FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Neuroblastoma Cancer Radiogenomics

Mafalda Malafaia Oliveira

Integrated Masters in Bioengineering

Supervisor: Hélder Oliveira

Co-Supervisor: Tania Pereira

Co-Supervisor: Helena Torrão

September 18, 2022

Neuroblastoma Cancer Radiogenomics

Mafalda Malafaia Oliveira

Integrated Masters in Bioengineering

September 18, 2022

Resumo

O Neuroblastoma (NB) é o tipo de tumor extracraniano mais comum em casos pediátricos. Devido à sua baixa taxa de incidência e ao seu comportamento heterogéneo, o seu prognóstico e escolha do melhor tratamento contsitui um desafio para a patologia. Adicionalmente, a taxa de sobrevivência de um paciente diagnosticado com NB é altamente variável, motivando para a necessidade de haver a estratificação de pacientes de acordo com o seu risco. A amplificação do oncogene MYCN está correlacionada com NB de alto risco, sendo a deteção deste biomarcador crucial para a seleção do tratamento e previsão do seu prognóstico. O protocolo clínico atual para a deteção deste biomarcador recorre a procedimentos invasivos como a biópsia.

A área radiogenómica estuda a deteção de biomarcadores relevantes a partir de fenótipos em imagens médicas, tendo sido bem sucessida para vários tipos de cancro. Após uma revisão da literatura, abordagens radiogenómicas para a deteção do estado de MNA, embora superficiais, obtiveram resultados satisfatórios na previsão deste biomarcador a partir de abordagens clássicas de Machine Learning (ML). Uma abordagem radiogenómica permite desenvolver um procedimento não invasivo para a previsão de MNA.

O trabalho proposto nesta dissertação aborda o desenvolvimento de abordagens radiogenómicas para detetar o estado do biomarcador MNA. O conjunto de dados utilizado inclui exames de Tomografia Computadorizada (TC) e informação clínica relativa a 46 pacientes, assim como o estado de MNA respetivo. São propostas quatro abordagens para este estudo, com o objetivo de implementar os métodos mais utilizados em publicações anteriores. A abordagem final inclui uma componente de interpretabilidade, onde métodos explicáveis são implementados para avaliar a robustez da metodologia desenvolvida. Os modelos gerados obtiveram desempenhos satisfatórios utilizando métodos semelhantes para comparação: o modelo radiómico apresentou uma AUROC de 0,84±0,06; o modelo semântico obteve uma AUROC de 1,00±0,00; a abordagem multimodal atingiu uma AUROC de 0,98±0,01; por fim, o modelo de Deep Learning (DL) atingiu uma AUROC de 0,94±0,04.

Abstract

Neuroblastoma (NB) is the most common extracranial tumor in pediatric cases. Due to its low incidence rate and behaviour heterogeneity, it constitute a challenging pathology in terms of prognosis and treatment options. Furthermore, the outcome of patients diagnosis with NB is highly variable, rising the need for risk stratification of the patients. The amplification of the MYCN oncogene is knowingly correlated with high-risk NB, being the detection of this biomarker crucial for treatment selection and survival prediction. The current clinical protocol for MNA detection includes invasive procedures, such as biopsy.

Radiogenomics addresses the detection of important biomarkers through imaging phenotypes, being successful in several types of cancer. As literature review is concerned, Radiogenomic approaches for MNA detection in NB patients has been superficially investigated, being able to predict the biomarker status through classical Machine Learning (ML) approaches. A Radiogenomics approach allows a non-invasive procedure for MNA prediction.

The proposed work tackles the challenge of developing a Radiogenomics approach to detect MNA status. The utilized dataset contains CT scans and clinical information of 46 patients, along with the correspondent MNA status. Four different approaches are proposed for this task, with the aim of covering the most common and state-of-the-art methods from previous publications. A final approach includes an interpretability component, where explainable methods are utilized to assess the robustness of the developed pipeline. The trained models were able to achieve satisfactory performances with the same baseline pipeline: the radiomic model presented a AUROC of 0.84±0.06; the semantic model obtained a AUROC of 1.00±0.00; the multi-modal approach reach a AUROC of 0.98±0.01; the Deep Learning (DL) model reached a AUROC of 0.94±0.04.

Aknowledgements

I would like to thank everyone without whom this work could not have been accomplished.

First of all, I am really grateful for having my supervisors, Hélder, Tânia, and Francisco, with me in this journey. Thanks to them, I acquired this passion for research and decided to pursue my studies and take a doctoral degree. And thank you to all the VCMI students, who were always available to help with anything I needed.

Secondly, I would like to thank my parents for all the support and insistence in doing my best at all times. I would especially like to thank my sister for always putting up with me during this period.

Finally, I thank all my friends for encouraging me and helping with everything. And a special acknowledgment goes to João and Pedro, that spent way too much time being bothered by me, always with a smile on their face.

Mafalda Malafaia Oliveira

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Objectives . . . 2
1.3 Contributions . . . 3
1.4 Document Structure . . . 3

2 Literature Review **5**
2.1 Cancer Radiogenomics . . . 5
2.1.1 Neuroblastoma Cancer Radiogenomics . . . 9
2.2 Multi-modality . . . 13
2.3 Interpretability . . . 15
2.4 Summary . . . 17

3 Data Description **19**
3.1 Dataset . . . 19
3.1.1 CT scans . . . 19
3.1.2 Patient Clinical Information . . . 20
3.1.3 MNA status . . . 20
3.2 Data Pre-processing . . . 22
3.2.1 Image Pre-Processing . . . 22
3.2.2 Region of Interest (ROI) Extraction . . . 25
3.2.3 Radiomic Features Extraction . . . 25
3.2.4 Tabular Data Pre-processing . . . 26
3.3 Summary . . . 26

4 MNA Status Prediction with ML Approaches **28**
4.1 Experimental Design . . . 28
4.1.1 Performance Evaluation . . . 29
4.2 Radiomic Approach . . . 29
4.2.1 Materials and Methods . . . 29
4.2.2 Results and Discussion . . . 31
4.2.3 Limitations . . . 35
4.3 Semantic Approach . . . 35
4.3.1 Materials and Methods . . . 35
4.3.2 Results and Discussion . . . 36
4.3.3 Limitations . . . 38
4.4 Multi-modal Approach . . . 39
4.4.1 Materials and Methods . . . 39

4.4.2 Results and Discussion . . . 40
4.4.3 Limitations . . . 42
4.5 Deep Learning Approach . . . 43
4.5.1 Materials and Methods . . . 43
4.5.2 Results and Discussion . . . 46
4.5.3 Limitations . . . 49
4.6 Summary . . . 49

5 Conclusions and Future Work **52**

References **54**

List of Figures

2.1 Architecture of DL model: convolutional layers with kernel size 3×3 and 1×1, batch normalization and pooling layers. From Wang et al. (2019). 8
2.2 Architecture of DL model: DenseNet-based deep learning classifier. From Xu et al. (2022). 9
2.3 Architecture of DL model: two selection steps, initialized with recognition of tumor regions in the input images through a state-of-the-art based CNN structure, and followed by deep filter responses extraction from the last convolutional layer through Fisher vector encoding. The prediction is obtained through an SVM. From Li et al. (2017). 10
2.4 Conceptual diagram of the traditional and comprehensive approaches based on CADs, illustrated with the use-case of lung cancer. From Pereira et al. (2020). . . 13
2.5 Pipeline developed for EGFR mutation status classification based on ensemble methods to combine patient clinical information, radiomic features, and semantic features. As illustrated in the diagram, five methods are studied to join the different modalities of features and make the final prediction. From Malafaia et al. (2021). 14
2.6 Architecture diagram of the developed DL model. Image and clinical data suffer transformations in order to be combined with fusion techniques. From Chen et al. (2022a). 15
2.7 Example of explanations for COVID-related models from several studies. From Hryniewska et al. (2021). 17

3.1 Demonstration of a CT slice (left) and the correspondent annotation of the tumor segmentation from a clinical expert (right). From the private utilized dataset in joint collaboration with HSJ and IPO. 20
3.2 The annotation provided by the clinician, representative of the tumor segmentation, is transformed into a binary mask through a filling process achieved with closing and opening methods. 23
3.3 Illustration of Pixel Resampling Process. The original image, with size 512×512 pixels, is resampled in order to have a pixel spacing of 0.50mm between adjacent pixels. The new size, with lower resolution, is 386×386 pixels. 24
3.4 Illustration of Image Size Standardization Process for images size lower than 224×224 pixels using padding methods. 25
3.5 Illustration of Image Size Standardization Process for images size higher than 224×224 pixels using cropping methods. 26
3.6 Illustration of Intensity Normalization Process. The original intensity range is normalized with a sigmoid function in order to standardize the scale of intensities for the whole dataset. 27

3.7 Illustration of ROI selection and centralization. The pixel resampled, size adjusted and normalized images are transformed in order to obtain exclusively the tumor, being the remaining pixels changed to zero intensity. Then, the ROI is relocated to the center of the image. 27
3.8 Illustration of pre-processing methods on the utilized dataset for the proposed work. 27

4.1 Experimental Design utilized to study the several approaches of the proposed work. An initial train-test split was made, being the training samples utilized to train the both by fine-tuning the hyper-parameters of the models. The best models of each approach were selected and tested with the testing data samples through standard performance metrics. 29
4.2 Suggested pipeline for Radiomic Approach. The training radiomic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics. 30
4.3 Suggested pipeline for Semantic Approach. The training semantic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics. 36
4.4 Suggested pipeline for the Multi-modal Approach. The training semantic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics. 39
4.5 Overview of the developed pipeline for the DL approach. 43
4.6 Attempted input images for the DL Approach. 44
4.7 Demonstration of expected results when utilizing the described XAI methods using three different test samples. The first sample is classified as "goldfinch" and the second one is labeled as "orange", being predicted by the pre-trained VGG16 network. The last sample belongs to the LIDC dataset, being predicted by a lung cancer classifier. For each image, the explanations generated by the 3 XAI methods in question are displayed, as well as an overlay of the explanation on the original image. 45
4.8 Network behaviour over epochs concerning Train and Validation Losses (on the left), and Performance Metrics (on the right). The best model is saved according to the lowest validation loss value, as indicated in the graphs by a red dotted line. 47
4.9 Illustrative test sample with generated explanations over 4 different train-validation splits. For every run the input is presented, as well as the explanation generated by Saliency Maps (SM), Integrated Gradients (IG) and LRP. 50

List of Tables

2.1 Overview of published studies regarding classification tasks with used radiogenomic approaches. 7

2.2 Overview of published studies regarding the prediction of MNA status in CT images of NB patients. 12

3.1 Patient Clinical Information Description. Each type of tabular information provided by the private dataset is enumerated in the following table. Additionally, the type of variable and possible values are also described. 21

4.1 Performance Results for the studied pipelines. Each approach is evaluated in terms of AUROC, Balanced Accuracy, F1 score, Precision and Recall. 32

4.2 Detailed description concerning the hyper-parameter tuning process for the machine learning models. Best values relate to the set of hyper-parameters that achieved the highest performance with the test set, whilst most frequent values refer to the set of hyper-parameters that were chosen by the search method more frequently. 34

4.3 Performance Results for the studied pipelines regarding the Semantic Approach. Each approach is evaluated in terms of AUROC, Balanced Accuracy, F1 score, Precision and Recall. 37

4.4 Detailed description concerning the hyper-parameter tuning process for the machine learning models. Best values relate to the set of hyper-parameters that achieved the highest performance with the test set, whilst most frequent values refer to the set of hyper-parameters that were chosen by the search method more frequently. 38

4.5 Performance Results for the studied classifiers with the Multi-modal Dataset. . . 41

4.6 Performance Results for the Multimodal Approach. Radiomic and Semantic Models were also trained in the same 10 runs for baseline comparison purposes. . . . 42

4.7 Average Performance Results for the DL Approach. 48

4.8 Overview of average Robustness Results for the DL approach. The results are indicative of the similarity between explanations generated by the predictive model with different train-validation splits and the same test sample. The metrics were computed for the three explainable methods: Saliency Maps (SM), Integrated Gradients (IG) and LRP. The reference column (ref) indicates the perfect score for each metric, which would reflect on equal explanations for the same image. . . . 48

5.1 Overview of Performance Results for the proposed approaches concerning MNA prediction. 52

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUROC	Area Under the Receiver-Operator Characteristics
CAD	Computer-Assisted Detection
ccRCC	clear cell Renal Cell Carcinoma
CLF	Classification
CNN	Convolutional Neural Network
CT	Computerized Tomography
DA	Data Augmentation
DL	Deep Learning
EGFR	Epidermal Growth Factor Receptor
FISH	Fluorescence in situ hybridization technique
FC	Feature Construction
FS	Feature Selection
GBM	Gradient Boosting Machine
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLSZM	Gray Level Size Zone Matrix
GLRLM	Gray Level Run Length Matrix
GTV	Gross Tumor Volume
HSJ	Hospital de São João
HU	Hounsfield Units
HVS	Human Visual System
ICC	Interclass Correlation Coefficient
IDRF	Image-Defined Risk Factor
INRGSS	International Neuroblastoma Risk Group Staging System
IPO	Instituto Português de Oncologia
IQA	Image Quality Assessment
KFS	Koehrsen's Feature Selection
LASSO	Least Absolute Shrinkage and Selection Operator
LIME	Local Interpretable Model Agnostic Explanatinons
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
MNA	MYCN amplification
MRI	Magnetic Ressonance Imaging
mRMR	minimum Redundancy Maximum Relevance
MS-SSIM	Multi-Scale Structural SIMilarity
NB	Neuroblastoma

NSCLC	Non-Small Cell Lung Cancer
PET	Positron Emission Tomography
PCA	Principal Component Analysis
PDAC	Pancreatic Ductal Adenocarcinoma
RF	Random Forest
RMSE	Root Mean Squared Error
ROI	Region Of Interest
ROSE	Random Oversampling Examples
SHAP	SHAPley Additive exPlanations
SIOPEN	International Society of Pediatric Oncology European Neuroblastoma
SMOTE	Synthetic Minoriyt Over-sampling Technique
SSIM	Structural SIMilarity
SVM	Support Vector Machine
US	Ultrasound
VOI	Volume Of Interest
XAI	eXplainable Artificial Intelligence
XGB	eXtreme Gradient Boosting

Chapter 1

Introduction

Neuroblastoma (NB) is the leading malignant tumor in children under the age of 1 worldwide, constituting around 30% of all diagnosed cancers among European and US infants, that by the age of two decreases by half. Due to its sporadic nature, the challenge of studying the aetiology of this pathology remains, being few the factors associated with the diagnosis and prognosis of NB (Heck et al., 2009).

NB arises from malignancy of neural crest cells that endure differentiation and migration to form the sympathetic nervous system, leading to tumor formation in various locations along the latter, with higher frequency on the adrenal glands, but also in the abdomen, chest or pelvis. The disease characterization follows the International Neuroblastoma Classification System to predict prognosis (Heck et al., 2009).

Due to its clinical behavior variability, the range of possible outcomes is wide, since it can follow three possible outcomes: spontaneous regression or maturation to a benign ganglioneuroma, that usually present high survival rates (80-96%), or rapid progression to life-threatening aggressiveness, which have a survival rate lower that 50% and may include patients that do not respond to therapy (Huang and Weiss, 2013). The survival of patients may rely on several factors, which include the age of diagnosis and the stage and biological profile of the disease. Moreover, a low survival probability is shown to be correlated with an advanced age of diagnosis (later than 15 months), advanced stage of the disease, and the presence of specific biomarkers, namely the myelocytomatosis viral related oncogene, NB derived (MYCN) oncogene (Heck et al., 2009), present in 20% of neuroblastoma cases, and in 40% of high stage tumors (Marrano et al., 2017).

1.1 Motivation

Since NB is characterized by its significant variability of the subset of neuroblastic tumors, the clinical protocol includes the adjustment of therapy in compliance with the predicted biological behavior of the individual tumor, which is determined according to several factors that include not only the age of the patient and the tumor stage, but also histological properties, tumor cell DNA content and the MYCN oncogene, being the latter highly correlated with aggressive tumor behaviour and, consequently, poor prognosis, regardless of age and stage of the tumor (Ambros et al., 2009). Therefore, the MYCN amplification (MNA) constitutes a key prognostic factor, essential for the evaluation and risk stratification of this pathology (Brisse et al., 2017).

Currently, the MNA detection protocol requires the study of a tumor sample through open or needle biopsy. These procedures are invasive, being not only painful for the patients, but also possible of causing complications, such as hemorrhage, bowel obstruction, and infections (Wu et al., 2021). Moreover, the study of a mass by analyzing a small portion of the latter can lead to non-representative solutions of the totality of the tumor, especially taking into account the heterogeneity of NB, namely for MYCN copy number in tumor cells (Marrano et al., 2017). Thus, the need arises to develop a non-invasive MNA detection method that is able to perform the analysis of the whole tumor, aiding doctors in a more accurate implementation of the best course of treatment for the patient, along with all the currently utilized diagnostic and staging tools.

Imaging techniques currently play a critical role in the clinical decision-making process in oncology, mainly in diagnosing and staging phases, presenting several tumoral characteristics (origin, volume, shape and local extension) that can provide relevant insights regarding the prognosis (Brisse et al., 2017). Computerized Tomography (CT) is one of the recommended techniques for evaluation of NB due to its fast acquisition time, therefore avoiding the need for sedation (Di Giannatale et al., 2021). Moreover, imaging phenotypes have shown correlations with the neuroblastoma tumor genomic profile (Brisse et al., 2017), motivating to the implementation of a radiomic analysis, which is a trending domain of advanced image analysis through mathematical computation algorithms, able to extract a large amount of features from a standard radiological image (Di Giannatale et al., 2021). Radiomics derived data has been shown predictive capability of specific gene expression patterns in different tumors. This methodology, named Radiogenomics, provides an alternative method to assess the genotype of the patient from phenotype features that can be acquired by medical imaging techniques (Bodalal et al., 2019; Rutman and Kuo, 2009). This type of information could allow physicians to find correlations with oncology-related prognostic factors, constituting a predictive biomarker.

In a nutshell, the detection of biomarkers is now a hallmark for cancer treatment selection and outcome prediction, which inherently leads to the importance of genetic profiling. Radiogenomics allows for the development of non-invasive, computer-assisted techniques to fulfill the said need (Bodalal et al., 2019). Moreover, a radiogenomic approach allows for a more comprehensive analysis of the tumor, since the medical exam provides a three-dimensional capture of the tumor and, consequently, of the intra tumural heterogeneity, contributing to overcome the sampling bias by having a more broad representation of the lesion, not mentioning the ultimate advantage of non-invasiveness (Chen et al., 2021).

1.2 Objectives

The aim of this dissertation is to develop a non-invasive binary detection system for MNA using imaging phenotypes identified in CT images that can contribute to the decision-making process in the field of precision medicine. In order to achieve the best predictive system with the available data resources, two distinct studies were conducted: (1) a machine learning approach using tabular data, namely automatically extracted from the CT images and from clinical annotations relevant to the usual protocol; (2) a deep learning approach considering the CT images as input and assessing the robustness of the model using interpretability methods.

Firstly, a study on the potential of the available data using classical machine learning techniques will be performed: standard, commonly used methods for medical applications will be trained with two different types of data, radiomic (features automatically extracted from the CT. images) and semantic (clinical annotations and patient information that are relevant for the prognosis assessment in the clinincal routine), generating two different models. Lastly, an ensemble approach will have place to evaluate the complementarity of the two different models to predict MNA.

