FEDERAL UNIVERSITY OF ALAGOAS
INSTITUTE OF COMPUTING
GRADUATE PROGRAM IN INFORMATICS

Masters dissertation

Multimodality CT/MRI Radiomics for Lung Nodule Malignancy Suspiciousness Classification

Anthony Emanoel de Albuquerque Jatobá
aeaj@ic.ufal.br

Advisor:

Prof. Dr. Marcelo Costa Oliveira

MACEIÓ, OCTOBER OF 2021

Anthony Emanoel de Albuquerque Jatobá

Multimodality CT/MRI Radiomics for Lung Nodule Malignancy Suspiciousness Classification

Dissertation presented to the Graduate Program in Informatics at the Federal University of Alagoas as a requirement for obtaining the Master's Degree in Informatics.

Advisor: Prof. Dr. Marcelo Costa Oliveira

Maceió

October of 2021

Universidade Federal de Alagoas
Divisão de Tratamento Técnico
Catalogação na Fonte
Bibliotecário: Marcelino de Carvalho Freitas Neto - CRB-4 – 1767

J32m Jatobá, Anthony Emanoel de Albuquerque.
Multimodality CT/MRI Radiomics for Lung Nodule Malignancy Suspiciousness Classification / Anthony Emanoel de Albuquerque Jatobá.
-. -- Maceió: 2021.
45 f.: il.

Orientação: Marcelo Costa Oliveira
Mestrado em Informática - Universidade Federal de Alagoas. Instituto de Computação, Maceió, 2021.
Bibliografia: f. 35-38.
Apêndice: f. 39-45.

1. Processamento de imagem assistida por computador. 2. Neoplasias 3. Características Radiômicas I.Título.

CDU:004.932

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

ANTHONY EMANOEL DE ALBUQUERQUE JATOBA

MULTIMODALITY CT/MRI RADIOMICS FOR LUNG NODULEMALIGNANCY SUSPICIOUSNESS CLASSIFICATION

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 29 de outubro de 2021.

Banca Examinadora:

Documento assinado digitalmente
gov.br
MARCELO COSTA OLIVEIRA
Data: 01/11/2021 11:48:47-0300
Verifique em https://verificador.iti.br

_____
Prof. Dr. MARCELO COSTA OLIVEIRA
UFAL – Instituto de Computação
Orientador

Documento assinado digitalmente
gov.br
THALES MIRANDA DE ALMEIDA VIEIRA
Data: 01/11/2021 12:27:07-0300
Verifique em https://verificador.iti.br

_____
Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA
UFAL – Instituto de Computação
Examinador Interno

Documento assinado digitalmente
gov.br
PAULO MAZZONCINI DE AZEVEDO MARQUES
Data: 03/11/2021 10:24:43-0300
Verifique em https://verificador.iti.br

_____
Prof. Dr. PAULO MAZZONCINI DE AZEVEDO MARQUES
USP - Universidade de São Paulo
Examinador Externo

Resumo

O câncer de pulmão é o tipo mais frequente e letal de câncer e o seu diagnóstico precoce é crucial para a sobrevivência do paciente. A tomografia computadorizada (TC) é o padrão-ouro para o rastreio da doença, mas estudos recentes têm demonstrado o potencial da ressonância magnética (RM) no diagnóstico de nódulos pulmonares, bem como da combinação de imagens médicas multimodalidade. Neste estudo, foi avaliado se a combinação de características radiômicas de imagens de TC e RM de pacientes de câncer pulmonar contribui para classificações mais precisas da suspeita de malignidade de nódulos. Para atingir tal objetivo, foi realizado o registro de exames de TC e RM de 47 pacientes, segmentação dos nódulos em cada modalidade, extração de características radiômicas dos nódulos, e classificação usando XGBoost, avaliando métricas de desempenho dos modelos em 30 iterações. O mesmo experimento foi realizado para quatro conjuntos de características: 1) somente de TC; 2) somente de RM; 3) concatenação de TC e RM; 4) fusão de TC e RM. Nossos resultados indicam que a estratégia de fusão de imagens pode levar a ganhos de desempenho significativos, em um teste dos postos sinalizados de Wilcoxon, sobre os modelos de modalidades individuais, com AUC média de *0.794*, mas a concatenação de características não se provou uma abordagem adequada para lidar com imagens multimodalidade, uma vez que a AUC média de *0.770* não indicou ganhos de desempenho. Além disso, foi observado que RM, com AUC média de *0.770*, apresentou desempenho significativamente superior à TC, com *0.755*, encorajando estudos em RM como modalidade para o acompanhamento do câncer de pulmão. Por fim, a análise das características reforçou a relevância da morfologia de um nódulo, como seu tamanho e esfericidade, além de características de textura que quantificam a complexidade e homogeneidade do ambiente intratumoral.

Palavras-chave: Imagens médicas multimodalidade; Câncer de pulmão; Características Radiômicas.

Abstract

Lung cancer is the most common and deadly form of cancer, and its early diagnosis is decisive to the patient's survival. Computed Tomography (CT) is the gold-standard imaging modality for lung cancer management, but recent studies have shown the potential of Magnetic Resonance Imaging (MRI) in lung cancer diagnosis and how combining multimodality medical images can yield better outcomes. In this study, we evaluated whether the combination of CT and MRI scans from lung cancer patients can leverage a more precise malignancy suspiciousness classification. For such, we registered paired CT and MRI scans from 47 patients, segmented the nodules in each modality, extracted radiomics features, and performed an experiment with an XGBoost classifier, evaluating models' performance metrics across 30 trials. The same experiment was performed with four sets of features: 1) CT-only; 2) MRI-only; 3) CT and MRI features; 4) CT/MRI fused images. Our results indicate that the image fusion approach can yield significant AUC performance gains over the single modalities models, with an average AUC of *0.794*, but feature concatenation is not an adequate strategy for dealing with multimodality data, as its average AUC of *0.770* didn't indicate any improvement over the single modalities. Additionally, we observed that MRI, with an average AUC of *0.770*, has shown significantly better performance than CT, with *0.754*, encouraging further studies in MRI as a lung cancer management image modality. Finally, the analysis on the importance of radiomics features reinforced the relevance of features that reflects on morphological characteristics of a nodule, such as its dimension and roundness, as well as texture features that relate to the intratumoral environment, measuring its complexity and homogeneity.

Keywords: Multimodality medical imaging; Lung cancer; Radiomics.

List of Figures

List of Tables

Acronyms

AUC	Area Under the ROC Curve
CADe	Computed-Aided Detection
CADx	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
HU	Hounsfield Units
LDCT	Low-dose Computed Tomography
MRI	Magnetic Resonace Imaging
NGTDM	Neighboring Gray-tone Difference Matrix
NSCLC	Non-Small Cell Lung Cancer
PET	Positron emission tomography
ROC	Receiver Operating Characteristics
ROI	Region of Interest
SHAP	SHapley Additive exPlanations

STS Soft-tissue sarcomas

SVM Support Vector Machines

T Tesla

VOI Volume of Interest

Contents

1 Introduction

Globally, lung cancer is the most common cause of cancer incidence and mortality. The Global Cancer Observatory (GCO) estimated 2.1 million new cases and 1.8 million deaths in 2018 (SIEGEL; MILLER; JEMAL, 2018). The moment of detection is a determinant factor for the prognosis; e.g., when a nodule is detected in its early stages, survival rates can reach up to 90%, in contrast to only 15% when diagnosed in its final stages (KNIGHT et al., 2017a). Therefore, early diagnosis is a decisive factor for patients' treatment and survival (SIEGEL; MILLER; JEMAL, 2018; KNIGHT et al., 2017b).

Currently, early detection of lung nodules is supported by screening programs using computed tomography (CT). However, despite being the gold standard in lung cancer screening, CT still presents some shortcomings; since the exam requires a considerable radiation dose, performing periodic examinations can be undesirable because of the risks of radiation-induced cancers (OHNO, 2014; KNIGHT et al., 2017b). In this context, low-dose computed tomography (LDCT) comes up as an alternative that requires a lower radiation dose, whose safety was backed up by several trials (PASTORINO et al., 2012; RAMPINELLI et al., 2017). Nevertheless, the reduced radiation dose from LDCT provides noisier images that can result in a high rate of false positives, leading to unnecessary invasive procedures (LI et al., 2018; LI et al., 2019).

In the last decades, technical advancements in sequencing, scanners, and coils made magnetic resonance imaging (MRI) a viable modality for the management of patients with chest diseases such as lung cancer (YI et al., 2008; SOMMER et al., 2014; KOENIGKAM-SANTOS et al., 2015). Apart from not exposing the patient to radiation, MRI presents particular advantages over CT and LDCT, such as superior soft-tissue contrast, which allows for better characterization of nodules (YI et al., 2008; BECKETT; MORIARITY; LANGER, 2015). Still, thoracic MRI is not as established as CT, requiring further clinical trials and protocol development to ascertain its actual potentials. Nonetheless, preliminary works display the benefits of using MRI as a complement to the standard screening protocol with CT (OHNO et al., 2018; FRANCISCO et al., 2019).

As a result of recent developments in both hardware and software, multimodality medical

imaging has been progressively applied in research and clinical practice (WEI et al., 2019). The intuition behind multimodality imaging is that distinct modalities can provide complementary information of a disease, allowing for better characterization and decision support (WEI et al., 2019). Moreover, this field holds potential for computer-aided diagnosis (CADx) research, as several works suggest that the combination of different imaging modalities can bring better results in computer-aided segmentation (GUO et al., 2019), therapy planning (VAIDYA et al., 2012), and prognosis (WEI et al., 2019) when compared to single-modality imaging.

While the literature on lung cancer CT-based CADx systems is vast, few studies have addressed the applicability of multimodality medical images for this type of cancer (YANG et al., 2018). Furthermore, CT/MRI is not a typical combination of imaging modalities for lung cancer assessment, making it beneficial to study the combination of these images' capabilities in CADx systems design (YANG et al., 2018). Lastly, there are still gaps in the knowledge on MRI as a lung cancer management image modality (FRANCISCO et al., 2019).

In this study, we intended to determine whether the combination of radiomics features extracted from lung nodules in paired CT and MRI scans can yield better classification results than its individual modalities. As a secondary objective, we compared the performance of two different approaches to combine multimodality information.

The work's main contribution is in showing that combining multimodality images can provide a more precise lung cancer malignancy suspiciousness classification. Our results pointed that a simple pixel average image fusion is enough to provide significant gains in most classification metrics compared to the best single modalities models. On the other hand, concatenating the features into a larger dataset has shown to be a poor strategy for dealing with multimodality data. A secondary contribution is in showing that MRI models presented an advantage over CT models. This improvement is exciting because using MRI for lung cancer management can mitigate issues such as radiation exposure and adverse reactions to contrast materials commonly used in CT.

The remainder of this dissertation is structured in the following chapters:

- **Chapter 2 - Theoretical Framework**: this chapter covers the main theoretical concepts related to this work, such as medical imaging modalities, radiomics, and multi-

modality medical imaging.

