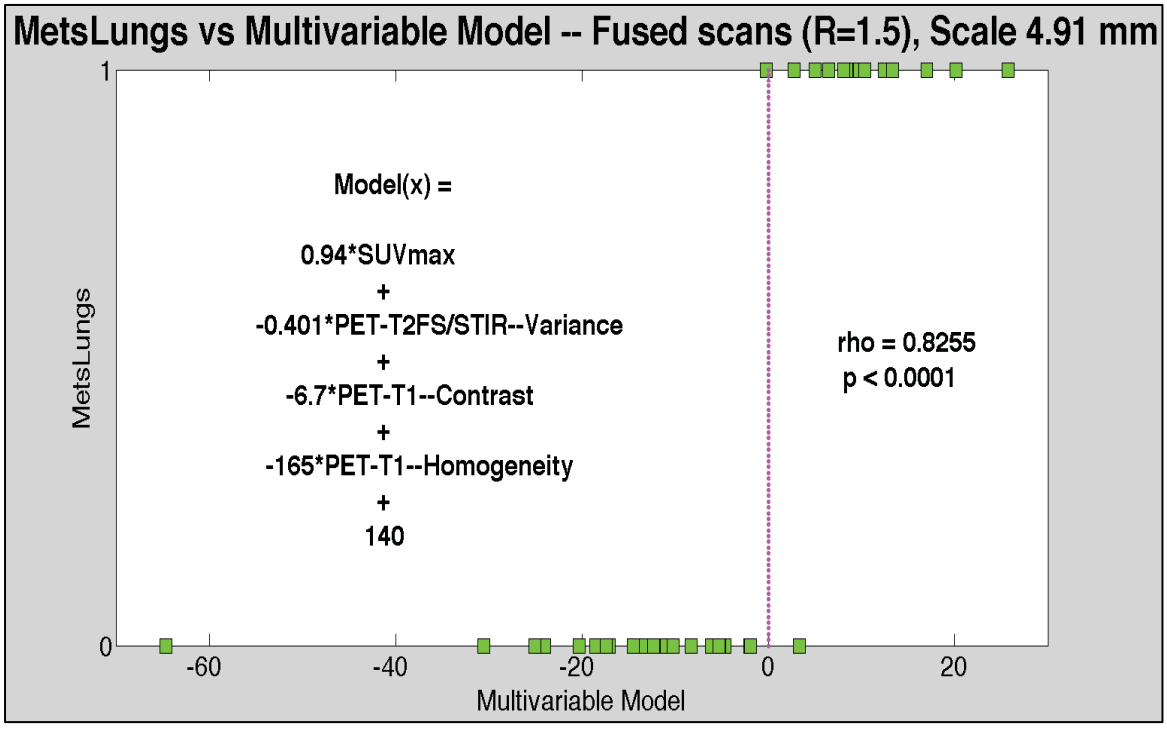

Figure 4.14. Plot of best multivariable model from fused scans.

4.3.3 Summary

This sub-section presents a summary of the last 2 sections. A direct comparison between multivariable results obtained with separate and fused scans is shown. In addition, the results are compared to the prediction performance of SUV_{max} alone. AUC and MCC were calculated for SUV_{max} by using the same method described in section 3.8.2. MCC was calculated with a threshold of SUV_{max} of 15. The performance metrics obtained for SUV_{max} alone are: $\text{AUC} = 0.880 \pm 0.002$ and $\text{MCC} = 0.685 \pm 0.004$.

Figure 4.15 incorporates the prediction performance results of SUV_{max} and of the optimal prediction conditions of separate and fused scans as a function of model order.

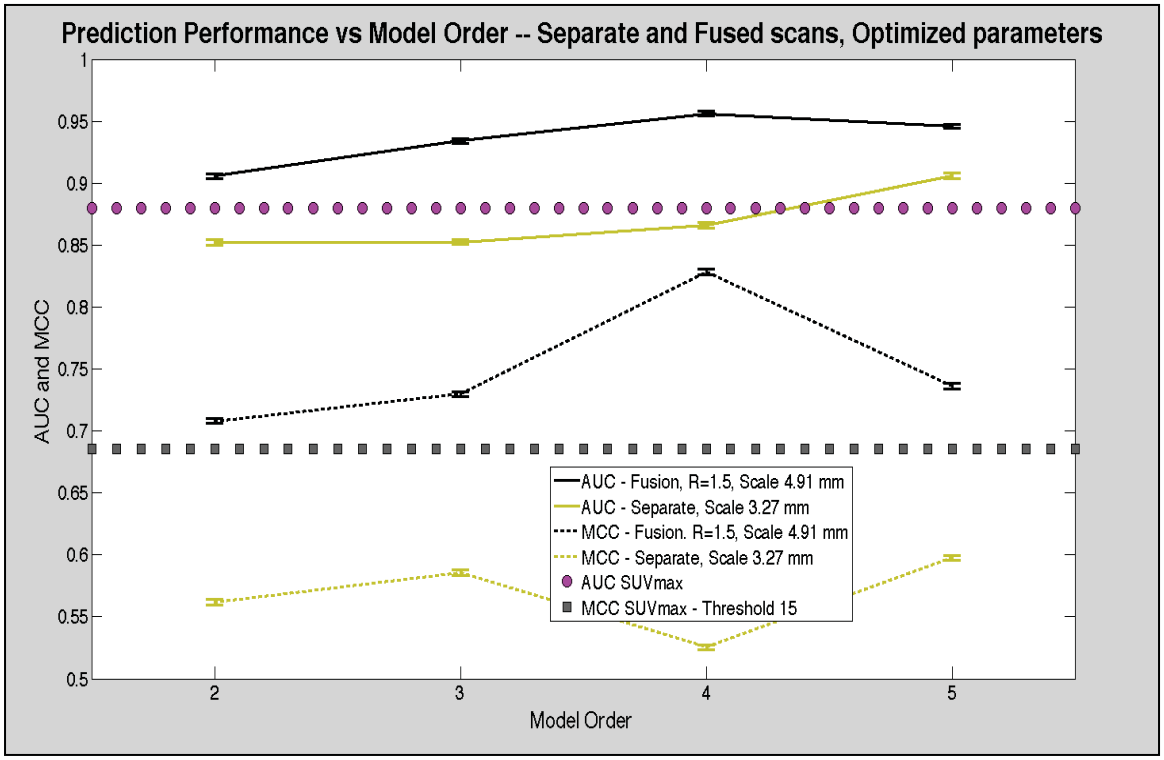

Figure 4.15. Prediction performance of SUV_{max} and of separate and fused scans (optimized conditions).

From the results of Figure 4.15, it is clear that the prediction performance of fused scans is superior to separate scans and SUV_{max}. However, the prediction performance of separate scans does not appear to be superior to SUV_{max}.

Overall, the best prediction performance is obtained with a linear combination of SUV_{max} and 3 texture features extracted from fused scans $(R = 1.5)$ at scale 4.91 mm. The optimal multivariable model of interest in this work and its corresponding prediction results are summarized in Figure 4.16. The goal of this work was to identify the general model shown in that figure and which takes the form of a linear combination of features as defined in Equation 2.14. Note that all optimal parameters (model order, variables, regression coefficients, R ratio, scale) and associated performance metrics (AUC, MCC) were identified from experiments performed using bootstrap training and testing sets. In other words, the simulation of unseen data was used to identify the parameters optimized for prediction performance. This final model then needs to be tested on an independent

patient cohort in order to validate its potential in predicting lung metastases at diagnosis of STS of the extremities.

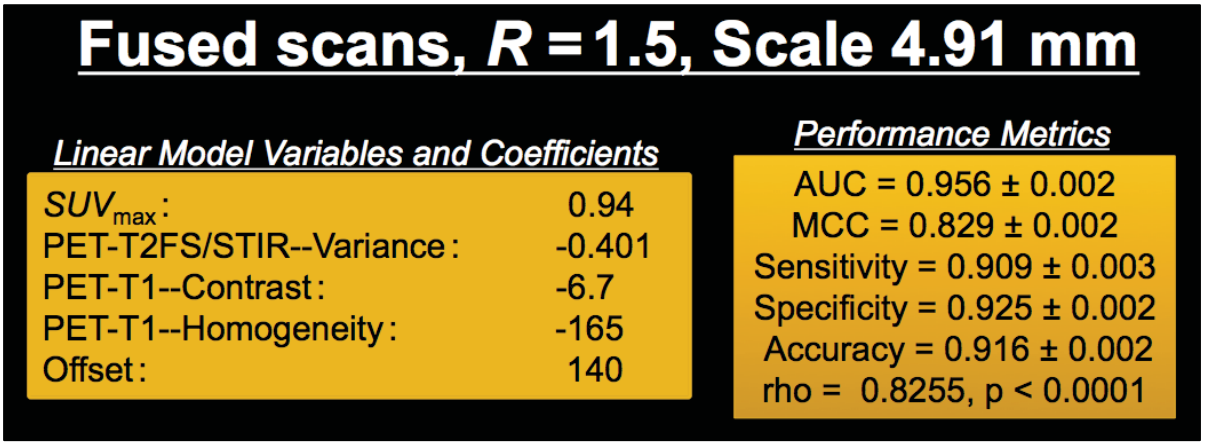

Figure 4.16. Optimal multivariable model of interest in this work and corresponding results.

CHAPTER 5: DISCUSSION

5.1 Overview

In this work, a new methodology has been proposed for the identification of a general model of imaging features allowing for the prediction of lung metastases in STS cancer of the extremities. The *best model* identified in this work (Figures 4.14 and 4.16) is made up of the combination of three texture features extracted from fused FDG-PET and MR scans as well as the SUV_{max} metric extracted from FDG-PET scans. The use of texture features from fused FDG-PET/MR scans as prognostic factors of tumours constitutes a new technique, and this study revealed its promising role. We think that the methodology developed in this work could be generalized to other cancers and outcomes, but it is obvious that the results obtained in this work are specific to lung metastases from STS of the extremities and to the different methods that were used. For example, the type of MR sequences used in this study has an influence on the final results (Appendix B acknowledges the different influence of T2FS and STIR sequences on prediction results). The heterogeneity of the scan protocols from the retrospective STS sample is also undesirable for future prospective applications, as one needs to minimize texture feature variations emerging from inter-acquisition differences [89] [92]. More sophisticated image pre-processing procedures such as those employed in the MR literature ([131] [132]) could be implemented in order to minimize this effect and to provide a common normalization scheme for FDG-PET and MR imaging modalities. On the other hand, heterogeneous scan protocols have the advantage of providing results that can be more easily generalized to the clinical environment. Furthermore, other specific methods such as the quantization scheme, the choice of texture features or the feature selection algorithm, have an impact on the identification of the best model. Throughout this section, the implication of the different methods used in this work in order to obtain the optimal parameters and their corresponding univariate and multivariable results is discussed.

In section 4.1, the results obtained for the optimization of the quantization algorithm and the number of gray levels were presented. These preliminary experiments

were performed in order to identify a quantization scheme to be used for the rest of the work (sections 4.2 and 4.3). They were carried out by examining the behaviour of the different quantization schemes on the entire dataset and by understanding the inherent bias of some metrics (defined in section 3.8.1) that were computed. Based on the results shown in Tables 4.1 to 4.3, the choice was made to use Lloyd-Max quantization algorithm and a number of gray levels (N_g) of 16 for FDG-PET, MR and fused FDG-PET/MR scans. Unfortunately, the high predictive power of the best model might be dependent on this initial choice. More importantly, with another choice of quantization scheme, another best model might have been found with better results. Ideally, the experiments performed in sections 4.2 and 4.3 should be repeated using different quantization parameters. In order to obtain the best predictions, an optimal algorithm and N_g could be found for every single texture feature prior to multivariable analysis. However, even if we can suspect the quantization scheme to have an effect on the predictive power of tumour outcomes, we believe that other parameters such as those discussed in the next paragraphs should be given more attention. As a matter of fact, Tixier et al. [99] demonstrated that there were no statistically significant differences on the prediction of the response to chemoradiotherapy in esophageal cancer by performing GLCM-based texture analysis with 16, 32, 64 or 128 gray levels in FDG-PET images.