The second portion of dissertation will target the use of deep learning to predict MNA using the available images per say, exploiting the consideration of different portions of the images as region of interest (ROI) to feed the model as input. After obtaining a model with satisfactory performance and behavior, a robustness study will be performed on the developed network by utilizing post-hoc explainable methods to generate maps of pixel contributions to the prediction.

1.3 Contributions

This dissertation presents the following contributions:

- A study on Artificial Intelligence (AI) systems to predict MNA status.
- Utilization of Machine Learning (ML) methods to explore the correlation between radiomic features and MNA status.
- Utilization of ML methods to explore the correlation between semantic features and MNA status.
- Utilization of ML ensemble methods to explore the combination of radiomic and semantic features to predict MNA status.
- Utilization of DL method to explore the correlation between CT slices of NB tumors and MNA status.
- Utilization of XAI methods as a tool to assess the robustness of models trained with different data splits.

1.4 Document Structure

This dissertation include 5 Chapters:

- Chapter 1 - Introduction: brief introduction of the problem to be studied, as well as some motivation for the proposed solutions.
- Chapter 2 - Literature Review: detailed overview of the state-of-the-art methods with relevance for the problem in question, namely Radiogenomic Approaches, Multi-modality and Interpretability.
- Chapter 3 - Data Description: detailed description of the utilized dataset, as well as the data preparation process before using as input in the models.
- Chapter 4 - MNA Status Prediction with ML Approaches: description of the 4 developed approaches, including methodology, results and discussion.
- Chapter 5 - Conclusions and Future Work: overview of the developed work, overall conclusions and future work considerations.

Chapter 2

Literature Review

The present chapter concerns an overview of relevant studies regarding Radiogenomics approaches in several types of cancer (2.1), namely in the target use-case of this dissertation - Neuroblastoma Cancer and the detection of an important prognostic biomarker. A collection of two relevant trends of AI-based systems is also covered in two distinguished sections: Multi-modality (2.2) and Interpretability (2.3).

2.1 Cancer Radiogenomics

Gene expression profiling provides an additional understanding regarding tumorogenic processes, assisting in diagnosis, staging, prognosis and treatment response prediction in cancer patients (Rutman and Kuo, 2009). These biomarkers can also be clinically relevant for the assessment of risk profiling and target therapy treatments, but their detection currently relies on the surgical procurement of tissue, and invasive procedure that allows for the analysis of a tumor sample that may not be representative (Bera et al., 2022).

The potentiality of Artificial Intelligence (AI) for radiology applications has motivated the research on several predictive tasks on cancer imaging analysis (Reyes et al., 2020), since radiographic imaging incorporates the routine of clinical care and studies have shown associations between radiological tumor phenotypes and gene expression signatures, providing information concerning tumor sub-type and molecular biology (Rutman and Kuo, 2009). This AI-enabled biomarker prediction approach, known as Radiogenomics, arises as an alternative solution to detect biomarkers with the potential of wildly improving the medical standard protocol procedure, due to its detection through routine clinical radiology scans, non-invasiveness and non-tissue destructiveness, small detection time-frame, easy serialization, low costs and no changes in the current clinical workflow. Additionally, a radiogenomic approach allows a global analysis of the tumor, overcoming the heterogeneity problem when sampling the tumor (Bera et al., 2022).

This methodology of biomarker prediction can be obtained with two main approaches: handcrafted radiomic and deep learning (DL) techniques. The radiomic approach uses quantitative measurement features that can be obtained with off-the-shelf radiomics libraries and are fed into a machine learning (ML) pipeline. DL approaches develop neural networks to generate new representations of the image data that can be synthesized into a biomarker prediction (Bera et al., 2022). Recent advances in the field allowed the achievement of promising results in several cancer types and biomarkers. Table 2.1 describes some of the most relevant investigations found related to this research topic.

Table 2.1: Overview of published studies regarding classification tasks with used radiogenomic approaches.

Publication	**Cancer Type**	**Image Source**	**Input**	**Output**	**Model**	**Best Results AUC**
Li et al. (2017)	Glioma	MRI	Image Data	IDH1 mutation status	[DL] CNN with normalized last convolutional layers	0.9207
Wang et al. (2019)	Lung Cancer (NSCLC)	CT	Image Data	EGFR mutation status	[DL] Sub-Network 1 (natural images) Sub-Network 2 (CT images)	0.81
Kocak et al. (2019)	Kidney Cancer (ccRCC)	CT	Radiomic Features	PBRMI mutation status	[DL] ANN [ML] RF	(ANN) 0.925 (RF) 0.987
Iwatate et al. (2020)	Pancreatic Cancer (PDAC)	CT	Radiomic Features	p53 mutation status PD-L1 expression status	[ML] XGBoost	(p53) 0.795 (PD-L1) 0.683
Morgado et al. (2021)	Lung Cancer (NSCLC)	CT	Radiomic Features	EGFR mutation status	[ML] PCA 70% + SVM	0.737
Chen et al. (2022b)	Breast Cancer	PET CT	Radiomic Features	HER2 expression status	[ML] XGBoost	0.76
Xu et al. (2022)	Breast Cancer	US	Image Data	HER2 expression status	[DL] End-to-end 3-block DenseNet	0.84

Morgado et al. (2021) proposed a machine learning (ML) pipeline with the experimentation of several feature selection and ML methods on the Non-Small Cell Lung Cancer (NSCLC) Radiogenomics Dataset (Bakr et al., 2018) to predict the EGFR mutation status, an important biomarker for target therapy in lung cancer patients. The input consisted on radiomic features extracted from the total volume of the lung containing the nodule on CT scans. The best results were achieved with Principal Component Analysis (PCA) as feature selection method, reaching similar performances with several commonly used classifiers, namely Elastic Net and linear Support Vector Machine (SVM), showing the latter slightly higher results. Overall, linear models presented the best behavior for the classification task in question.

Wang et al. (2019) proposed a DL architecture to predict the EGFR mutation status on two private datasets of adenocarcinoma CT scans, a subtype of NSCLC. A region of interest (ROI) containing the tumor for all slices of the patient was given as input to the network, being the EGFR mutation status probability the average of each patient probabilities. The pipeline of this work is demonstrated in Figure 2.1, where two sub-networks are presented. The sub-network 1 is equivalent to the first 20 layers in DenseNet (Huang et al., 2017), having been pre-trained with natural images. The sub-network 2 is trained with the dataset of the study.

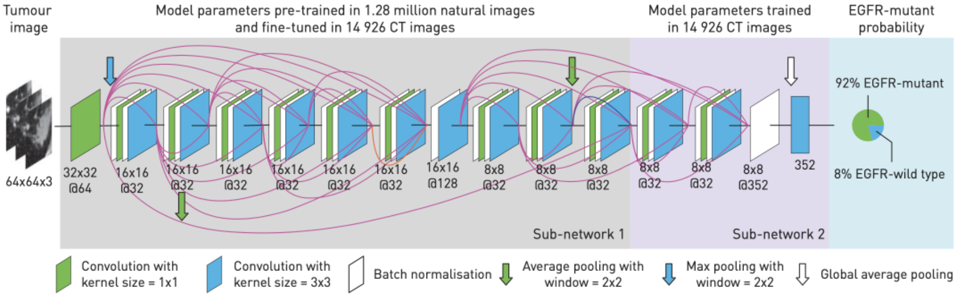
Tumour image
Model parameters pre-trained in 1.28 million natural images and fine-tuned in 14 926 CT images
Model parameters trained in 14 926 CT images
EGFR-mutant probability
64x64x3
32x32 @64
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @32
16x16 @128
8x8 @32
8x8 @32
8x8 @32
8x8 @32
8x8 @32
8x8 @352
352
92% EGFR-mutant
8% EGFR-wild type
Sub-network 1
Sub-network 2
Convolution with kernel size = 1x1
Convolution with kernel size = 3x3
Batch normalisation
Average pooling with window = 2x2
Max pooling with window = 2x2
Global average pooling

Figure 2.1: Architecture of DL model: convolutional layers with kernel size 3×3 and 1×1, batch normalization and pooling layers. From Wang et al. (2019).

Chen et al. (2022b) proposed a ML pipeline with the experimentation of several ML methods on a private dataset of breast cancer PET and CT scans to predict the HER2 expression status, important for target therapy purposes. The input of the pipeline was an integration of radiomic features extracted from both PET and CT scans of the volume of interest (VOI) containing the gross tumor volume (GTV). The best results were achieved using eXtreme Gradient Boosting (XGBoost).

Xu et al. (2022) proposed a DL architecture to predict the HER2 expression status of a private dataset of breast cancer cross-section ultrasound (US) images. The input of the network comprised a cropped, resized ROI of the image, being the pipeline of the proposed work presented in Figure 2.2. The proposed network contains 2 types of dense-block, having the dense-block 14 layers, whilst the dens-block 2 has 32 layers. Both types have shortcut connections from one layer to the subsequent ones.

Iwatate et al. (2020) proposed a ML pipeline with the experimentation of XGBoost to predict both p53 mutation status and PD-L1 expression status, two biomarkers with clinical relevance for

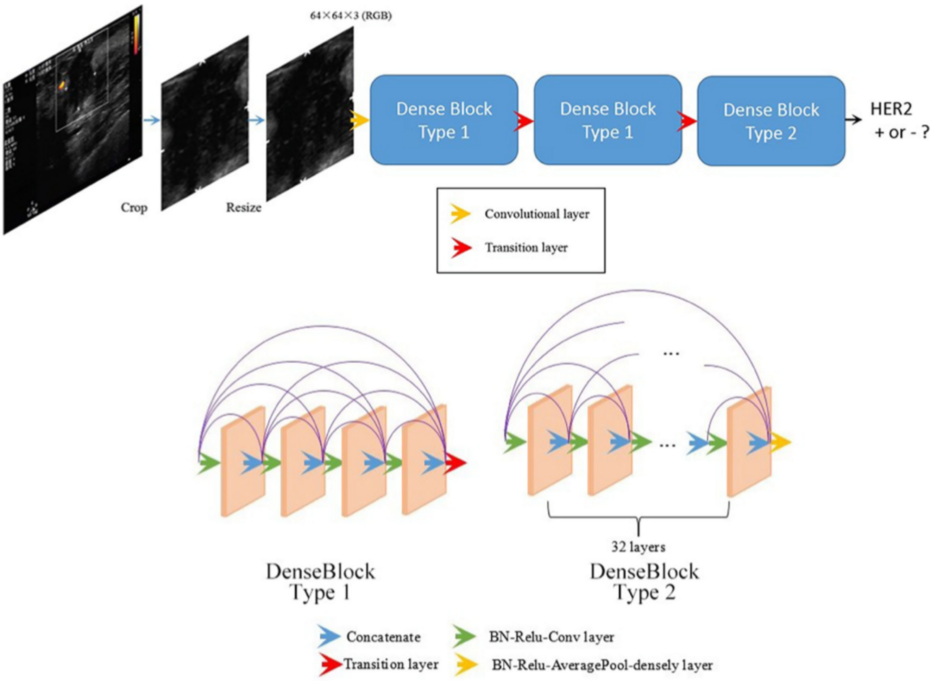

Figure 2.2: Architecture of DL model: DenseNet-based deep learning classifier. From Xu et al. (2022).

pancreatic cancer for being a tumor suppressor gene and a poor prognostic factor, respectively. The utilized dataset was private and contained CT scans from patients with pancreatic ductal adenocarcinoma (PDAC). The pipeline was fed radiomic features extracted from the VOI of the segmented tumor.

Kocak et al. (2019) proposed two different pipelines for the prediction of PBRMI mutation status, a biomarker associated with poor survival rate in clear cell renal cell carcinoma (ccRCC), a subtype of kidney cancer. The utilized dataset was obtained from The Cancer Genome Atlas - Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) database (Akin et al., 2016; Clark et al., 2013), being the input of the pipeline radiomic features extracted from the 3D VOI of the tumor. The radiomic features were fed to two distinct algorithms: Artificial Neural Network (ANN) and Random Forest (RF), being the performance superior in the classical machine learning approach - RF.

Li et al. (2017) proposed a DL approach for the prediction of the IDH1 mutation status in glioma patients through a private MRI imaging dataset. The used pipeline is illustrated in Figure 2.3.

2.1.1 Neuroblastoma Cancer Radiogenomics

As demonstrated in the previous section, utilizing radiogenomics approaches with the analysis of CT-based radiomic features has shown promising results for predicting specific target biomarkers in several tumors. More recently, this technique was also applied to Neuroblastoma (NB), being

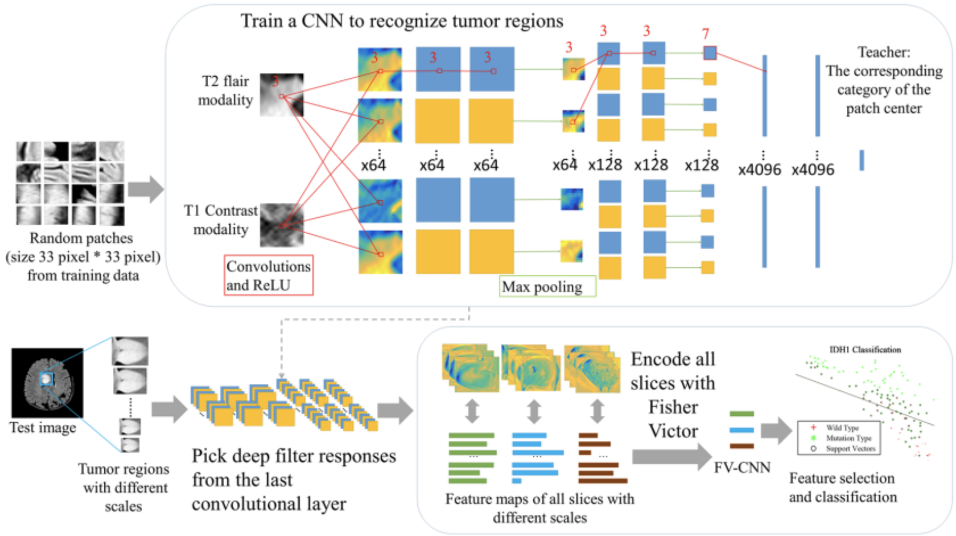
Train a CNN to recognize tumor regions
T2 flair modality
T1 Contrast modality
Convolutions and ReLU
Max pooling
3
3
3
3
3
3
7
x64
x64
x64
x64
x128
x128
x128
x4096
x4096
Teacher: The corresponding category of the patch center
Random patches (size 33 pixel * 33 pixel) from training data
Test image
Tumor regions with different scales
Pick deep filter responses from the last convolutional layer
Encode all slices with Fisher Victor
Feature maps of all slices with different scales
FV-CNN
IDH1 Classification
Wild Type
Mutation Type
Support Vectors
Feature selection and classification

Figure 2.3: Architecture of DL model: two selection steps, initialized with recognition of tumor regions in the input images through a state-of-the-art based CNN structure, and followed by deep filter responses extraction from the last convolutional layer through Fisher vector encoding. The prediction is obtained through an SVM. From Li et al. (2017).

all the publications to the best of our knowledge related to MNA detection in CT scans and using classical ML techniques for the predictive model. Table 2.2 describes some of the most relevant investigations found related to this research topic.

Wu et al. (2021) proposed a simple radiomic pipeline to return the MNA prediction. The features were extracted from the ROI of the tumor and selected using the Least Absolute Shrinkage and Selection Operator (LASSO) regression method. Data Imbalance was addressed with data augmentation, being used the SMOTE library (Chawla et al., 2002). A radiomics score (rad-score) was implemented with a linear combination of the features after selection.

Di Giannatale et al. (2021) used logistic regression (LogReg) to radiomic features extracted from CT images and selected with two feature selection methods: Boruta algorithm and Pearson correlation analysis.

Chen et al. (2021) used four ML algorithms to predict MNA with radiomic features from the ROI containing the tumor: Logistic Regression, Support Vector Machine, Bayes and Random Forest. The features are selected before being fed to the methods by utilizing Interclass Correlation Coefficient (ICC); minimum Redundancy Maximum Relevance (mRMR) and LASSO methods.

More recently, Tan et al. (2022) proposed a MNA detetcion ML approach utilizing 3D radiomic fetaures from the VOI containing the tumor. The features were selected by computing and analysing the correlation matrix of the latter with a cut-off threshold of 0.70 correlation. Data imbalance was addressed with an oversampling algorithm - Random Oversampling Examples (ROSE) (Lunardon et al., 2014). The XGBoost algorithm was used to train the predictive model.

As previous work, Pereira et al. (2022) used a Random Forest-based classifier for MNA status prediction utilizing 2D Radiomic Features extracted from CT slices. The proposed work utilized a private dataset that will be utilized for this dissertation, being further described in the next chapter. Thus, this publication will be the baseline standard for the developed pipelines.

Table 2.2: Overview of published studies regarding the prediction of MNA status in CT images of NB patients.

Publication	**Methods**	**Input**	**Best Results AUC**
Wu et al. (2021)	Features extracted from three-phase CT images Feature Selection with LASSO regression methodology SMOTE to address data imbalance Rad-score from linear combination of features	2D Radiomic features of ROI	[Rad-score] 0.92
Di Giannatale et al. (2021)	Features extracted from CT images Feature Selection with Boruta algorithm and Pearson correlation analysis Predictive model: LogReg	2D Radiomic features of ROI	[LogReg] 0.813
Chen et al. (2021)	Features extracted from three-phase CT images Feature selection with ICC, mRMR, and LASSO methods Predictive models: LogReg, SVM, Bayes, and RF	2D Radiomic features of ROI	[LogReg] 0.909 [SVM] 0.909 [RF] 0.851 [Bayes] 0.729
Tan et al. (2022)	Feature extracted from CT images Feature Selection with correlation matrix (T>0.7) and mRMR Data Imbalance with oversampling Predictive model: XGBoost	3D Radiomic features of VOI	[XGB] 0.880
Pereira et al. (2022)	Feature extracted from CT images Predictive model: Random Forest	2D Radiomic features of ROI	[RF] 0.69

2.2 Multi-modality

Clinical routine includes various phases, namely screening, diagnosis, treatment selection and prognosis assessment, during which several types of medical data are generated, such as clinical data, laboratory data and medical imaging data (Jiang et al., 2017). The totality of the acquired information is taken into consideration in clinical practice to make a final decision, being however a complex approach, possibly limited by the inherent subjectivity of the human decision-making process. Computer-aided solutions based on AI techniques have proven ability of providing supportive assistance in the clinical routine (Aljaaf et al., 2015).

Imaging data extracted from Computed Tomography (CT) scans have been able to achieve promising results when utilized for training AI-based methods for tasks such as the classification of malignancy risk (Silva et al., 2020) and for the characterization of cancer genotypes (Pinheiro et al., 2020). Furthermore, the use of annotated semantic features has also shown to provide relevant information for cancer characterization (Pinheiro et al., 2020; Gevaert et al., 2017). The incorporation of clinical data, genomic profiling and medical imaging may leverage the prognosis predictive capability, which will improve clinical decision-making (Bi et al., 2019).