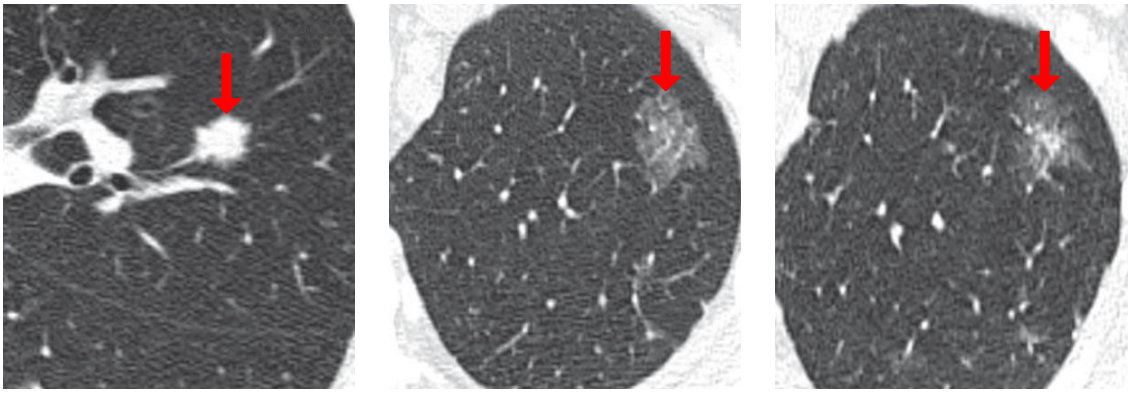
- **Chapter 3 - Related Work**: this chapter presents an overview of the multimodality medical images machine learning research field.
- **Chapter 4 - Material and Methods**: this chapter describes the steps taken into our experiments, including the source of data, participants' information, data preprocessing pipeline, multimodality fusion approaches, and model development and validation.
- **Chapter 5 - Results and Discussion**: this chapter reports the results obtained through the experiments and its interpretation.
- **Chapter 6 - Conclusion**: this chapter concludes this work by summarising the main observed results, its limitations, and future works.

2 Theoretical Framework

2.1 Lung Nodules in Medical Images

The Fleischner Society defines a lung nodule as an approximately rounded opacity in radiography or CT imaging, well or poorly defined, measuring up to 30mm in diameter (HANSELL et al., 2008). Rounded lesions measuring more than 30mm are designated lung masses and should be considered indicative of lung cancer until histologically proven otherwise (LOVERDOS et al., 2019).

Lung nodules are categorized into three different types according to their attenuation in CT imaging, as shown in Figure 1: (1a) solid nodules, the most common, characterized by uniform soft-tissue attenuation; (1b) ground-glass nodules, nonuniform in appearance with a hazy attenuation of lung parenchyma not obscuring the underlying structures, and (1c) part-solid nodules, comprising both solid and ground-glass attenuation characteristics (HANSELL et al., 2008).



(a) Solid lung nodule. (b) Ground-glass lung nodule. (c) Part-solid lung nodule.

Figure 1 – Different lung nodules types.

2.2 Computed Tomography

Computed tomography is a medical imaging technique that combines multiple X-ray measurements to produce cross-sectional (tomographic, from the greek *tomos*, "slice") images of a scanned object. Figure 2 exhibits an overview of a CT scanner operation. The

patient is placed well-centered and usually in a supine position over the table. Then, the CT computer controls table motion by using precision motors and moves the patient along the bore, where the data is effectively collected. Inside the gantry, an X-ray tube performs a full 360-degree rotation emitting a fan of X-ray beams that are received by a detector array that rotates along the same axis. Modern CT scanners possess more detectors arrays in the z-axis, improving acquisition time and image quality.

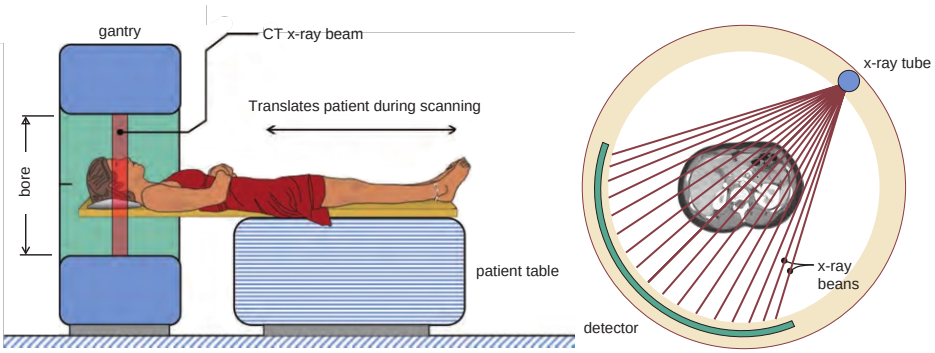


Figure 2 – CT schematics. (RASHI, 2020)

X-ray imaging is acquired by emitting a pulse of X-radiation through the body of the patient. A fraction of the radiation interacts with the patient tissues, while the remaining pass through the body, reaching the detector. The result is an X-ray distribution that reflects the tissues' unique attenuation properties, such as bone, soft tissue, and air inside the patient (BUSHBERG; BOONE, 2011). This distribution is measured in Hounsfield Units (HU), which standardizes the values of each tissue in a scale, which is shown in Figure 3.

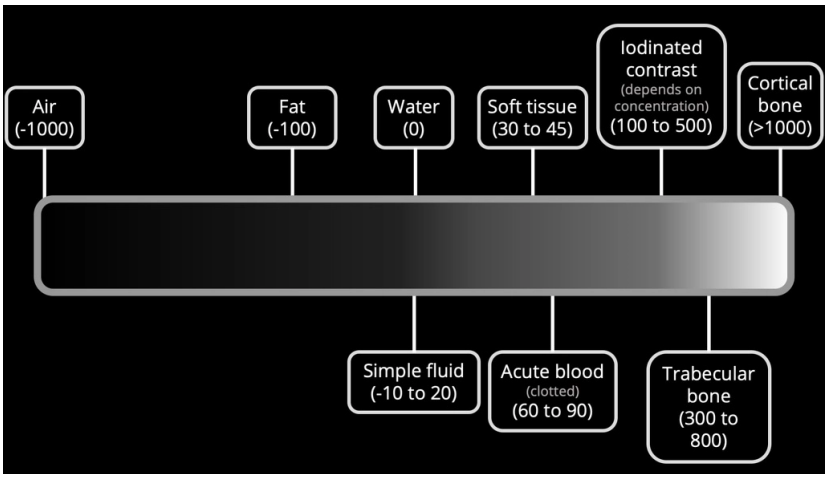


Figure 3 – Hounsfield Scale (HU) simplified. (RASHI, 2020)

2.3 Magnetic Resonance Imaging

Magnetic resonance is a medical imaging procedure that generates detailed tomographic images of the organs inside the body. MRI uses a powerful magnet to generate a strong magnetic field (1.5T and 3T are typical values) that align the protons of the patient's body with that field. A radiofrequency current is pulsed through the patient, causing the protons to misalign with the magnetic field. When the radiofrequency pulse is turned off, the sensors can detect the energy which is released as the protons realign with the magnetic field, whose time varies according to the chemical nature of the molecule, allowing for tissue differentiation (BUSHBERG; BOONE, 2011). Figure 4 shows an MRI scanner and its main components.

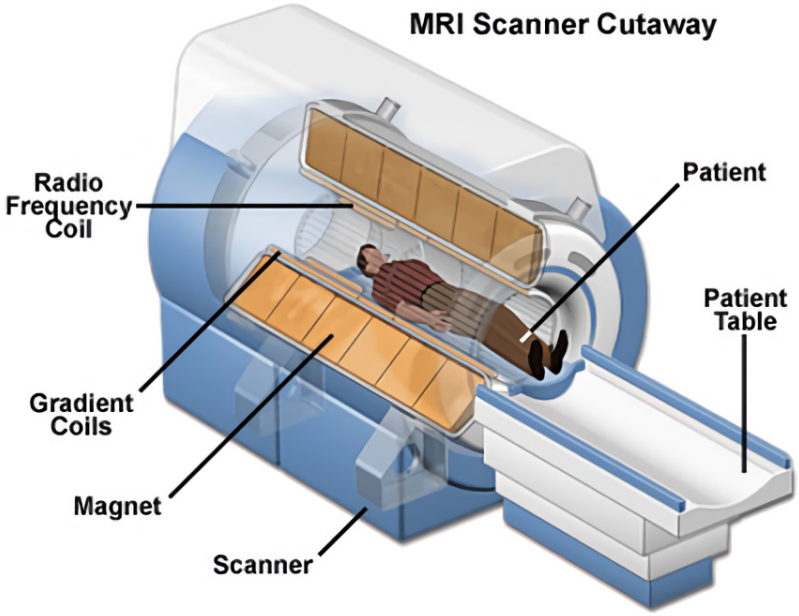
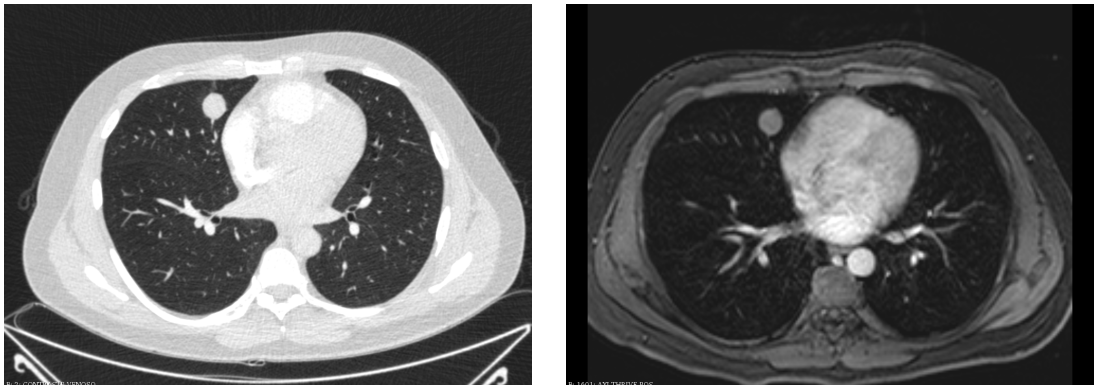


Figure 4 – MRI scanner components. (RASHI, 2020)

The MRI acquisition protocol can be fine-tuned to better reflect specific tissues properties. For example, T1-weighted images can be acquired through the T1 relaxation process (magnetization in the same direction as the magnetic field) by allowing magnetization to recover before measuring the MR signal. In these settings, MRI can better reflect tissues such as fat.

CT and MRI images may differ significantly, as each modality is sensitive to different tissue properties. CT images present lower soft tissue detail, as denser tissues such as bones block the X-rays. MRI customarily uses the hydrogen nucleus spin to capture the imaging

signal, as this atom is abundant in water and fat, allowing this modality to present richer soft tissue contrast (BUSHBERG; BOONE, 2011). Figure 5 compares the same anatomical region as seen in CT (5a) and T1 MRI (5b).



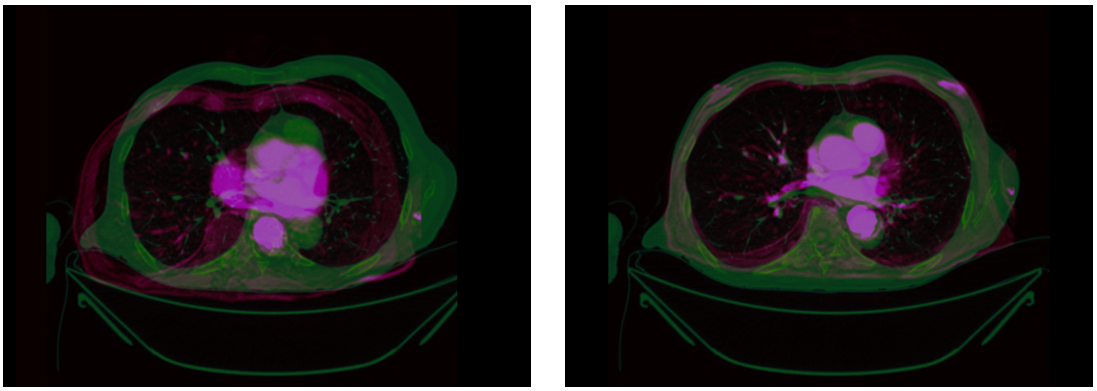
(a) CT image. (b) T1 MRI.