In section 4.2, investigation of the univariate correlation of imaging features with lung metastases was performed. First, it was shown that the $SUV_{\max}$ metric represents a useful predictor of lung metastases in STS cancer of the extremities ($rho = 0.6382$, $p < 0.0001$). This result is in agreement with other studies ([51] [52] [53]) assessing the potential of this metric to depict the aggressiveness of STS tumours. Next, the univariate correlation of texture features with lung metastases was computed. Let us recall that textural analysis was performed with a distance parameter $d = 1$. No other distance parameters were investigated and its influence has also not yet been explored in similar works ([97] [99] [100]) that study GLCM-based texture features as potential predictors of tumour outcomes. However, our study has the particularity that it simulated the effect of different distance parameters by examining textures extracted from FDG-PET, MR and fused FDG-PET/MR scans resampled at different scales. This procedure allowed to

identify a scale at which textures provide best discrimination between the two classes of patients. We suggest that this optimization step should be performed prior to classification/prediction tasks in all future similar studies. Also, this procedure should be divided into two independent FDG-PET and MR steps in the case of separate scans analysis. In reality, GLCM-based textural measurements represent a convolution of the response coming from the quantization algorithm, the parameter distance d, the number of gray levels N_g, the scale at which the extraction is performed and the parameter θ. Ideally, these 5 parameters should be optimized simultaneously in order to better capture the imaging patterns of tumours. The influence of the angle parameter was not taken into account in this work, nor was it in similar studies ([97] [99] [100]) as it is a common practice to average the co-occurrence contributions of all neighbours around a center voxel. In future studies, the parameter θ could be optimized using Gabor wavelets ([133] [134]) in order to find the orientation at which textural properties are most dominant. Nevertheless, the methodology used in this work still provides useful insights on the potential of texture features as predictors of lung metastases in STS cancer of the extremities. From Tables 4.4 and 4.5 as well as Figures 4.2 and 4.3, it can be seen that texture features extracted from fused scans have stronger correlation with lung metastases than those extracted from separate scans. This result suggests that the rearrangement of gray levels in fused scans provides advantageous new textural properties not present in separate scans, although FDG-PET texture features appear dominant over MR ones. In fact, the band-pass enhancement of wavelet coefficients allowed a better definition of objects forming textures in the spatial domain, which ultimately lead to better discrimination of the two classes of patients. As expected, an optimal R ratio greater than 1 was found. At one point, the discriminative power of textures went down with increasing R due to excessive addition of noise into the images. We believe that the band-pass enhancement technique should also be generalized to separate scans (apply DWT to FDG-PET and MR scans and reconstruct them separately with different R ratios). In this manner, we could verify if the enhanced textural properties of fused scans come from the fusion of FDG-PET and MR spatial characteristics of from the enhancement of wavelet band-pass coefficients. Furthermore, it was shown in our work that none of the univariate GLCM-based texture features offered better correlation with lung metastases than $SUV_{\max}$.

This emphasizes the need to explore the potential of other texture features for tumour outcome prediction in future work. Finally, let us recall that the best average univariate correlation results (separate and fused scans: 1.64 mm) were not obtained at the same scale as the best multivariable prediction performance (separate scans: 3.27 mm, fused scans: 4.91 mm). As a matter of fact, it has been demonstrated in multivariable modeling theory that "a variable that is completely useless by itself can provide significant improvement when taken with others" [135]. The optimization of texture features in terms of prediction performance from the optimization of FDG-PET/MR imaging protocols would therefore be a difficult task.

Then, section 4.3 presented the results form the multivariable analysis. The experiments converged to the identification of a general model of the form of Equation (2.14) for the prediction of lung metastases in STS cancer of the extremities, hence reaching the major objective of this work. All optimal parameters (model order, variables, regression coefficients, R ratio, scale) were identified from experiments performed using bootstrap training and testing sets. As shown in Figures 4.6 and 4.7, multivariable analysis performed on separate scans identified a model of order 5 with best prediction performance at scale 3.27 mm ($rho = 0.8021$ with $p < 0.0001$, $\mathrm{AUC} = 0.906 \pm 0.002$, $\mathrm{MCC} = 0.598 \pm 0.002$). As shown in Figures 4.13 and 4.14, multivariable analysis performed on fused scans identified a model of order 4 with best prediction performance at scale 4.91 mm with $R = 1.5$ ($rho = 0.8255$ with $p < 0.0001$, $\mathrm{AUC} = 0.956 \pm 0.002$, $\mathrm{MCC} = 0.829 \pm 0.002$). These results cannot be compared with the few similar works ([97] [98] [99] [100]) in the literature since these studies were not performed on STS patients. Generally, small sample size was an issue to generate conclusive results in previous works and it is a limitation in our study as well. However, the refinement of our methodology compared to similar works (FDG-PET/MR fusion, evaluation of prediction performance on bootstrap testing sets, etc.) provides new tools for better identification and robust validation of prediction models. As a matter of fact, it was clearly demonstrated in our work that models built from fused scans provide superior prediction performance than those from separate scans. These results likely reflect the stronger univariate correlation obtained with fused scans as shown in section 4.2. In addition, the

reduction of the dimensionality of the feature set of fused scans surely provided more stable simulations. More importantly, the fusion of FGD-PET/MR scans allowed to group some features that were important in both modalities and as a consequence, perhaps more texture information that is relevant to lung metastases prediction was incorporated into the linear models for the same model order (thus without being subject to overfitting). To verify the latter assertion, let us recall that the best model of separate scans (Figure 4.7) is composed of the texture variables *T1--SumMean*, *PET--Homogeneity*, *T1--Homogeneity* and *T1--Contrast*, whereas the best model of fused scans is composed of the variables *PET-T2FS/STIR--Variance*, *PET-T1--Contrast* and *PET-T1--Homogeneity*. For example, it can be seen that the *Homogeneity* texture is an important feature for both FDG-PET and T1 modalities. Although some information might have been lost in the fusion process, the grouping of the *Homogeneity* features in fused scans diminished the redundancy of the linear model and at the same time allowed another texture variable to be incorporated into the modeling for the same model order. However, from another point of view, the unpredictable form of the AUC and MCC curves of separate scans in Figure 4.6 highlights a major limitation of our methodology. Normally, bell-shape curves similar to Figure 4.13 for fused scans should have been found. The fact that the optimization of scale was not performed independently for FDG-PET and MR scans might be one of the causes of this abnormality. The feature selection algorithm based on the maximization of the *Wald test* in the logistic regression procedure might also be one of the causes of this abnormality. Indeed, the *Wald test* is known to produce unstable results on data with small sample size and relatively large feature set dimensionality [112]. As a consequence, the algorithm might not have picked up the dominant features of separate scans allowing for best prediction performance. Other feature selection algorithms such as the *likelihood ratio test* ([112] [113]) in the logistic regression procedure should be tested and results compared with the current method. Besides, it is obvious that the combination of the imaging features in a linear fashion using logistic regression might not be optimal. In future work, other non-linear learning algorithms such as neural networks or kernelized support vector machines (SVM) should be explored.

Furthermore, it is interesting to investigate the significance of the features forming the best model from a biological perspective. More specifically, the signs of Spearman's

coefficients from the univariate association of features with lung metastases should be more instructive on the biological meaning of tumour textures than the signs of regression coefficients of the best model. The regression coefficients computed from a logit transformation are hard to interpret and can lead to misinterpretation. We thus refer the reader to Table 4.5 for the following discussion. As expected, SUV_{max} has a positive correlation with lung metastases; high glucose uptake in tumours translates into high metabolic activity and is likely a sign of aggressiveness and metastatic potential. Next, *PET-T2FS/STIR--Variance*, *PET-T1--Contrast* and *PET-T1--Homogeneity* have respectively negative, negative and positive correlations with lung metastases. These three results are counter-intuitive with the hypothesis that tumour heterogeneity translates into tumour aggressiveness. Now referring to the interpretation of GLCM-based textures provided in Table 2.1, these results respectively implies that: 1) the lower the variations in the distribution of pairs of gray levels of PET-T2FS/STIR tumour volumes 2) the lower the local variations of intensities in PET-T1 tumour volumes 3) the more uniform (gray levels similarity) the PET-T1 tumour volumes are ➔ the higher is the risk of lung metastases in STS of the extremities. By visually inspecting the co-registered separate scans of the two classes of tumours, it was noticed that tumours with lung metastases often contains regions with uniform intensities in T2FS/STIR and T1 scans that corresponds with uniform low-uptake regions in FDG-PET scans. These low-uptake regions were most of the time present in the inner portion of tumours and thus most likely represent necrotic and/or hypoxic areas. The presence of these inner low-uptake uniform regions suggests that the tumour is rapidly increasing in size and might be more at risk to metastasize. This could explain the counter-intuitive signs of Spearman's coefficients of some texture features. In future work, texture analysis should be performed separately on sub-regions of STS in order to have a better comprehension of their importance in the assessment of tumour aggressiveness.

Finally, it should be verified that the variables forming the best model are not highly inter-correlated. In fact, the inclusion of redundant features is not expected to improve the prediction accuracy on future patient cohorts, as it would unnecessarily model repetitious characteristics of the training data. To verify if such conditions occur, Spearman's correlation coefficients were computed between the variables forming the

best model, that is to say, the vector of $SUV_{\max}$ values for the $n = 35$ patients and the vectors of texture features $\mathbf{f}_j = \{x_{ij} \in \mathbb{R} : i = 1,2,...,n\}$ that were extracted from the whole dataset of fused scans at scale 4.91 mm with $R = 1.5$, where $j = 1$ refers to *PET-T2FS/STIR--Variance*, $j = 2$ refers to *PET-T1--Contrast* and $j = 3$ refers to *PET-T1--Homogeneity*.

Table 5.1. Spearman's correlation coefficients between variables of the best model

Feature	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$
$SUV_{\max}$	rho = -0.0700, $p = 0.6884$	rho = -0.3787, $p = 0.0255$	$rho = 0.5008$, $p = 0.0025$
$\mathbf{f}_1$	–	$rho = 0.0367$, $p = 0.8340$	rho = -0.1081, $p = 0.5350$
$\mathbf{f}_2$	–	–	rho = -0.9476, p <0.0001

$\mathbf{f}_1$: *PET-T2FS/STIR--Variance*, $\mathbf{f}_2$: *PET-T1--Contrast*, $\mathbf{f}_3$: *PET-T1--Homogeneity*

It can be seen from Table 5.1 that the variables *PET-T1--Contrast* and *PET-T1--Homogeneity* are highly inter-correlated. This is a serious limitation of the best model and further work is needed to find features useful for the prediction of lung metastases that are less inter-correlated. A dimensionality reduction procedure such as *principal component analysis* ([136]) should be carried out prior to the multivariable modeling process in order to build a feature set containing linearly uncorrelated variables. For example, Clausi [103] showed that the GLCM-based texture feature *Correlation* (defined in [101]) is almost uncorrelated with the rest of *Haralick features*. It was verified that the *Correlation* texture feature extracted on PET-T1 fused scans at scale 1.64 mm with $R = 1.5$ reached $rho = 0.6031$ with $p = 0.0001$. In future work, this texture feature will be included in the feature set. Other features incorporating the shape patterns of STS tumours should also be incorporated in our analysis. As a matter of fact, O'Sullivan et al. [137] recently provided an interesting update of their initial studies ([55] [56]) on the analysis of spatial heterogeneity of FDG-PET uptake in sarcoma tumours. The study showed that significant improvements in terms of prediction of patient survival and tumour progression could be

obtained by using a new method of statistical analysis of tumour profiles. In the future, we will investigate if the prognostic value of these shape patterns could be generalized to the prediction of lung metastases.

5.2 Uncertainty analysis

It would now be interesting to illustrate how the best model identified in this work could be used clinically for the prediction of lung metastases in STS cancer of the extremities. A patient diagnosed with STS of the extremities would present in a hospital and undergo FDG-PET and MR scans. A single value of the form of $g(\mathbf{x}_i)$ could then be obtained out of the scans by following the methodology presented in this work. Using Equation (2.15), this value can be transformed into the posterior probability of observing outcome $y_i = 1$ (MetsLungs) given the input $\mathbf{x}_i$. This probability could provide useful insights to physicians in order to assess the risk of developing lung metastases. However, there is a need to identify a decision threshold with which the physician could consider with sufficient confidence that the patient is likely to develop lung metastases. Inherently, this decision threshold could dictate if the traditional treatment is to be adapted due to potential metastases. It has been shown that the form of the linear combination of features of Equation (2.14) includes an offset β_0 such that $g(\mathbf{x}_i) = 0$ represents the decision threshold of interest. However, we need to investigate the confidence interval of the measurement $g(\mathbf{x}_i) = 0$ and consequently of the decision-making, a concept that was not examined in studies involving multivariable analysis for tumour outcome prediction similar to this one ([97] [100]).