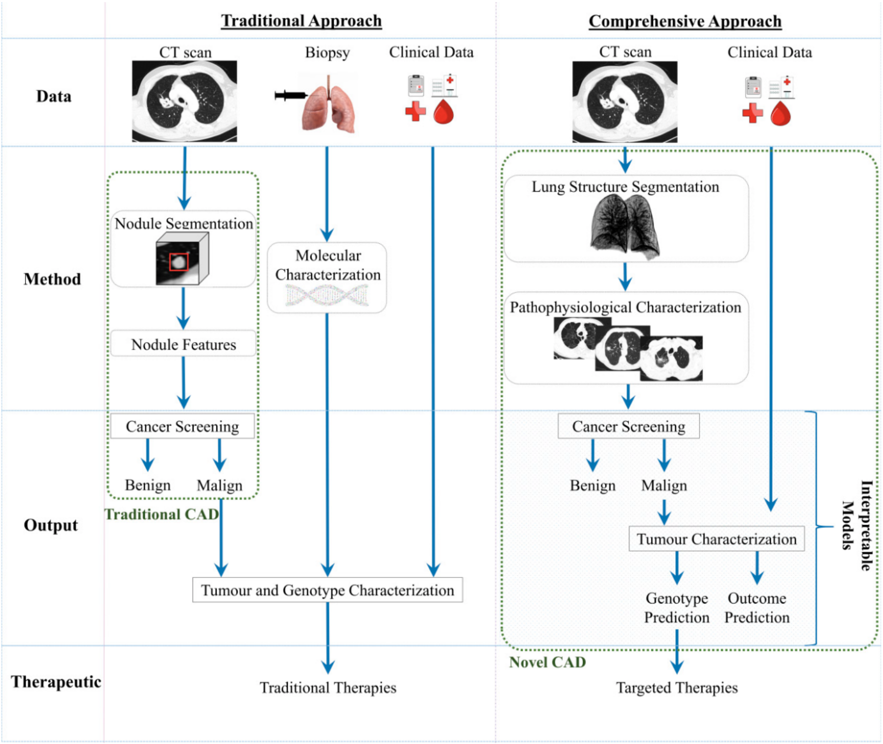

Figure 2.4: Conceptual diagram of the traditional and comprehensive approaches based on CADs, illustrated with the use-case of lung cancer. From Pereira et al. (2020).

A multi-modal approach provides the opportunity of exploring relationships between data modalities (Bi et al., 2019), which can aid in the identification of specific biomarkers that are not usually detected on an individual input modality. The development of a multi-modal comprehensive system will allow for the combination of information from several data sources by claiming that the whole is greater than the sum of its parts (Fiandaca et al., 2017). However, due to the high degree of variability within the gathered data, there is the need for a representative arrangement of the data that covers all the heterogeneities, in order to create a reliable representation of the affected population (Pereira et al., 2020), as illustrated in Figure 2.4.

Recently, several approaches have used multi-modal datasets to achieve higher scores. Most state-of-the-art pipelines include fusing features from different medical imaging techniques, namely CT scans and PET scans (Chen et al., 2022b), magnetic resonance imaging (MRI) and ultrasound (Gayet et al., 2016).

In Malafaia et al. (2021), an machine learning (ML)-based exploratory study is performed on the Non-Small Cell Lung Cancer (NSCLC) Radiogenomics public dataset (Bakr et al., 2018) to predict the mutation status of the Epidermal Growth Factor Receptor (EGFR) biomarker. The used methodologies consist on exploring the contributions of ensemble methods when applied to the combination of clinical data, semantic data, and radiomic data. The proposed pipeline can be found in Figure 2.5.

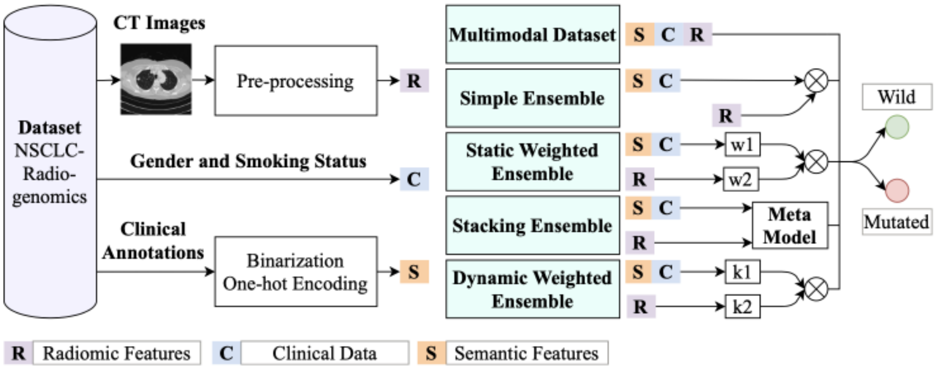

Figure 2.5: Pipeline developed for EGFR mutation status classification based on ensemble methods to combine patient clinical information, radiomic features, and semantic features. As illustrated in the diagram, five methods are studied to join the different modalities of features and make the final prediction. From Malafaia et al. (2021).

The five used methods were studied and tuned in order to outperform the performance of the best single modality learner: the semantic model, being used XGBoost for all the ML predictive models. The first approach, Multimodal Dataset, consisted on training the predictive model with all data modalities. The following approach, Simple Ensemble, consisted on the utilization of a simple cascade model that fed the predictions of the semantic model along with the radiomic features to make the final prediction. The third approach, Static Weighted Ensemble, computed the output classification by performing a weighted average of the predictions from both radiomic and semantic data. A Stacking Ensemble approach was also used, in which a simple linear meta-model outputs the final predictions, being fed as features the predictions of the learners in question. The last approach, Dynamic Weighted Ensemble, focused on introducing some novelty to an weighted average of the predictions, by dynamically assigning the weights of each learner according to the confidence on its prediction. To achieve that, a confidence measurement was defined as the distance of a sample on a specific predictive outcome to the value 0.50. Despite not being able to outperform the stronger learner, this study describes several commonly used approaches that may increase the predictive ability of the model by broaden the spectrum of utilized data modalities, however being extremely dependent on the data availability and single modality predictive capability.

On now a deep learning (DL) perspective, DL multi-modal approaches mainly rely on the utilized technique for the fusion of features. In Chen et al. (2022a), a DL fusion approach is proposed to combine medical images with clinical data in skin cancer patients. The proposed pipeline in Figure 2.6 illustrates the use of a feature extractor to transform the medical image into a feature vector, and one-hot encoding and transformation of clinical data to apply feature fusion and an attention mechanism to predict skin cancer. Despite the elaborate pipeline, the authors face the main challenge of multi-modality by not being able to outperform the single-modality model with feature fusion. This demonstrates that multi-modal approaches can cause overlapping and using redundant information due to the complexity of the fusion features.

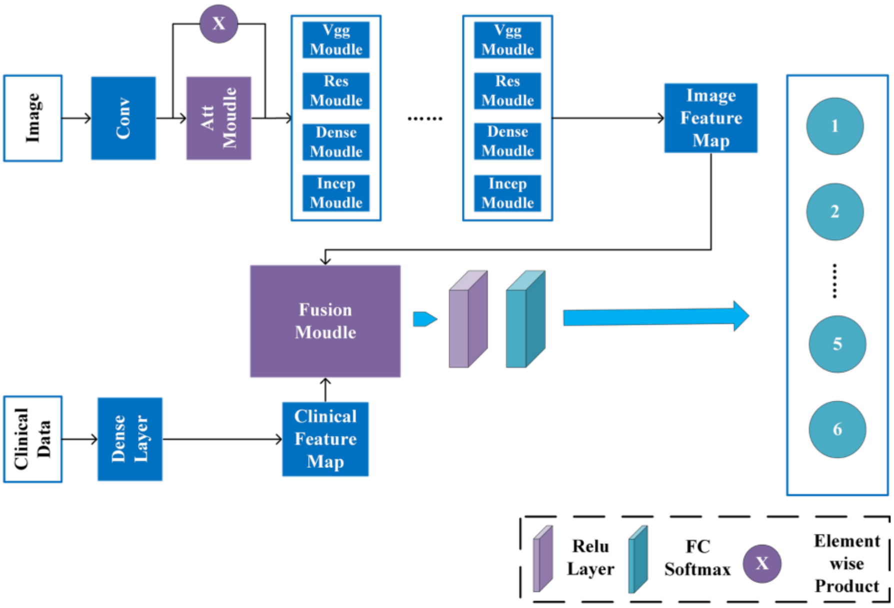

Figure 2.6: Architecture diagram of the developed DL model. Image and clinical data suffer transformations in order to be combined with fusion techniques. From Chen et al. (2022a).

2.3 Interpretability

The study of AI-based systems to applications in the medical domain have achieved top results, namely in radiology applications. The developed solutions, however, consist of complex, black-box models that present inherent difficulties for human comprehension of their reasoning. Consequently, the trust on the decisions of these systems is compromised, principally for high-stake decisions such as in a clinical practice environment (Reyes et al., 2020). Therefore, the development of systems that are transparent and in which humans trust is crucial. Additionally, within machine learning (ML) approaches, deep learning (DL) methods are considered the least interpretable due to the inherent mathematical complexity, leading to the lack of reasoning for the prediction and, hence, the lack of trust in these models (Reyes et al., 2020). To allow these frameworks to be potentially used in the medical domain, it is essential to ensure trustworthiness and reliability to clinicians, enforcing the need of transparency and easy comprehension for humans incorporated in these approaches.

Explainable AI (XAI) refers to techniques or methods that aim to find connections between the input and the prediction of the black-box; hence, looking to provide some reasoning to the decision and its reliability. Perceptive interpretability consists of XAI methods with a focus on generating interpretations that can be easily perceived by humans, despite not actually 'unblack-boxing' the algorithm (Tjoa and Guan, 2020). Visual Explanations are the most commonly used XAI methodologies in deep learning image analysis approaches (van der Velden et al., 2022), namely in radiology image-based predictive models, where the trust on a Computer-Assisted Detection (CAD) system can increase substantially by presenting the areas of a medical image with higher contribution to the prediction, along with the prediction itself (Reyes et al., 2020).

A large portion of the most utilized XAI methods in the medical domain are post-hoc models, which consist of methods external to the already trained predictive model, performing evaluations on the predictions without altering the model itself. These are off-the-shelf agnostic methods that can be found in open-source libraries, such as Pytorch Captum (Kokhlikyan et al., 2020). An example of a post model approach can be found in Knapič et al. (2021), where two popular post-hoc methods, local interpretable model agnostic explanations (LIME), and SHAPley Additive exPlanations (SHAP) were compared in terms of understandability for humans in the predictive model with the same medical image dataset.

Moreover, gradient-based methods have been extensively used to explain COVID-related deep learning models using medical image (Hryniewska et al., 2021), due to the need to produce more information about the classification models to be used in the clinical routine. An illustration of these studies is presented in Figure 2.7. These models have been studied and used in several other studies, namely for lung classification tasks, demonstrated in Lei et al. (2020).

As previous work is concerned, Malafaia et al. (under peer-review) proposed a pipeline for a lung nodule classifier that uses XAI gradient-based methods in order to assess the robustness of the predictive model over different data splits. Explanations are generated for all test samples and all trained models, being compared the explanations from the same sample with image similarity metrics.

The concern for interpretability is increasing, especially in the medical field, where there are high stakes and responsibilities in the CAD systems that are used. However, the research in the area of interpretable models is still in progress, despite the recent rise in the development of this approach. The increase in research efforts of interpretable CAD systems is already noticeable,

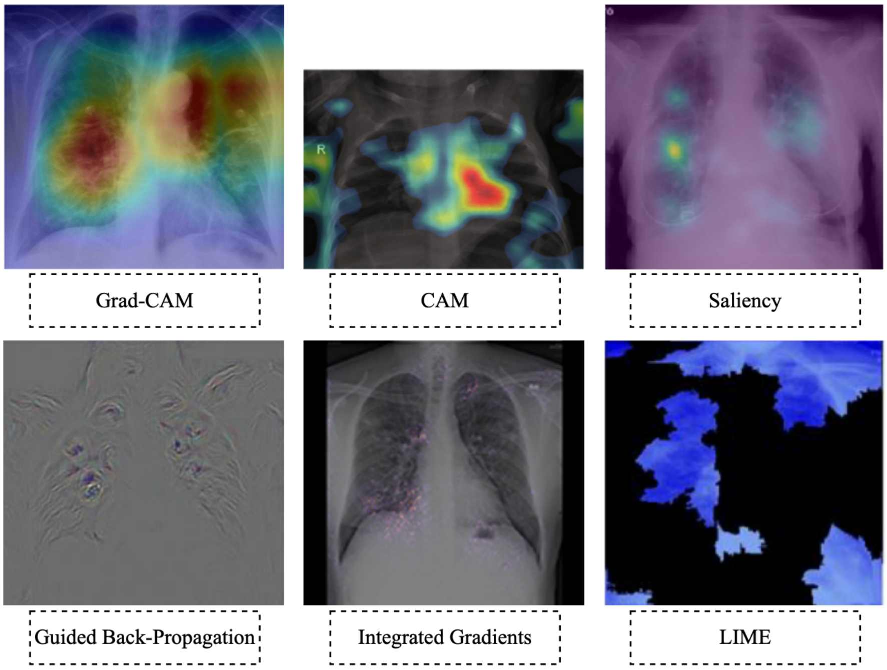

Figure 2.7: Example of explanations for COVID-related models from several studies. From Hryniewska et al. (2021).

mainly regarding the verification and explanation of the predicted decision, rather than the unravelling of the black-box (Tjoa and Guan, 2020). The methods may show future potential, not only in providing trustworthy explanations to physicians, but also assuring the reliability and consistency of the developed models.

2.4 Summary

The studies enumerated in this chapter were selecting considering three main subjects: the development of radiogenomics approaches, namely for neuroblastoma cancer; the combination of several data modalities; the incorporation of interpretability to the developed black-box models, namely using DL techniques. Considering the overall analysis of the reviewed issues and the considered methodologies, there are several key aspects one should consider:

- Radiogenomics approaches have shown promising results in several types of cancer, both using classical ML or DL techniques;
- Studies concerning the detection of the MNA biomarker are limited, being, to the best of our knowledge, related to classical ML techniques and demonstrating a high correlation between MNA detection and imaging phenotypes;

- The implementation of more than one data modality to an AI-based system can provide complementary information and increase the predictive capacity of the framework. However, the employment of different modalities to the model is a critical task that may not be successful due to data redundancy;
- Interpretability is a key aspect to future deployment of the model into a CAD system; there are several aspects of the functioning of the developed pipeline that may be studied with XAI, namely the reasoning behind a prediction made by the developed pipeline. Visual explanations, namely gradient-based methods, are suitable for human understanding and provide insightful information regarding the areas of higher importance in a input image.

Chapter 3

Data Description

The following chapter illustrates the database that was utilized for this work, on top of the required steps to adjust the raw medical data to properly fit the model development stage. Hence, two sections were included: Dataset 3.1, where a description of the utilized data, inclusion criteria and available information is presented; Data Pre-processing 3.2, which entails the used methodologies on the presents types of data that lead to the features to be utilized in the AI models.

3.1 Dataset

For the proposed work, a private dataset was collected with the collaboration of Hospital de São João (HSJ) and Instituto Português de Oncologia (IPO). The named dataset consisted of multimodal data from 46 patients with diagnosed Neuroblastoma cancer between 2005 and 2020, that included CT scans and correspondent tumor segmentation masks of some slices, patient clinical information, and the MNA status (16 MNA positive patients / 30 MNA negative patients). The course of treatment chosen for all patients was based on previous or ongoing protocols or trials of the International Society of Pediatric Oncology European Neuroblastoma (SIOPEN).

Inclusion criteria for this study consisted on: age of diagnosis under 18 years old; confirmed neuroblastoma through an histopathological report; MYCN amplification detection; availability of CT studies at the time of diagnosis before any intervention such as biopsy, radiotherapy, chemotherapy or surgery; availability of patient clinical information.

3.1.1 CT scans

The available medical exams for the 46 patients concern the primary tumor location, which may include several anatomic compartments, namely abdomen, chest or pelvis. Regarding the image acquisition protocol, CT scans present a highly variable number of slices per exam, with a slice thickness range of 0.5-5.0 mm per slice and pixel spacing in (x,y) directions of 0.2383-0.9766 mm.

Regarding the segmentation masks of the tumor, experts manually segmented the mass in five slices of the exam, which were the ones utilized for this study. The remaining slices without

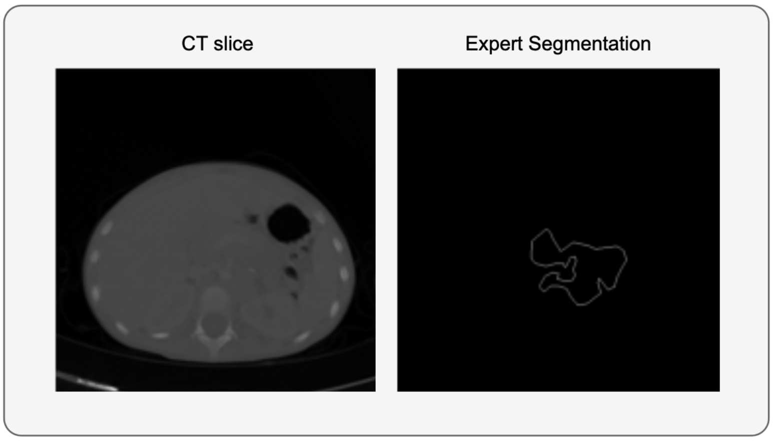

Figure 3.1: Demonstration of a CT slice (left) and the correspondent annotation of the tumor segmentation from a clinical expert (right). From the private utilized dataset in joint collaboration with HSJ and IPO.

annotations are discarded. In Figure 3.1, an illustration of the CT slice from the exam of a patient is presented, along with the following manual segmentation.

Additionally, for every exam there was information available regarding whether the CT scan was done with or without contrast.

3.1.2 Patient Clinical Information

Clinical information concerned data with clinical relevance for the prognosis of the patient, thus having potential to aid in the prediction of the MYCN amplification status.

Regarding patient data itself, the sex of each patient was available, as well as the age of diagnosis and information regarding the tumor, namely its anatomical location and type of mass. Furthermore, a description of the image-defined risk factors (IDRFs) was also provided, along with the International Neuroblastoma Risk Group Staging System (INRGSS). In addition, genetic alterations were also listed for all patients due to potential genetic predisposition.

3.1.3 MNA status

The MYCN amplification status was measured through Fluorescence in situ hybridization technique (FISH), being consideres amplified in case at least a fourfold increase in the MYCN expression was verified towards the reference probe. This requirement followed the regulations from the INRG Biology Committee (Ambros et al., 2009). This information will be used as the target label for which the designed frameworks will be trained.

Table 3.1: Patient Clinical Information Description. Each type of tabular information provided by the private dataset is enumerated in the following table. Additionally, the type of variable and possible values are also described.