Figure 5 – A comparison between CT and different MRI sequences.

2.4 Image Registration

Image registration consists of finding the spatial transform that maps points from one image to the corresponding points in another image (MAINTZ; VIERGEVER, 1998). Medical image registration has many applications. For instance, repeated images of a subject are often used to obtain temporal information of diseases, and image registration is utilized to align these acquisitions and allow for the detection of more subtle changes (YOO, 2004). Moreover, registration is a valuable tool for correlating information obtained from different imaging modalities (MAINTZ; VIERGEVER, 1998; BASHIRI et al., 2019). Figure 6 illustrates the registration process by presenting CT (in cyan) and MRI (in magenta) overlapped slices in the transverse plane before and after performing rigid image registration.

A registration problem is classified according to the spatial transformation used to map the points of one or more moving images to a fixed image. The transformations may be rigid, when only translations and rotations are allowed; affine, when skewing and scaling transformations are also allowed; and elastic, whose transformations can define free-form mappings between the images (YOO, 2004).



(a) Before registration. (b) After registration.

Figure 6 – Overlapped CT (in cyan) and MRI (in magenta) slices in the transverse plane before and after performing image registration.

Registration can be seen as an optimization problem, where the objective is to minimize a cost function that maps the quality of the alignment between the images. A typical cost function is the Mean Square Error (MSE) (Equation 1), which accounts for the difference in pixel intensities between the corresponding pixels X in the fixed image and Y in the moving image.

$$MSE(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{1}$$

However, when it is intended to register images from different modalities, the raw pixel values are not a useful measure of registration quality. In this case, it is preferable to account for the statistical distribution of pixel values, using a metric such as Mutual Information (2), where X and Y are random variables defining the fixed and moving image intensities, respectively; p(x,y) is the joint probability of X and Y; and $p(x)$ and $p(y)$ are the marginal probability of X and Y, respectively.

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

Mutual Information measures how much one random variable tells us about another, and can be thought of as the reduction in uncertainty about one random variable, given knowledge of another. Higher mutual information values indicate a larger reduction in uncertainty; a value of zero indicates that the two random variables are independent.

2.5 Nodule Segmentation

Segmentation consists of subdividing an image into distinct regions or objects. In lung cancer management, the goal is to divide the image into regions that represent normal anatomy and those that are abnormal, such as a nodule or its subregions (GILLIES; KINAHAN; HRICAK, 2016). Accurate segmentation of a nodule region plays an important role in image interpretation, analysis, and measurement. Nevertheless, lung nodule segmentation is challenging because many tumors may have indistinct borders (such as ground-glass opacity nodules) or be attached to adjacent lung structures, such as the lung wall or mediastinum.

A critical question is whether segmentation will be performed manually, thus having closer to ground truth volumes, or automatically, allowing for better reproducibility and design of automated CADe and CADx systems (GILLIES; KINAHAN; HRICAK, 2016). An example of a semi-automatic segmentation approach is the GrowCut algorithm (ZHU et al., 2014). This algorithm requires the specialist to manually draw seed regions of the tissues to be segmented in each anatomical plane. These labels are then propagated based on principles from cellular automata to classify all the voxels as foreground (nodule) or background (the remaining of the image).

2.6 Radiomics

Radiomics is the field that aims to convert digital medical images into high-dimensional data for improved decision support, based on the premise that biomedical images contain information that reflects underlying pathophysiology and that these relationships can be revealed via quantitative image analyses (GILLIES; KINAHAN; HRICAK, 2016). Quantitative image features may include intensity, shape, size, volume, and texture, offering information on tumor phenotype and habitat that is distinct from that provided by clinical reports and laboratory test results.

Although radiomics can be used in a series of applications, it is most well developed in oncology because of support from the National Cancer Institute (NCI), Quantitative Imaging Network (QIN), and other initiatives. Possible applications are improved decision support (e.g., diagnosis, prognosis assessment, therapy response) and precision medicine (e.g., therapy planning) (PAREKH; JACOBS, 2019).

2.7 Multimodal Machine Learning and Medical Imaging

As humans, we can perceive the world surrounding us through various modalities - such as image, sound, textures, and odors. In general terms, a modality consists of a way in which something can be experienced, for example, through our senses. This definition can also be applied to machine learning research: a problem is multimodal when its inputs include multiple modalities. Therefore, multimodality machine learning aims to develop models that can process and correlate information from multiple modalities (BALTRUŠAITIS; AHUJA; MORENCY, 2018).

Specifically, the research field of multimodal medical images aims to combine distinct medical images for diagnostic, therapeutic, and prognostic purposes. Using more than one image modality can provide separate yet complementary anatomical (e.g., X-ray and CT) and functional (e.g., PET and MRI) information of a patient, transforming the way the body can be studied but also presenting several challenges due to the diversity of biophysical and biochemical mechanisms (WEI et al., 2019). Baltrušaitis et al. presents five core challenges that surround this research field (BALTRUŠAITIS; AHUJA; MORENCY, 2018):

1. **Representation:** how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities.

2. **Translation:** how to map information from one modality to another, as not only is the data heterogeneous, but the relationship between modalities is often subjective.

3. **Alignment:** identify the direct relations between elements and subelements from different modalities.

4. **Fusion:** to join information from multiple modalities to perform a prediction.

5. **Co-learning:** to transfer knowledge between modalities, their representation, and their predictive models.

In this work, we had to tackle specifically the challenges of representation, alignment, and fusion, detailed in the Feature Extraction (4.4), Image Registration (4.2) and Multimodality Strategies (4.5) subsections.

2.8 Multimodal Image Feature Fusion Strategies

Multimodal image fusion is an essential step in multimodality model design, comprehending the strategy according to which image information is combined. Guo *et al.* (GUO et al., 2019) proposes an algorithmic abstraction for image fusion strategies that cover most supervised multimodal medical imaging analysis. The abstraction divides the possible strategies into fusion at the decision-making, feature, and classifier level, according to the moment where information is combined into the predictive model pipeline. Although this abstraction is proposed for Deep Learning models, the concepts are also valid in radiomics-based systems.

Fusion at Decision-Making Level

This strategy is the most intuitive one: each modality is used to learn a single-modality classifier. Then, the single-modality classifiers' output is combined into a multimodality decision in a process referred to as "voting", although the decision scheme may vary.

Fusion at Feature Level

In this strategy, multimodality images are used together to learn a unified feature set to support the learning of a classifier. A common practice is fusing the images from multiple modalities into a new image, requiring proper registration of those images.

Fusion at Classifier Level

In this strategy, each modality's images are used as separate inputs to learn individual feature sets that will support the learning of a multimodal classifier. Single modality feature learning and classifier learning can be conducted in an integrated framework (e.g., using a CNN with one input for each modality) or separately (e.g., extracting features from single-modality images and then training a multimodality classifier).

2.9 Gradient Trees Boosting and XGBoost

Gradient tree boosting is a technique that composes ensembles of decision trees and calculates the final prediction by summing up the score in the corresponding leaves of each tree. XGBoost is a scalable machine learning system for tree boosting available as an open-source package widely used and displayed state-of-the-art performance on many standard classification benchmarks (RAMRAJ et al., 2016), competitions (NIELSEN, 2016), and real-world applications (HE et al., 2014). Being a decision tree-based model, this method has a built-in feature selection mechanism and can be seen as actively taking the bias-variance tradeoff into account when fitting models. They start with a low variance, high bias model and gradually decrease bias by reducing the size of neighborhoods where it seems most necessary. Consequently, model development is significantly simplified, allowing the specialist to dedicate more effort to improve the quality of data (NIELSEN, 2016).

2.10 Model Interpretation

Another advantage of tree-based models is their wide support for model interpretation approaches, including feature importance analysis. A simple way of investigating feature importance is by calculating the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value, the more important the feature.

Recently, another approach to model interpretation is becoming popular: SHAP (SHapley Additive exPlanations) is a game-theoretic approach proposed by Lundberg et al. to explain the output of any machine learning model in regards to the contribution of each feature (LUNDBERG; LEE, 2017).

2.11 Hyperparameter Optimization

Nowadays, a vast number of machine learning methods are available, ranging from tree-based models over neural networks and ensemble models. A common characteristic of those methods is that they can be parametrized by a set of hyperparameters which can modify cer-

tain aspects of the learning algorithm and must be set appropriately by the user to maximize the usefulness of the learning approach and performance (CLAESEN; MOOR, 2015).

Hyperparameter optimization was traditionally performed manually or following rules-of-thumb (CLAESEN; MOOR, 2015). However, those approaches leave to be desired in terms of reproducibility and are impractical in situations where the number of parameters is large. To solve those issues, it has become necessary to develop automated approaches to hyperparameter optimization. Currently, a variety of automated hyperparameter approaches are available, ranging from iterating over a grid of values (BERGSTRA; BENGIO, 2012) to probabilistic models (FEURER; HUTTER, 2019). A common baseline for hyperparameter optimization is the random search, which is still reliable and competitive in a series of applications (LI; TALWALKAR, 2020).

3 Related Work

Due to recent technological progress, multimodality imaging has been progressively applied in clinical practice and research. Initially, multimodality imaging applications were essentially combining anatomical and functional images to improve diagnostic accuracy or target definition. More recently, the fusion of various images, such as CT, positron emission tomography (PET), and MRI, has become more common, making new applications possible (VAIDYA et al., 2012; VALLIÈRES et al., 2015; GUO et al., 2019). Employing machine learning techniques across multimodality images has shown improved results than single modality modeling for prognostic and prediction of clinical outcomes, holding great potential for precision medicine (WEI et al., 2019).

Vaidya *et al.* (VAIDYA et al., 2012) used pre-treatment PET and MRI images from 27 patients diagnosed with non-small-cell lung carcinoma (NSCLC) to evaluate post-radiotherapy tumor progression in terms of local and loco-regional recurrence. Thirty-two handcrafted features were extracted from both PET and MRI, including statistical descriptors of each modality, total lesion glycolysis of PET images, intensity volume histogram (I_X and V_X features), and texture features using the co-occurrence matrix. The predictive value of these metrics was evaluated using Spearman's rank correlation coefficient (rs) and multivariate logistic regression. A two-parameter model using PET and MRI attributes yielded a gain of performance of 30% for loco-regional and 7% for local failure compared to single modality models, holding promise as an approach to allow for more individualized treatments.

Desseroit *et al.* (DESSEROIT et al., 2016) used features from PET and MRI images from 116 patients for developing a nomogram (a device for graphical function calculation used in several medical fields) for stratification of patients with NSCLC stage I-III. The tumors VOI were segmented using FLAB algorithm for PET, and 3D Slicer for MRI images, extracting texture features from each modality. The authors discovered that the features PET entropy and MRI zone percentage could be used to build a nomogram with a higher stratification power than staging alone for patients with stage II and III disease.

Vallières *et al.* (VALLIÈRES et al., 2015) combined PET and MRI features for evaluation of lung metastasis risk in soft tissue sarcomas (STS). Nine non-texture and forty-one

texture features were extracted from the tumor region of separate (PET, T1, and T2) and fused (PET/T1 and PET/T2) scans from 51 patients with STSs of the extremities. Image fusion was performed using the wavelet transform, and the influence of different extraction parameters on the predictive value of textures was investigated. The model consisted of a logistic regression classifier that used four texture features from the fused PET/T1 and PET/T2 scans, reaching an AUC of 0.984 ± 0.002 in bootstrapping validation. PET features presented a higher predictive value than MRI; however, the addition of MRI information to PET significantly improved performance.