First, an estimation of the uncertainty of $g(\mathbf{x}_i)$ due to contouring variations is investigated. It was previously mentioned that 16 patients of the dataset had visible edema that could be clearly identified in the vicinity of their tumours. For these patients, two types of contours were drawn: one incorporating the visible edema (denoted as *Edema*) and one excluding it (denoted as *Mass*). The whole set of results presented in this work were acquired using the *Mass* contour. We now evaluate the uncertainty of $g(\mathbf{x}_i)$ due to contouring variations by investigating how the response of the best model shown in Figures 4.14 and 4.16 changes using the *Edema* contours. First, Table 5.2 reports the

tumour size (longest diameter) of all patients calculated in MATLAB® from the *Mass* and *Edema* contours. Table 5.2 also reports the value x_{ij} of the 3 texture features of the model of interest extracted from the two different contours for all patients $i = 1,2,...,n$, where $n = 35$. Again, $j = 1$ refers to *PET-T2FS/STIR--Variance*, $j = 2$ refers to *PET-T1--Contrast* and $j = 3$ refers to *PET-T1--Homogeneity*. Then, the absolute difference between the texture features extracted from the *Mass* and *Edema* contours for the i^{th} patient is defined as:

$$\Delta x_{ij} = \left| x_{ij}[Edema] - x_{ij}[Mass] \right|, \quad \text{for } j = 1,2,3 \tag{5.1}$$

Table 5.2. Contouring effect on texture features of best model

Patient number (*i*)	**Size Mass (mm)**	**Size Edema (mm)**	**ΔSize (mm)**	x_{i1} **Mass**	x_{i1} **Edema**	Δx_{i1}	x_{i2} **Mass**	x_{i2} **Edema**	Δx_{i2}	x_{i3} **Mass**	x_{i3} **Edema**	Δx_{i3}
1	113	176	62	78.7	80.2	1.4	7.86	7.21	0.64	0.445	0.457	0.012
2	163	–	0	42.0	–	0	10.15	–	0	0.402	–	0
3	153	–	0	80.6	–	0	4.71	–	0	0.510	–	0
4	67	–	0	79.1	–	0	11.98	–	0	0.392	–	0
5	280	–	0	72.6	–	0	8.11	–	0	0.448	–	0
6	303	327	23	79.7	79.6	0.1	6.46	6.95	0.49	0.471	0.461	0.010
7	153	–	0	83.3	–	0	7.26	–	0	0.452	–	0
8	53	–	0	100.4	–	0	10.60	–	0	0.395	–	0
9	54	–	0	109.4	–	0	5.51	–	0	0.476	–	0
10	126	–	0	70.0	–	0	9.18	–	0	0.414	–	0
11	189	241	51	89.5	93.1	3.6	5.05	4.77	0.28	0.511	0.517	0.005
12	70	94	24	83.9	89.9	6.0	5.63	5.92	0.29	0.476	0.480	0.004
13	101	–	0	97.9	–	0	10.27	–	0	0.423	–	0
14	131	208	78	105.5	84.8	20.7	7.72	7.41	0.30	0.442	0.456	0.014
15	144	159	15	75.9	72.2	3.7	5.85	4.34	1.51	0.484	0.519	0.034
16	119	–	0	36.8	–	0	4.85	–	0	0.539	–	0
17	123	–	0	129.5	–	0	14.30	–	0	0.364	–	0
18	158	–	0	95.8	–	0	7.26	–	0	0.445	–	0
19	193	–	0	72.8	–	0	7.53	–	0	0.438	–	0
20	181	263	82	93.6	101.2	7.6	6.02	5.05	0.97	0.476	0.502	0.026

21	95	–	0	91.0	–	0	9.21	–	0	0.407	–	0
22	212	–	0	75.0	–	0	4.86	–	0	0.528	–	0
23	98	–	0	97.1	–	0	6.18	–	0	0.505	–	0
24	186	–	0	96.8	–	0	4.83	–	0	0.523	–	0
25	240	247	7	59.8	55.1	4.7	4.37	3.80	0.57	0.519	0.534	0.015
26	96	192	96	102.7	99.3	3.4	5.21	5.36	0.16	0.484	0.501	0.017
27	144	286	143	92.9	87.7	5.2	5.95	4.66	1.29	0.489	0.524	0.035
28	155	238	83	103.7	96.9	6.8	4.40	3.83	0.57	0.506	0.523	0.016
29	84	–	0	76.6	–	0	5.03	–	0	0.483	–	0
30	49	55	6	85.2	105.5	20.3	8.32	8.07	0.24	0.437	0.439	0.002
31	122	–	0	58.6	–	0	5.58	–	0	0.488	–	0
32	161	297	136	70.8	52.2	18.6	9.44	6.24	3.20	0.418	0.468	0.050
33	110	223	113	41.4	60.0	18.6	7.53	5.68	1.85	0.467	0.490	0.023
34	68	70	2	115.5	107.8	7.7	4.10	3.71	0.40	0.514	0.530	0.015
35	89	166	77	54.0	74.3	20.3	4.01	4.47	0.46	0.586	0.555	0.031
Mean	137	165	29	82.8	83.0	***4.2***	7.01	6.71	***0.38***	0.467	0.474	***0.009***
SD (1σ)	61	76	44	20.9	19.9	***6.9***	2.45	2.51	***0.68***	0.049	0.049	***0.013***
SEM ($2\sigma/\sqrt{n}$)	21	26	15	7.1	6.7	***2.3***	0.83	0.85	***0.23***	0.016	0.017	***0.004***

x_{i1}: *PET-T2FS/STIR--Variance*, x_{i2}: *PET-T1--Contrast*, x_{i3}: *PET-T1--Homogeneity*
SD: *Standard deviation*
SEM: *Standard error of the mean*

Table 5.2 also reports the estimated mean $\overline{\Delta x_j}$ of Δx_{ij} over all patients, the standard deviation (SD) of the Δx_{ij} distribution as well as the standard error of the mean (SEM) of the Δx_{ij} distribution on a 95.5% confidence interval. Hence, under 95.5% confidence, the true mean of Δx_{ij} lies in the range $[\overline{\Delta x_j} - \text{SEM}_j, \overline{\Delta x_j} + \text{SEM}_j]$. In order to account for the worst-case scenario as normally required before clinical implementation of a new protocol, the uncertainty δ_j on x_{ij} measurements is defined as:

$$\delta_j = \overline{\Delta x_j} + \text{SEM}_j, \quad \text{for } j = 1,2,3 \tag{5.2}$$

The contribution of contour variations on the uncertainty of $g(\mathbf{x}_i)$ defined as $\varepsilon_{contour}$ is obtained via the following error propagation scheme:

$$\varepsilon_{contour} = \sqrt{(\frac{\partial g(\mathbf{x}_i)}{\partial x_{ij=1}})^2(\delta_1)^2 + (\frac{\partial g(\mathbf{x}_i)}{\partial x_{ij=2}})^2(\delta_2)^2 + (\frac{\partial g(\mathbf{x}_i)}{\partial x_{ij=3}})^2(\delta_3)^2}$$

$$\varepsilon_{contour} = \sqrt{(-0.401)^2(4.2+2.3)^2 + (-6.7)^2(0.38+0.23)^2 + (-165)^2(0.009+0.004)^2} \quad (5.3)$$

$$\varepsilon_{contour} \approx 5.3$$

The uncertainty value of 5.3 on $g(\mathbf{x}_i)$ is constant across all values of $g(\mathbf{x}_i)$ as it applies to every point of Figure 5.1. However, in this figure, $\varepsilon_{contour}$ is inserted only in the context of the decision threshold $g(\mathbf{x}_i) = 0$ for the best model identified in this work. It can be seen that contour variations have a significant impact on the precision of the model. However, we shall keep in mind that *Mass* and *Edema* are two extreme contours and that an observer is likely to draw a contour in-between. Figure 5.1 also illustrates that 6 patients without lung metastases and 3 patients with lung metastases are in the uncertainty zone around $g(\mathbf{x}_i) = 0$. This emphasizes the need to identify a model with better separation of the two classes of patients. In any case, no conclusion on the outcome of a given patient should be drawn if its corresponding model value $g(\mathbf{x}_i)$ is within the contouring uncertainty range.

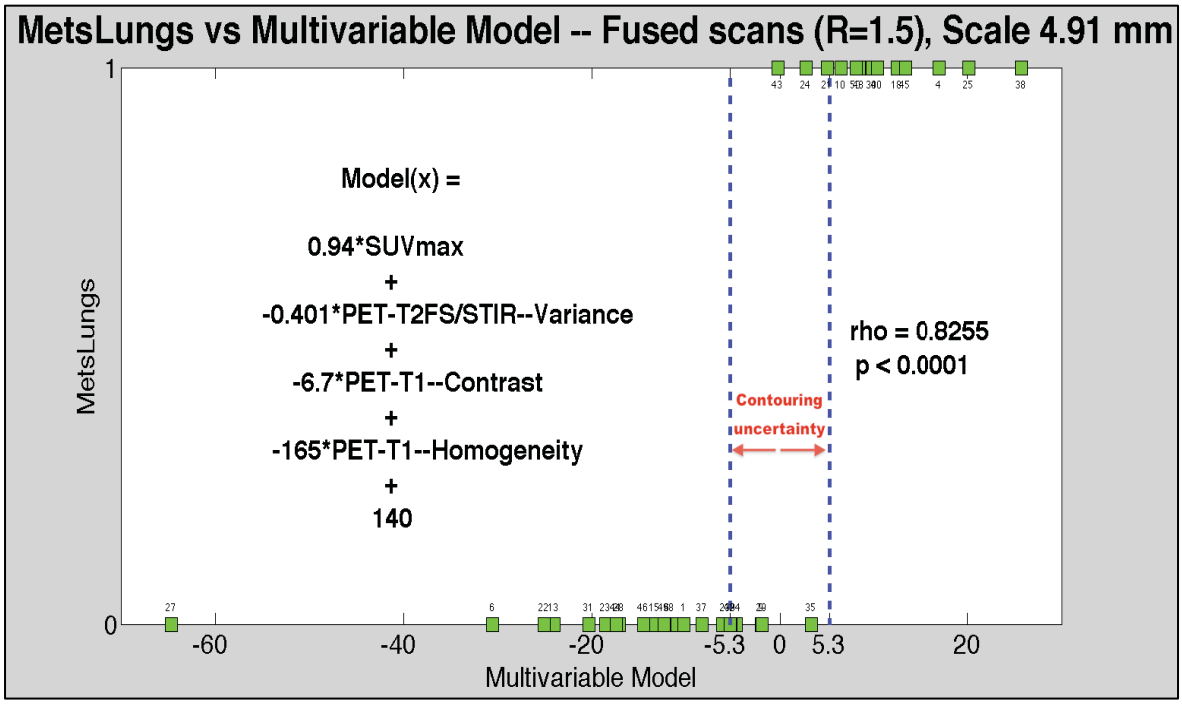

Figure 5.1. Plot of best model with contouring uncertainty around decision threshold.

Note that the uncertainty on $SUV_{\max}$ measurements was not considered in Equation (5.3). From a contouring point of view, this has no effect on the uncertainty of the model as it was verified that, for all patients, the voxel with maximal SUV was contained within the *Mass* contour. Recently, Burger et al. [138] evaluated the repeatability of $SUV_{\max}$ measurements using dynamic FDG-PET imaging on a dataset of 20 patients with different cancer types (including 1 sarcoma patient). The authors estimated the mean absolute change in $SUV_{\max}$ measurements to be close to 0.5 and the standard deviation of the absolute change to be close to 1. The contribution of $\delta_{SUV_{\max}}$ on the uncertainty of $g(\mathbf{x}_i)$ can thus be estimated on a 95.5% confidence interval using the latter results:

$$\begin{aligned} \varepsilon_{SUV_{\max}} &= \left| \frac{\partial g(\mathbf{x}_i)}{\partial x_{ij=SUV_{\max}}} \right| \delta(SUV_{\max}) \\ \varepsilon_{SUV_{\max}} &= (0.94) \cdot (0.5 + \frac{2*1}{\sqrt{20}}) \\ \varepsilon_{SUV_{\max}} &\approx 0.9 \end{aligned} \tag{5.4}$$

Again, this uncertainty is constant across all values of $g(\mathbf{x}_i)$. The results from Equations (5.3) and (5.4) show that the uncertainty associated to contouring variations dominates over the uncertainty of $SUV_{\max}$ measurements. As a matter of fact, the quadrature addition of $\varepsilon_{SUV_{\max}}$ and $\varepsilon_{contour}$ yields an overall uncertainty $\varepsilon_{contour+SUV_{\max}} = 5.4$. In future work, the repeatability of $SUV_{\max}$ measurements should be assessed on extremity STS tumours only in order to provide a more reliable value of $\delta_{SUV_{\max}}$. On the other hand, the prediction performance of another metric close to $SUV_{\max}$ but that has better repeatability could also be tested. In this respect, Burger et al. [138] proposed to use the average of the 10 hottest voxels of the tumour instead of $SUV_{\max}$. This metric is in theory more reliable than $SUV_{\max}$ since a single voxel measurement can be more easily altered by noise and image artefacts.

Next, registration errors of FDG-PET and MR volumes also bring an uncertainty on texture measurements. This type of uncertainty was not evaluated in this work due to the difficulty to identify ground-truth voxel correspondence. One approach for estimating the registration uncertainty would be to simulate precise shifts in all directions of the

FDG-PET scans with respect to the MR scans. For every shift, new resulting FDG-PET/MR fused volumes would be obtained. Then, the repeatability of texture measurements could be assessed in a similar manner as in the case of the contouring uncertainty. Although this would require exhaustive work, this type of analysis is desirable for a better assessment of the precision of the model $g(\mathbf{x}_i)$.

Lastly, the uncertainty of $g(\mathbf{x}_i)$ due to standard errors of the coefficient estimates that is inherent to logistic regression is investigated. In the case of p independent variables with values $\{x_{i0}=1, x_{i1}, x_{i2}..., x_{ip}\}$ and coefficient estimates $\{\hat{\beta}_j : j=0,1,...,p\}$, the estimated variance of $g(\mathbf{x}_i)$ due to the estimated variances of the regression coefficients for the i^{th} patient is defined as ([112]):

$$\widehat{\text{Var}}[\hat{g}(\mathbf{x}_i)] = \sum_{j=0}^{p} x_{ij}^2 \widehat{\text{Var}}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_{ij}\, x_{ik}\, \widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k), \quad \text{for } i=1,2,...,n \qquad (5.5)$$

The covariance matrix of the regression coefficients of the best model was obtained with the DREES software and is summarized in Table 5.3. However, instead of analyzing the covariance of regression coefficients found in bootstrap training sets, the logistic regression coefficients were computed one time on the entire dataset in order to get the best possible fit (result: $rho = 0.8372,\ p < 0.0001$).