Data	*Type of Variable*	*Possible Values*
Age of Diagnosis	Discrete	In months
Sex	Binary	0 - Female 1 - Male
Anatomical Compartment	Categorical	0 - Adrenal Glands 1 - Retroperitoneum 2 - Posterior Mediastinum 3 - Cervical 4 - Pelvis
Cross Middle Line	Binary	0 - No 1 - Yes
Morphology	Binary	0 - Single Mass 1 - Several Masses
Calcium	Binary	0 - Absent 1 - Present
IDFRs	Categorical	0 - Tumor encasing the aorta and/or vena cava 1 - Tumor encasing branches of superior mesenteric artery 2 - Tumor encasing origin of the celiac axis and/or of the superior mesenteric artery 3 - Tumor encasing brachial plexus roots 4 - Tumor encasing the iliac vessels 5 - Tumor encasing subclavian vessels and/or vertebral and/or carotid artery 6 - Tumor infiltrating duodeno-pancreatic block 7 - Tumor infiltrating mesentery 8 - Tumor infiltrating the porta hepatis 9 - Tumor compressing the trachea and/or principal bronchi 10 - Tumor invading renal pedicles 11 - Tumor crossing the sciatic notch 12 - Lower mediastinal tumor, infiltrating the costo-vertebral junction 13 - Intraspinal tumor extension 14 - Kidney 15 - Liver 16 - Pericardium 17 - Diaphragm 18 - Two body compartments
Number of IDFRs	Categorical	No. of listed IDRFs
INRGSS	Categorical	0 - L1: Local-regional tumor without IDRFs 1 - L2: Local-regional tumor with one or more IDRFs 2 - M: Distant metastatic disease 3 - MS: Metastatic disease in children younger than 18 months
Segmental Chromosomal Alterations	Binary	0 - No 1 - Yes
11q Deletion	Binary	0 - No 1 - Yes
1p36 Deletion	Binary	0 - No 1 - Yes
17q Duplication	Binary	0 - No 1 - Yes
Chromosome 1 Trisomy	Binary	0 - No 1 - Yes

3.2 Data Pre-processing

Each data type was pre-processed according to its characteristics. For image processing, the CT scans were converted from the original format to image data, undergoing several phases to standardize all images. Regarding tabular data, simple pre-processing steps were used to make all types of variables equivalent for the studied approaches.

3.2.1 Image Pre-Processing

Given the CT scans available for each patient, only 5 slices of each had annotations concerning the mask of the segmented tumor. Thus, only 5 slices were utilized for each patient. The original data format was DICOM, which contained all the relevant information, including the CT slices, tumor masks and pixel spacing.

3.2.1.1 DICOM Objects to Image Data Conversion

As mentioned above, it was necessary to convert the medical exam, stored in DICOM format, to an image array, in order to implement the necessary processing steps. For the conversion step, the open source package, *pydicom* (Mason, 2011), developed to work with data elements of DICOM data. Each *.dcm* file is organized by study (*SOPInstanceUID*), series (*SeriesSequence*) and slice image (*ImageSequence*), being provided both the image and, if available, a non-image file, correspondent to the mask annotation. Additionally, there is an abundance of meta-data available within the file, containing various information regarding the protocol of the medical exam that can also be helpful for the study, namely *Pixel Spacing* with tag (6000), *Rescale Intercept* with tag (0028,1052) and *Rescale Slope* with tag (0028,1053).

The tags of slice images of each patient with correspondent annotations were selected for conversion. The target images and annotations were saved into image arrays of, in most cases, 512×512 pixels with the *Pixel Array* and *Overlay Array* attributes, respectively. Additionally to the image and tumor segmentation, pixel spacing was saved to implement Pixel Resampling.

The annotations from the segmentation of the tumor were processed with a series of simple image processing steps, namely closing and opening methods, in order to fill the annotation into a binary mask representative of the Region of Interest (ROI). An illustrative example of the process in question is displayed in Figure 3.2.

3.2.1.2 Pixel Resampling

Due to inconsistencies in scanning protocols, CT scans from different patients present different spacing between adjacent pixels, being a common practice to resample all images in order to achieve a constant pixel spacing. For this task, the pixel spacing range of 0.2383-0.9766 mm was extracted from the *Pixel Spacing* attribute in the *.dcm* files. A new pixel spacing of 0.50 mm was chosen in order to minimize the loss information of images with small pixel spacing (<0.50 mm) and, on the other hand, minimize the addition of redundant information through interpolation

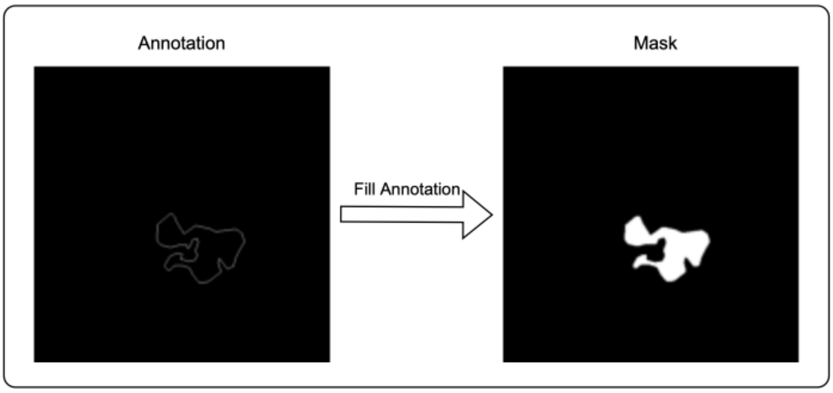

Figure 3.2: The annotation provided by the clinician, representative of the tumor segmentation, is transformed into a binary mask through a filling process achieved with closing and opening methods.

on images with large pixel spacing (>0.50 mm). This procedure was implemented using a *zoom* factor given by the ratio between the original pixel spacing and the new pixel spacing (0.50 mm), which was applied to change the resolution of both the image and the correspondent tumor mask. Due to the change of resolution, the size of the images was changed accordingly, as illustrated by Figure 3.3. This process guaranteed a similar representation of the same anatomical structures for the whole image dataset.

3.2.1.3 Image Size

After resampling all images in order to have the same pixel spacing, the resolution of the altered images was changes, consequently modifying the size of each picture. Furthermore, some exams presented non-squared dimensions. Thus, a size standardization process was implemented to all images in order to be squared and with constant image size. The chosen dimensions for image size were 224×224 pixels, since its a standard size, used, for example, in the ImageNet Dataset (Deng et al., 2009), that simultaneously is able to include the tumor size in the totality of the utilized dataset.

For cases in which the image size was smaller than the desired size, a padding was used with the intensity value of the border of the image in question, as illustrated in Figure 3.4. For images with size bigger than 224×224 after resampled, the surplus pixels were removed from the edges of each image through a cropping process, as demonstrated in Figure 3.5. The image size standardization process was applied to both images and respective tumor masks.

3.2.1.4 Intensity Normalization

CT scans use multiple X-Ray projections that pass through the tissues with different density values, being the detector reached with various levels of energy that provide contrast imaging of the different tissues. This density values are expressed with Hounsfield (HU) scale values. As demonstrated in Equation 3.1, the formula for HU computation varies with the linear attenuation

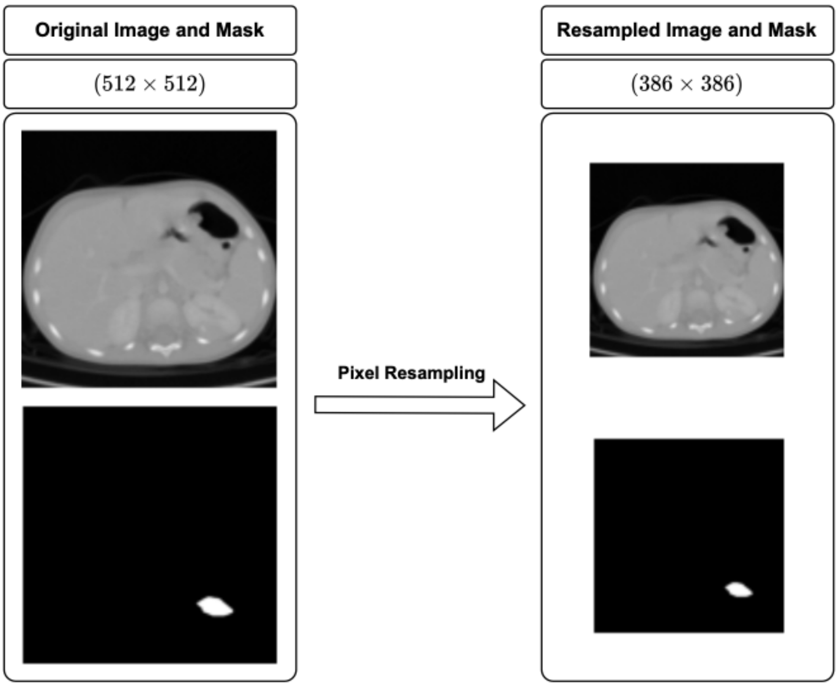

Figure 3.3: Illustration of Pixel Resampling Process. The original image, with size 512×512 pixels, is resampled in order to have a pixel spacing of 0.50mm between adjacent pixels. The new size, with lower resolution, is 386×386 pixels.

coefficient of water, μ_{water}, the linear attenuation coefficient of air, μ_{air}, and the linear attenuation coefficient of the substance μ (Kalra, 2018).

$$HU = 100 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{3.1}$$

The result of the equation 3.1 is based on a range in which the radiodensity of water is 0 (HU) and the radiodensity of air is -1000 HU, at standard pressure and standard conditions. Thus, higher HU values concern regions with higher densities, represented with brighter pixels in the image (Kalra, 2018).

In order to obtain the intensities in HU units, the pixel values were converted utilizing the attributes *Rescale Intercept* and *Rescale Slope*. However, after analysing the range of intensities amongst images from different exams, disparities were found which lead to discredit the veracity of these values.

To overcome this problem, a simple approach was implemented in order to scale the whole dataset according to a single reference image, chosen by visually assessment of the contrast between different tissues and the tumor itself. Thus, a non-linear sigmoid normalization was implemented following Equation 3.2, representing α the arithmetic mean of the input intensity range, and β the standard deviation of the input intensity range.

$$I_{norm} = (max_{new} - min_{new}) \frac{1}{1 + \exp^{-\frac{I - \beta}{\alpha}}} + min_{new} \tag{3.2}$$

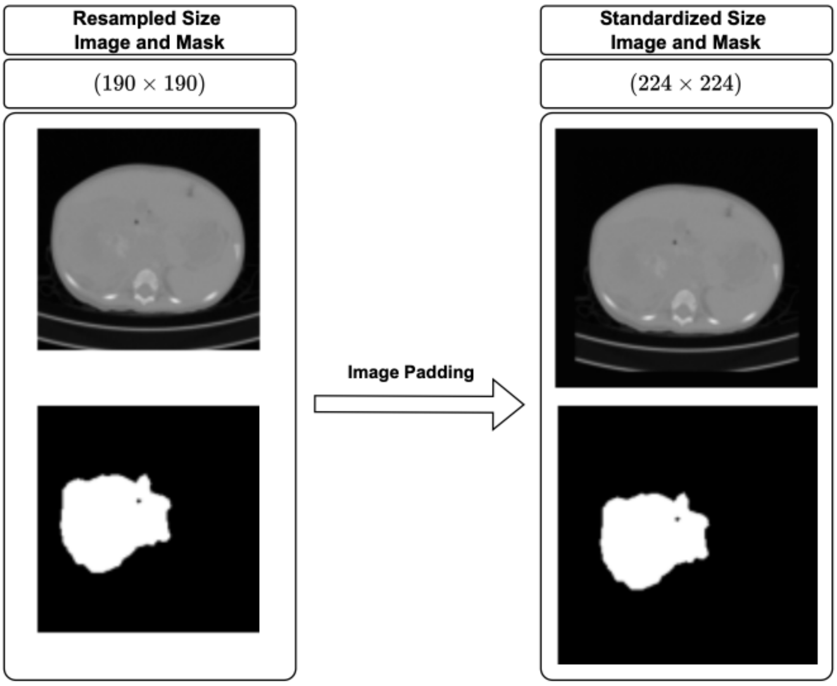

Figure 3.4: Illustration of Image Size Standardization Process for images size lower than 224×224 pixels using padding methods.

An illustration of the process of reference normalization is presented in Figure 3.6, where the intensity range of an image is normalized according to the defined image reference.

3.2.2 Region of Interest (ROI) Extraction

Taking advantage of the manual segmentation of the tumor for all images, and since the radiomic features were extracted exclusively in the ROI, one possible image input for DL models can consist on the portion of the images containing the tumor. Thus, images were pre-processed in order to obtain a transformed image containing only the tumor. In order to facilitate the learning process, and simultaneously as an attempt to overcome the variability of size and shape of the tumor, the ROIs were centered in the images, being the center of mass of the nodules constantly in the center of the image, as illustrated in Figure 3.7.

3.2.3 Radiomic Features Extraction

Radiomic features were extracted from the CT images after conversion to image array, pixel resampling, image size standardization and intensity normalization. The feature extraction was used with the open-source package *PyRadiomics* (Van Griethuysen et al., 2017). A total of 869 features were obtained utilizing several filters organized by classes: First Order Statistics, 2D Shape-based, Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Gray Level Dependence Matrix (GLDM). All the feature classes but shape were computed both from the original image and derived images from applied

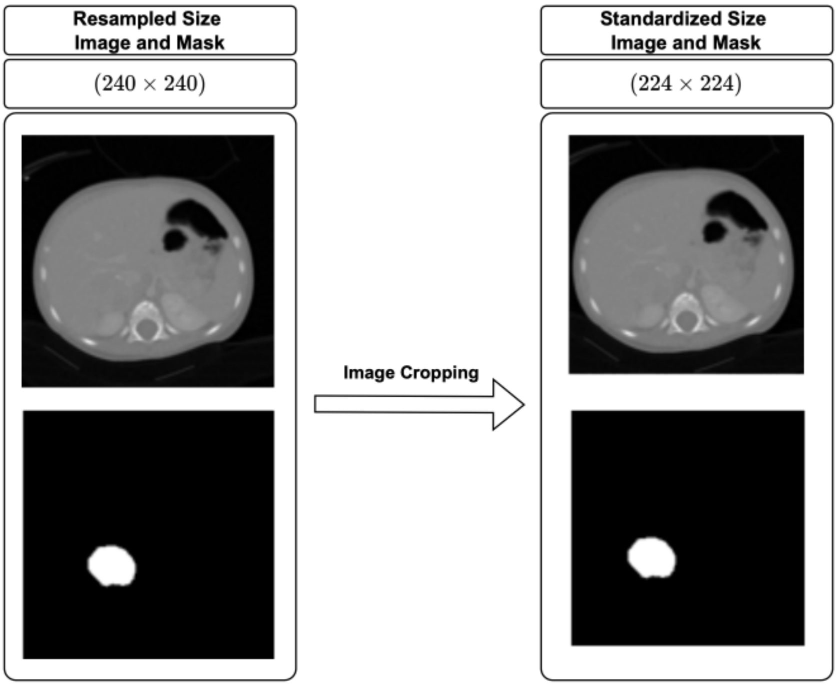

Figure 3.5: Illustration of Image Size Standardization Process for images size higher than 224×224 pixels using cropping methods.

Laplacian-of-Gaussian and Wavelet filters. These features were based exclusively on the region of interest (ROI), in these case concerning the manually delineated tumor masks.

3.2.4 Tabular Data Pre-processing

With the goal of standardize the different types of tabular features to be equally representative in a machine learning model, discrete features were transformed into ranges of values, in order to convert them to categorical features. Then, all categorical features were binarized through the one-hot encoding process. The final number of semantic features was 42.

3.3 Summary

Figure 3.8 summarizes the detailed description of the pre-processing methods utilized for the utilized dataset. From the previous techniques, three types of data become available to use and train machine learning (ML) and deep learning (DL) models: pre-processed CT images, including exclusively or not the ROI; Radiomic features, representative of CT imaging properties, but expressed in tabular data; Semantic features, binarized with one-hot encoding.

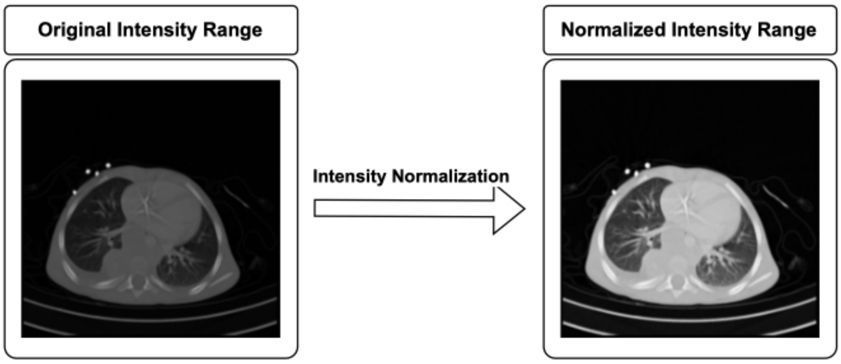

Figure 3.6: Illustration of Intensity Normalization Process. The original intensity range is normalized with a sigmoid function in order to standardize the scale of intensities for the whole dataset.

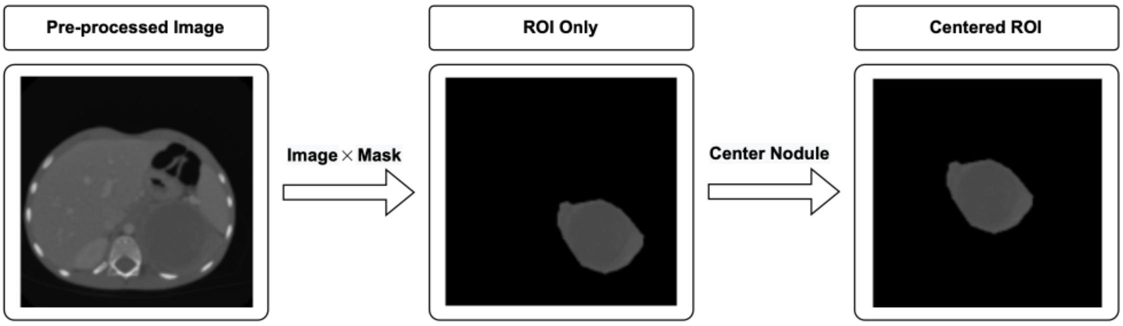

Figure 3.7: Illustration of ROI selection and centralization. The pixel resampled, size adjusted and normalized images are transformed in order to obtain exclusively the tumor, being the remaining pixels changed to zero intensity. Then, the ROI is relocated to the center of the image.

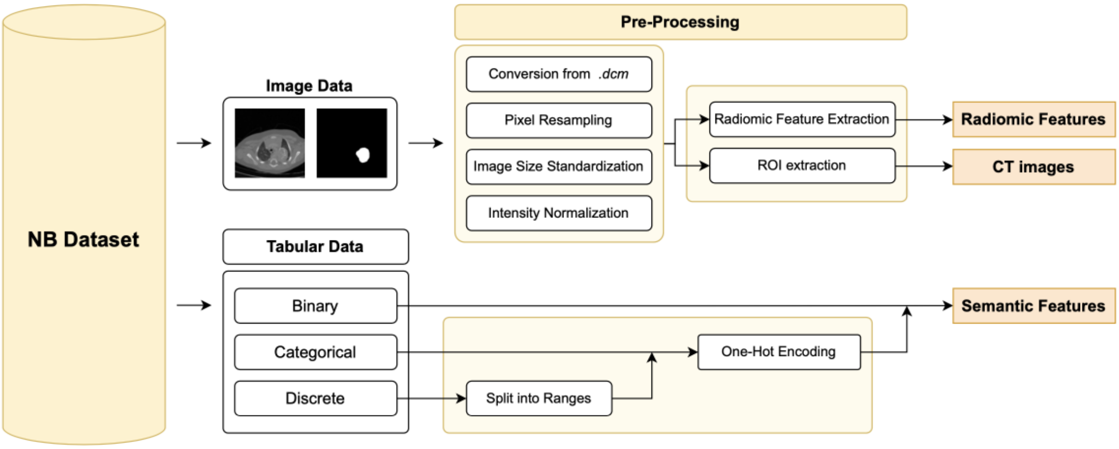

Figure 3.8: Illustration of pre-processing methods on the utilized dataset for the proposed work.