Mu *et al.* (MU et al., 2018) investigated PET and MRI images for the prediction of immunotherapy response in 64 NSCLC patients. The authors extracted 195 features from the original images and 1,235 features from images fused with multiple methodologies. The best model was a Support Vector Machine (SVM) classifier using 87 fused features and 13 single-modality features, yielding an AUC of 0.82, an improvement of 0.14 in AUC compared to only single-modality features.

Guo *et al.* (GUO et al., 2019) proposes an algorithmic architecture for supervised multimodal image analysis and categorizes feature fusion at Feature Level, Classifier Level, and Decision-Making Level. The authors designed a deep convolutional neural network system for image segmentation to contour lesions of STS using multimodal images of CT, MRI, and PET. The network trained with multiple modalities presented superior performance to networks trained with single modal images. According to the author, performing image fusion within the network layers is generally better than fusing the features at the network output.

Li *et al.* (LI et al., 2019) proposes a deep learning method to fuse multimodality information for tumor segmentation in PET/MRI. The solution consists of a 3D fully convolutional network to produce a probability map for MRI segmentation, followed by a fuzzy variational model to incorporate the probability map and the PET intensity for an accurate multimodality tumor segmentation. The experimental results demonstrated that this model is suitable for small datasets and can outperform the existing deep learning-based multimodality segmentation methods, with a dice similarity index of 0.86 ± 0.05.

Mu *et al.* (MU et al., 2020) proposes a multimodality PET/CT deep-learning-based model to assist NSCLC therapy planning, by combining those modalities into a model to detect biomarkers that can guide patients' therapy. The developed deep learning score was

significantly correlated with different mutations and was evaluated as a non-invasive method for precise quantification of EGFR mutation status in NSCLC patients, which is promising to identify NSCLC patients sensitive to certain treatments.

Studies on multimodality imaging for lung cancer assessment are relatively recent and scarce. To the best of our knowledge, no works were developed regarding lung nodules classification in CT/MRI sequences. Because MRI is not widely used in clinical practice and its capacity is still being investigated, research is severely impaired by this lack of data. Thus, our work intends to fill those gaps in the knowledge by performing an investigation on how these image modalities can be integrated into a more precise lung cancer diagnosis support.

4 Material and Methods

Figure 7 presents an overview of our methods, which can be broken into Data Preparation, Radiomic Feature Extraction, and Classification. First, the dataset was consolidated, and viable study cases were selected to compose our CT and MRI image sets (Section 4.1); Next, CT and MRI images were registered (Section 4.2), and the nodules in each modality were segmented (Section 4.3). Those steps were important to support our Image Fusion strategy, which fused CT and MRI images and segmentation into a new set (Section 4.5). After our data was properly prepared, we were able to extract radiomics features from individual and fused images using the pyradiomics library (Section 4.4). These sets of data supported our classification experiments (Section 4.6), which consisted of training XGBoost models to classify nodules as benign or malignant, recording output such as classification metrics, feature importances, and SHAP values to support analysis.

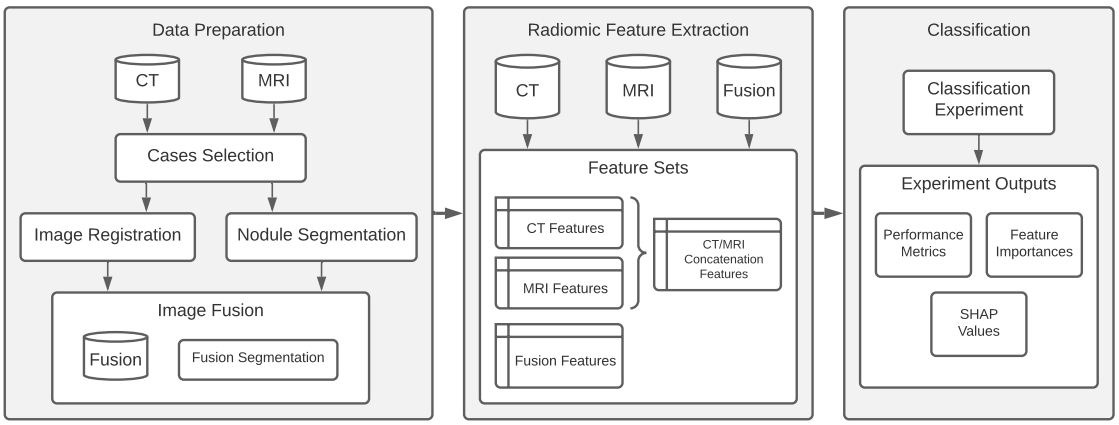
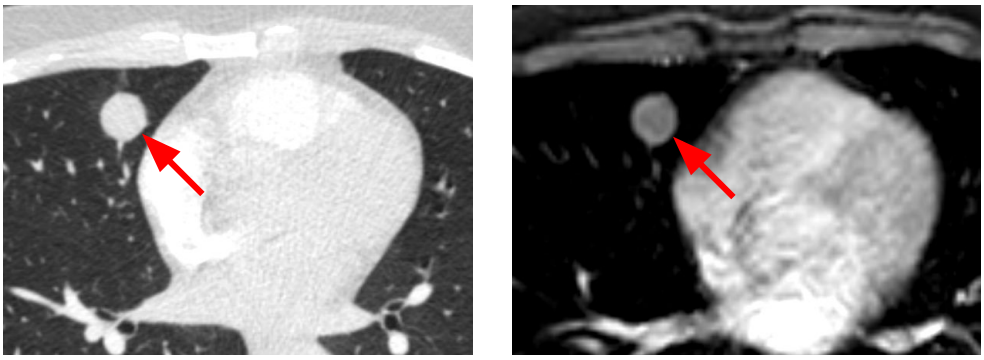


Figure 7 – Overview of the Methods.

4.1 Source of Data

For this retrospective study, we acquired image information from lung cancer patients observed in the Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (HCFMRP/USP) between October 2017 and November 2019, which

had both CT and MRI chest images. Figure 8 shows a lung nodule as seen in CT (8a) and MRI (8b).



(a) CT image. (b) MRI image.

Figure 8 – Lung nodule as seen in each imaging modality.

Our research started from a set of 65 patients, but after a meticulous analysis from two experient radiologists, we discarded the cases which presented: 1) lung masses; 2) undefined diagnosis; and 3) severe image artifacts in any modality. After this filter, we ended up with images from 47 patients whose demographic information is summarized in Table 1. The hospital's institutional research board approved this prospective study with process number 3733/2017 and all patients' informed consent. All exams were anonymized to ensure patients' privacy.

Table 1 – Demographic Information of Study Population

Gender	Male	25
	Female	22
Age	Min-max	18-80
	Median	58
	Mean	57.95
	Standard Deviation	12.77
Diagnosis	Malignant	24
	Benign	23

CT imaging was acquired with a CT scanner (Philips, Brilliance CT Big Bore) and MRI using a 1.5T device (Phillips, Achieva 1.5T). The sequences were obtained with the patients in the supine, head-first position and using the deep inspiration breath-hold technique. The clinical chest MRI protocol included the T1 post-contrast (T1) sequence that approximates

post-contrast CT images with proper spatial and contrast resolution. Table 2 summarizes the DICOM image acquisition parameters.

4.2 Image Registration

The first step before performing image fusion is the image registration, consisting of finding the spatial transform that maps points from one image to the corresponding points in another image (MAINTZ; VIERGEVER, 1998), in our case, CT and MRI images were registered in two steps.

To provide an initial space alignment between the modalities, we used an automatic rigid registration using SimpleITK library (version 2.1.1) (LOWEKAMP et al., 2013). The registration was set up to use CT as fixed image and MRI as moving image, using the Mattes Mutual Information as an optimization metric with *50* histogram bins and a sampling rate of *1%*, a linear interpolator, and a gradient descent optimizer with a learning rate of *1.0* and *400* max iterations. Those settings were chosen empirically, trying to balance registration quality and registration running time.

This initial registration was considered satisfactory under visual inspection. However, since the registration method is applied globally, trying to align the whole chest, the alignment between the nodules was often not adequate since even a slight difference in air volume in the lungs may cause a displacement in nodule position between the images. One possible solution is to apply elastic registration to adjust this difference, however, it could alter the morphology of the nodule, which is one of the prime factors for diagnosis. Consequently, we agreed to fine-tune the register performing an additional manual registration step con-

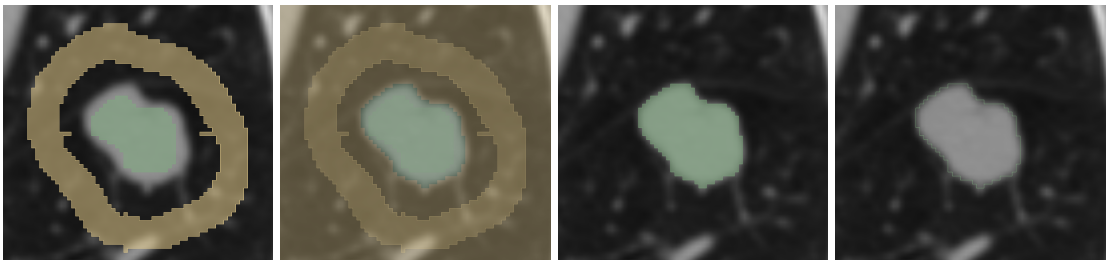
Table 2 – DICOM Image Parameters.

	Image Modality	
Attribute	Computed Tomography	Magnetic Resonance Imaging
SliceThickness	1mm	4mm
SpacingBetweenSlices	1mm	2mm
Pixel Size	0.873mm	1.607mm
Spatial Resolution	512x512	224x224
Bits per Pixel	12	12

sisting of translations in the three axes, assisted by the 3D Slicer (version 4.11.20200930) transforms tool and resampling using linear interpolation.

4.3 Nodule Segmentation

Next, the nodules were segmented to support radiomics features extraction. For such, we used the semi-automatic segmentation algorithm FastGrowCut, available as a plugin for 3D Slicer platform (ZHU et al., 2014). Figure 9 summarizes the algorithm steps, that require the radiologist to draw seed regions of the tissues to be segmented in each anatomical plane (9a). These labels are then propagated to classify all the voxels as either foreground or background (9b), obtaining a segmentation mask (9c) and the nodule outline (9d).



(a) Seeds (b) Grown regions (c) Segmentation mask (d) Final result

Figure 9 – Lung nodule and the segmentation steps taken.

We applied a grayscale lung windowing to highlight the lung tissues by setting the window in 1,400 and level in -500 Hounsfield unit (HU) in the CT images. A radiologist set the level and width values for the MRI sequences to 800 and 2,000, respectively.

4.4 Feature Extraction

The radiomics study area consists of turning medical images into high-dimensional data for improved decision support and precision medicine by extracting a large number of hand-crafted features from a volume of interest (VOI) (GILLIES; KINAHAN; HRICAK, 2016). These features can capture characteristics of this region that otherwise would be challenging or impossible to be discerned even by experienced professionals (HATT et al., 2017).

We extracted a series of radiomics features from each lesion using the open-source library pyradiomics (version 3.0.1) (GRIETHUYSEN et al., 2017). At the time of this work,

pyradiomics supported the following feature classes: First Order Statistics; Shape-based; Gray Level Co-occurrence Matrix (GLCM); Gray Level Run Length Matrix (GLRLM); Gray Level Size Zone Matrix (GLSZM); Neighboring Gray-tone Difference Matrix (NGTDM); and Gray Level Dependence Matrix (GLDM).