Table 5.3. Covariance matrix of the regression coefficients of best model

$\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k)$	$SUV_{\max}$	β_1	β_2	β_3	β_0
$SUV_{\max}$	0.0102	–	–	–	–
β_1	-0.0008	0.0028	–	–	–
β_2	0.0096	-0.0241	0.5562	–	–
β_3	-0.5975	-0.4888	27.34	1759	–
β_0	0.1490	0.1687	-14.82	-965.7	540.8

β_1 : PET-T2FS/STIR--Variance, β_2 : PET-T1--Contrast, β_3 : PET-T1--Homogeneity

From Equation (5.5), it can be seen that $\widehat{\text{Var}}[\hat{g}(\mathbf{x}_i)]$ depends on the values $\mathbf{x}_i$ of the variables in the model and is different for all data points $g(\mathbf{x}_i)$. A lower bound on the estimated standard error of $\hat{g}(\mathbf{x}_i)$ can be obtained if we consider the specific case where

$\mathbf{x}_i = \{x_{ij} = 0 : j = 1,2,...,p\}$ with $x_{i0} = 1$ such that $\widehat{\mathrm{Var}}[\hat{g}(\mathbf{x}_i)]$ reduces to only the estimated variance of the offset estimate $\hat{\beta}_0$:

$$\begin{aligned} \widehat{\mathrm{SE}}[\hat{g}(\mathbf{x}_i)] &> \sqrt{\widehat{\mathrm{Var}}[\hat{\beta}_0]} \\ \widehat{\mathrm{SE}}[\hat{g}(\mathbf{x}_i)] &> \sqrt{540.8} \approx 23.3 \end{aligned} \tag{5.6}$$

A very large uncertainty inherent to logistic regression is associated to $\widehat{\mathrm{SE}}[\hat{g}(\mathbf{x}_i)]$. This uncertainty is much more significant than the other ones presented in this sub-section and constitutes the principal limiting factor to the precision of the best model. A better goodness-of-fit is needed to improve the precision of $g(\mathbf{x}_i)$ measurements. Although the relation is not explicit, a larger patient cohort is needed to achieve a better goodness-of-fit of the model.

5.3 Permutation tests

The theory behind permutation tests has emerged from two pioneering works by Fisher [139] and Pitman [140] in the 1930's. Permutations tests are a form of significance tests that has for goal to assess the significance of the sampling distribution of any test statistic when the null hypothesis (H_0) is true. This is commonly done by calculating the test statistic under all possible permutations of the labels of the observed data points. If the number of permutations is too extensive to be computed, a large number of random permutations are performed. The observed statistic with the true labels is then compared to the distribution of the test statistic when H_0 is true (permutation of labels) to assess if the observed effect could be attributed to the randomness introduced in selecting the sample or if it reflects an effect present in the population.

In our case, we need to verify if the statistical results ($\mathrm{AUC} = 0.956$, $\mathrm{MCC} = 0.829$) of the best model can be attributed to the randomness introduced in the selection of the patient cohort used in this study. More importantly, we also need to verify if the dimensionality of the set of variables {$SUV_{\max}$ + 12 scan-textures} has the effect of an overdetermined system in the feature selection part of the experiments. If this is the case, a set of variables could most of the time be found and model the data such that

similar results to those from the best model would be obtained with randomly assigned outcomes. Hence, the null hypothesis can be defined as:

> $\boldsymbol{H_0}$: *The effect observed using the best model on the sample of 35 patients can be ascribed to chance and is not present in the population of STS of the extremities. Other combination of features can produce the same effect.*

The data must then be resampled many times such that it is consistent with the null hypothesis. This is carried out by randomly shuffling the real outcome vector (ROV) in which the i^{th} position represents the outcome of the i^{th} patient (0: No MetsLung*s*, 1: MetsLungs). This will create a shuffled outcome vector (SOV) in which the 0's and 1's are randomly assigned new positions without replacement. The exact methodology used to create a SOV is:

1. For the first position of the SOV, randomly choose one position in the ROV. Assign to the first position of the SOV the value found at the chosen position of the ROV.
2. Delete the previously chosen value in the ROV.
3. Repeat step 1 and 2 for SOV positions 2 to 35.

The proportion of the two classes of patients therefore does not change in the SOV. Considering that the ROV contains 13 1's and 22 0's, the maximal number of differing positions between the SOV and ROV is 26. Ideally, every possible permutation ($35! \approx 1.033*10^{40}$) would be explored. However, this is not computationally possible.

Then, in order to test the validity of H_0, the following experiment was performed using the 13 features extracted from fused scans at scale 4.91 mm with $R = 1.5$:

1. Randomly shuffle the ROV for 1000 iterations.
2. For every SOV (iteration), apply bootstrap technique to divide the dataset into training and testing sets for 1000 bootstrap samples. For every bootstrap run, select 4 variables in the training set using the forward selection algorithm from

DREES. Calculate the percentage frequency of occurrence of group of variables. Pick up the most frequent group of variables for that SOV.

3. For every SOV (iteration), apply bootstrap technique to divide the dataset into training and testing sets for another 1000 bootstrap samples. For every bootstrap run, use the most frequent group of variables found in step 2 and calculate its corresponding regression coefficients in the training set. Use the regression coefficient found in the training set to calculate a vector of linear combination values $\mathbf{g}_{test} = \{g(\mathbf{x}_i) : i \in \text{SOV testing set}\}$. Use $\mathbf{g}_{test}$ and the outcome vector $\mathbf{y}_{test} = \{y_i : i \in \text{SOV testing set}\}$ in order to calculate performance metrics in the testing set, namely AUC and MCC. Average the performance metrics of the 1000 bootstrap runs.
4. Analyze the overall performance of the 1000 SOVs using the mean performance metrics computed in step 3.

For the rest of this section, the displayed AUCs and MCCs correspond to the average of the 1000 bootstrap samples applied to a particular SOV (step 3). The uncertainties correspond to the standard error of the mean on a 95.5% confidence interval. First, the maximum, mean and minimum AUC and MCC results obtained out of the 1000 iterations are presented in Table 5.4:

Table 5.4. Summary of permutation tests

Metric	AUC	MCC
Max	0.945 ± 0.006	0.67 ± 0.02
Mean	0.646 ± 0.006	0.194 ± 0.009
Min	0.362 ± 0.009	-0.17 ± 0.01
Observed (true)	0.956 ± 0.002	0.829 ± 0.002

The iteration yielding the maximal AUC was the same as the iteration yielding the maximal MCC. At that iteration, 10 positions were differing between the ROV and the SOV. Next, Figure 5.2 (top) presents the complete AUC and MCC distributions (1000

iterations) in terms of the number of positions differing between the ROV and every SOV. Figure 5.2 (bottom) also presents histograms of the AUC and MCC distributions built using 50 bins.

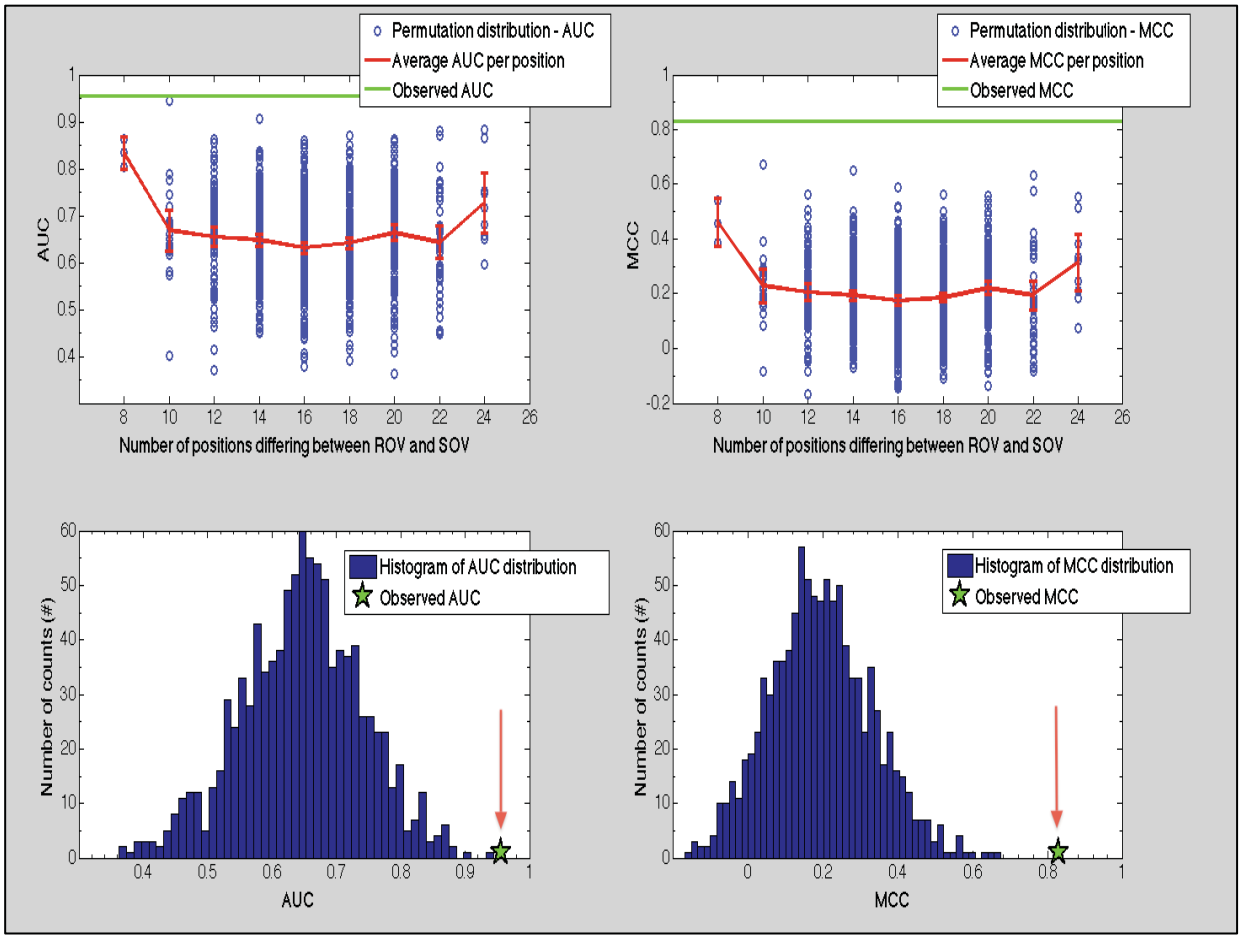

Figure 5.2. Complete results from permutation tests. Top-left: Permutation distribution of AUC, Top-right: Permutation distribution of MCC, Bottom-left: Histogram of AUC permutation distribution, Bottom-right: Histogram of MCC permutation distribution.

From a statistical point of view, Table 5.4 and Figure 5.2 show that the null hypothesis can be rejected. None of the iteration yielded an AUC or MCC higher that the observed values, so the probability that H_0 is true is estimated to be $p = (0+1)/(1000+1) < 0.001$. The results give strong evidence that the effect observed on the sample data did not occur by chance and that the effect is most likely present in the population of STS of the extremities. Furthermore, the design of the experiment let the algorithm find the variables that were the most important for every SOV. Even with this

degree of freedom, the observed values still perform better than all permutation tests (although the validity of the feature selection algorithm could again be discussed here). Ideally, every permutation test would have been optimized for model order, scale and *R* ratio prior to the calculation of AUC and MCC. This optimization step represents an exhaustive but necessary work that would probably have increased the results of the permutation tests. However, one simple test can provide insights on whether this optimization step would have brought significant improvements to the permutations tests. Let us recall Figure 4.12 which displays scale and *R* ratio optimization for fused scans for a model order of 4. That figure shows that the lowest results were obtained at scale 1.64 mm and $R = 3$, where $\mathrm{AUC} = 0.8488$ and $\mathrm{MCC} = 0.4885$. If we consider these values as the observed statistics for comparison with the latter permutation tests, we can come up with new estimates of the significance of the null hypothesis. In terms of AUC, 15 iterations yielded higher values than the new observed statistic such that the *p*-value can be estimated as $p = (15+1)/(1000+1) = 0.016 < 0.05$. In terms of MCC, 20 iterations yielded higher values than the new observed statistic such that the *p*-value can be estimated as $p = (20+1)/(1000+1) = 0.021 < 0.05$. Let us point out that this is a one-tail test as we are trying to find values greater than the observed statistics. Hence, for both AUC and MCC, the null hypothesis could be rejected with a confidence level higher than the commonly used 95% threshold.

The validity of the method presented in this study still needs to be questioned solely based on the maximal AUC obtained from permutation tests. This result ($\mathrm{AUC} = 0.945 \pm 0.006$) is very close from the observed statistic ($\mathrm{AUC} = 0.956 \pm 0.002$) and moreover, it was obtained with a different model than the best model identified in this work. As a matter of fact, for the specific case of the SOV that yielded the maximal AUC and MCC, the model variables chosen from the initial 1000 bootstrap samples prior to the 1000 bootstrap samples used for prediction performance evaluation are: $SUV_{\max}$, *PET-T1--Homogeneity*, *PET-T1--SumMean* and *PET-T2FS/STIR--Homogeneity*. This might reflect the use of an over-determined system in our analysis. It also emphasizes the need to increase the number of patients in future studies in order to clearly establish the best combination of variables found in this work as potential predictors of lung metastases in

STS of the extremities. However, the robustness of the best model cannot be totally rejected for two reasons. First, 2 out of 4 variables from the best permutation test are the same as the best model. Secondly, the AUC difference between the maximal result and the second maximal result of permutations tests is 0.0369. This is a huge difference in comparison to the mean difference between two consecutive values of the AUC sorted vector. The mean difference between two consecutive values of the AUC sorted vector is 0.0006 with 2 standard deviations of 0.004. As such, the AUC result obtained with the best SOV is clearly an outlier of the overall results obtained with the other 999 SOVs. This result could therefore be attributed to chance, as it is consistent with the corresponding p-value of the experiment ($p = 1/1001 < 0.001$). It is also interesting to note in Figure 5.2 that for a given number of positions differing between the ROV and SOV, many different SOVs yielding different AUC values can arise. This suggests that some features are dominant on patients of a certain class of outcome (0 or 1) in the STS sample. If many patients with dominant features are assigned to different classes on a particular SOV, the prediction power will diminish, and the inverse is also true.