Chapter 4

MNA Status Prediction with ML Approaches

The following chapter illustrates the development of predictive models for MNA status classification that were a target in this study. The chapter is distributed in 5 sections: overall experimental design, followed by all studied approaches for comparability purposes; detection of MNA using radiomic features extracted from CT images and classical ML methods; detection of MNA using clinical semantic data and classical ML methods; detection of MNA using both radiomic and semantic features with a multi-modal approach; detection of MNA using CT images and a DL approach.

4.1 Experimental Design

In order to promote comparability amongst different approach, the experimental design, described in Figure 4.1, was utilized for the following sections: Section 4.2, concerning radiomic features extracted from CT images are utilized as input to traditional ML methods; Section 4.3, concerning semantic features based on patient clinical annotations, utilized as input to traditional ML methods; Section 2.2, concerning the combination of the two previous approaches to implement a multi-modal pipeline; Section 4.5, concerning the use of CT images in a DL-based architecture. Firstly, out of the total of 46 patients and 230 slices, 6 patients and the corresponding 30 slices were separated from the remaining dataset for evaluation purposes. Each approach was target of 10 independent runs, obtaining 10 trained models for each studied algorithm, in order to assess the variability within models from different data splits. The final performance of each approach consisted on the average results of the totality of runs, in order to address the data variance over different train-validation splits when training the models. The test set was utilized to assess the predictive ability of the developed models with various performance metrics.

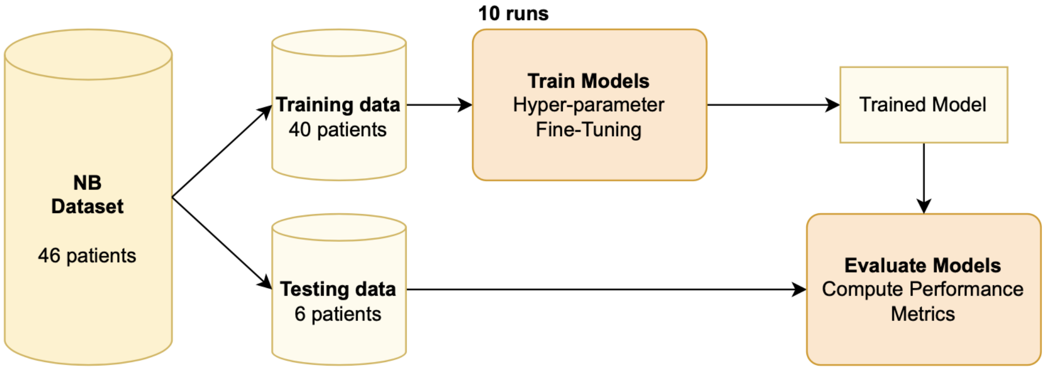

Figure 4.1: Experimental Design utilized to study the several approaches of the proposed work. An initial train-test split was made, being the training samples utilized to train the both by fine-tuning the hyper-parameters of the models. The best models of each approach were selected and tested with the testing data samples through standard performance metrics.

4.1.1 Performance Evaluation

In order to evaluate the performance of each developed model and compare the developed approaches, standard performance metrics were computed with the predictions of the test samples and the correspondent true labels. The results provide objective measures of the predictive ability of the model. The utilized metrics for the performance evaluation stage were: Balanced Accuracy, Area Under the Receiver-Operator Characteristics (AUROC) Curve, F1 score, Recall and Precision.

4.2 Radiomic Approach

Radiomic features extracted from a total of 230 CT slices (5 slices per patient), as described in Chapter 3, Section 3.2.3, were utilized as input for traditional machine learning methods. Several methods were studied regarding data augmentation, feature selection and classification, in order to construct the pipeline with best results.

4.2.1 Materials and Methods

The methods utilized for this approach are illustrated in Figure 4.2. The pipeline utilized the previously pre-processed data, in this case the extracted radiomic features, following the training process and, finally, the evaluation process.

4.2.1.1 Data Augmentation

The utilized radiomic features were extracted from the CT slices of the NB available dataset, after pre-processing the images, through the process describe in Section 3.2.1. However, due to the low amount of samples and the notorious imbalance between classes ($\approx 35\%/65\%$), the data

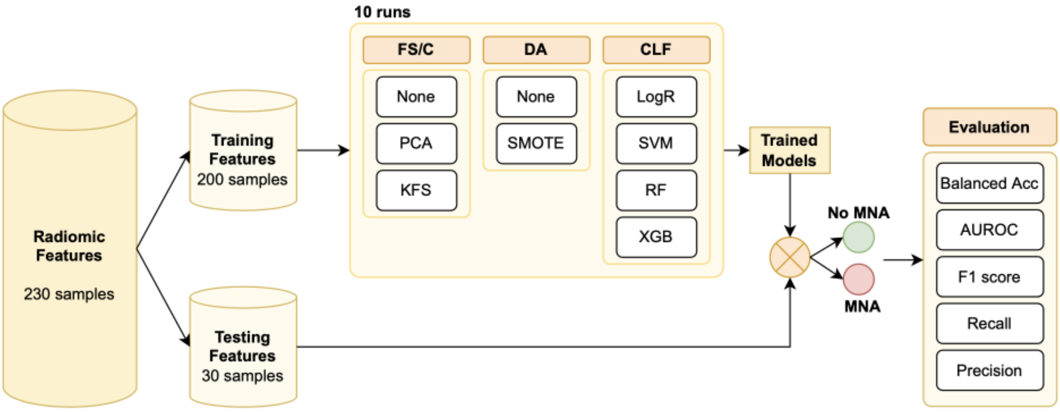

Figure 4.2: Suggested pipeline for Radiomic Approach. The training radiomic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics.

was oversampled using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). The final ratio between classes after creating synthetic samples of the minority class was, approximately, 50/50.

4.2.1.2 Dimensionality Reduction

The number of features that are used to trained the model is directly correlated with computational costs while training the models. Hence, Feature Selection and Feature Construction allow a decrease of data dimensionality, either by excluding redundant or misleading features from the feature space, or by transforming the feature space into a more compact one, respectively. As Feature Selection is concerned, the Koehrsen's Feature Selection (KFS) package (Koehrsen) was used. This open-source toolkit provides several methods for the selection of features, being the applicable utilized for our study. The first step included identifying and removing collinear features, with a correlation threshold of 0.95. Afterwards, a Gradient Boosting Machine (GBM) learning model is utilized to identify and remove both zero and low importance features that do not constitute a cumulative importance of 0.98. In order to reduce variance, the model is run 10 times using early stopping with a validation set in order to avoid overfitting. Regarding Feature Construction methods, Principal Component Analysis (PCA) was utilized since it is widely applied for Feature Selection in ML tasks with datasets containing continuous variables.

4.2.1.3 Classification Models

In order to predict the MNA status with the provided dataset after assessing Data Augmentation and Feature Selection or Feature Construction, four traditional ML algorithms were utilized to obtain the best predictive ability: Logistic Regression (LogR), Support Vector Machines (SVM), Random Forest (RF) and eXtreme Gradient Boosting (XGB). These methods are widely used in classical ML approaches, being proven to show interesting results in literature, as described in Chapter 2, Section 2.1.1. Several hyper-parameters for each algorithm were trained, as listed in Table 4.2.

The hyper-parameters fine-tuning process was achieved through a hyper-parameter search method, *GridSearchCV* (Scikit-learn), which performs an exhaustive search over the parameters of a certain estimator using a 5-fold cross-validation process.

4.2.2 Results and Discussion

In this Section the best hyper-parameters of the trained models as well as the performance results of the studied pipelines are presented and discussed.

4.2.2.1 Classification Results

The average performance results for the four classification models are presented in Table 4.1. Several performance metrics, described in Subsection 4.1.1, were computed for the predictions of the best models in each run, being the final result the mean and standard deviation values. These results were also compared with the baseline results (Pereira et al., 2022), consisting on a Random Forest classifier with an average AUROC of 0.69±0.16.

After analyzing the obtained results, the SVM classifier presents the overall best results in terms of performance, with a AUROC value of 0.84±0.06, a Balanced Accuracy of 0.76±0.04 and a F1 Score of 0.68±0.05. With this performance results, one can consider the developed SVM model as capable of class separation, due to a high AUROC value directly correlated with the ability of distinguishing between positive and negative samples. Furthermore, a high Balanced Accuracy indicates a good ratio of correct predictions whilst addressing the imbalance of the dataset. Despite not presenting the best precision (0.71±0.07) and recall (0.65±0.05) values, its results are satisfactory, providing information regarding true positive ratio over all predictive positives and all labelled positives, respectively. Since the F1 score represents the harmonic mean of both the previous metrics, the SVM classifier presented the higher value in the metric in question (0.68±0.05), being an indicator of a better balance between precision and recall scores.

Regarding Data Augmentation, the utilization of SMOTE to balance the dataset showed different behaviours according to the classifier. Considering Logistic Regression, SMOTE was significant for the improvement of performance, not only in terms of AUROC but also and especially in terms of recall and precision, inherently related to the ratio of true positives. Random Forest also improved with data augmentation techniques, although not showing great improvements in precision, recall and F1 score. On the other hand, the addition of synthetic positive samples to balance the dataset did not increase the predictive ability of the classifiers SVM and XGB.

Concerning now the study on Feature Construction and Feature Selection methods, PCA and KFS, respectively, were used to all classifiers with and without data augmentation as an attempt to improve the results. However, none of the methods caused a significant improvement on the performance results. Furthermore, the results of pipelines including PCA showed a significant

Table 4.1: Performance Results for the studied pipelines. Each approach is evaluated in terms of AUROC, Balanced Accuracy, F1 score, Precision and Recall.

			AUROC	**Balanced Accuracy**	**F1 Score**	**Precision**	**Recall**
Logistic Regression	No SMOTE	No FS	0.51±0.03	0.50±0.00	0.0±0.0	0.0±0.0	0.0±0.0
		PCA	0.42±0.15	0.51±0.03	0.05±0.14	0.05±0.14	0.05±0.15
		KFS	0.50±0.01	0.50±0.01	0.0±0.0	0.0±0.0	0.0±0.0
	SMOTE	No FS	0.69±0.07	0.62±0.06	0.52±0.05	0.47±0.06	0.58±0.04
		PCA	0.69±0.01	0.60±0.01	0.48±0.01	0.46±0.02	0.50±0.0
		KFS	0.40±0.02	0.35±0.01	0.32±0.01	0.24±0.01	0.50±0.0
SVM	No SMOTE	No FS	**0.84±0.06**	**0.76±0.04**	**0.68±0.05**	0.71±0.07	0.65±0.05
		PCA	0.63±0.08	0.53±0.07	0.38±0.07	0.40±0.10	0.39±0.10
		KFS	0.67±0.06	0.50±0.06	0.41±0.05	0.34±0.05	0.53±0.05
	SMOTE	No FS	0.77±0.07	0.72±0.04	0.62±0.05	0.66±0.09	0.59±0.05
		PCA	0.73±0.23	0.70±0.22	0.63±0.27	0.61±0.30	**0.67±0.23**
		KFS	0.63±0.12	0.49±0.06	0.39±0.04	0.33±0.05	0.50±0.0
Random Forest	No SMOTE	No FS	0.50±0.02	0.51±0.07	0.14±0.18	0.26±0.38	0.10±0.12
		PCA	0.56±0.06	0.50±0.04	0.13±0.18	0.13±0.17	0.14±0.21
		KFS	0.54±0.05	0.58±0.05	0.30±0.12	**0.88±0.24**	0.19±0.08
	SMOTE	No FS	0.64±0.06	0.52±0.04	0.13±0.12	0.48±0.42	0.08±0.07
		PCA	0.72±0.04	0.70±0.08	0.56±0.14	0.83±0.18	0.45±0.15
		KFS	0.64±0.03	0.61±0.03	0.41±0.09	0.67±0.08	0.31±0.10
XGB	No SMOTE	No FS	0.61±0.0	0.68±0.0	0.56±0.0	0.63±0.0	0.5±0.0
		PCA	0.62±0.03	0.66±0.01	0.54±0.01	0.58±0.03	0.50±0.0
		KFS	0.64±0.0	0.73±0.0	0.63±0.0	0.83±0.0	0.50±0.0
	SMOTE	No FS	0.61±0.0	0.68±0.0	0.56±0.0	0.63±0.0	0.5±0.0
		PCA	0.59±0.02	0.59±0.03	0.49±0.03	0.42±0.04	0.59±0.03
		KFS	0.78±0.0	0.68±0.0	0.56±0.0	0.63±0.0	0.5±0.0

increase in terms of standard deviation over the 10 runs, indicating a certain instability in the capacity of the models to detect MNA status. This can be related to the stochastic component of the PCA technique, due to its unsupervised character and also for not maintaining the original features, but creating new ones. Regarding KFS, its utilization did not show great improvements in terms of performance, with the exception of Random Forest models, in which both KFS and PCA influenced positively the performance of this classifier.

Finally, several studied approaches were able to overcome the baseline performance of AUROC (0.69±0.16). Despite using the same dataset, there was an investment in data preparation and pre-processing (Section 3) in order to provide standardized CT images for the radiomic feature extraction. Furthermore, additional data augmentation and dimensionality reduction techniques were studied, as well as other classical ML classifiers.

Overall, the SVM classifier presented the best behaviour amongst the studied classifiers, not improving performance with Data Augmentation nor Feature Selection or Construction. However, this techniques did increase the predictive ability of other classifiers, namely the utilization of SMOTE with Logistic Regression, PCA with Random Forest and KFS with XGB.

4.2.2.2 Best Hyper-parameters

As previously described in Subsection 4.2.1, for each studied algorithm, a search method was used to perform a cross-validated grid search over a range of each parameter and find the best hyper-parameters for each classifier. The best models of each run were selected to make predictions of the test samples and evaluate their performance. Table 4.2 illustrates the hyper-parameter fine-tuning process of this approach, analyzing both the most frequent set of hyper-parameters and the best estimator over the 10 different runs, by assessing the AUROC performance. For the XGB classifier, the range of several parameters was continuous and, therefore, obtaining a great variety of best values. Thus, the best hyper-parameters for this algorithm were chosen exclusively according to a set of values that achieved the top AUROC.

After analyzing the best and most frequent hyper-parameter values, it is noticeable that the set of parameters with best achieving performance is often different from the most frequent values, which is an indicator that the best model is highly variable according to the random train-validation splits used by the search algorithm. This may be a direct consequence of the low amount of available data samples and its inherent high variability, which leads to different models with different training sets. However, the XGB classifier is able to present the top performance for the classifier over different runs despite the variability of chosen hyper-parameters, being able to maintain a constant predictive ability.

Furthermore, the definition of hyper-parameters is a challenging task, due to its specificity towards the task and the dataset itself. The used fine-tuning approach addresses the challenge by studying several ranges of parameters for several runs, in order to understand which subset of values is able to achieve the best performance.

Table 4.2: Detailed description concerning the hyper-parameter tuning process for the machine learning models. Best values relate to the set of hyper-parameters that achieved the highest performance with the test set, whilst most frequent values refer to the set of hyper-parameters that were chosen by the search method more frequently.

		Possible Values	Best Values						Most Frequent Values					
			No SMOTE			SMOTE			No SMOTE			SMOTE		
			No FS	PCA	KFS	No FS	PCA	KFS	No FS	PCA	KFS	No FS	PCA	KFS
LogR	C	0.01, 0.1, 1, 10	0.01	10	0.01	10	10	1	0.01	0.01	0.1	10	10	10
	penalty	L1, L2	L2	L1	L1	L1	L2	L2	L1	L1	L2	L1	L1	L2
SVM	C	0.01, 0.1, 1, 10, 100	100	0,1	100	100	10	100	100	1	100	100	100	100
	kernel	linear, poly, rbf	rbf	poly	rbf	rbf	rbf	rbf	rbf	poly	rbf	linear	rbf	rbf
RF	max_depth	5, 20, 50, 100, 250, None	5	250	None	5	None	50	250	100	None	100	100	50
	max_leaf_nodes	3, 4, 5, None	None	5	3	None	5	None	3	4	3	None	None	None
	n_estimators	50, 100, 250, 500	100	250	500	50	250	250	500	500	500	50	500	500
XGB	colsample_bytree	0.3 : 1	0.92	0.43	0.63	0.57	0.66	0.98						
	gamma	0.5 : 2.5	1.83	1.68	2.37	1.43	0.79	1.65						
	learning rate	0.03 : 0.33	0.25	0.32	0.12	0.07	0.26	0.30						
	max_depth	2 : 6	2	4	5	4	5	3						
	min_child_weight	1 : 5	1	2	3	2	2	2						
	n_estimators	100 : 1100	814	340	751	349	844	646						
	subsample	0.4 : 1	0.75	0.87	0.87	0.87	0.79	0.96						

4.2.3 Limitations

The proposed work using the radiomic features to predict MNA studied several state-of-the-art pipelines utilizing traditional ML methods, and including Data Augmentation, Feature Selection, and Feature Construction methods to help improve the performance of the trained models. However, this study still presented some limitations, being possible to further understand and perhaps improve the obtained results. Firstly, a more exhaustive fine-tuning process not only regarding the classification hyper-parameters, but also the parameters of PCA and KFS, could lead to performance improvements since it would reflect on a more personalized fit to the dataset and the classification task. Furthermore, the increase of running trials would provide further information regarding the stability of models and the variability within the training samples.

4.3 Semantic Approach

Semantic approach consisted on utilizing semantic features, previously pre-processed in Chapter 3, Section 3.2.4, and utilize the same classical machine learning methods than in the previous approach (Section 4.2), as well as data augmentation and feature selection and construction methods. Contrarily to the Radiomic Data, Semantic Data concerns patients clinical information and not slice-related annotations. Therefore, the semantic features have 46 samples, one per patient, constituting an even greater challenge to obtain satisfactory and reliable results with such a small amount of data samples.

4.3.1 Materials and Methods

With the purpose of building similar pipelines for both approaches using tabular data, the same methods were utilized for the semantic data. Figure 4.3 illustrates the procedure for training semantic models. The input of the pipeline consisted on the previously pre-processed clinical annotations provided by experts, all transformed in order to be categorical and binarized.

4.3.1.1 Data Augmentation

As mentioned in the previous approach, a data augmentation technique was studied in order to balance the classes in the dataset. Since semantic features are categorical and not continuous such as radiomic features, SMOTEN, a different version of SMOTE, was utilized.

4.3.1.2 Dimensionality Reduction

Comparatively to the radiomic pipeline, PCA and KFS were utilized as Feature Construction and Fetaure Selection methods to reduce the number of features. Several number of components were studied with PCA, and KFS consisted solely in utilizing a correlation threshold of 0.80 to remove collinear features, due to the significantly low number of semantic features and its categorical nature.