Shape-based features take into account the morphological characteristics of the nodule (e.g., nodules with spiculated borders are suspicious for malignancy, while well-defined round nodules are usually benign) (FERREIRA; OLIVEIRA; AZEVEDO-MARQUES, 2018). First-order statistics describe the gray-level frequency distribution within the region of interest, which can be obtained from the histogram of pixel intensities. The remaining features contain texture information, taking into account local intensity-spatial distribution. Those features' performances are not affected by tumor position, orientation, size, and brightness (WEI et al., 2019). After the extraction, we obtained a set of 120 features for each lesion in each imaging modality, divided into 19 first order statistics, 14 shape-based, 24 GLCM, 16 GLRLM, 16 GLSZM, 5 NGTDM, and 14 GLDM features.

4.5 Multimodality Strategies

As for our fusion strategies, we evaluated two distinct approaches, which differ in relation to the moment into the pipeline in which the information is combined.

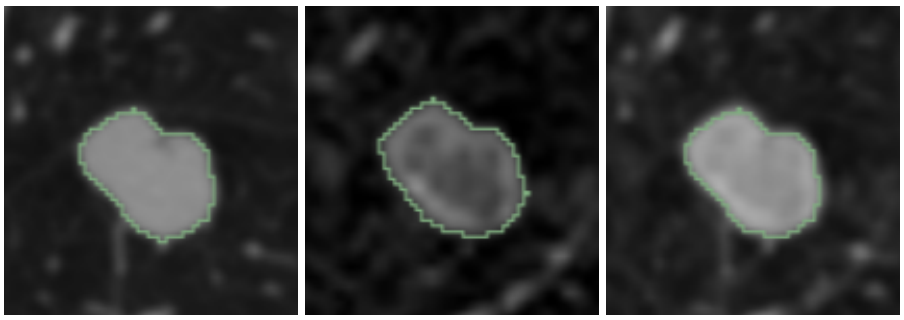
CT/MRI Concatenation Approach

The first approach consists of concatenating the radiomics features obtained in each modality before feeding them to the model. This approach may be classified as classifier-level fusion in the classification proposed by Guo et al. (GUO et al., 2019). This feature set contains 240 radiomics features.

Image Fusion Approach

This approach consists of combining two modalities into a new image. For such, we calculated the average pixel-wise intensity value through the scans (TIRUPAL; MOHAN; KUMAR, 2020). This approach may be classified as feature-level fusion in the classification

proposed by Guo et al. (GUO et al., 2019). To define the segmentation to be used for feature extraction in the fused image, we calculated the intersection of the segmentations in each modality, because this area is guaranteed to contain information of the nodule in both modalities, as Figure 10 shows. In the end, this feature set contains 120 radiomic features.



(a) CT (b) MRI (c) Fused image

Figure 10 – Lung and its segmentation in each modality and fused image.

4.6 Classification

To evaluate the predictive performance of each set of radiomics features, we performed an experiment consisting of a machine learning pipeline composed of hyperparameter optimization and classification for each dataset. Each experiment was performed 30 times to obtain the mean and deviation in performance for each metric. Models' parameters, metrics, and outputs were tracked with the aid of MLFlow platform (ZAHARIA et al., 2018).

Classifier

To perform the nodules classification, we used the XGBoost classifier (CHEN; GUESTRIN, 2016). XGBoost is a tree boosting system widely used to achieve state-of-the-art results for many machine learning challenges. Although deep learning models, such as convolutional neural networks and deep belief networks, are state of the art in lung nodule classification, the size of our dataset is insufficient to take advantage of those models (PAREKH; JACOBS, 2019). In this case, models such as XGBoost are more suitable for addressing the problem of classification in the context of smaller tabular datasets, with the advantage of allowing the radiologist to interpret the role of each feature in the classifier's output (GILLIES; KINAHAN; HRICAK, 2016).

Hyperparameter Optimization

For this work, we performed hyperparameter optimization using a randomized search approach using the area under the receiver AUC as the optimization metric and performing 50 iterations in each trial (CLAESEN; MOOR, 2015). The search space explored is shown in Table 3.

Table 3 – Hyperparameter Optimization Search Spacee

Parameter	Search Space
Number of estimators	[200, 1000]
Learning rate	[0.2, 0.6]
Max depth	{3, 4, 5, 6, 7, 8, 9}
Subsample ratio of the training instances	{0.6, 0.8, 1.0}
Subsample ratio of columns	{0.6, 0.8, 1.0}

Metrics and Evaluation

As our pipeline includes hyperparameter optimization, it is necessary to perform a robust validation to ensure no data leakage occurs between our train and test folds, leading to unreliable results due to overfitting. Thus, we performed validation using a stratified 5-fold nested cross-validation (CAWLEY; TALBOT, 2010), performing the optimizations mentioned above only on the train folds. To assess model performance, we measured a set of metrics well-established in CADx systems design: AUC, sensitivity (Equation 3), and specificity (Equation 4), as well as accuracy (Equation 3) and precision (Equation 3), where TP is the number of correctly classified positive instances; TN is the number of correctly classified negative instances; FP is the number of incorrectly classified positive instances; and FN is the number of incorrectly classified negative instances (FAWCETT, 2006). AUC was chosen as optimization and is our main metric for comparing models.

$$\textit{sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\textit{specificity} = \frac{TP}{TN + FP} \tag{4}$$

$$accuracy = \frac{TP + TN}{TP + FN + FP + FN} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

To assess the statistical significance of the classifiers' performance difference across different sets of data, we performed a Wilcoxon signed-rank test (W-test) at a significance level of 5% (DEMŠAR, 2006).

Finally, to allow for interpretation of features' contribution in the results, we recorded the models' feature importance and SHAP values for each trial, as well as its average results across the experiments in each modality.

5 Results and Discussion

5.1 Classification performance

Table 4 summarizes the classification results, with the average performance and standard deviation for each metric across the experiment trials. The best overall AUC performance was obtained by the Image Fusion approach, with an average value of 0.794, against 0.755 obtained on the Computed Tomography, and 0.770 for both MRI and CT/MRI Concatenation feature sets. The difference in AUC performance was significant against CT (p-value: $4.63\text{e}{-4}$), MRI ($1.79\text{e}{-2}$), and CT/MRI Concatenation ($6.40\text{e}{-3}$).

Models based on the Image Fusion approach also have shown superior performance in terms of sensitivity, with an average value of 0.741, versus 0.658 in CT, 0.722 in MRI, and 0.706 in CT/MRI Concatenation. In this metric, the difference was verified significant against CT ($5.46e-04$) and CT/MRI Concatenation ($9.16e-03$). However, we could not reject the null hypothesis for MRI ($7.46e-02$). Similarly, the advantage in accuracy, with a value of 0.694, versus 0.668 in CT and CT/MRI Concatenation, and 0.684 in MRI, was significant against CT ($1.50e-02$), CT/MRI Concatenation ($1.26e-02$), but not for MRI ($2.80e-01$).

Because AUC was our metric of choice, the Image Fusion approach has been shown as the best alternative. In CADx research, a good sensitivity is often desirable, as it reflects the positive prediction power, and this approach has also shown performant in regards to this metric. Nevertheless, the performance of the MRI-based models is remarkable, as we could not find significant differences against the Image Fusion-based models in some classification metrics.

Table 4 – Average and standard deviation in performance for each feature set.

Feature set	AUC	Sensitivity	Specificity	Accuracy	Precision
Computed Tomography	0.755±0.036	0.658±0.080	**0.681±0.065**	0.668±0.045	**0.709±0.054**
Magnetic Resonance Imaging	0.770±0.052	0.722±0.056	0.646±0.084	0.684±0.049	0.698±0.062
CT/MRI Concatenation	0.770±0.048	0.706±0.070	0.631±0.056	0.668±0.039	0.686±0.044
Image Fusion	**0.794±0.036**	**0.741±0.073**	0.648±0.061	**0.694±0.044**	0.707±0.051

Our second multimodality approach, the CT/MRI Concatenation, has not presented improvement over the single modalities. In fact, the average AUC value of 0.770 was the same as seen in MRI models, while every other metric value is slightly smaller than its MRI counterparts. This is most likely a manifestation of the *curse of dimensionality*: as the number of features increases, different issues with the data may emerge, requiring a more robust feature selection strategy (ALTMAN; KRZYWINSKI, 2018). Furthermore, this is in accordance with Wei *et al.*, who argue that the direct combination of features extracted separately might not make full use of the underlying biological correlation (WEI et al., 2019).

Finally, the single modalities feature sets contain an interesting finding, as MRI-based models have shown a superior AUC of 0.770, when compared to CT, with 0.755, and this difference was shown significant by our hypothesis test (p-value: $3.60e-02$). A similar fact occurred with sensitivity, with 0.722 in MRI versus 0.658 in CT ($6.06e-03$). However, the difference in accuracy, with 0.684 for MRI and 0.668 for CT is not clear ($6.53e-02$); and CT even has the advantage in specificity: with 0.681 versus 0.646 in CT and a significant difference ($2.50e-02$), although models can be fine-tuned to favor sensitivity or specificity. These results are significant because CT is the standard modality for lung nodule classification and throws the spotlight on MRI as a viable imaging modality for chest disease management.

5.2 Radiomic Features

Computed Tomography

Figure 11 shows the top 15 features in terms of average importance for the CT-based models. The most important feature was *Variance*, a first-order feature that measures the spread of gray level values about the mean. Other first-order features in relatively high positions on the ranking are *10Percentile*, *Energy*, and *Mean*.

Next, we have GLCM features, such as *ClusterProminence* and *Maximal Correlation Coefficient* (MCC) figuring high in the list, as well as *Autocorrelation*, *ClusterShade*, and *MaximumProbability* with a lesser importance. GLCM features characterize the texture of an image, and those results are an indicative that the intratumoral environment in CT images contains important information to define a nodules' malignancy.

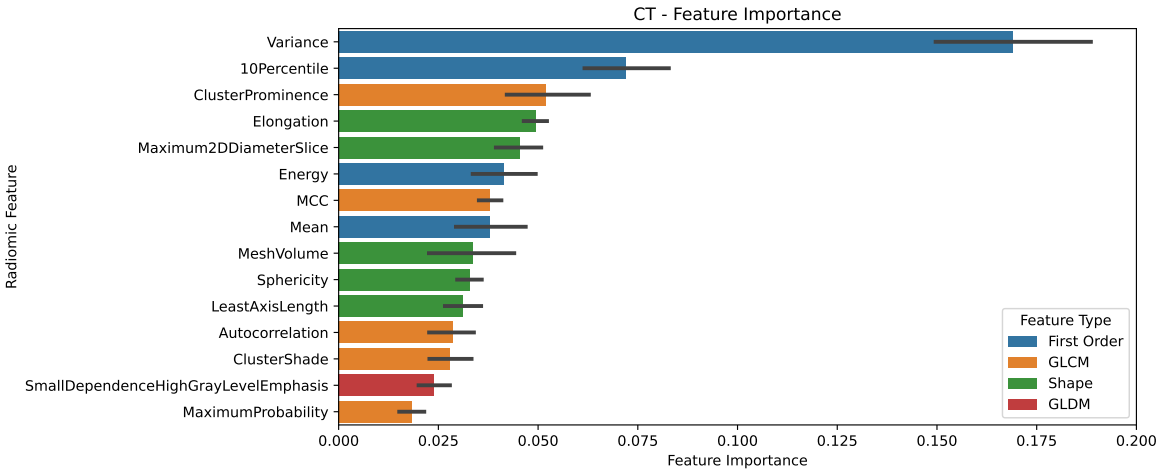


Figure 11 – CT Radiomic Feature Importances

Shape-based features also appear frequently, reflecting nodules' morphology, namely *Elongation*, *Maximum2DDiameterSlice*, *Sphericity*, and *LeastAxisLength*; as well as its dimension with *MeshVolume*.