To summarize, the permutations tests performed in this section give evidence that the strong predictive power of the best model observed on the sample patient cohort used in this study did not occur by chance and that the effect is most likely present in the population of STS of the extremities. This is a concept that was not previously examined in previous studies involving multivariable analysis for tumour outcome prediction ([97] [100]) similar to this one. However, a larger patient cohort is needed to fully assess the potential of the best model for the prediction of lung metastases in STS of the extremities.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

In this work, new imaging techniques were explored and developed to assess tumour aggressiveness at the diagnosis of STS cancer of the extremities. More specifically, we developed a novel approach based on image fusion to evaluate the potential of texture features extracted from fused FDG-PET/MR scans for the prediction of tumour outcomes. We have identified a 4^{th} order linear model combining three texture features from fused FDG-PET/MR scans and the SUV_{max} metric to be used for the prediction of lung metastases in STS cancer of the extremities. This model was verified to possess the best parsimonious characteristics and was validated using bootstrapping and permutation tests. The results revealed that FDG-PET and MR imaging features can act as strong prognostic factors of STS tumours of the extremities and can provide insights about their underlying biology. Our study demonstrates that the fusion of FDG-PET/MR image texture features using the discrete wavelet transform holds promise in the assessment of STS tumour heterogeneity. Texture features extracted from fused scans revealed to be better predictors of lung metastases in STS cancer of the extremities than those extracted from separate scans. However, FDG-PET texture features were also shown to be generally superior to MR texture features. Despite the fact that none of the texture features explored in this work had better predictive power than SUV_{max}, multivariable modeling of imaging features demonstrated significant improvements in terms of prediction performance in bootstrap testing sets.

Although the results are promising, the methodology of this work needs some refinements in order to robustly validate the prediction properties of the imaging model. The optimization of the quantization scheme, of texture measurements and of the feature selection algorithm still needs to be addressed more extensively. The uncertainty analysis of the imaging model revealed that the errors inherent to logistic regression modeling represent the main limitations to its precision. Ultimately, a larger patient cohort is needed to validate the robustness of the imaging model prior to prospective studies. Other texture features also need to be explored in order to build an optimal feature set prior to

multivariable modeling. Other than *Haralick features*, fractals and features from run-length matrices could be used as good alternative candidates.

In future work, the aforementioned robustness issues will be addressed. A more in-depth analysis of the wavelet transform possibilities for image fusion and texture measurements will also be performed, as the power of multiresolution wavelet decomposition has not been fully exploited in this work. Wavelets can also be used as a denoising technique, so the potential prediction improvements due to noise removal prior to texture analysis will also be examined. The optimization of FDG-PET and MR imaging protocols for the enhancement of the dominant texture features of tumours is another interesting avenue to explore. Ultimately, we hope that our current and future methodologies could be generalized to other types of cancer and that they could lead to improvements in the personalization of treatments since this should give a better chance to patients to overcome this deadly disease.

APPENDIX

A Histopathology of primary STS of the extremities at the MSKCC

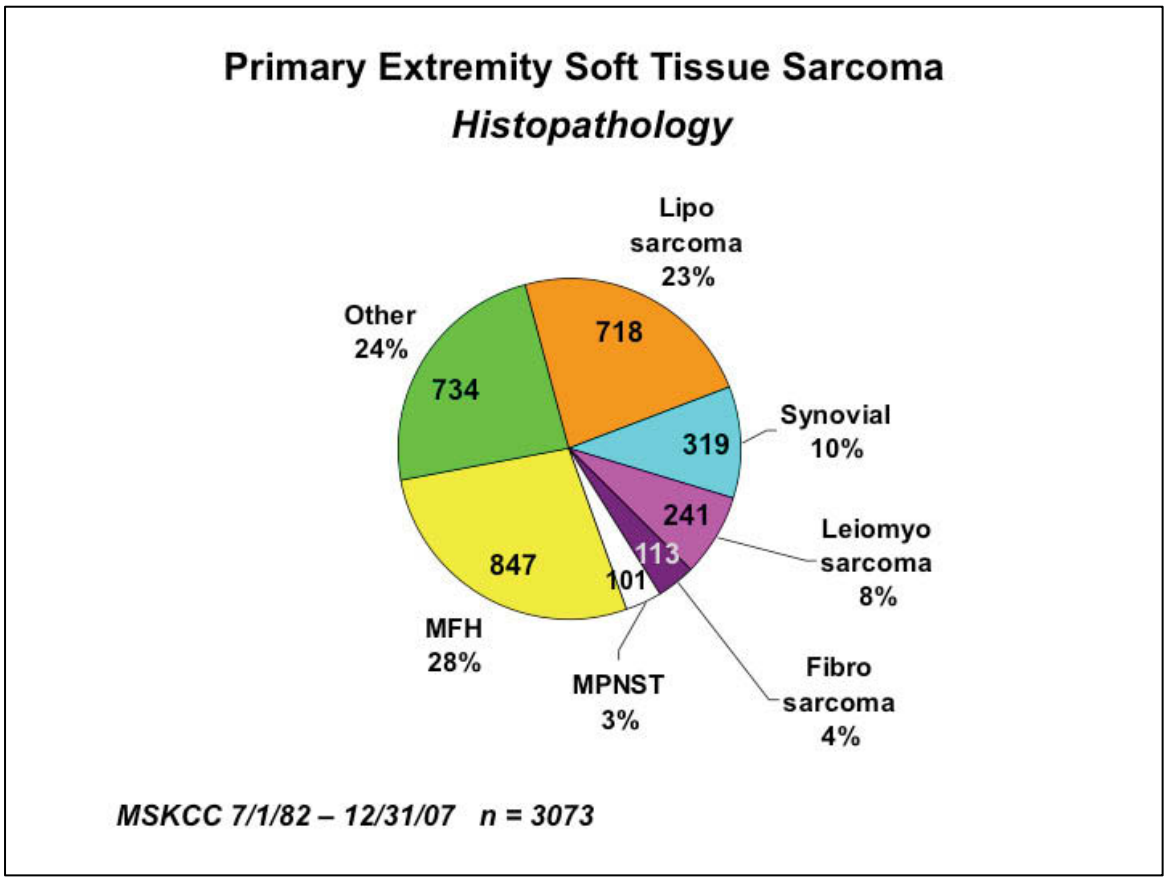

Figure A.1. Histopathology of primary STS of the extremities at the MSKCC. This distribution is similar to the one at the MUHC used in our study. The image was obtained by email correspondence and reproduced with the permission of Murray F. Brennan from the Memorial Sloan-Kettering Cancer Center (MSKCC).

B Comparison of T2FS and STIR texture features

This section explores the difference of texture features extracted from T2FS and STIR scans and the impact on Spearman's correlation with lung metastases.

From Table 3.4, we report 20 patients for whom both types of fat suppression scans were available. Texture features from both scans were extracted on these 20 patients and the average of the absolute percentage differences over the 20 patients for each feature was computed. The percentage difference of the absolute Spearman's

correlation coefficients (*rho*) of each feature vector with the outcome vector over the 20 patients was also computed. Results are reported in Table B.1.

Table B.1. Difference between T2FS and STIR texture features

Texture Feature	***rho* % difference** **[(T2FS-STIR)/T2FS]**	**Average absolute % differences** **[(T2FS-STIR)/T2FS]**
Energy	-105	0.8
Contrast	-10	3.3
Entropy	-29	0.1
Homogeneity	-40	0.2
SumMean	-12	3.2
Variance	-17	5.3
Average ($\mu \pm \sigma$)	-35 ± 36	2.2 ± 2.1

Tables B.1 shows that the difference between texture features of T2FS and STIR scans is minimal. However, their correlation with the endpoint of lung metastases is still significantly affected, as the average of absolute differences of Spearman's coefficients is shown to be 35%. Furthermore, STIR scans provides better correlation with the endpoint of interest for all texture features. In order to see the relative effect of the difference between T2FS and STIR texture features, results similar to Table B.1 are reported for T2FS and T1 texture features in Table B.2. The same process is also repeated in Table B.3 for STIR and T1 texture features.

Table B.2. Difference between T2FS and T1 texture features

Texture Feature	***rho* % difference** **[(T2FS-T1)/T2FS]**	**Average absolute % differences** **[(T2FS-T1)/T2FS]**
Energy	57	42.7
Contrast	29	16.0
Entropy	34	5.6
Homogeneity	67	8.1
SumMean	61	11.1
Variance	26	19.2
Average ($\mu \pm \sigma$)	46 ± 18	17 ± 13

Table B.3. Difference between STIR and T1 texture features

Texture Feature	***rho* % difference** **[(STIR-T1)/STIR]**	**Average absolute % differences** **[(STIR-T1)/STIR]**
Energy	79	43.9
Contrast	35	18.7
Entropy	49	5.6
Homogeneity	76	7.9
SumMean	65	13.8
Variance	37	23.3
Average ($\mu \pm \sigma$)	57 ± 19	19 ± 14

Some important results can be seen from Tables B.1 to B.3. First, only a 2% difference between texture features from T2FS and STIR scans (on average) brings the difference to 35% between their associated average absolute Spearman's correlation coefficients with respect to lung metastases. In contrast, a difference of about 18% between T1 and T2FS/STIR (T2FS or STIR) textures features brings a difference of about 51% between their associated average absolute Spearman's correlation coefficients with respect to the same endpoint. In other words, the difference between T2FS and STIR texture features is much lower than between T1 and T2FS/STIR. However, the difference between T2FS and STIR Spearman's correlation coefficients is not negligible. Also, as previously mentioned from Table B.1, STIR textures features provide better correlation with lung metastases than T2FS texture features and this result is emphasized in Table B.2 and Table B.3.

The multivariable models built in this study must contain variables (e.g., texture features) that are uncorrelated as much as possible such that they are independent of each other. From a pattern recognition point of view, it is preferred to minimize redundancy and the space dimensionality of the feature set in order to improve the prediction accuracy [103]. The final test performed in this section thus attempts to determine which of T2FS or STIR textures features have better independence properties with T1. For the 20 patients described at the beginning of this section, Spearman's correlation coefficients were computed between all respective texture feature vectors of the 3 types of scans. Results are presented in Table B.4.

Table B.4. Spearman's correlation between T1, T2FS and STIR texture features

Texture Feature	T2FS-STIR	T2FS-T1	STIR-T1
Energy	*rho* = 0.7654 (p = 0.0001)	*rho* = 0.4090 (p = 0.0745)	*rho* = 0.3729 (p = 0.1061)
Contrast	*rho* = 0.6902 (p = 0.0010)	*rho* = 0.5278 (p = 0.0182)	*rho* = 0.5684 (p = 0.0101)
Entropy	*rho* = 0.7699 (p = 0.0001)	*rho* = 0.4737 (p = 0.0364)	*rho* = 0.4391 (p = 0.0542)
Homogeneity	*rho* = 0.8481 (p < 0.0001)	*rho* = 0.6992 (p = 0.0008)	*rho* = 0.5398 (p = 0.0154)
SumMean	*rho* = 0.5925 (p = 0.0069)	*rho* = -0.2541 (p = 0.2783)	*rho* = -0.1729 (p = 0.4643)
Variance	*rho* = 0.5639 (p = 0.0108)	*rho* = -0.1940 (p = 0.4108)	*rho* = -0.1053 (p = 0.6580)
Average absolute *rho* ($\mu \pm \sigma$)	0.7050 ± 0.1106	0.4263 ± 0.1849	0.3664 ± 0.1907

Results from Table B.4 show, as expected, that the correlation between T2FS and STIR texture features is much higher than the correlation between T1 and T2FS/STIR texture features. This result in itself could justify the use in this work of the texture features from both T2FS and STIR scans in an equal manner. However, from the results presented in this section, STIR texture features appear to be better predictors of lung metastases than T2FS texture features. Also, Table B.4 shows that the correlation between STIR and T1 texture features is smaller than the correlation between T2FS and T1 texture features. In the future, the priority will be given to STIR texture features even at the expense of a lower sample size, although it does not guarantee that it will yield better prediction performance than the method used in this work (priority to T2FS). Further investigation with both types of fat-suppression scans on a larger patient cohort is needed. From another point of view, the method used in this work has the advantage of being generalized to both types of fat-suppression scans since a significant amount of both T2FS and STIR texture features were used to build prediction models. This can be important in clinical situations where only one or the other type of fat-suppression scan is available.

C Wavelet basis function analysis

This section quantifies the prediction power of fused scans constructed with the entire set of different wavelet basis functions available in the *Image Fusion* module of the MATLAB® Wavelet ToolBox™. This procedure was performed as follows:

1. Fuse FDG-PET and MR scans using the different wavelet basis functions with $R = 1.5$ by following the methodology described in section 3.4.2.
2. Extract texture features from the two types of fused scans (PET-T1 and PET-T2FS/STIR) at scale 3.27 mm using Lloyd-Max quantization algorithm and 16 gray levels by following the methodology described in sections 3.5 and 3.6.
3. Evaluate the correlation of all texture features and SUV_{max} (13 variables) with lung metastases by following the methodology described in section 3.7.
4. Average the absolute Spearman's coefficients (*rho*) of variables with $p < 0.1$ for every wavelet basis function.

Results are presented in Table C.1.