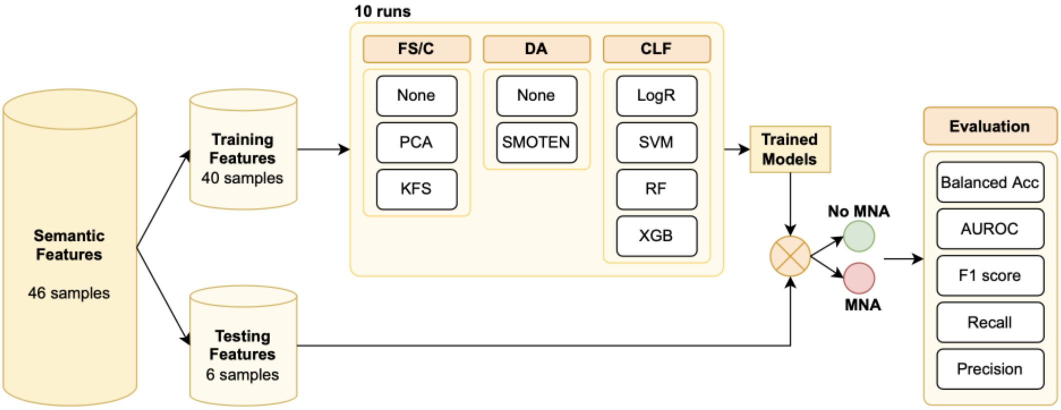

Figure 4.3: Suggested pipeline for Semantic Approach. The training semantic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics.

4.3.1.3 Classification Models

The same classification algorithms were studied for comparison purposes. These algorithms are described in Subsection 4.2.1.3, including the utilized hyper-parameters and the fine-tuning process.

4.3.2 Results and Discussion

In this Section the best hyper-parameters of the trained models as well as the performance results of the studied pipelines are presented and discussed.

4.3.2.1 Classification Results

Similarly to the previous approach, the classification performance results from the 10 runs of the pipeline were averaged in order to have a final performance value for each methodology.

After an overview of the results, it is noticeable that, due to the small number of test samples, a certain instability in the results is shown, being several results equal to the minimum or maximum possible values, and some displaying high standard deviation values. Furthermore, despite generally high AUROC and Balanced Accuracy values, F1 score, Precision and Recall display lower values overall. In this case, these last metrics are descriptive of the true positive ratio, which for medical use-cases is extremely important since positive patients must not be misdiagnosed. In MNA detection, if a classifier presented a precision and recall of 0.00 and, consequently, a F1 score of 0.00, it indicates that all MNA positive cases were misdiagnosed and, therefore, all test samples were classified as MNA negative. In this case, the model showed no separability between classes, not serving the task in question by mislabelling the whole minority class.

Table 4.3: Performance Results for the studied pipelines regarding the Semantic Approach. Each approach is evaluated in terms of AUROC, Balanced Accuracy, F1 score, Precision and Recall.

			AUROC	**Balanced Accuracy**	**F1 Score**	**Precision**	**Recall**
Logistic Regression	No SMOTEN	No FS	1.00±0.00	0.65±0.20	0.33±0.42	0.40±0.49	0.30±0.40
		PCA	1.00±0.00	0.58±0.11	0.20±0.31	0.30±0.46	0.15±0.23
		KFS	1.0±0.00	0.58±0.11	0.20±0.031	0.30±0.46	0.15±0.23
	SMOTEN	No FS	1.00±0.00	0.98±0.05	0.96±0.08	0.93±0.13	1.00±0.00
		PCA	0.56±0.29	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		KFS	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
SVM	No SMOTEN	No FS	0.19±0.38	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		PCA	0.49±0.35	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		KFS	0.50±0.43	0.48±0.05	0.00±0.00	0.00±0.00	0.00±0.00
	SMOTEN	No FS	0.84±0.11	0.53±0.05	0.10±0.20	0.10±0.20	0.10±0.20
		PCA	0.85±0.19	0.63±0.13	0.33±0.33	0.50±0.50	0.25±0.25
		KFS	0.75±0.38	0.63±0.17	0.35±0.37	0.43±0.47	0.35±0.39
Random Forest	No SMOTEN	No FS	1.00±0.00	0.53±0.08	0.07±0.20	0.10±0.30	0.05±0.15
		PCA	0.93±0.06	0.55±0.10	0.13±0.27	0.20±0.40	0.10±0.20
		KFS	1.0±0.00	0.68±0.11	0.47±0.31	0.70±0.46	0.35±0.23
	SMOTEN	No FS	1.00±0.00	0.80±0.19	0.67±0.37	0.80±0.40	0.60±0.37
		PCA	0.94±0.06	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		KFS	1.00±0.00	0.75±0.00	0.67±0.00	1.00±0.00	0.50±0.00
XGB	**No SMOTEN**	No FS	1.00±0.00	0.50±0.00	0.00±0.00	0.00±0.00	0.00±0.00
		PCA	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
		KFS	0.88±0.00	0.75±0.00	0.67±0.00	1.00±0.00	0.50±0.00
	SMOTEN	No FS	1.00±0.00	0.75±0.00	0.67±0.00	1.00±0.00	0.50±0.00
		PCA	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
		KFS	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**

Notwithstanding the good performance of several used pipelines, XGB was the classifier that presented a more stable performance not only over different runs, but regardless of data augmentation and dimensionality reduction techniques, predicting correctly all test samples using PCA without SMOTEN, PCA with SMOTEN, and KFS with SMOTEN. Logistic Regression using KFS as Feature Selection and SMOTEN for Data Augmentation was also able to accurately predict all test samples.

Regarding Data Augmentation, SMOTEN improved the predictive ability of several classifiers, especially in terms of F1-score, Precision and Recall. Models with dimensionality reduction also showed an overall higher predictive ability, although not as significant as with data augmentation.

4.3.2.2 Best Hyper-parameters

After performing a grid-search technique to search the best model for each classifier over 10 different runs, the hyper-parameters of the chosen models were analysed in order to assess the combination with best results. Contrarily to the previous approach, the best models over the 10 runs were assessed exclusively taking into account the most common grid of parameters per classifier. The change in methodology was due to the decrease of the total number of samples, since the semantic information concerns each patient (46 samples) and not each slice (230 samples).

Table 4.4: Detailed description concerning the hyper-parameter tuning process for the machine learning models. Best values relate to the set of hyper-parameters that achieved the highest performance with the test set, whilst most frequent values refer to the set of hyper-parameters that were chosen by the search method more frequently.

			Most Frequent Values					
		Possible Values	No SMOTE			SMOTE		
			No FS	PCA	KFS	No FS	PCA	KFS
LogR	C	0.01, 0.1, 1, 10	0.01	0.01	0.1	1	0.01	1
	penalty	L1, L2	L2	L2	L2	L1	L2	L1
SVM	C	0.01, 0.1, 1, 10, 100	0.1	0.01	100	1	0.1	10
	kernel	linear, poly, rbf	linear	linear	poly	linear	linear	poly
RF	max_depth	5, 20, 50, 100, 250, None	250	None	None	5	250	250
	max_leaf_nodes	3, 4, 5, None	3	4	5	None	5	None
	n_estimators	50, 100, 250, 500	100	100	100	50	500	50
XGB	colsample_bytree	0.3 : 1						
	gamma	0.5 : 2.5						
	learning rate	0.03 : 0.33						
	max_depth	2 : 6						
	min_child_weight	1 : 5						
	n_estimators	100 : 1100						
	subsample	0.4 : 1						

After analyzing the most common set of hyper-parameters for each approach, there were some patterns. Firstly, regarding Logistic Regression, a L2 regularization was constant for several approaches, whilst varying the C parameter. In SVM pipelines, linear kernels also were most often chosen as best hyper-parameters indicating that a linear solution was more suitable for most approaches. In RF models, the hyper-parameters presented a higher variability, also due to the fact of having a larger range of possible values. XGB not only presented several grids of best hyper-parameters, since most range of values were continuous. Furthermore, in all runs, XGB models were able to achieve the same performance, indicating that all chosen hyper-parameters were able to equally predict the test samples.

Once again, it is important to reinforce that the definition of the best set of parameters for a certain model is not a standard process, displaying an open solution space with several possible methodologies.

4.3.3 Limitations

The proposed study on predicting MNA status utilizing semantic features was done using the same traditional ML pipelines as in the previous approach, described in section 4.2. Data augmentation techniques showed significant improvements in the predictive ability of the models. However, the size of the dataset directly influences not only the performance of the developed models, as well as the reliability in its interpretation, since the number of test samples is too low to draw consistent conclusions.

4.4 Multi-modal Approach

After studying both the radiomic and semantic approaches, a multi-modal approach is proposed in order to build a classifier with higher performance results than the radiomic approach and more robust predictive ability than the semantic approach. Therefore, several ensemble methods are proposed in order to combine both sets of features in order to obtain a stronger learner. This work follows previous work (Malafaia et al., 2021), utilizing the same methods but with adaptation towards the task in question.

4.4.1 Materials and Methods

To follow the previous work (Malafaia et al., 2021), 5 different approaches were studied to combine radiomic and semantic features to predict MNA status. This approaches were previously described in Chapter 2, Section 2.2, and are summarized in Figure 4.4.

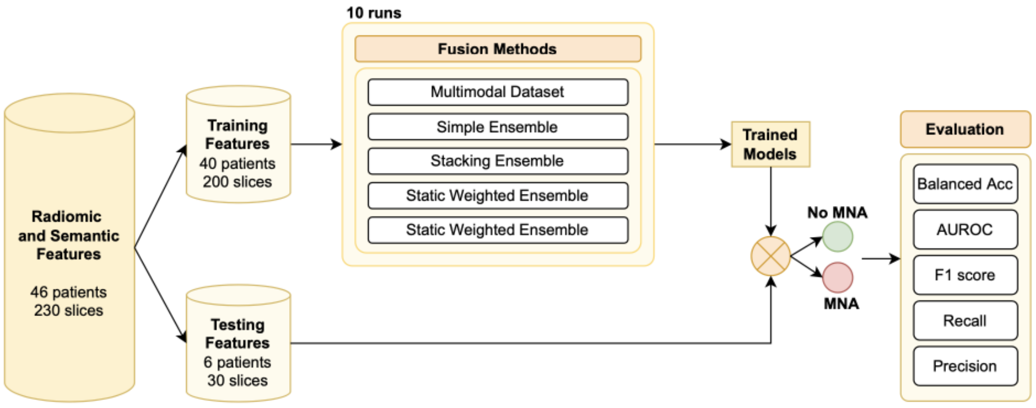

Figure 4.4: Suggested pipeline for the Multi-modal Approach. The training semantic features are utilized to train several models, with various Feature Selection/Construction (FS/C), Data Augmentation (DA) and Classification (CLF) methods. The trained models are utilized to get MNA predictions for the test samples, which are then used for computing the Evaluation metrics.

4.4.1.1 Fusion Methods

The 5 studied predictive ensemble techniques were used as an attempt to explore the most common solutions to combine several modalities of data and build a more robust model to predict MNA status. This was achieved through the combination of two learners - Radiomic Model and Semantic Model - with the same samples but different types of information, being able to use the whole dataset into one approach. The utilized approaches were the following:

- *Multimodal Dataset:* Both semantic and radiomic features were concatenated in order to represent a single, multi-model, dataset. This data was utilized to train models using the same techniques as in the two previous approaches, described in Sections 4.2 and 4.3. The best model was selected and then compared to the remaining fusion approaches.

- *Simple Ensemble:* A cascade pipeline was used where the semantic model was trained, being its predictions given as a feature along with the radiomic features. The utilized classifiers were the best classifiers in the previous approaches: SVM without Data Augmentation and Dimensionality Reduction for the Radiomic Model; XGB with SMOTE and KFS for the Semantic Model.

- *Stacking Ensemble:* A simple linear meta-model, in this case Logistic Regression, is trained, giving the predictions of the semantic and radiomic models as features. The meta-model returns the final probability of the MNA status.

- *Static Weighted Ensemble:* In order to obtain the final MNA status prediction, a static weight is given to the predictions of each model. Both radiomic and semantic models were previously trained. The weight of each prediction was manually adjusted.

- *Dynamic Weighted Ensemble:* This approach was developed as an attempt of improving the performance of the model by dynamically assigning a weight to each prediction, according to the trust of the model in the result. Thus, a confidence value was computed, being defined as the absolute difference between the probability of a sample being positive and the value 0.50. The higher this value, the less confidence the learner has in the prediction, being the weight lower. This parameter was then utilized to perform a weighted average of both predictions with respect to the ratio of confidences between both sub-models. The described algorithm followed Equation 4.1, in which d_{sem} and d_{rad} refer to the confidence values of each sub-model, and α is a manually adjusted constant.

$$y_{prob} = \alpha \frac{d_{sem}}{d_{rad}} y_{sem} + (1-\alpha) \frac{d_{sem}}{d_{rad}} y_{rad} \tag{4.1}$$

4.4.2 Results and Discussion

4.4.2.1 Classification Results

The performance results of this stage of the purpose work were split into to portions: the first one, where a procedure very similar to the previous approaches is used, but with a multi-modal dataset of the concatenated radiomic and semantic features; the second one, where the best classifier from the previous phase is utilized to train the multi-modal dataset, and all the five fusion methods are train over 10 runs along with the radiomic and semantic models, as baseline.

In Table 4.5, the average and standard deviation results of the first stage of this approach are presented. When utilizing the multi-modal dataset, Logistic Regression is the classifier with an overall best performance, presenting an average AUROC of 0.98±0.01 with Data Augmentation and Feature Construction methods. For those reasons, this was the classifiers chosen for the next stage. It is noticeable that SMOTENC, a version of SMOTE for both continuous and categorical data, improves the performance results of most classifiers.

Table 4.5: Performance Results for the studied classifiers with the Multi-modal Dataset.

			AUROC	**Balanced Accuracy**	**F1 score**	**Precision**	**Recall**
Logistic Regression	no SMOTENC	No FS	0.47±0.03	0.51±0.02	0.04±0.07	0.20±0.40	0.02±0.04
		PCA	0.44±0.05	0.49±0.05	0.05±0.08	0.30±0.46	0.03±0.09
		KFS	0.53±0.11	0.50±0.000	0.08±0.16	0.07±0.13	0.10±0.20
	SMOTENC	No FS	0.82±0.01	**0.80±0.02**	**0.73±0.02**	0.67±0.05	0.81±0.00
		PCA	**0.98±0.01**	0.75±0.01	0.66±0.01	0.50±0.01	**1.00±0.00**
		KFS	0.73±0.01	0.64±0.02	0.51±0.02	0.53±0.05	0.50±0.00
SVM	No SMOTENC	No FS	0.61±0.15	0.59±0.10	0.48±0.13	0.39±0.07	0.65±0.27
		PCA	0.55±0.08	0.59±0.04	0.40±0.09	0.54±0.13	0.37±0.14
		KFS	0.43±0.09	0.36±0.06	0.33±0.03	0.25±0.03	0.50±0.00
	SMOTENC	No FS	0.57±0.16	0.62±0.07	0.49±0.11	0.46±0.05	0.61±0.27
		PCA	0.50±0.21	0.55±0.08	0.32±0.17	0.38±0.06	0.36±0.34
		KFS	0.30±0.15	0.36±0.06	0.28±0.07	0.22±0.05	0.39±0.11
Random Forest	No SMOTENC	No FS	0.56±0.01	0.57±0.02	0.23±0.07	**1.00±0.00**	0.13±0.05
		PCA	0.80±0.09	0.53±0.05	0.14±0.20	0.28±0.39	0.16±0.27
		KFS	0.43±0.05	0.53±0.12	0.39±0.13	0.39±0.16	0.40±0.10
	SMOTENC	No FS	0.55±0.03	0.58±0.05	0.32±0.13	0.62±0.09	0.23±0.12
		PCA	0.76±0.10	0.74±0.05	0.65±0.08	0.81±0.16	0.55±0.07
		KFS	0.56±0.06	0.65±0.08	0.54±0.08	0.62±0.16	0.50±0.00
XGB	No SMOTENC	No FS	0.43±0.00	0.48±0.00	0.30±0.00	0.30±0.00	0.30±0.00
		PCA	0.67±0.01	0.72±0.01	0.62±0.01	0.81±0.05	0.50±0.00
		KFS	0.49±0.00	0.48±0.00	0.38±0.00	0.31±0.00	0.50±0.00
	SMOTENC	No FS	0.50±0.00	0.50±0.00	0.32±0.00	0.33±0.00	0.30±0.00
		PCA	0.83±0.01	0.65±0.00	0.53±0.00	0.56±0.00	0.50±0.00
		KFS	0.43±0.00	0.45±0.00	0.37±0.00	0.29±0.00	0.50±0.00

Table 4.6: Performance Results for the Multimodal Approach. Radiomic and Semantic Models were also trained in the same 10 runs for baseline comparison purposes.

Fusion Method	**AUROC**	**Balanced Accuracy**	**F1 score**	**Precision**	**Recall**
Radiomic Only	0.78±0.06	0.70±0.10	0.63±0.08	0.64±0.08	0.60±0.07
Semantic Only	1.00±0.00	0.75±0.00	0.67±0.00	1.00±0.00	0.50±0.00
Multimodal Dataset	**0.98±0.01**	**0.74±0.02**	**0.66±0.01**	**0.49±0.02**	**1.00±0.00**
Simple Ensemble	0.76±0.07	0.68±0.04	0.58±0.06	0.65±0.10	0.53±0.05
Stacking Ensemble	0.63±0.10	0.53±0.06	0.13±0.021	0.15±0.24	0.12±0.20
Static Average	0.53±0.05	0.74±0.03	0.65±0.06	1.00±0.00	0.48±0.06
Dynamic Average	0.51±0.01	0.74±0.03	0.65±0.06	1.00±0.00	0.48±0.06

After selecting the best pipeline for the multi-modal dataset, the 5 ensemble approaches were trained over the same 10 runs for comparability purposes. Although none of the methods being able to over-perform the semantic model, the multi-modal dataset approach was able to reach a close performance from the baseline, obtaining a higher recall value, which inherently means that all true positives were correctly predicted. Furthermore, the semantic model is evaluated with 6 test samples, one per patient, whilst the multi-modal dataset includes information per patient and per slice, counting with 30 test samples. Regarding the remaining fusion methods, once can state a general decrease of performance. It is possible to correlate the discrepancy between the two types of combined models and their performance with this decrease.

4.4.3 Limitations

After performing an exploratory study on the most common ML methods for combining different feature modalities, the approach of simply joining all features and use traditional ML methods seemed to have shown the best performance results. However, when combining two learners with different types of information, one can infer that if the predictive ability of one is significantly lower than the other, the combination of the two models may arise in a decrease of overall performance, as noticeable in the majority of the utilized methods. This suggests that the predictive ability of the radiomic approach has room for improvement and, once its performance is increase with a better pipeline, the multi-modal results may improve significantly.

Furthermore, the Dynamic Weighted Ensemble method brings an interesting perspective on how the models should be combined, by dynamically giving more importance to the learner that, at a specific test sample, showed more confidence in the prediction. It may be interesting to further develop this last methodology and study whether its possible to increase its predictive ability.

Nonetheless, a multi-modal approach is a target of great interest for medical applications, since the clinical protocol allows for the availability of several different types of data that should be considered for the classification task.

4.5 Deep Learning Approach

The previous sections focused on utilized traditional ML methods to assess the correlation between imaging phenotypes and/or patient clinical information with the prediction of MNA status, showing a high correlation and a great predictive ability with the developed models. As a final approach, a simple Deep Learning (DL) approach was developed, where transfer learning is utilized to build a CNN network that receives as input CT slices and predicts the MNA status. Moreover, a robustness analysis is performed on the developed architecture using as tool XAI methods. The used experimental design follows Figure 4.1 from Section 4.1, in order for the results to be comparable to the previous approaches. The proposed study is based on previous work (Malafaia et al., under peer-review), following the utilized methods.