Figure 12 presents the average SHAP values for the 10 which had the most impact on models' output. Each point is a sample, and its color represents the feature value in a scale from blue (lower values) to red (higher values). The x-axis is the SHAP value; higher positive values contributed more to a positive outcome (malignant nodule), and higher negatives to a negative outcome (benign).

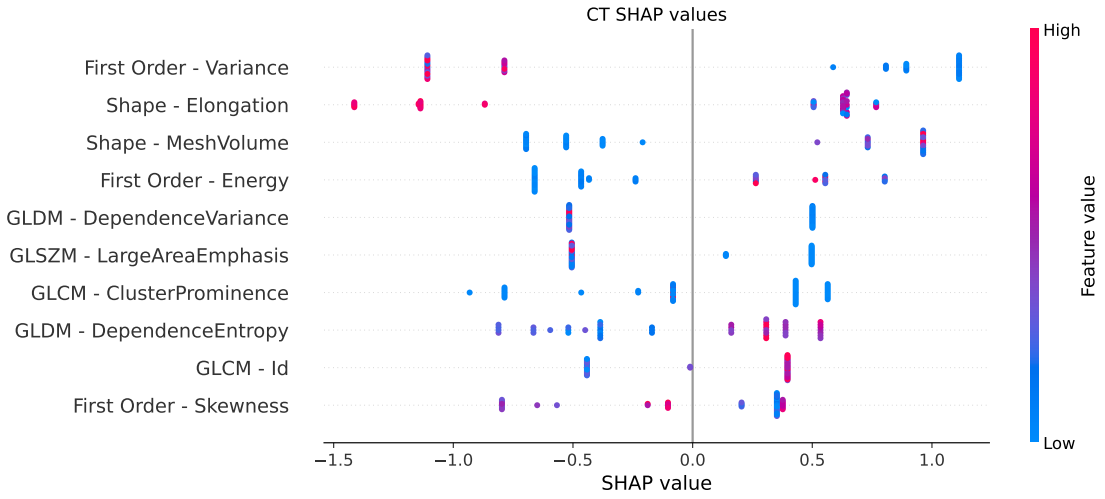


Figure 12 – CT Radiomic Features SHAP values

The features in this list roughly reflect the most important features in the previous analysis. First-order features such as *Variance* have a clear pattern, with lower variance val-

ues contributing to positive outputs, while the opposite happens with *Energy*. Looking into shape-based features, higher values of *Elongation*, which reflects less sphere-like nodules, contributed to positive outputs; also, lower *MeshVolume* (smaller nodules) contributed to negative outputs. Those characteristics are well-known indicatives of a lung nodules' malignancy and are contained in the Fleischner Society guidelines (MACMAHON et al., 2017).

Texture features such as GLCM, GLDM, and GLSZM also have a considerable impact on models' output. *LargeAreaEmphasis* (less coarse texture) contributed to positive outcomes, and higher *Inverse Difference* (*ID*) (more uniform gray levels), also contributed to positive outcomes. Those features may account for ground-glass or part-solid nodules, which are more often malignant and have hazy textures in CT (MACMAHON et al., 2017).

Magnetic Resonance Imaging

Figure 13 presents the 15 most important features for MRI-based models.

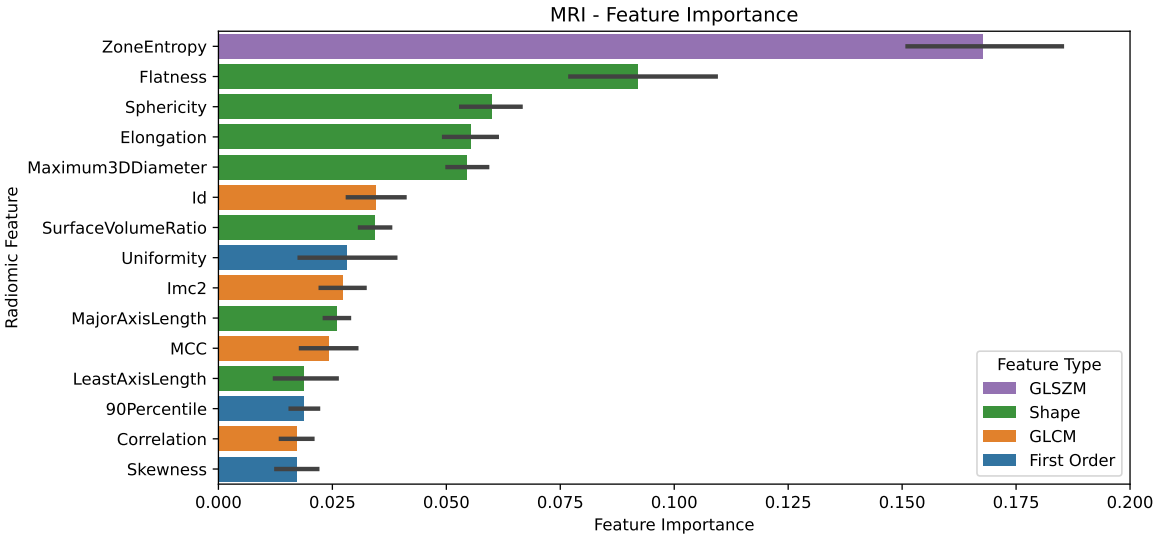


Figure 13 – MRI Radiomic Feature Importances

A noticeable difference to CT is the presence of a GLSZM feature, namely *ZoneEntropy* figuring first in the ranking. This feature measures the uncertainty in the distribution of gray level zones sizes within the ROI, with higher values indicating more heterogeneity in the texture patterns. Next, shape-based features figure in the ranking with considerable importance. The most important seem to be related to the roundness of the nodule, for instance, *Flatness*, *Sphericity*, *Elongation* and *SurfaceVolumeRatio* measures, among other aspects, if a nodule

is more or less sphere-like. The remaining shape-based features account for nodules' size characteristics, such as *Maximum3DDiameter*, *MajorAxisLength*, and *LeastAxisLength*.

GLCM features also figure in the ranking, albeit in smaller degree when compared to CT, accounting for the homogeneity (*ID*) and complexity (*Imc2*, *MCC*, *Correlation*) of the nodule. Finally, another difference compared to CT is the relative absence of first-order features in the ranking. An inherent limitation of MRI is that its images have arbitrary intensity units, i.e., the images' gray levels have no physiological meaning, as opposed to the CT HU scale, making image quantification with histogram-based features ineffective. Those features were likely left out by the models' for this reason.

Figure 5b summarises the SHAP values for the MRI samples.

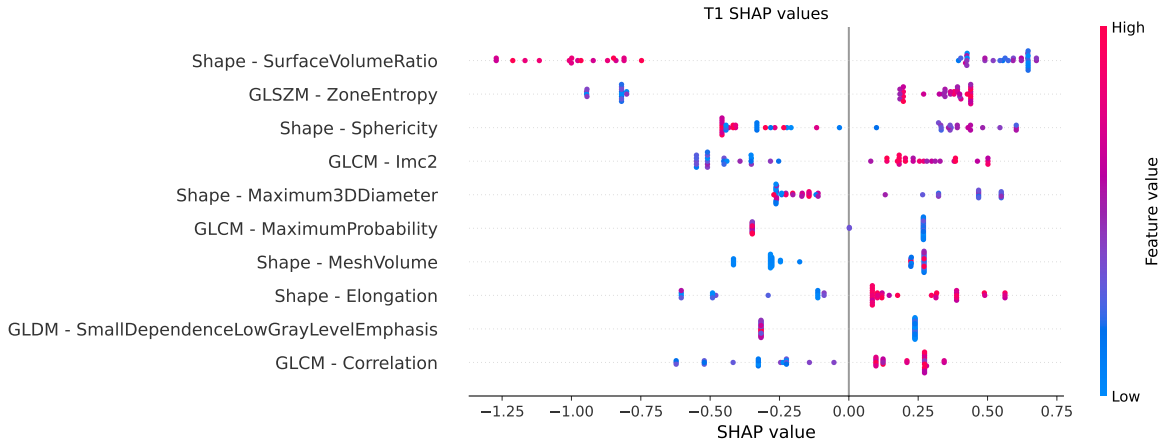


Figure 14 – MRI Radiomic Feature SHAP Values

Here, most features are also figured in the importance list. Like in CT, shape-based features that reflects on the nodules' dimension, namely *SurfaceVolumeRatio* and *MeshVolume* have significant contribution to the model output, as well as features that measure nodules' roundness, such as *Sphericity* and *Elongation*. *ZoneEntropy*, which was the most important in the previous analysis also figure here, with more complex textures contributing to positive outputs, and vice versa. Other texture features, such as *Imc2*, *Correlation*, and *MaximumProbability*, have significant contributions to model output.

CT/MRI Feature Concatenation

Figure 15 shows the 15 most important features for the CT/MRI Concatenation-based models, and the modality that each feature belongs.

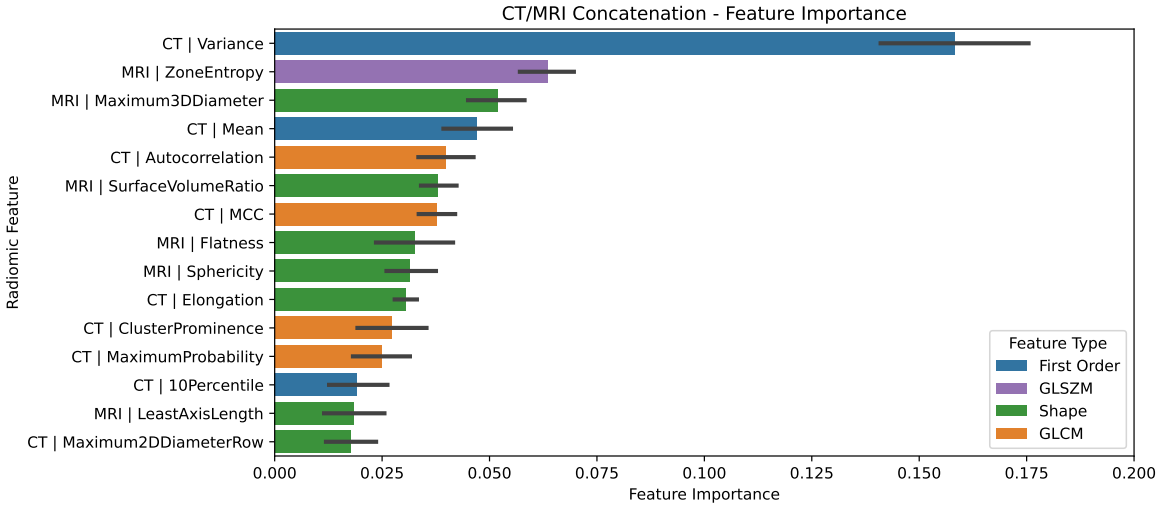


Figure 15 – CT/MRI Concatenation Radiomic Feature Importances

The list is a mixture of CT and MRI's most important features. Roughly speaking, CT contributed with first-order and GLCM features, while MRI contributed with a series of shape-based features. However, as seen in Section 5.1, this combination was not enough to leverage gains in performance.

Figure 16 summarises the average SHAP values for the CT/MRI Concatenation samples, also showing a mixture of CT and MRI's individually most important features, and its interpretation is mostly a throwback to the individual modalities.