Table C.1. Spearman's correlation of fused scans features (built from different wavelet basis functions) with lung metastases

Wavelet basis	Average of absolute *rho* (variables with $p < 0.1$)
sym8	0.4326
rbio2.4	0.4257
coif1	0.4241
rbio4.4	0.4214
sym4	0.4212
coif2	0.4209
bior6.8	0.4208
bior3.1	0.4204
sym6	0.4201
bior2.6	0.4148
db8	0.4143
haar	0.4133
db1	0.4133
bior1.1	0.4133
rbior1.1	0.4133
bior2.4	0.4132

bior2.2	0.4130
bior4.4	0.4129
bior1.3	0.4128
rbio3.1	0.4124
db7	0.4116
coif4	0.4108
coif3	0.4108
db9	0.4096
rbio2.2	0.4096
rbio1.5	0.4094
coif5	0.4089
rbio2.8	0.4087
rbio1.3	0.4081
rbio5.5	0.4076
rbio2.6	0.4073
bior2.8	0.4072
bior1.5	0.4068
rbio6.8	0.4058
db2	0.4047
sym2	0.4047
sym7	0.4042
db6	0.4028
dmey	0.4016
bior3.5	0.4011
sym5	0.4005
db4	0.4002
bior3.3	0.3993
rbio3.9	0.3990
bior5.5	0.3989
bior3.9	0.3944
db3	0.3917
sym3	0.3917
bior3.7	0.3916
db10	0.3914
rbio3.3	0.3888
rbio3.5	0.3880
db5	0.3876
rbio3.7	0.3815

Table C.1 shows that the best predictive power is obtained with the wavelet basis function *sym8*.

D Effect of the inclusion of patients with lung metastases at presentation

This section presents results for the following situation: separate scans, Lloyd-Max quantization algorithm, 16 gray levels and scale 3.27 mm. The correlation of all features with lung metastases was investigated using 3 different patient subsets:

Test1: All patients from Table 3.1.

Test2: Exclusion of patients 11 and 20 (lung metastases at diagnosis of primary STS).

Test 3: Exclusion of patients 11, 20, 24 and 30 (any metastasis at diagnosis of primary STS).

Table D.1. Effect of the inclusion of patients with lung metastases at diagnosis of primary STS

Feature	Test 1	Test 2	Test 3
SUV_{max}	*rho* = 0.6382 ($p < 0.0001$)	*rho* = 0.6008 ($p = 0.0002$)	*rho* = 0.6030 ($p = 0.0003$)
PET--Homogeneity	*rho* = 0.4391 ($p = 0.0083$)	*rho* = 0.3848 ($p = 0.0270$)	*rho* = 0.3920 ($p = 0.0292$)
PET--Contrast	*rho* = -0.3806 ($p = 0.0241$)	*rho* = -0.3308 ($p = 0.0601$)	*rho* = -0.3392 ($p = 0.0619$)
PET--Entropy	*rho* = -0.3279 ($p = 0.0545$)	*rho* = -0.2565 ($p = 0.1496$)	*rho* = -0.2714 ($p = 0.1398$)
PET--Energy	*rho* = 0.2986 ($p = 0.0814$)	*rho* = 0.2498 ($p = 0.1619$)	*rho* = -0.2714 ($p = 0.1398$)
PET--SumMean	*rho* = -0.2869 ($p = 0.0947$)	*rho* = -0.2565 ($p = 0.1496$)	*rho* = -0.2186 ($p = 0.2374$)
PET--Variance	*rho* = -0.2752 ($p = 0.1096$)	*rho* = 0.2498 ($p = 0.1619$)	*rho* = 0.2111 ($p = 0.2544$)
T1--Contrast	*rho* = -0.2283 ($p = 0.1871$)	*rho* = -0.2565 ($p = 0.1496$)	*rho* = -0.2412 ($p = 0.1911$)
T2FS/STIR--SumMean	*rho* = -0.2283 ($p = 0.1871$)	*rho* = -0.2363 ($p = 0.1856$)	*rho* = -0.2186 ($p = 0.2374$)
T1--Energy	*rho* = -0.1991 ($p = 0.2516$)	*rho* = -0.1283 ($p = 0.4768$)	*rho* = -0.0829 ($p = 0.6574$)
T1--Entropy	*rho* = 0.1874 ($p = 0.2812$)	*rho* = 0.1215 ($p = 0.5005$)	*rho* = 0.0829 ($p = 0.6574$)
T2FS/STIR--Variance	*rho* = -0.1698 ($p = 0.3295$)	*rho* = -0.1958 ($p = 0.2749$)	*rho* = -0.1809 ($p = 0.3301$)
T2FS/STIR--Contrast	*rho* = -0.1405 ($p = 0.4207$)	*rho* = -0.1080 ($p = 0.5496$)	*rho* = -0.1357 ($p = 0.4668$)
T1--Variance	*rho* = 0.1405 ($p = 0.4207$)	*rho* = 0.1013 ($p = 0.5750$)	*rho* = 0.0829 ($p = 0.6574$)
T1--SumMean	*rho* = 0.1112 ($p = 0.5246$)	*rho* = 0.0675 ($p = 0.7089$)	*rho* = 0.0678 ($p = 0.7169$)

T2FS/STIR--Homogeneity	*rho* = 0.1054 (p = 0.5468)	*rho* = 0.0945 (p = 0.6008)	*rho* = 0.1206 (p = 0.5181)
T2FS/STIR--Entropy	*rho* = 0.0761 (p = 0.6639)	*rho* = 0.0405 (p = 0.8229)	*rho* = 0.0226 (p = 0.9039)
T1--Homogeneity	*rho* = 0.0468 (p = 0.7893)	*rho* = 0.0945 (p = 0.6008)	*rho* = 0.1206 (p = 0.5181)
T2FS/STIR--Energy	*rho* = -0.0410 (p = 0.8152)	*rho* = -0.0135 (p = 0.9406)	*rho* = -0.0135 (p = 0.9406)
Average of absolute *rho*	0.2274	0.1993	0.1956

E Supplement of univariate correlation results

E.1 Separate scans

Table E.1 presents the average correlation results in the case of separate scans at all scales. All situations were tested using 16 gray levels and Lloyd-Max quantization algorithm. The values in Table E.1 represent the average of absolute Spearman's correlation coefficients with lung metastases for all texture features from separate scans (SUV_{max} not included in the calculation).

Table E.1. Univariate correlation of separate scans with lung metastases in terms of scale

Scale	Average of absolute rho
1.64 mm	0.2589
2.45 mm	0.2231
3.27 mm	0.2046
2.18 mm	0.2039
6.54 mm	0.1913
4.91 mm	0.1890
4.36 mm	0.1828

E.2 Fused scans

Table E.2 presents the average correlation results in the case of fused scans for all combinations of scales and R ratios. All situations were tested using 16 gray levels and Lloyd-Max quantization algorithm. The values in Table E.2 represent the average of absolute Spearman's correlation coefficients with lung metastases for all texture features from fused scans (SUV_{max} not included in the calculation).

Table E.2. Univariate correlation of fused scans with lung metastases in terms of scale and *R* ratio

Scale	R	Average of absolute rho
1.64 mm	1.5	0.4474
1.64 mm	1	0.4284
1.64 mm	0.67	0.3952
4.36 mm	1	0.3928
3.27 mm	1.5	0.3913
2.45 mm	1.5	0.3864
2.18 mm	0.67	0.3835
2.18 mm	1	0.3820
4.36 mm	1.5	0.3762
3.27 mm	0.5	0.3757
3.27 mm	0.33	0.3737
2.18 mm	1.5	0.3708
2.18 mm	0.5	0.3703
1.64 mm	0.5	0.3664
3.27 mm	0.67	0.3659
2.45 mm	1	0.3615
1.64 mm	2	0.3557
3.27 mm	1	0.3557
6.54 mm	1.5	0.3459
4.91 mm	0.33	0.3445
2.45 mm	0.67	0.3435
2.45 mm	2	0.3425
1.64 mm	0.33	0.3420
6.54 mm	1	0.3367
2.18 mm	0.33	0.3337
4.36 mm	2	0.3332
2.18 mm	2	0.3328
4.36 mm	0.5	0.3303
4.91 mm	1.5	0.3293
2.45 mm	0.5	0.3293
4.36 mm	0.67	0.3181
4.91 mm	0.67	0.3157
2.45 mm	0.33	0.3098
4.91 mm	0.5	0.3045
6.54 mm	2	0.3025
6.54 mm	0.67	0.3020
2.45 mm	3	0.2971
6.54 mm	0.5	0.2971
4.91 mm	1	0.2971
4.36 mm	3	0.2957
2.18 mm	3	0.2937
3.27 mm	2	0.2874
4.36 mm	0.33	0.2830

1.64 mm	3	0.2727
6.54 mm	3	0.2693
4.91 mm	2	0.2552
3.27 mm	3	0.2288
4.91 mm	3	0.2249
6.54 mm	0.33	0.2137

BIBLIOGRAPHY

[1] A. C. Paulino and B. S. Teh, *PET-CT in Radiotherapy Treatment Planning*. Philadelphia, PA: Saunders/Elsevier, 2008.

[2] D. E. Heron, R. S. Andrade, S. Beriwal, and R. P. Smith, "PET-CT in Radiation Oncology: The Impact on Diagnosis, Treatment Planning, and Assessment of Treatment Response," *American Journal of Clinical Oncology: Cancer Clinical Trials,* vol. 31, pp. 352-362, 2008.

[3] S. Dresel, *PET in Oncology*. Berlin: Springer, 2008.

[4] R. Tucker, M. Coel, J. Ko, P. Morris, G. Druger, and P. McGuigan, "Impact of Fluorine-18 Fluorodeoxyglucose Positron Emission Tomography on Patient Management: First Year's Experience in a Clinical Center," *Journal of Clinical Oncology,* vol. 19, pp. 2504-2508, 2001.

[5] V. Kalff, R. J. Hicks, M. P. MacManus, D. S. Binns, A. F. McKenzie, R. E. Ware, A. Hogg, and D. L. Ball, "Clinical Impact of ^{18}F Fluorodeoxyglucose Positron Emission Tomography in Patients with Non-Small-Cell Lung Cancer: A Prospective Study," *Journal of Clinical Oncology,* vol. 19, pp. 111-118, 2001.

[6] H. Schoder, J. Meta, C. Yap, M. Ariannejad, J. Rao, M. E. Phelps, P. E. Valk, J. Sayre, and J. Czernin, "Effect of Whole-Body ^{18}F-FDG PET Imaging on Clinical Staging and Management of Patients with Malignant Lymphoma," *Journal of Nuclear Medicine,* vol. 42, pp. 1139-1143, 2001.

[7] H. Jadvar, D. L. Johnson, and G. M. Segall, "The Effect of Fluorine-18 Fluorodeoxyglucose Positron Emission Tomography on the Management of Cutaneous Malignant Melanoma," *Clinical Nuclear Medicine,* vol. 25, pp. 48-51, 2000.

[8] H. Zaidi, *Quantitative Analysis in Nuclear Medicine Imaging*. Boston, MA: Springer, 2006.

[9] J. M. Ollinger and J. A. Fessler, "Positron-Emission Tomography," *IEEE Signal Processing Magazine,* vol. 14, pp. 43-55, 1997.

[10] S. R. Cherry, J. A. Sorenson, and M. E. Phelps, *Physics in Nuclear Medicine*, 3rd ed. Philadelphia, PA: Saunders, 2003.

[11] R. A. Powsner and E. R. Powsner, *Essential Nuclear Medicine Physics*, 2nd ed. Malden, MA: Blackwell Publishing, 2006.

[12] S. Yu, "Review of ^{18}F-FDG synthesis and quality control," *Biomedical Imaging and Intervention Journal,* vol. 2, 2006.

[13] O. Warburg, F. Wind, and E. Negelein, "The metabolism of tumors in the body," *Journal of General Physiology,* vol. 8, pp. 519-530, 1927.

[14] E. K. J. Pauwels, M. J. Ribeiro, J. H. M. B. Stoot, V. R. McCready, M. Bourguignon, and B. Mazière, "FDG Accumulation and Tumor Biology," *Nuclear Medicine and Biology,* vol. 25, pp. 317-322, 1998.

[15] R. A. Hawkins, C. Hoh, J. Glaspy, Y. Choi, M. Dahlbom, S. Rege, C. Messa, E. Nietszche, E. Hoffman, L. Seeger, J. Maddahi, and M. E. Phelps, "The Role of Positron Emission Tomography in Oncology and Other Whole-Body Applications," *Seminars in Nuclear Medicine,* vol. 22, pp. 268-284, 1992.