4.5.1 Materials and Methods

Methods utilized for the proposed approach are described in Figure 4.5. In this section the utilized data is described, as well as the MNA status classification model and the explainable methods. Each portion of the pipeline is detailed in the following subsections.

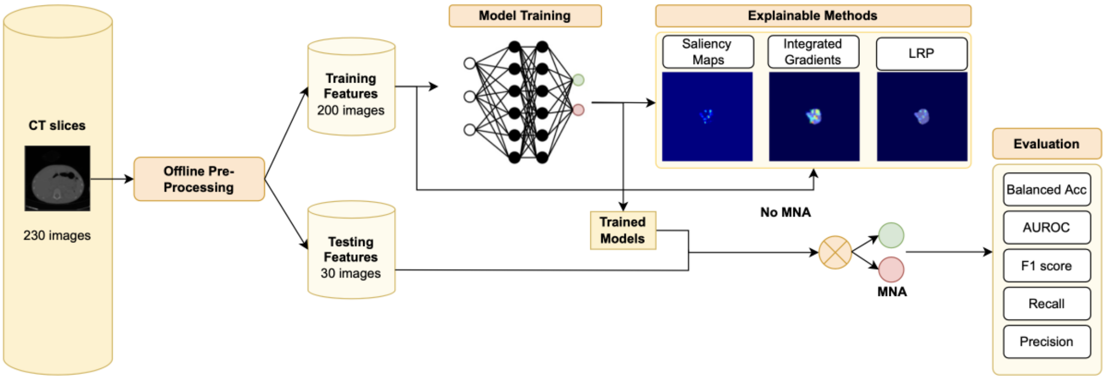

Figure 4.5: Overview of the developed pipeline for the DL approach.

4.5.1.1 Data Preparation

As previously described in Chapter 3.2.1, the input for the model consisted on pre-processed CT slices. In order to assess which image input would get better results, 3 types of images, exemplified in Figure 4.6 were attempted: total CT slice; CT slice just containing the manually segmented tumor (ROI); CT slice with the tumor centered in the image (centered ROI). Within the three inputs, the model is provided different information: with the whole input image, the network has pixel data exterior to the tumor; the ROI will exclusively provide the pixels related to the tumor, which may allow the network to focus on the most relevant region of the image; the centered ROI may facilitate the learning process of the network, by being constant the center of mass location of the tumor and, thus, providing spacial features relevant for the classification.

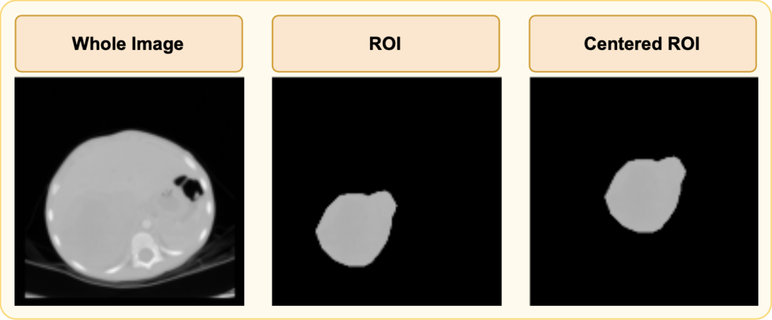

Figure 4.6: Attempted input images for the DL Approach.

For data storage and image manipulation purposes, the pre-processing for this approach was made online, transforming the images within the pipeline. Data Augmentation was used in order to increase variability within the dataset, namely Random Horizontal Flips, Gaussian Blurs and Random Rotations.

4.5.1.2 Classification Architecture

The proposed DL architecture consisted on a ResNet50 (He et al., 2016) pre-trained with the ImageNet Dataset (Deng et al., 2009). Several other pre-trained Convolutional Neural Networks (CNNs) were attempted, being the one with higher perfomance results utilized.

Several hyper-parameters were tuned, namely the number of frozen layer, the learning rate, the batch size and the optimizer. A weighted loss was used to address the data imbalance.

This architecture was utilized for 10 different runs, arising 10 models, once again to assess the variability of the trained models when given different data splits. The developed models were assessed in terms of performance with the metrics utilized in the previous sections.

4.5.1.3 Explainable Methods

Visual explanations are easily understandable by humans, since its possible to analyze which regions of the input image contributed the most for the prediction made by the classifier. For this study, 3 gradient-based methods were used to generate visual explanations of the test samples over the 10 runs. This explanations consisted on heatmaps that translate the score of each pixel for the final classification.

The utilized methods were: Saliency Maps (Simonyan et al., 2013), Integrated Gradients (Sundararajan et al., 2017) and Layer-wise Relevance Propagation (LRP) (Bach et al., 2015). The latter constitute off-the-shelf post-hoc methods that can be found in open-source libraries such us *Pytorch Captum* (Kokhlikyan et al., 2020). To illustrate the expected output of these methods, Figure 4.7 illustrates a representation of the prediction of 2 images from ImageNet belonging to two different classes and predicted by a VGG16 pre-trained network (Pal, 2016). Additionally, a prediction of a lung nodule image from the LIDC dataset (Armato III et al., 2011) with a lung classifier is also demonstrated, as a more similar example to our use-case.

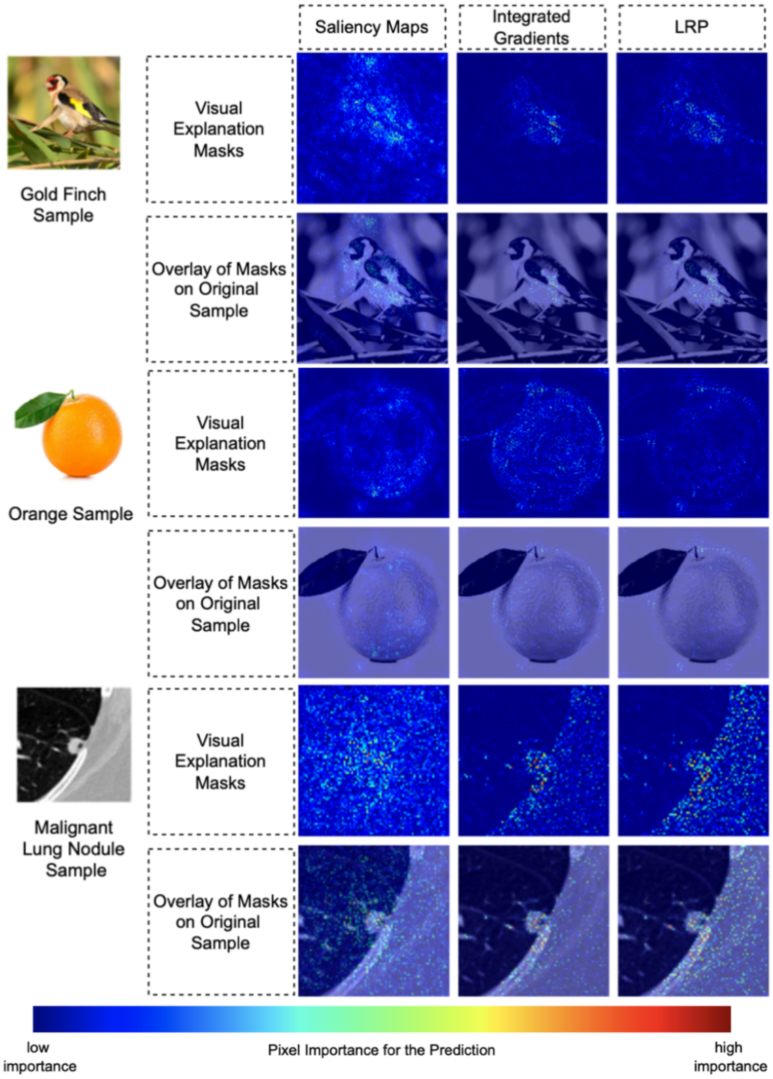
Saliency Maps
Integrated Gradients
LRP
Gold Finch Sample
Visual Explanation Masks
Overlay of Masks on Original Sample
Orange Sample
Visual Explanation Masks
Overlay of Masks on Original Sample
Malignant Lung Nodule Sample
Visual Explanation Masks
Overlay of Masks on Original Sample
low importance
Pixel Importance for the Prediction
high importance

Figure 4.7: Demonstration of expected results when utilizing the described XAI methods using three different test samples. The first sample is classified as "goldfinch" and the second one is labeled as "orange", being predicted by the pre-trained VGG16 network. The last sample belongs to the LIDC dataset, being predicted by a lung cancer classifier. For each image, the explanations generated by the 3 XAI methods in question are displayed, as well as an overlay of the explanation on the original image.

4.5.1.4 Robustness Assessment

The robustness assessment stage of this stage consisted on a comparison amongst the explanations from the same test sample and XAI method over different runs, with the aim of assessing the similarity between explanations generated from different predictive models for the same input. The more similar the explanations from the sample sample, the more consistent the proposed architecture was over different runs, since the generated predictive models would attribute higher importance scores to the same image regions.

In order to asses the coherence of the network for the same test sample over different train-validation data splits, the heatmap explanations were compared two by two according to quantitative image similarity metrics. Quality Image Assessment Techniques (IQA), namely full-reference metrics, allow the comparison between two images, being utilized traditional methods, such as Root Mean Squared Error (RMSE), as well as methods that attempt to approximate the perceptual quality assessment of the Human Visual System (HVS), namely the Structural SIMilarity (SSIM) index and the Multi-Scale Structural SIMilarity (MS-SSIM) index. Euclidean distance was also used due to its popularity in image similarity metrics.

4.5.2 Results and Discussion

4.5.2.1 Best Architecture

Several parameters and changes were made to the pipeline in order to achieve the best performance, which was evaluated with the AUROC value.

Firstly, regarding the input of the network, the three types of images were tested, being the performance better when the input was the centered ROI. A comparison between the Loss and Performance curves over the number of epochs with the whole image and the centered ROI is presented in Figure 4.8, where the stability of the model is clearly higher when using just the centered ROI instead of the whole image. When using the whole image as input, the model starts overfitting around epoch number 50. On the other hand, when giving exclusively the centered ROI as input, the validation loss acocompanies the train losse decrease, in this case until around epoch 170. Moreover, when feeding the whole image, both the AUROC and the Balanced Accuracy values are very unstable, not improving as the model is training. When giving the centered ROI as input to the network, the performance metrics increase with the number of epochs.

After choosing the most appropriate input for the model, the hyper-parameters were selected for the pre-trained ResNet50 according to the AUROC value computed with the test set over the 10 runs. The best learning rate was 0.00001, with a batch size of 32. The maximum number of epochs was set to 300, with an early stopping criteria of not improving the performance over 30 consecutive epochs. The chosen optimizer was Adam, and a Weighted Cross Entropy Loss was utilized. Furthermore, an adaptive layer defreezing was implemented to the pre-trained network, being the last convolutional layer defrozen at epoch number 80, and the previous to the last one at epoch number 150.

4.5.2.2 Classification Results

After the selection of the best hyper-parameters for the task, 10 models were trained and tested, being the performance computed, as presented in Table 4.7.

The obtained performance with this approach was able to reach high performances, achieving a AUROC value of 0.94±0.04 in the test set. The Recall value of 0.98±0.04 indicates a very high ratio of true positives over the total existent positives, indicating that the model accurately detects the MNA positive samples. The Precision, however, is lower than expected, indicating that some

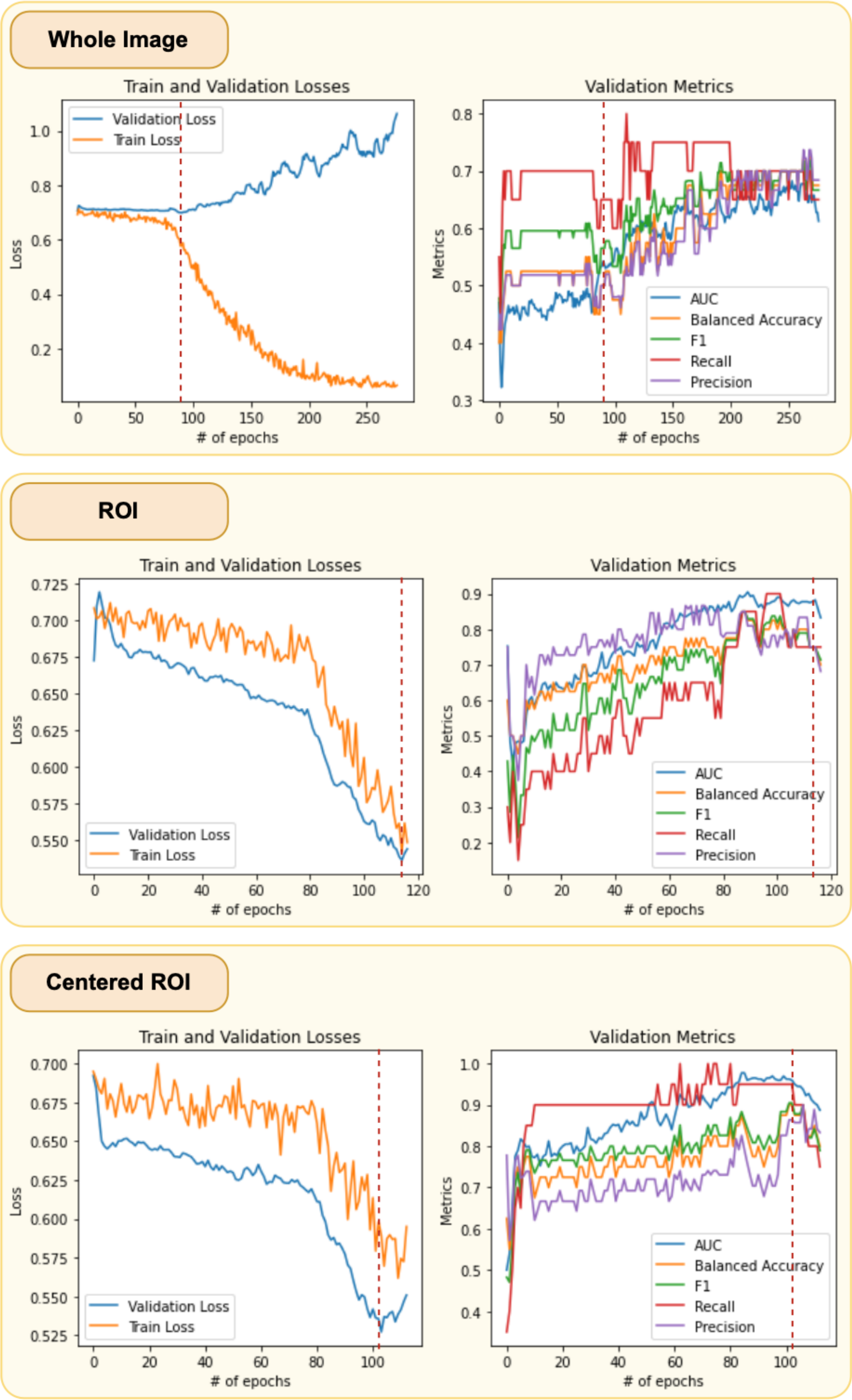

Figure 4.8: Network behaviour over epochs concerning Train and Validation Losses (on the left), and Performance Metrics (on the right). The best model is saved according to the lowest validation loss value, as indicated in the graphs by a red dotted line.

Table 4.7: Average Performance Results for the DL Approach.

	AUROC	**Balanced Accuracy**	**F1 score**	**Precision**	**Recall**
Validation Set	0.89±0.05	0.80±0.07	0.81±0.07	0.74±0.07	0.87±0.06
Test Set	0.94±0.04	0.74±0.08	0.66±0.08	0.51±0.10	0.98±0.04

negative samples was mislabelled as positive. However, for clinical use-cases, a compromise in terms of precision may occur if recall presents high values, since it guarantees that most positive cases are being detected.

4.5.2.3 Robustness Assessment

Similarity metrics were computed based on the obtained explanations from Saliency Maps, Integrated Gradients and LRP methods. Results were computed for every two heatmaps of the same test sample, being the results then average to get a overall comparison of similarity between samples, as presented in Table 4.8.

After analyzing the coherence results among explanations from the same sample, indirect conclusions regarding the assessment of the predictive model can be drawn. The explainable methods were utilized to assess the congruence between heatmaps of importance scores for the model to make predictions when providing different train-validation data sets. The obtained results show values of NRMS close to zero, as well as low L2 distances between maps, which are indicators of similar explanations for the same input. Furthermore, SSIM and MS-SSIM results are close to the reference, 1.00, indicating structural similarity amongst the compared images.

However, the results for Saliency Maps generated explanations show worst values when comparing to the other methods. This can be due to the fact that Saliency Maps is a simpler, less detailed method that generates generally more scattered and noisy explanations, causing worse similarity results. This can be visually observed in Figure 4.9, where it is noticeable that the explanation provided by Saliency Maps is scattered beside the tumor region, activating regions of

Table 4.8: Overview of average Robustness Results for the DL approach. The results are indicative of the similarity between explanations generated by the predictive model with different train-validation splits and the same test sample. The metrics were computed for the three explainable methods: Saliency Maps (SM), Integrated Gradients (IG) and LRP. The reference column (ref) indicates the perfect score for each metric, which would reflect on equal explanations for the same image.

metric	**ref**	**SM**	**IG**	**LRP**
NRMS ↓	0.00	0.05±0.01	0.03±0.01	0.02±0.00
L2 ↓	0.00	1258±151	385±185	193±61
SSIM ↑	1.00	0.61±0.07	0.95±0.05	0.98±0.02
MS-SSIM ↑	1.00	0.77±0.07	0.91±0.06	0.94±0.04

the image that have null intensity. On the other hand, it is also clear that Integrated Gradients and LRP show more specific regions that are always within the ROI boundaries.

Overall, this study showed that even with different data splits the trained models show some degree of consistency regarding the regions of the input that contribute the most to the prediction.

4.5.3 Limitations

After the development of a simple DL learning pipeline with a explainability component to assess the robustness of the classification model, the obtained performance results showed a high predictive capability and high consistency between explanations from the same test sample. However, the utilized architecture consisted on a pre-trained, well-known and commonly used network, ResNet50, thus having the opportunity of personalizing the architecture to further improve the performance results.

Moreover, despite the coherence results having shown a correlation between explanations of the same image, the utilized XAI methods consisted exclusively of post-hoc models, due to their simple utilization and ability to use over the trained models. These methods focus on explaining the prediction with concern to the input image, not interpreting the rationale of the architecture itself. This may be a limitation to the extent of explainability they are able to provide.

Finally, by studying and developing DL approaches, the need for larger amounts of data is clear, in order to build a more robust pipeline that is able to learn from a representative amount of data.

4.6 Summary

In this chapter, an extensive study on state-of-the-art approach for the prediction of MNA status utilizing different types of data was in order. Four different approach were developed and detailed, as well as the classification results for each methodology. For comparison purposes, all techniques were trained using the same pipeline and the same test cohort, running 10 models for each approach and averaging its results in order to also assess consistency and address the dataset variability. Despite using different data modalities, all data concerned the same patients.