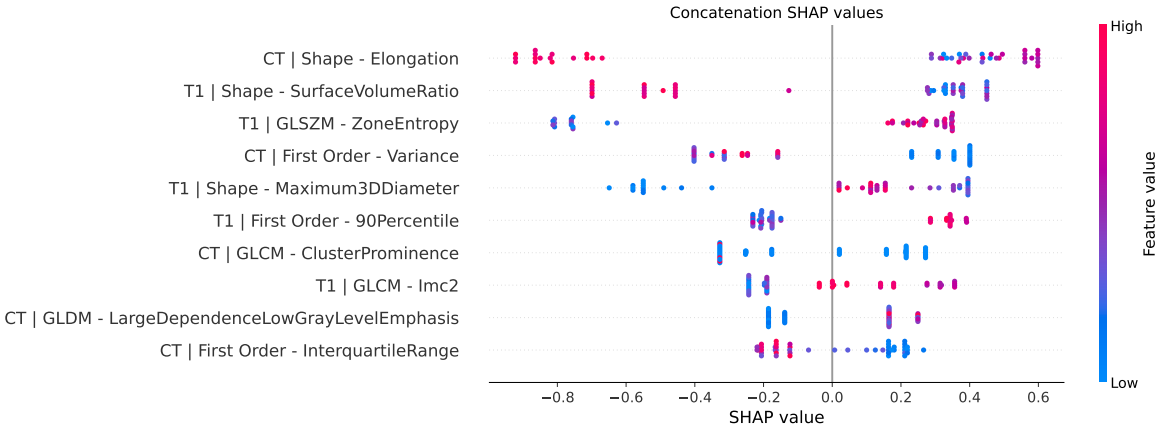


Figure 16 – CT/MRI Concatenation Radiomic Feature SHAP Values

Image Fusion

Figure 17 shows the 15 most important features in terms of average importance for the Image Fusion-based models.

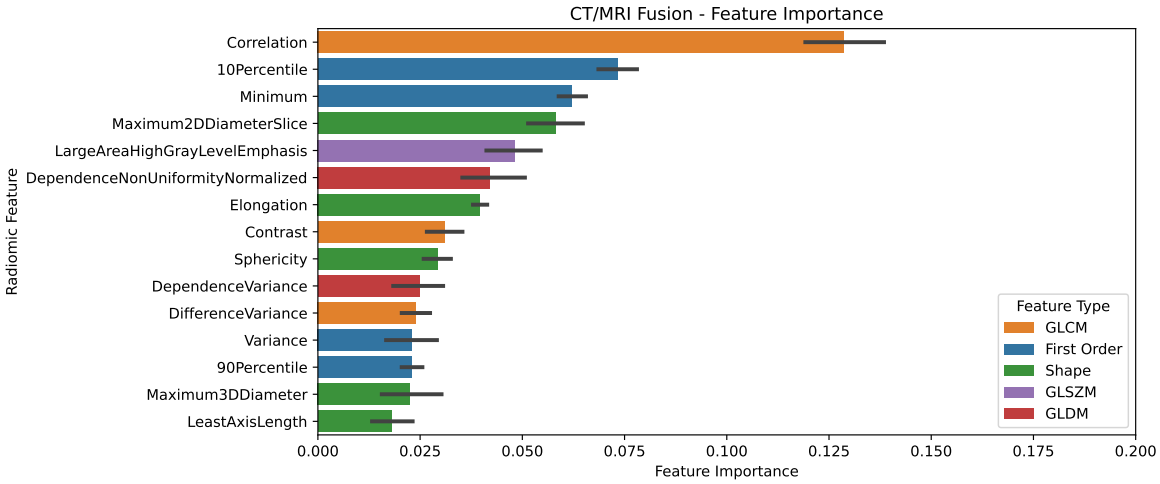


Figure 17 – CT/MRI Fusion Radiomic Feature Importances

The features that figure in the importance ranking are different from the ones seen previously. *Correlation* has not figured in CT and MRI importances, and measures whether the intensity values are correlated with the voxels, reflecting intensity patterns in the perinodular and intranodular regions. Other texture features are *Contrast* and *DifferenceVariance*, which are different measures of texture complexity. Lastly, two GLDM features also figure in the list: *DependenceNonUniformityNormalized* and *DependenceVariance*, accounting for texture homogeneity. This suggests that the fused image possesses texture characteristics that are distinct from CT and MRI. First-order features also figure in high positions of the list, namely *10Percentile* and *Minimum*, which did not appear in the individual CT and MRI modalities and likely emerged from the Image Fusion strategy.

Finally, some well-known shape-based features are also there, accounting for nodules' size (*Maximum2DDiameterSlice*, *Maximum3DDiameter*, *LeastAxisLength*) and roundness (*Elongation* and *Sphericity*), therefore, it is reasonable to think that the strategy for combining the distinct segmentations is still able to reflect those characteristics.

Figure 18 shows the average SHAP values for the CT/MRI Concatenation samples.

The SHAP values roughly mirror feature importances. *Correlation* figures first in the list, with an apparent large separation between lower and higher values leading to negative

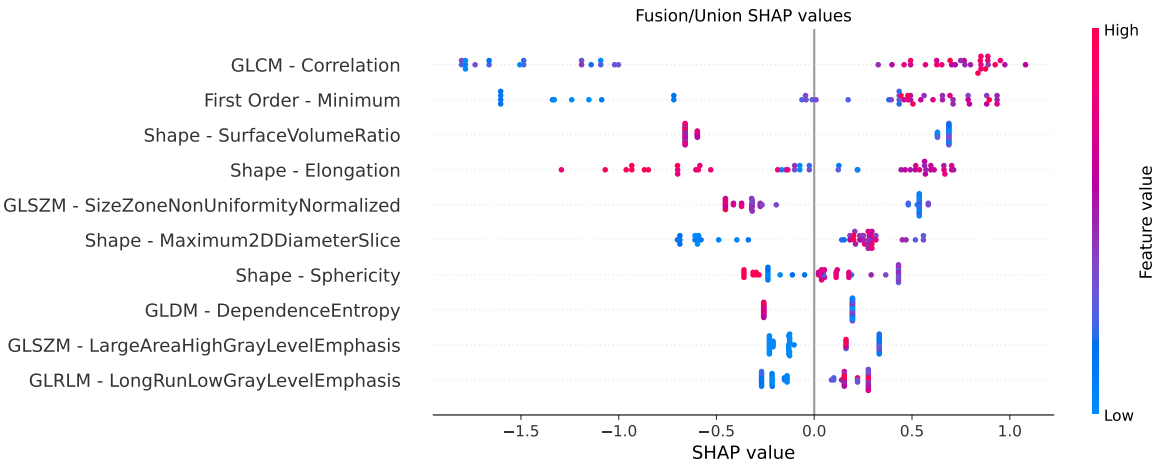


Figure 18 – CT/MRI Fusion Radiomic Feature SHAP Values

and positive outputs, including a higher weight to the negative ones, as the absolute values of negative SHAP values are larger than its positive counterparts. This also happens on some level to the *Minimum* feature.

The behavior of shape-based features recall what was seen previously, with values indicating rounder nodules' (*SurfaceVolumeRatio*, *Sphericity*) leading to negative outputs, although *Elongation* does not presents a clear pattern; and *Maximum2DDiameterSlice* linking smaller nodules to negative outputs. Similarly, texture features related to nodules' texture homogeneity (*SizeZoneNonUniformityNormalized* and *LargeAreaHighGrayLevelEmphasis*) and coarseness (*DependenceEntropy* and *LongRunLowGrayLevelEmphasis*) also have significant contribution to positive and negative outputs.

Analyzing the importance of each individual feature in the construction of the models and their contribution in classifying a sample as benign or malignant, some particularities were noticed: CT-based models had a significant number of first-order features, which may be a result of the physiological information measured in Hounsfield Units, while this category of features barely appeared in the MRI lists. Furthermore, both modalities relied on texture information, mainly related to texture complexity and/or homogeneity, in specific GLCM features, suggesting that the intratumoral environment is a strong indicator of a nodules' malignancy. Finally, shape-based features are also figured in both modalities, in accordance with lung cancer management guidelines which use a nodules' size and roundness as a parameter to deciding the next steps in patients' treatment.

In regards to the multimodality approaches, we noticed that in the feature concatenation

approach, CT and MRI features appeared mixed in the rankings, but as mentioned previously, this combination was not enough to yield any gains in AUC performance, although it also did not affect this metric negatively. On the image fusion approach, it was apparent that from the image fusion, new metrics emerged as relevant to the models' output, in first-order and texture features, in addition to the always present shape-based features that measure a nodules' size and roundness.

6 Conclusion

This study aimed to evaluate whether using radiomics features from combined CT and MRI scans can increase classification performance when compared to the individual modalities. Overall, we found that a simple pixel-wise average image fusion is enough to provide significant gains in most classification metrics compared to the best performant single modalities models. On the other hand, merely concatenating the features into a larger dataset has shown to be a poor strategy for dealing with multimodality data, as we did not see an increase in performance in comparison to the best single modality models. This observation may be indicative of the importance of the strategy according to which features are combined.

As a secondary result, we noticed that MRI-based models exhibited an advantage over CT-based ones. This improvement is exciting because using MRI for lung cancer management can mitigate issues such as radiation exposure and adverse reactions to contrast materials commonly used in CT.

Finally, our research does not come without limitations. First, we recognize that our data size is limited in size, thus, certain aspects of this single study may not be generalized, although our observations are in line with similar research on the field. Second, we barely scratched the surface of multimodality fusion, as the number of strategies to which those modalities can be combined is virtually infinite. Third, a more in-depth analysis of features' importance on model development could be conducted, in particular with the participation of radiologists.

As future works, we intend to take this work further through the following roadmap: 1) increase the study population by retrieving more recent exams that are compatible with our dataset; 2) adding other MRI sequences, such as T2; 3) evaluating more image fusion strategies, such as wavelet and deep learning-based ones; 4) perform a more in-depth analysis of the nodules' relevant characteristics to the models' output.

Bibliography

ALTMAN, N.; KRZYWINSKI, M. The curse (s) of dimensionality. *Nat Methods*, v. 15, n. 6, p. 399–400, 2018.

BALTRUŠAITIS, T.; AHUJA, C.; MORENCY, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 41, n. 2, p. 423–443, 2018.

BASHIRI, F. S. et al. Multi-modal medical image registration with full or partial data: A manifold learning approach. *Journal of Imaging*, Multidisciplinary Digital Publishing Institute, v. 5, n. 1, p. 5, 2019.

BECKETT, K. R.; MORIARITY, A. K.; LANGER, J. M. Safe use of contrast media: what the radiologist needs to know. *Radiographics*, Radiological Society of North America, v. 35, n. 6, p. 1738–1750, 2015.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, v. 13, n. 2, 2012.

BUSHBERG, J. T.; BOONE, J. M. *The essential physics of medical imaging*. [S.l.]: Lippincott Williams & Wilkins, 2011.

CAWLEY, G. C.; TALBOT, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, JMLR. org, v. 11, p. 2079–2107, 2010.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

CLAESEN, M.; MOOR, B. D. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006. ISSN 15337928.

DESSEROIT, M.-C. et al. Development of a nomogram combining clinical staging with 18 f-fdg pet/ct image features in non-small-cell lung cancer stage i–iii. *European journal of nuclear medicine and molecular imaging*, Springer, v. 43, n. 8, p. 1477–1485, 2016.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FERREIRA, J. R.; OLIVEIRA, M. C.; AZEVEDO-MARQUES, P. M. de. Characterization of pulmonary nodules based on features of margin sharpness and texture. *Journal of digital imaging*, Springer, v. 31, n. 4, p. 451–463, 2018.

FEURER, M.; HUTTER, F. Hyperparameter optimization. In: *Automated machine learning*. [S.l.]: Springer, Cham, 2019. p. 3–33.