[16] J. W. Keyes Jr, "SUV: Standard Uptake or Silly Useless Value?," *Journal of Nuclear Medicine,* vol. 36, pp. 1836-1839, 1995.
[17] R. Boellaard, "Standards for PET Image Acquisition and Quantitative Data Analysis," *Journal of Nuclear Medicine,* vol. 50, pp. 11S-20S, 2009.
[18] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt Jr., and J. M. Boone, *The Essential Physics of Medical Imaging*, 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2002.
[19] D. B. Plewes and W. Kucharczyk, "Physics of MRI: A Primer," *Journal of Magnetic Resonance Imaging,* vol. 35, pp. 1038-1054, 2012.
[20] H. Zaidi and A. D. Guerra, "An outlook on future design of hybrid PET/MRI systems," *Medical Physics,* vol. 38, pp. 5667-5689, 2011.
[21] B. J. Pichler, A. Kolb, T. Nägele, and H. P. Schlemmer, "PET/MRI: Paving the Way for the Next Generation of Clinical Multimodality Imaging Applications," *Journal of Nuclear Medicine,* vol. 51, pp. 333-336, 2010.
[22] H. Zaidi, M. L. Montandon, and A. Alavi, "The Clinical Role of Fusion Imaging Using PET, CT, and MR Imaging," *Magnetic Resonance Imaging Clinics of North America,* vol. 18, pp. 133-149, 2010.
[23] E. A. Ortiz, "Quantitative Functional MRI Based Evaluation of Caffeine's Effects on Brain Physiology," Master of Science, Medical Physics Unit, McGill University, 2011.
[24] R. B. Buxton, *Introduction to Functional Magnetic Resonance Imaging : Principles and Techniques*. Cambridge: Cambridge University Press, 2002.
[25] D. Hoa. (June 15, 2012). *MRI sequences: Spin echo*. Available: http://www.imaios.com/en/e-Courses/e-MRI/MRI-Sequences/Spin-echo
[26] J. M. Gomori, G. A. Holland, R. I. Grossman, W. B. Gefter, and R. E. Lenkinski, "Fat suppression by section-select gradient reversal on spin-echo MR imaging. Work in progress," *Radiology,* vol. 168, pp. 493-495, 1988.
[27] R. C. Semelka, W. Chew, H. Hricak, E. Tomei, and C. B. Higgins, "Fat-Saturation MR Imaging of the Upper Abdomen," *American Journal of Roentgenology,* vol. 155, pp. 1111-1116, 1990.
[28] G. M. Bydder and I. R. Young, "MR imaging: Clinical Use of the Inversion Recovery Sequence," *Journal of Computer Assisted Tomography,* vol. 9, pp. 659-675, 1985 1985.
[29] G. M. Bydder, R. E. Steiner, and L. H. Blumgart, "MR imaging of the liver using short TI inversion recovery sequences," *Journal of Computer Assisted Tomography,* vol. 9, pp. 1084-1089, 1985.
[30] D. G. Nishimura, *Principles of Magnetic Resonance Imaging*. San Francisco, CA: Stanford University, 1996.
[31] M. A. Bernstein, K. F. King, and X. J. Zhou, *Handbook of MRI Pulse Sequences*. Amsterdam: Elsevier Academic Press, 2004.
[32] S. M. Schuetze, "Imaging and Response in Soft Tissue Sarcomas," *Hematology/Oncology Clinics of North America,* vol. 19, pp. 471-487, 2005.
[33] J. N. Cormier and R. E. Pollock, "Soft Tissue Sarcomas," *CA: A Cancer Journal for Clinicians,* vol. 54, pp. 94-109, 2004.
[34] R. Siegel, D. Naishadham, and A. Jemal, "Cancer Statistics, 2012," *CA: A Cancer Journal for Clinicians,* vol. 62, pp. 10-29, 2012.

[35] M. F. Brennan, "Soft tissue sarcoma: Advances in understanding and management," *Surgeon,* vol. 3, pp. 216-223, 2005.

[36] V. T. DeVita, S. Hellman, and S. A. Rosenberg, *Cancer : Principles & Practice of Oncology*, 7th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2005.

[37] K. G. Billingsley, J. J. Lewis, D. H. Y. Leung, E. S. Casper, J. M. Woodruff, and M. F. Brennan, "Multifactorial Analysis of the Survival of Patients with Distant Metastasis Arising from Primary Extremity Sarcoma," *Cancer,* vol. 85, pp. 389-395, 1999.

[38] J. J. Lewis and M. F. Brennan, "Soft Tissue Sarcomas," *Current Problems in Surgery,* vol. 33, pp. 826-872, 1996.

[39] K. G. Billingsley, M. E. Burt, E. Jara, R. J. Ginsberg, J. M. Woodruff, D. H. Y. Leung, and M. F. Brennan, "Pulmonary Metastases From Soft Tissue Sarcoma: Analysis of Patterns of Disease and Postmetastasis Survival," *Annals of Surgery,* vol. 229, pp. 602-612, 1999.

[40] A. Rehders, S. B. Hosch, P. Scheunemann, N. H. Stoecklein, W. T. Knoefel, and M. Peiper, "Benefit of Surgical Treatment of Lung Metastasis in Soft Tissue Sarcoma," *Archives of Surgery,* vol. 142, pp. 70-75, 2007.

[41] R. Komdeur, H. J. Hoekstra, E. Van Den Berg, W. M. Molenaar, E. Pras, E. G. E. De Vries, and W. T. A. Van Der Graaf, "Metastasis in soft tissue sarcomas: Prognostic criteria and treatment perspectives," *Cancer and Metastasis Reviews,* vol. 21, pp. 167-183, 2002.

[42] C. Nanni and S. Fanti, "FDG-PET and PET/CT for Evaluating Soft Tissue Sarcomas," *PET Clinics,* vol. 5, pp. 341-347, 2010.

[43] M. W. Kattan, D. H. Y. Leung, and M. F. Brennan, "Postoperative Nomogram for 12-Year Sarcoma-Specific Death," *Journal of Clinical Oncology,* vol. 20, pp. 791-796, 2002.

[44] A. Carneiro, P.-O. Bendahl, J. Engellau, H. A. Domanski, C. D. Fletcher, P. Rissler, A. Rydholm, and M. Nilbert, "A Prognostic Model for Soft Tissue Sarcoma of the Extremities and Trunk Wall Based on Size, Vascular Invasion, Necrosis, and Growth Pattern," *Cancer,* vol. 117, pp. 1279-1287, 2011.

[45] C. W. D. Tzeng, J. K. Smith, and M. J. Heslin, "Soft Tissue Sarcoma: Preoperative and Postoperative Imaging for Staging," *Surgical Oncology Clinics of North America,* vol. 16, pp. 389-402, 2007.

[46] L. G. Shapeero, D. Vanel, K. L. Verstraete, and J. L. Bloem, "Fast magnetic resonance imaging with contrast for soft tissue sarcoma viability," *Clinical Orthopaedics and Related Research,* pp. 212-227, 2002.

[47] C. S. P. van Rijswijk, M. J. A. Geirnaerdt, P. C. W. Hogendoorn, J. L. Peterse, F. van Coevorden, A. H. M. Taminiau, R. A. E. M. Tollenaar, B. B. R. Kroon, and J. L. Bloem, "Dynamic contrast-enhanced MR imaging in monitoring response to isolated limb perfusion in high-grade soft tissue sarcoma: initial results," *European Radiology,* vol. 13, pp. 1849-1858, 2003.

[48] D. Vanel, S. Bonvalot, J. M. Guinebretière, P. Petrow, C. Dromain, and H. Caillet, "MR imaging in the evaluation of isolated limb perfusion: a prospective study of 18 cases," *Skeletal Radiology,* vol. 33, pp. 150-156, 2004.

[49] P. Clarkson and P. C. Ferguson, "Primary Multidisciplinary Management of Extremity Soft Tissue Sarcomas," *Current Treatment Options in Oncology,* vol. 5, pp. 451-462, 2004.

[50] G. C. Toner and R. J. Hicks, "PET for sarcomas other than gastrointestinal stromal tumors," *Oncologist,* vol. 13, pp. 22-26, 2008.

[51] M. H. M. Schwarzbach, A. Dimitrakopoulou-Strauss, F. Willeke, U. Hinz, L. G. Strauss, Y. M. Zhang, G. Mechtersheimer, N. Attigah, T. Lehnert, and C. Herfarth, "Clinical Value of [18-F] Fluorodeoxyglucose Positron Emission Tomography Imaging in Soft Tissue Sarcomas," *Annals of Surgery,* vol. 231, pp. 380-386, 2000.

[52] J. F. Eary, E. U. Conrad, J. D. Bruckner, A. Folpe, K. J. Hunt, D. A. Mankoff, and A. T. Howlett, "Quantitative [F-18]fluorodeoxyglucose positron emission tomography in pretreatment and grading of sarcoma," *Clinical Cancer Research,* vol. 4, pp. 1215-1220, 1998.

[53] J. F. Eary, F. O'Sullivan, Y. Powitan, K. Chandhury, C. Vernon, J. D. Bruckner, and E. U. Conrad, "Sarcoma tumor FDG uptake measured by PET and patient outcome: a retrospective analysis," *European Journal of Nuclear Medicine,* vol. 29, pp. 1149-1154, 2002.

[54] H. El-Zeftawy, S. I. Heiba, S. Jana, G. Rosen, S. Salem, J. F. Santiago, and H. M. Abdel-Dayem, "Role of Repeated F-18 Fluorodeoxyglucose Imaging in Management of Patients with Bone and Soft Tissue Sarcoma," *Cancer Biotherapy and Radiopharmaceuticals,* vol. 16, pp. 37-46, 2001.

[55] F. O'Sullivan, S. Roy, and J. Eary, "A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data," *Biostatistics,* vol. 4, pp. 433-448, 2003.

[56] F. O'Sullivan, S. Roy, J. O'Sullivan, C. Vernon, and J. Eary, "Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET," *Biostatistics,* vol. 6, pp. 293-301, 2005.

[57] K. M. Skubitz and D. R. D'Adamo, "Sarcoma," *Mayo Clinic Proceedings,* vol. 82, pp. 1409-1432, 2007.

[58] E. Bastiaannet, B. Groen, P. L. Jager, D. C. P. Cobben, W. T. A. van der Graaf, W. Vaalburg, and H. J. Hoekstra, "The value of FDG-PET in the detection, grading and response to therapy of soft tissue and bone sarcomas; a systematic review and meta-analysis," *Cancer Treatment Reviews,* vol. 30, pp. 83-101, 2004.

[59] G. H. Heppner and B. E. Miller, "Tumor heterogeneity: biological implications and therapeutic consequences," *Cancer Metastasis Reviews,* vol. 2, pp. 5-23, 1983.

[60] D. Shibata, "Heterogeneity and Tumor History," *Science,* vol. 336, pp. 304-305, 2012.

[61] T. Tsuruo and I. J. Fidler, "Differences in Drug Sensitivity among Tumor Cells from Parental Tumors, Selected Variants, and Spontaneous Metastases," *Cancer Research,* vol. 41, pp. 3058-3064, 1981.

[62] W. K. A. Yung, J. R. Shapiro, and W. R. Shapiro, "Heterogeneous Chemosensitivities of Subpopulations of Human Glioma Cells in Culture," *Cancer Research,* vol. 42, pp. 992-998, 1982.

[63] H. Z. Hill, G. J. Hill, C. F. Miller, F. Kwong, and J. Purdy, "Radiation and Melanoma: Response of B16 Mouse Tumor Cells and Clonal Lines to in Vitro Irradiation," *Radiation Research,* vol. 80, pp. 259-276, 1979.

[64] J. T. Leith, D. L. Dexter, J. K. Dewyngaert, E. M. Zeman, M. Y. Chu, P. Calabresi, and A. S. Glicksman, "Differential Responses to X-irradiation of Subpopulations of Two Heterogeneous Human Carcinomas in Vitro," *Cancer Research,* vol. 42, pp. 2556-2561, 1982.

[65] B. Nielsen, F. Albregtsen, and H. E. Danielsen, "Statistical Nuclear Texture Analysis in Cancer Research: A Review of Methods and Applications," *Critical Reviews in Oncogenesis,* vol. 14, pp. 89-164, 2008.

[66] D. L. Longo, "Tumor Heterogeneity and Personalized Medicine," *New England Journal of Medicine,* vol. 366, pp. 956-957, 2012.

[67] M. C. Asselin, J. P. B. O'Connor, R. Boellaard, N. A. Thacker, and A. Jackson, "Quantifying heterogeneity in human tumours using MRI and PET," *European Journal of Cancer,* vol. 48, pp. 447-455, 2012.

[68] A. Pugachev, S. Ruan, S. Carlin, S. M. Larson, J. Campa, C. C. Ling, and J. L. Humm, "Dependence of FDG uptake on tumor microenvironment," *International Journal of Radiation Oncology Biology Physics,* vol. 62, pp. 545-553, 2005.

[69] E. Henriksson, E. Kjellen, P. Wahlberg, T. Ohlsson, J. Wennerberg, and E. Brun, "2-Deoxy-2-[^{18}F]Fluoro-D-glucose Uptake and Correlation to Intratumoral Heterogeneity," *Anticancer Research,* vol. 27, pp. 2155-2159, 2007.

[70] A. P. Pathak, "Magnetic resonance imaging of tumor physiology," *Methods in molecular medicine.,* vol. 124, pp. 279-297, 2006.

[71] A. R. Padhani, G. Liu, D. Mu-Koh, T. L. Chenevert, H. C. Thoeny, T. Takahara, A. Dzik-Jurasz, B. D. Ross, M. Van Cauteren, D. Collins, D. A. Hammoud, G. J. S. Rustin, B. Taouli, and P. L. Choyke, "Diffusion-Weighted Magnetic Resonance Imaging as a Cancer Biomarker: Consensus and Recommendations," *Neoplasia,* vol. 11, pp. 102-125, 2009.

[72] A. R. Padhani, "Diffusion Magnetic Resonance Imaging in Cancer Patient Management," *Seminars in Radiation Oncology,* vol. 21, pp. 119-140, 2011.

[73] J. Yokota, "Tumor progression and metastasis," *Carcinogenesis,* vol. 21, pp. 497-503, 2000.