The first experiment consisted on utilizing radiomic features extracted from CT slices and traditional ML methods in order to predict MNA status. The obtained results showed that the utilization of an SVM model without Data Augmentation nor Dimensionality Reduction was the pipeline with the highest performance, achieving and AUROC of 0.84±0.06.

The second experiment utilized semantic data with the same classical ML algorithms to predict the MNA status. The best semantic model was achieved with both XGB and Logistic Regression, being SMOTEN helpful for performance improvement, as well as PCA and KFS. The chosen best model was XGB with SMOTEN for data augmentation and KFS as feature selector, with an AUROC performance of 1.00±0.00. Despite the great results, it is important to notice that the semantic model only provides one sample per patient, whilst the remaining models utilize 5

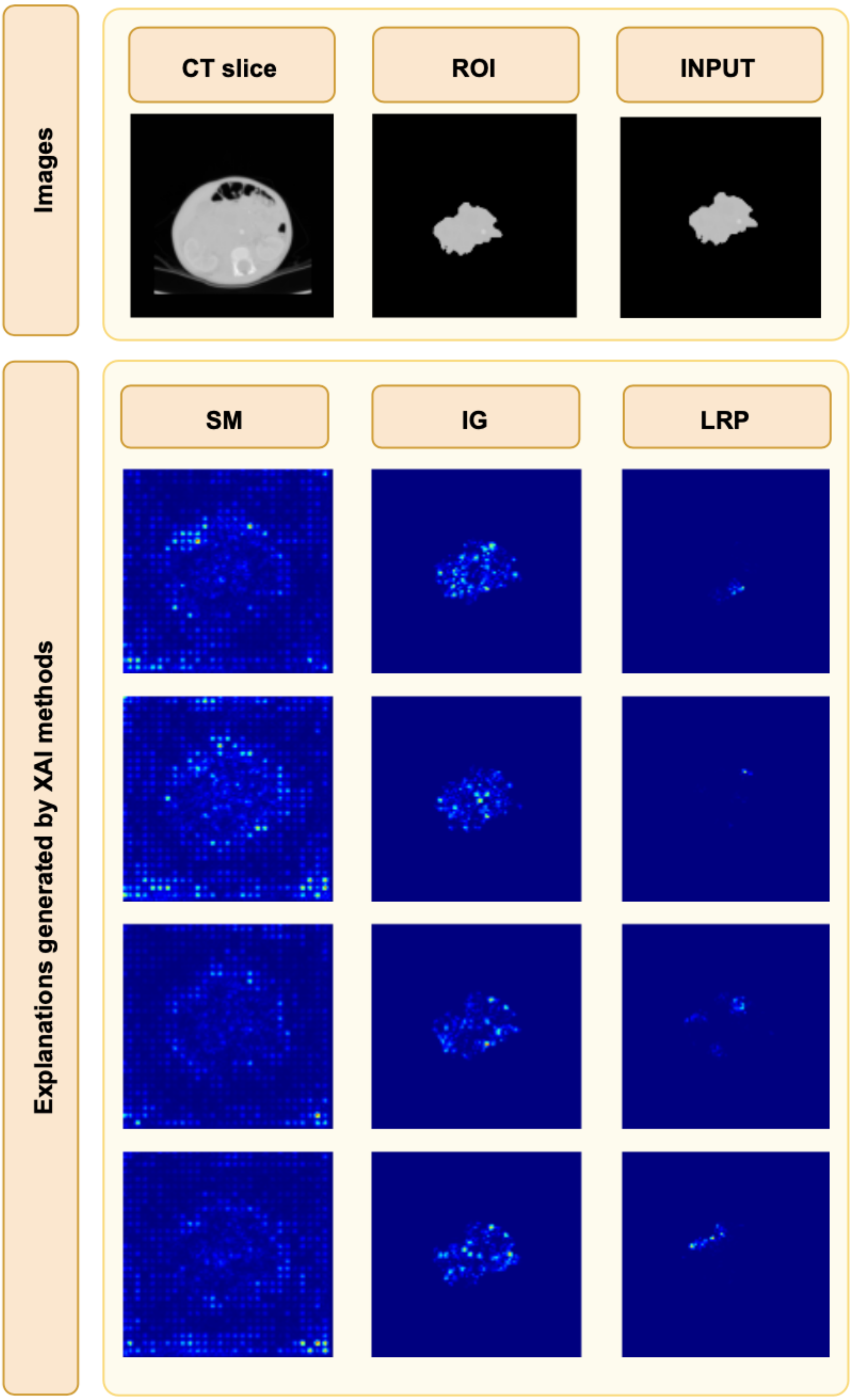

Figure 4.9: Illustrative test sample with generated explanations over 4 different train-validation splits. For every run the input is presented, as well as the explanation generated by Saliency Maps (SM), Integrated Gradients (IG) and LRP.

samples per patient, due to the utilization of imaging features. Therefore, the test cohort consists exclusively of 6 patients, being the results not comparable to the other approaches.

With the goal of joining both radiomic and semantic features to a single pipeline, a multi-modal approach was studied and implemented in the third experiment. Five different fusion methods were utilized to predict MNA status utilizing the semantic and radiomic models, being the best results achieved by the concatenation of both fetaure modalities into one single multi-modal dataset. The best ML algorithm for this dataset was Logistic Regression, utilizing SMOTENC for data augmentation and PCA for feature construction, achieving a AUROC value of 0.98±0.01.

The final experiment used DL techniques to predict the MNA status. A partially pre-trained ResNet50 was utilized as classifier, being the input of the network the centered ROI of each slice, containing the tumor. The performance was evaluated utilized the test set, achieving an average AUROC value of 0.94±0.04. After assessing the predictive ability of the model, its robustness was evaluated by utilizing XAI methods to generate importance heatmaps for all test samples over different runs. All explanations from the same sample were compared utilized image similarity metrics, obtaining satisfactory results that show correlation between explanations. Thus, models trained with different data splits consider similar importance for the same regions of the image, guaranteeing some degree of coherence and consistency to the model.

Overall, the developed models showed a promising behaviour in predicting the MNA status, being able to achieve great performance results.

Chapter 5

Conclusions and Future Work

The proposed work evaluates several approaches with the goal of predicting MNA status, an important biomarker in Neuroblastoma (NB) cases for patient risk stratification. To achieve that goal, several state-of-the-art Machine Learning (ML) and Deep Leaning (DL) methods were studied and utilized, in order to evaluate its predict ability towards the application of this study.

The utilized dataset for this work incorporated a total of 46 patients, with 5 CT slices each, comprising data from two different sources. Furthermore, the heterogeneity of this type of cancer includes tumors in several anatomical compartments, being the available images very different amongst themselves. The low amount of data and its variability constitute 2 great challenges in this type of tasks, despite the attempt of standardizing all data in the pre-processing stage.

All studied approaches were able to demonstrate a strong correlation between imaging phenotypes, patient clinical information and the amplification of the MYCN oncogene (MNA). High performance scores were obtained for all approaches, as summarized in Table 5.1, being able to over-perform the baseline approach (Pereira et al., 2022). Despite the good results, the size of the dataset does not guarantee that the data is representative, which may influence the performance with external samples. Furthermore, the semantic model has a significantly lower amount of data, which may translate in a less robust model.

Regarding the robustness assessment of the last approach, this study could be of great interest not only for this approach, but also for the remaining ones. In order for the developed models to be utilized in real-life applications, there is a rising need to understand the behaviour of the model and be able to explain to physicians the decisions made by the developed system. Therefore, this

Table 5.1: Overview of Performance Results for the proposed approaches concerning MNA prediction.

MNA Detection Approach	**No. of Test Samples**	**AUROC**	**Balanced Accuracy**	**F1 Score**	**Precision**	**Recall**
Baseline Radiomic Approach	30	0.69±0.16	—	—	—	—
Radiomic Approach	30	0.84±0.06	0.76±0.04	0.68±0.05	0.71±0.07	0.65±0.05
Semantic Approach	6	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Multi-modal Approach	30	0.98±0.01	0.74±0.02	0.66±0.01	0.49±0.02	1.00±0.00
Deep Learning Approach	30	0.94±0.04	0.74±0.08	0.66±0.08	0.71±0.10	0.98±0.04

analysis is just an introductory study on the importance of ensuring the reliability, transparency and consistency of the developed pipelines, and should be further extended in the broad spectrum of interpretability and explainability. The generated explanations may be also used for clinical interpretation, by interpreting the given heatmaps and provide insights to the correspondent clinical meaning apprehended by the algorithm.

Overall, the proposed work allowed for a exploratory study on the state-of-the-art methods in Artificial Intelligence for Medical Applications, namely in Radiogenomics. The used algorithms showed great potential for the detection of MNA status, being a great milestone for further investigation. As Future Work is concerned, the utilization of more datasets is crucial for the evolution of this study, along with the development of more personalized algorithms for the task itself, namely in the DL domain and the fusion of multi-modality and interpretability with the latter. In order for the pipeline to be applicable to pratical cases, the development of aggregation methods of the slice-wise results into patient-level prediction is necessary, providing a single prediction for each patient.

References

O Akin, P Elnajjar, M Heller, R Jarosz, B Erickson, S Kirk, and J Filippini. Radiology data from the cancer genome atlas kidney renal clear cell carcinoma [tcga-kirc] collection. *The Cancer Imaging Archive*, 2016.

Ahmed J Aljaaf, Dhiya Al-Jumeily, Abir J Hussain, Paul Fergus, Mohammed Al-Jumaily, and Khaled Abdel-Aziz. Toward an optimal use of artificial intelligence techniques within a clinical decision support system. In *2015 Science and Information Conference (SAI)*, pages 548–554. IEEE, 2015.

PF Ambros, IM Ambros, GM Brodeur, M Haber, J Khan, A Nakagawara, G Schleiermacher, Franki Speleman, R Spitz, WB London, et al. International consensus for neuroblastoma molecular diagnostics: report from the international neuroblastoma risk group (inrg) biology committee. *British journal of cancer*, 100(9):1471–1482, 2009.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5(1):1–9, 2018.

Kaustav Bera, Nathaniel Braman, Amit Gupta, Vamsidhar Velcheti, and Anant Madabhushi. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nature Reviews Clinical Oncology*, 19(2):132–146, 2022.

Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

Zuhir Bodalal, Stefano Trebeschi, Thi Dan Linh Nguyen-Kim, Winnie Schats, and Regina Beets-Tan. Radiogenomics: bridging imaging and genomics. *Abdominal radiology*, 44(6):1960–1984, 2019.

Hervé J Brisse, Thomas Blanc, Gudrun Schleiermacher, Véronique Mosseri, Pascale Philippe-Chomette, Isabelle Janoueix-Lerosey, Gaelle Pierron, Eve Lapouble, Michel Peuchmaur, Paul Fréneaux, et al. Radiogenomics of neuroblastomas: relationships between imaging phenotypes, tumor genomic profile and survival. *PloS one*, 12(9):e0185190, 2017.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Qian Chen, Min Li, Chen Chen, Panyun Zhou, Xiaoyi Lv, and Cheng Chen. Mdfnet: application of multimodal fusion method based on skin image and clinical data to skin cancer classification. *Journal of Cancer Research and Clinical Oncology*, pages 1–13, 2022a.

Xin Chen, Haoru Wang, Kaiping Huang, Huan Liu, Hao Ding, Li Zhang, Ting Zhang, Wenqing Yu, and Ling He. Ct-based radiomics signature with machine learning predicts mycn amplification in pediatric abdominal neuroblastoma. *Frontiers in oncology*, 11:1866, 2021.

Yiwen Chen, Ziyang Wang, Guotao Yin, Chunxiao Sui, Zifan Liu, Xiaofeng Li, and Wei Chen. Prediction of her2 expression in breast cancer by combining pet/ct radiomic analysis and machine learning. *Annals of Nuclear Medicine*, 36(2):172–182, 2022b.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Angela Di Giannatale, Pier Luigi Di Paolo, Davide Curione, Jacopo Lenkowicz, Antonio Napolitano, Aurelio Secinaro, Paolo Tomà, Franco Locatelli, Aurora Castellano, and Luca Boldrini. Radiogenomics prediction for mycn amplification in neuroblastoma: A hypothesis generating study. *Pediatric Blood & Cancer*, 68(9):e29110, 2021.

Massimo S Fiandaca, Mark Mapstone, Elenora Connors, Mireille Jacobson, Edwin S Monuki, Shaista Malik, Fabio Macciardi, and Howard J Federoff. Systems healthcare: a holistic paradigm for tomorrow. *BMC systems biology*, 11(1):1–17, 2017.

Maudy Gayet, Anouk van der Aa, Harrie P Beerlage, Bart Ph Schrier, Peter FA Mulders, and Hessel Wijkstra. The value of magnetic resonance imaging and ultrasonography (mri/us)-fusion biopsy platforms in prostate cancer detection: a systematic review. *BJU international*, 117(3): 392–400, 2016.

Olivier Gevaert, Sebastian Echegaray, Amanda Khuong, Chuong D Hoang, Joseph B Shrager, Kirstin C Jensen, Gerald J Berry, H Henry Guo, Charles Lau, Sylvia K Plevritis, et al. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports*, 7(1):1–8, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Julia E Heck, Beate Ritz, Rayjean J Hung, Mia Hashibe, and Paolo Boffetta. The epidemiology of neuroblastoma: a review. *Paediatric and perinatal epidemiology*, 23(2):125–143, 2009.

Weronika Hryniewska, Przemysław Bombiński, Patryk Szatkowski, Paulina Tomaszewska, Artur Przelaskowski, and Przemysław Biecek. Checklist for responsible deep learning modeling of medical images based on covid-19 detection studies. *Pattern Recognition*, 118:108035, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Miller Huang and William A Weiss. Neuroblastoma and mycn. *Cold Spring Harbor perspectives in medicine*, 3(10):a014415, 2013.

Yosuke Iwatate, Isamu Hoshino, Hajime Yokota, Fumitaka Ishige, Makiko Itami, Yasukuni Mori, Satoshi Chiba, Hidehito Arimitsu, Hiroo Yanagibashi, Hiroki Nagase, et al. Radiogenomics for predicting p53 status, pd-l1 expression, and prognosis with machine learning in pancreatic cancer. *British Journal of Cancer*, 123(8):1253–1261, 2020.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

Anil Kalra. Developing fe human models from medical images. In *Basic finite element method as applied to injury biomechanics*, pages 389–415. Elsevier, 2018.

Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, 3(3):740–770, 2021.

Burak Kocak, Emine Sebnem Durmaz, Ece Ates, and Melis Baykara Ulusan. Radiogenomics in clear cell renal cell carcinoma: machine learning–based high-dimensional quantitative ct texture analysis in predicting pbrm1 mutation status. *American Journal of Roentgenology*, 212 (3):W55–W63, 2019.

Will Koehrsen. Willkoehrsen/feature-selector: Feature selector is a tool for dimensionality reduction of machine learning datasets. URL `https://github.com/WillKoehrsen/feature-selector`.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Yiming Lei, Yukun Tian, Hongming Shan, Junping Zhang, Ge Wang, and Mannudeep K Kalra. Shape and margin-aware lung nodule classification in low-dose ct images via soft activation mapping. *Medical image analysis*, 60:101628, 2020.

Zeju Li, Yuanyuan Wang, Jinhua Yu, Yi Guo, and Wei Cao. Deep learning based radiomics (dlr) and its usage in noninvasive idh1 prediction for low grade glioma. *Scientific reports*, 7(1):1–11, 2017.

Nicola Lunardon, Giovanna Menardi, and Nicola Torelli. Rose: A package for binary imbalanced learning. *R journal*, 6(1), 2014.

Mafalda Malafaia, Tania Pereira, Francisco Silva, Joana Morgado, António Cunha, and Hélder P Oliveira. Ensemble strategies for egfr mutation status prediction in lung cancer. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3285–3288. IEEE, 2021.

Mafalda Malafaia, Francisco Silva, Inês Neves, Tania Pereira, and Hélder P. Oliveira. Robustness analysis of deep learning-based lung cancer classification using explainable methods. *IEEE Access*, under peer-review.

Paula Marrano, Meredith S Irwin, and Paul S Thorner. Heterogeneity of mycn amplification in neuroblastoma at diagnosis, treatment, relapse, and metastasis. *Genes, Chromosomes and Cancer*, 56(1):28–41, 2017.

Darcy Mason. Su-e-t-33: pydicom: an open source dicom library. *Medical Physics*, 38(6Part10): 3493–3493, 2011.

Joana Morgado, Tania Pereira, Francisco Silva, Cláudia Freitas, Eduardo Negrão, Beatriz Flor de Lima, Miguel Correia da Silva, António J Madureira, Isabel Ramos, Venceslau Hespanhol, et al. Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. *Applied Sciences*, 11(7):3273, 2021.

S Pal. Transfer learning and fine tuning for cross domain image classification with keras. *GitHub: transfer learning and fine tuning for cross domain image classification with Keras*, 2016.

Tania Pereira, Cláudia Freitas, José Luis Costa, Joana Morgado, Francisco Silva, Eduardo Negrão, Beatriz Flor de Lima, Miguel Correia da Silva, António J Madureira, Isabel Ramos, et al. Comprehensive perspective for lung cancer characterisation based on ai solutions using ct images. *Journal of Clinical Medicine*, 10(1):118, 2020.

Tania Pereira, Francisco Silva, Pedro Claro, Diogo Costa Carvalho, Sílvia Costa Dias, Helena Torrão, and Hélder P Oliveira. A random forest-based classifier for MYCN status prediction in neuroblastoma using ct images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3854–3857. IEEE, 2022.

Gil Pinheiro, Tania Pereira, Catarina Dias, Cláudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P Oliveira. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: Egfr and kras. *Scientific reports*, 10(1):1–9, 2020.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3), 2020.

Aaron M Rutman and Michael D Kuo. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *European journal of radiology*, 70(2):232–241, 2009.

Scikit-learn. Sklearn model selection GRIDSEARCHCV. URL `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV`.

Francisco Silva, Tania Pereira, Julieta Frade, José Mendes, Claudia Freitas, Venceslau Hespanhol, José Luis Costa, António Cunha, and Hélder P Oliveira. Pre-training autoencoder for lung nodule malignancy assessment using ct images. *Applied Sciences*, 10(21):7837, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Eelin Tan, Khurshid Merchant, Bhanu Prakash Kn, Arvind Cs, Joseph J Zhao, Seyed Ehsan Saffari, Poh Hwa Tan, and Phua Hwee Tang. Ct-based morphologic and radiomics features for the classification of mycn gene amplification status in pediatric neuroblastoma. *Child's Nervous System*, pages 1–9, 2022.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

Shuo Wang, Jingyun Shi, Zhaoxiang Ye, Di Dong, Dongdong Yu, Mu Zhou, Ying Liu, Olivier Gevaert, Kun Wang, Yongbei Zhu, et al. Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning. *European Respiratory Journal*, 53(3), 2019.

Haoting Wu, Chenqing Wu, Hui Zheng, Lei Wang, Wenbin Guan, Shaofeng Duan, and Dengbin Wang. Radiogenomics of neuroblastoma in pediatric patients: CT-based radiomics signature in predicting MYCN amplification. *European Radiology*, 31(5):3080–3089, 2021.

Zilong Xu, Qiwei Yang, Minghao Li, Jiabing Gu, Changping Du, Yang Chen, and Baosheng Li. Predicting her2 status in breast cancer on ultrasound images using deep learning method. *Frontiers in oncology*, 12:829041, 2022.