FRANCISCO, V. et al. Computer-aided diagnosis of lung cancer in magnetic resonance imaging exams. In: SPRINGER. *XXVI Brazilian Congress on Biomedical Engineering*. [S.l.], 2019. p. 121–127.

GILLIES, R. J.; KINAHAN, P. E.; HRICAK, H. Radiomics: images are more than pictures, they are data. *Radiology*, Radiological Society of North America, v. 278, n. 2, p. 563–577, 2016.

GRIETHUYSEN, J. J. V. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, AACR, v. 77, n. 21, p. e104–e107, 2017.

GUO, Z. et al. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, IEEE, v. 3, n. 2, p. 162–169, 2019. ISSN 2469-7311.

HANSELL, D. M. et al. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, Radiological Society of North America, v. 246, n. 3, p. 697–722, 2008.

HATT, M. et al. Radiomics in pet/ct: more than meets the eye? *Journal of Nuclear Medicine*, Soc Nuclear Med, v. 58, n. 3, p. 365–366, 2017.

HE, X. et al. Practical lessons from predicting clicks on ads at facebook. In: *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. [S.l.: s.n.], 2014. p. 1–9.

KNIGHT, S. B. et al. Progress and prospects of early detection in lung cancer. *Open biology*, The Royal Society, v. 7, n. 9, p. 170070, 2017.

KNIGHT, S. B. et al. Progress and prospects of early detection in lung cancer. *Open Biology*, v. 7, n. 9, 2017. ISSN 20462441.

KOENIGKAM-SANTOS, M. et al. Contrast-enhanced magnetic resonance imaging of pulmonary lesions: description of a technique aiming clinical practice. *European journal of radiology*, Elsevier, v. 84, n. 1, p. 185–192, 2015.

LI, L.; TALWALKAR, A. Random search and reproducibility for neural architecture search. In: PMLR. *Uncertainty in artificial intelligence*. [S.l.], 2020. p. 367–377.

LI, L. et al. Deep learning for variational multimodality tumor segmentation in pet/ct. *Neurocomputing*, Elsevier, 2019.

LI, S. et al. Predicting Lung Nodule Malignancies by Combining Deep Convolutional Neural Network and Handcrafted Features. 2018. Disponível em: <http://arxiv.org/abs/1809.02333>.

LI, S. et al. Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. *Physics in Medicine & Biology*, IOP Publishing, v. 64, n. 17, p. 175012, 2019.

LOVERDOS, K. et al. Lung nodules: A comprehensive review on current approach and management. *Annals of thoracic medicine*, Wolters Kluwer–Medknow Publications, v. 14, n. 4, p. 226, 2019.

LOWEKAMP, B. C. et al. The design of simpleitk. *Frontiers in neuroinformatics*, Frontiers, v. 7, p. 45, 2013.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems*. [S.l.: s.n.], 2017. p. 4768–4777.

MACMAHON, H. et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology*, Radiological Society of North America, v. 284, n. 1, p. 228–243, 2017.

MAINTZ, J. A.; VIERGEVER, M. A. A survey of medical image registration. *Medical image analysis*, Elsevier, v. 2, n. 1, p. 1–36, 1998.

MU, W. et al. Non-invasive decision support for nsclc treatment using pet/ct radiomics. *Nature communications*, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2020.

MU, W. et al. Radiomic biomarkers from pet/ct multi-modality fusion images for the prediction of immunotherapy response in advanced non-small cell lung cancer patients. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Medical Imaging 2018: Computer-Aided Diagnosis*. [S.l.], 2018. v. 10575, p. 105753S.

NIELSEN, D. *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* Dissertação (Mestrado) — NTNU, 2016.

OHNO, Y. New applications of magnetic resonance imaging for thoracic oncology. In: THIEME MEDICAL PUBLISHERS. *Seminars in respiratory and critical care medicine*. [S.l.], 2014. v. 35, n. 01, p. 027–040.

OHNO, Y. et al. Mri for solitary pulmonary nodule and mass assessment: current state of the art. *Journal of Magnetic Resonance Imaging*, Wiley Online Library, v. 47, n. 6, p. 1437–1458, 2018.

PAREKH, V. S.; JACOBS, M. A. Deep learning and radiomics in precision medicine. *Expert review of precision medicine and drug development*, Taylor & Francis, v. 4, n. 2, p. 59–72, 2019.

PASTORINO, U. et al. Annual or biennial ct screening versus observation in heavy smokers: 5-year results of the mild trial. *European Journal of Cancer Prevention*, LWW, v. 21, n. 3, p. 308–315, 2012.

RAMPINELLI, C. et al. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. *bmj*, British Medical Journal Publishing Group, v. 356, p. j347, 2017.

RAMRAJ, S. et al. Experimenting xgboost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, v. 9, p. 651–662, 2016.

RASHI, M. *MRI Components & Functions*. 2020. Disponível em: <https://snc2dmri.weebly.com/components--functions.html>.

SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, v. 68, n. 1, p. 7–30, 2018. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21442>.

SOMMER, G. et al. Lung nodule detection in a high-risk population: comparison of magnetic resonance imaging and low-dose computed tomography. *European journal of radiology*, Elsevier, v. 83, n. 3, p. 600–605, 2014.

TIRUPAL, T.; MOHAN, B. C.; KUMAR, S. S. Multimodal medical image fusion techniques–a review. *Current Signal Transduction Therapy*, v. 15, n. 1, p. 1–22, 2020.

VAIDYA, M. et al. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiotherapy and Oncology*, Elsevier Ireland Ltd, v. 102, n. 2, p. 239–245, 2012. ISSN 01678140. Disponível em: <http://dx.doi.org/10.1016/j.radonc.2011.10.014>.

VALLIÈRES, M. et al. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine and Biology*, IOP Publishing, v. 60, n. 14, p. 5471–5496, 2015. ISSN 13616560.

WEI, L. et al. Machine learning for radiomics-based multimodality and multiparametric modeling. *The quarterly journal of nuclear medicine and molecular imaging : official publication of the Italian Association of Nuclear Medicine (AIMN) [and] the International Association of Radiopharmacology (IAR), [and] Section of the Society of..*, v. 63, n. 4, p. 323–338, 2019. ISSN 18271936.

YANG, Y. et al. Deep learning aided decision support for pulmonary nodules diagnosing: A review. *Journal of Thoracic Disease*, v. 10, n. Suppl 7, p. S867–S875, 2018. ISSN 20776624.

YI, C. A. et al. Non–small cell lung cancer staging: efficacy comparison of integrated pet/ct versus 3.0-t whole-body mr imaging. *Radiology*, Radiological Society of North America, v. 248, n. 2, p. 632–642, 2008.

YOO, T. S. *Insight into images: principles and practice for segmentation, registration, and image analysis*. [S.l.]: AK Peters/CRC Press, 2004.

ZAHARIA, M. et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, v. 41, n. 4, p. 39–45, 2018.

ZHU, L. et al. An effective interactive medical image segmentation method using fast growcut. In: *MICCAI workshop on interactive medical image computing*. [S.l.: s.n.], 2014.

APPENDIX A – Radiomic Features

First Order Features

1. Energy
2. Total Energy
3. Entropy
4. Minimum
5. 10th percentile
6. 90th percentile
7. Maximum
8. Mean
9. Median
10. Interquartile Range
11. Range
12. Mean Absolute Deviation (MAD)
13. Robust Mean Absolute Deviation (rMAD)
14. Root Mean Squared (RMS)
15. Standard Deviation
16. Skewness

Shape-based (3D)

1. Mesh Volume
2. Voxel Volume
3. Surface Area
4. Surface Area to Volume Ratio
5. Sphericity
6. Compactness 1
7. Compactness 2
8. Spherical Disproportion
9. Maximum 3D diameter
10. Maximum 2D diameter (Slice)
11. Maximum 2D diameter (Column)
12. Maximum 2D diameter (Row)
13. Major Axis Length
14. Minor Axis Length
15. Least Axis Length
16. Elongation
17. Flatness

Shape-based (2D)

1. Mesh Surface
2. Pixel Surface

3. Perimeter
4. Perimeter to Surface ratio
5. Sphericity
6. Spherical Disproportion
7. Maximum 2D diameter
8. Major Axis Length
9. Minor Axis Length
10. Elongation

Gray Level Cooccurence Matrix (GLCM) Features

1. Autocorrelation
2. Joint Average
3. Cluster Prominence
4. Cluster Shade
5. Cluster Tendency
6. Contrast
7. Correlation
8. Difference Average
9. Difference Entropy
10. Difference Variance
11. Joint Energy
12. Joint Entropy

13. Informational Measure of Correlation (IMC) 1
14. Informational Measure of Correlation (IMC) 2
15. Inverse Difference Moment (IDM)
16. Maximal Correlation Coefficient (MCC)
17. Inverse Difference Moment Normalized (IDMN)
18. Inverse Difference (ID)
19. Inverse Difference Normalized (IDN)
20. Inverse Variance
21. Maximum Probability
22. Sum Average
23. Sum Entropy
24. Sum of Squares

Gray Level Size Zone Matrix (GLSZM) Features

1. Short Run Emphasis (SRE)
2. Long Run Emphasis (LRE)
3. Gray Level Non-Uniformity (GLN)
4. Gray Level Non-Uniformity Normalized (GLNN)
5. Run Length Non-Uniformity (RLN)
6. Run Length Non-Uniformity Normalized (RLNN)
7. Run Percentage (RP)
8. Gray Level Variance (GLV)

9. Run Variance (RV)
10. Run Entropy (RE)
11. Low Gray Level Run Emphasis (LGLRE)
12. High Gray Level Run Emphasis (HGLRE)
13. Short Run Low Gray Level Emphasis (SRLGLE)
14. Short Run High Gray Level Emphasis (SRHGLE)
15. Long Run Low Gray Level Emphasis (LRLGLE)
16. Long Run High Gray Level Emphasis (LRHGLE)

Gray Level Run Length Matrix (GLRLM) Features

1. Short Run Emphasis (SRE)
2. Long Run Emphasis (LRE)
3. Gray Level Non-Uniformity (GLN)
4. Gray Level Non-Uniformity Normalized (GLNN)
5. Run Length Non-Uniformity (RLN)
6. Run Length Non-Uniformity Normalized (RLNN)
7. Run Percentage (RP)
8. Gray Level Variance (GLV)
9. Run Variance (RV)
10. Run Entropy (RE)
11. Low Gray Level Run Emphasis (LGLRE)
12. High Gray Level Run Emphasis (HGLRE)

13. Short Run Low Gray Level Emphasis (SRLGLE)
14. Short Run High Gray Level Emphasis (SRHGLE)
15. Long Run Low Gray Level Emphasis (LRLGLE)
16. Long Run High Gray Level Emphasis (LRHGLE)

Neighbouring Gray Tone Difference Matrix (NGTDM) Features

1. Coarseness
2. Contrast
3. Busyness
4. Complexity
5. Strength

Gray Level Dependence Matrix (GLDM) Features

1. Small Dependence Emphasis (SDE)
2. Large Dependence Emphasis (LDE)
3. Gray Level Non-Uniformity (GLN)
4. Dependence Non-Uniformity (DN)
5. Dependence Non-Uniformity Normalized (DNN)
6. Gray Level Variance (GLV)
7. Dependence Variance (DV)
8. Dependence Entropy (DE)

9. Low Gray Level Emphasis (LGLE)
10. High Gray Level Emphasis (HGLE)
11. Small Dependence Low Gray Level Emphasis (SDLGLE)
12. Small Dependence High Gray Level Emphasis (SDHGLE)
13. Large Dependence Low Gray Level Emphasis (LDLGLE)
14. Large Dependence High Gray Level Emphasis (LDHGLE)