[74] I. J. Fidler, "Critical Factors in the Biology of Human Cancer Metastasis: Twenty-eighth G.H.A. Clowes Memorial Award Lecture," *Cancer Research,* vol. 50, pp. 6130-6138, 1990.

[75] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification," *Magnetic Resonance Imaging,* vol. 22, pp. 81-91, 2004.

[76] C. H. Chen, L. F. Pau, and P. S. P. Wang, *Handbook of Pattern Recognition and Computer Vision*, 4th ed. Singapore: World Scientific Publishing, 2010.

[77] M. Mirmehdi, X. Xie, and J. S. Suri, *Handbook of Texture Analysis*. London: Imperial College Press, 2008.

[78] T. R. Reed and J. M. H. Dubuf, "A Review of Recent Texture Segmentation and Feature Extraction Techniques," *CVGIP: Image Understanding,* vol. 57, pp. 359-372, 1993.

[79] C. Wang, Y. Guan, C. Zuo, F. Hua, L. Chen, W. Zhang, Y. Pan, Y. Yuan, and Y. Zhou, "Value of the texture feature for solitary pulmonary nodules and mass lesions based on PET/CT," *4th International Conference on Bioinformatics and Biomedical Engineering Proceedings*, 2010.

[80] H. Yu, C. Caldwell, K. Mah, and D. Mozeg, "Coregistered FDG PET/CT-Based Textural Characterization of Head and Neck Cancer for Radiation Treatment Planning," *IEEE Transactions on Medical Imaging,* vol. 28, pp. 374-383, 2009.

[81] H. Yu, C. Caldwell, K. Mah, I. Poon, J. Balogh, R. MacKenzie, N. Khaouam, and R. Tirona, "Automated Radiation Targeting in Head-and-Neck Cancer Using Region-Based Texture Analysis of PET and CT Images," *International Journal of Radiation Oncology Biology Physics,* vol. 75, pp. 618-625, 2009.

[82] G. V. Saradhi, G. Gopalakrishnan, A. S. Roy, R. Mullick, R. Manjeshwar, K. Thielemans, and U. Patil, "A framework for automated tumor detection in thoracic FDG PET images using texture-based features," *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro,* pp. 97-100, 2009

[83] M. E. Mayerhoefer, M. Breitenseher, G. Amann, and M. Dominkus, "Are signal intensity and homogeneity useful parameters for distinguishing between benign and malignant soft tissue masses on MR images? Objective evaluation by means of texture analysis," *Magnetic Resonance Imaging,* vol. 26, pp. 1316-1322, 2008.

[84] J. M. Garcia-Gomez, C. Vidal, L. Marti-Bonmati, J. Galant, N. Sans, M. Robles, and F. Casacuberta, "Benign/malignant classifier of soft tissue tumors using MR imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine,* vol. 16, pp. 194-201, 2004.

[85] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, and N. Betrouni, "Prostate cancer characterization on MR images using fractal features," *Medical Physics,* vol. 38, pp. 83-95, 2011.

[86] M. E. Mayerhoefer, W. Schima, S. Trattnig, K. Pinker, V. Berger-Kulemann, and A. Ba-Ssalamah, "Texture-Based Classification of Focal Liver Lesions on MRI at 3.0 Tesla: A Feasibility Study in Cysts and Hemangiomas," *Journal of Magnetic Resonance Imaging,* vol. 32, pp. 352-359, 2010.

[87] B. J. Woods, B. D. Clymer, T. Kurc, J. T. Heverhagen, R. Stevens, A. Orsdemir, O. Bulan, and M. V. Knopp, "Malignant-Lesion Segmentation Using 4D Co-Occurrence Texture Analysis Applied to Dynamic Contrast-Enhanced Magnetic Resonance Breast Image Data," *Journal of Magnetic Resonance Imaging,* vol. 25, pp. 495-501, 2007.

[88] D. Assefa, H. Keller, C. Ménard, N. Laperriere, R. J. Ferrari, and I. Yeung, "Robust texture features for response monitoring of glioblastoma multiforme on T1-weighted and T2-FLAIR MR images: A preliminary investigation in terms of identification and segmentation," *Medical Physics,* vol. 37, pp. 1722-1736, 2010.

[89] P. E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, "Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters," *Acta Oncologica,* vol. 49, pp. 1012-1016, 2010.

[90] F. Tixier, M. Hatt, C. C. Le Rest, A. Le Pogam, L. Corcos, and D. Visvikis, "Reproducibility of Tumor Uptake Heterogeneity Characterization Through

Textural Feature Analysis in ^{18}F-FDG PET," *Journal of Nuclear Medicine,* vol. 53, pp. 693-700, 2012.

[91] R. A. Lerski, L. R. Schad, R. Luypaert, A. Amorison, R. N. Muller, L. Mascaro, P. Ring, A. Spisni, X. Zhu, and A. Bruno, "Multicentre magnetic resonance texture analysis trial using reticulated foam test objects," *Magnetic Resonance Imaging,* vol. 17, pp. 1025-1031, 1999.

[92] M. E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig, "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study," *Medical Physics,* vol. 36, pp. 1236-1243, 2009.

[93] S. A. Waugh, R. A. Lerski, L. Bidaut, and A. M. Thompson, "The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms," *Medical Physics,* vol. 38, pp. 5058-5066, 2011.

[94] L. Alic, J. Veenland, M. Van Vliet, C. F. Van Dijke, A. M. M. Eggermont, and W. J. Niessen, "Quantification of heterogeneity in dynamic contrast enhanced MRI data for tumor treatment assessment," *3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings*, pp. 944-947, 2006.

[95] L. Harrison, P. Dastidar, H. Eskola, R. Järvenpää, H. Pertovaara, T. Luukkaala, P. L. Kellokumpu-Lehtinen, and S. Soimakallio, "Texture analysis on MRI images of non-Hodgkin lymphoma," *Computers in Biology and Medicine,* vol. 38, pp. 519-524, 2008.

[96] P. A. Eliat, D. Lechaux, A. Gervais, N. Rioux-Leclerc, F. Franconi, L. Lemaire, L. Dazord, V. Catros-Quemener, and J. D. De Certaines, "Is magnetic resonance imaging texture analysis a useful tool for cell therapy in vivo monitoring?," *Anticancer Research,* vol. 21, pp. 3857-3860, 2001.

[97] I. El Naqa, P. W. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W. L. Thorstad, and J. O. Deasy, "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes," *Pattern Recognition,* vol. 42, pp. 1162-1171, 2009.

[98] P. Maday, P. Khurd, L. Ladic, M. Schnall, M. Rosen, C. Davatzikos, and A. Kamen, "Imaging as a Surrogate for the Early Prediction and Assessment of Treatment Response through the Analysis of 4-D Texture Ensembles (ISEPARATE)," *Lecture Notes in Computer Science*, pp. 164-173, 2011.

[99] F. Tixier, C. C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J. P. Metges, L. Corcos, and D. Visvikis, "Intratumor Heterogeneity Characterized by Textural Features on Baseline ^{18}F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer," *Journal of Nuclear Medicine,* vol. 52, pp. 369-378, 2011.

[100] M. Vaidya, K. M. Creach, J. Frye, F. Dehdashti, J. D. Bradley, and I. El Naqa, "Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer," *Radiotherapy and Oncology,* vol. 102, pp. 239-245, 2012.

[101] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 3, pp. 610-621, 1973.

[102] D. Markel, "Automatic Segmentation of Lung Carcinoma using 3D Texture Features in Co-Registered 18-FGD PET/CT Images," Master of Science, Medical Biophysics, University of Toronto, 2011.
[103] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Canadian Journal of Remote Sensing,* vol. 28, pp. 45-62, 2002.
[104] S. Fu, B. Muralikrishnan, and J. Raja, "Engineering Surface Analysis With Different Wavelet Bases," *Journal of Manufacturing Science and Engineering, Transactions of the ASME,* vol. 125, pp. 844-852, 2003.
[105] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press, 1996.
[106] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms : A Primer*. Upper Saddle River, N.J: Prentice Hall, 1998.
[107] M. Kociolek, A. Materka, M. Strzelecki, and P. Szczypinski, "Discrete wavelet transform - derived features for digital image texture analysis," *Proceedings of Interational Conference on Signals and Electronic Systems*, pp. 163-168, 2001.
[108] G. Pajares and J. M. de la Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognition,* vol. 37, pp. 1855-1872, 2004.
[109] A. Ardeshir Goshtasby and S. Nikolov, "Image fusion: Advances in the state of the art," *Information Fusion,* vol. 8, pp. 114-118, 2007.
[110] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor Image Fusion Using the Wavelet Transform," *Graphical Models and Image Processing,* vol. 57, pp. 235-245, 1995.
[111] I. Aganj, C. Lenglet, E. Yacoub, G. Sapiro, and N. Harel, "A 3D Wavelet Fusion Approach for the Reconstruction of Isotropic-Resolution MR Images From Orthogonal Anisotropic-Resolution Scans," *Magnetic Resonance in Medicine,* vol. 67, pp. 1167-1172, 2012.
[112] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. New York, NY: Wiley, 2000.
[113] D. G. Kleinbaum, M. Klein, and E. R. Pryor, *Logistic Regression : A Self-Learning Text*, 2nd ed. New York, NY: Springer-Verlag, 2002.
[114] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman and Hall, 1989.
[115] C. R. Rao, *Linear Statistical Inference and its Applications*, 2nd ed. New York, NY: Wiley, 1973.
[116] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics,* vol. 7, pp. 1-26, 1979.
[117] R. Wehrens, H. Putter, and L. M. C. Buydens, "The bootstrap: a tutorial," *Chemometrics and Intelligent Laboratory Systems,* vol. 54, pp. 35-52, 2000.
[118] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall, 1993.
[119] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press, 1997.
[120] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," *Statistics in Medicine,* vol. 19, pp. 1141-1164, 2000.

[121] M. A. Gadd, E. S. Casper, J. M. Woodruff, P. M. McCormack, and M. F. Brennan, "Development and Treatment of Pulmonary Metastases in Adult Patients with Extremity Soft Tissue Sarcoma," *Annals of Surgery,* vol. 218, pp. 705-712, 1993.

[122] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging,* vol. 22, pp. 986-1004, 2003.

[123] J. O. Deasy, A. I. Blanco, and V. H. Clark, "CERR: A computational environment for radiotherapy research," *Medical Physics,* vol. 30, pp. 979-985, 2003.

[124] F. J. Anscombe, "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika,* vol. 35, pp. 246-254, 1948.

[125] H. Martens, A. K. Thybo, H. J. Andersen, A. H. Karlsson, S. Donstrup, H. Stodkilde-Jorgensen, and M. Martens, "Sensory Analysis for Magnetic Resonance-image Analysis: Using Human Perception and Cognition to Segment and Assess the Interior of Potatoes," *LWT - Food Science and Technology,* vol. 35, pp. 70-79, 2002.

[126] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory,* vol. IT-28, pp. 129-137, 1982.

[127] J. Max, "Quantizing for Minimum Distortion," *Ire Transactions on Information Theory,* vol. 6, pp. 7-12, 1960.

[128] A. K. Jain, *Fundamentals of Digital Image Processing*. London: Prentice-Hall International, 1989.

[129] A. Kaufman, D. Cohen, and R. Yagel, "Volume Graphics," *Computer,* vol. 26, pp. 51-64, 1993.

[130] I. El Naqa, G. Suneja, P. E. Lindsay, A. J. Hope, J. R. Alaly, M. Vicic, J. D. Bradley, A. Apte, and J. O. Deasy, "Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships," *Physics in Medicine and Biology,* vol. 51, pp. 5719-5735, 2006.

[131] J. G. Sied, A. P. Zijdenbos, and A. C. Evans, "A Nonparametric Method for Automatic Correction of Intensity Nonuniformity in MRI Data," *IEEE Transactions on Medical Imaging,* vol. 17, pp. 87-97, 1998.

[132] L. G. Nyul, J. K. Udupa, and X. Zhang, "New Variants of a Method of MRI Scale Standardization," *IEEE Transactions on Medical Imaging,* vol. 19, pp. 143-150, 2000.

[133] T. S. Lee, "Image Representation Using 2D Gabor Wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 959-971, 1996.

[134] S. Arivazhagan, L. Ganesan, and S. P. Priyal, "Texture classification using Gabor wavelets based rotation invariant features," *Pattern Recognition Letters,* vol. 27, pp. 1976-1982, 2006.

[135] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research: Special Issue on Variable and Feature Selection,* vol. 3, pp. 1157-1182, 2003.

[136] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY: Springer, 2002.

[137] F. O'Sullivan, E. Wolsztynski, J. O'Sullivan, T. Richards, E. U. Conrad, and J. F. Eary, "A Statistical Modeling Approach to the Analysis of Spatial Patterns of

FDG-PET Uptake in Human Sarcoma," *IEEE Transactions on Medical Imaging,* vol. 30, pp. 2059-2071, 2011.

[138] I. A. Burger, D. M. Huser, C. Burger, G. K. Von Schulthess, and A. Buck, "Repeatability of FDG quantification in tumor imaging: averaged SUVs are superior to SUV_{max}," *Nuclear Medicine and Biology,* vol. 39, pp. 666-670, 2012.

[139] R. A. Fisher, *The Designs of Experiments*. New York, NY: Hafner, 1935.

[140] E. J. G. Pitman, "Significance Tests Which May be Applied to Samples From any Populations," *Supplement to the Journal of the Royal Statistical Society,* vol. 4, pp. 119-130, 1937.