POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Explainable Machine Learning and Deep Learning models to predict immunotherapy response in NSCLC patients using CT scans

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Margherita Favali, 967212

Abstract: Immunotherapy (IO) has brought a significant revolution in the treatment of non-small cell lung cancer (NSCLC). Hence, reliable biomarkers are required to identify patients that are most likely to benefit from this therapy. Since available biomarkers, such as PD-L1, demonstrated limited predicted efficacy, there is an urgent need for novel models to improve predictive capabilities. Analyzing CT scans, using machine learning (ML) and deep learning (DL) techniques, offers a promising approach to extract features from medical images and construct predictive models. This study aims at developing two types of solutions to predict efficacy of IO in advanced NSCLC. The first is ML-based and utilizes six different ML classifiers, by determining the best-performing one, while the second employs an end-to-end DL pipeline. The evaluation is performed on two data modalities: real-world data (RWD) and features extracted from CT scans. Baseline CT scans and clinical data were retrospectively collected from a cohort of 375 patients with advanced NSCLC at Fondazione IRCCS Istituto Nazionale dei Tumori di Milano. These patients received any-line of IO, either alone or in combination with chemotherapy. The final objective of this study is two-fold. Firstly, compare performances with the use of CT scans alone versus integrating them with RWD. Secondly, to evaluate and compare the performance of ML and DL models in terms of predictive accuracy. The final step for both the solutions involves conducting an explainability analysis with the computation of SHapley Additive exPlanations (SHAP) values. The main findings of the present work suggest that DL approach, with an accuracy of 0.63, slightly outperforms ML, which achieved an accuracy of 0.61, in predicting IO response using only features derived from CT scans. However, when the two data modalities are combined, ML achieves higher performance, with an accuracy of 0.69, compared to DL, which achieved an accuracy of 0.64. These results suggest another interesting observation. When the two data modalities are combined, ML exhibit an increase in predictive performance and ability to predict clinical benefit from IO. In the context of the explainability analysis, in ML models trained on combination of RWD and CT scans features, SHAP values revealed that the ECOG PS (RWD) and Large Dependence Emphasis (CT scan feature) had the greatest impact on the predictions. In DL, SHAP values were assigned to image pixels, revealing that the network predominantly concentrated on the edges of the tumor region of interest (ROI). These initial achievements could be the base of the ultimate goal of developing novel tools for selection of ideal candidates for IO. By investigating future perspectives, this research may contribute to the development of innovative approaches that can be applied in clinical practice.

Advisor:
Prof. Alessandra Laura Giulia Pedrocchi

Co-advisors:
Vanja Miskovic, PhD
Arsela Prelaj, MD
Alessandro Quarta

Academic year:
2022-2023

Key-words: NSCLC, Immunotherapy, Machine Learning, Deep Learning, Explainability, CT scans

1. Introduction

According to estimates from the World Health Organization (WHO) cancer is the second leading cause of death globally [1]. Among neoplasms, lung cancer is the leading cause of cancer death among men in the United States [2] and Europe [3], accounting for an estimated 1,761,000 deaths worldwide in 2017 [4].
There are several causes that may develop this disease, including environmental agents (such as exposure to asbestos, radon, radiation, and air pollution) as well as genetic factors. However, smoking remains the primary risk factor. Consequently, screening is highly recommended for individuals at higher risk [5].
Lung cancer is classified in mainly two histological subtypes with different clinical behaviour: non-small cell lung carcinoma (NSCLC), accounting for 80-85% of all lung cancer cases [6], and small-cell lung carcinoma (SCLC). There are several types of NSCLC, depending on different kinds of cancer cells, exhibiting different growth patterns and methods of spreading [5]: Squamous cell carcinoma that forms in the thin, flat cells lining the inside of the lungs, Large cell carcinoma which may begin in several types of large cells and Adenocarcinoma which begins in the cells that line the alveoli and make substances such as mucus.
In numerous instances, the diagnosis of lung cancer occurs when the disease has already reached an advanced stage. Consequently, surgical intervention is often not feasible for such cases. Therefore, other options are chemotherapy, radiotherapy, targeted therapy and IO [5]. Even if traditional therapies like chemotherapy and radiotherapy provided benefit in terms of survival for lung cancer patients and are still incorporated in therapeutic algorithms, the prognosis remained poor with an estimated median overall survival (OS) of about 14 months in the metastatic setting [7]. In this context, IO has brought a significant revolution in the treatment of NSCLC. In fact, recent clinical studies demonstrated that IO, delivered either alone or in combination with other therapies, could improve survival outcomes of advanced NSCLC patients, with about 20% of patients still alive 5 years after diagnosis of metastatic disease [8].

IO aims at harnessing immune system in recognizing and attacking the tumor cells. There are several types of IO, one of the most used involves the Immune Checkpoint Inhibitors (ICIs), which are drugs that block the checkpoint proteins from binding with proteins on tumor cells and preventing that immune cells (T cells) are switched off. Programmed death-1 (PD-1) is a cell surface receptor that functions as a T cell checkpoint. Binding of PD-1 to its ligand, programmed death-ligand 1 (PD-L1) located in cancer cells, inhibits T cells from killing tumor cells. To break the binding and restart the immune system, ICIs drugs are used, especially in the case of PD-1/PD-L1 the main drugs are Nivolumab, Pembrolizumab, Atezolizumab, Avelumab and Durvalumab [9].
Over the last decade, ICIs have transformed the treatment of advanced malignancies, however, response rates can widely vary among patients [10]. Hence, biomarkers with high sensitivity and specificity are required to identify patients that are most or least likely to experience a sustained response to these therapies [11].

There are several biomarkers tested by Food and Drug administration (FDA) [12] until now, but none of them resulted as completely reliable.
PD-L1 was the first FDA-approved predictive biomarker for NSCLC in 2015 [12]. Testing the PD-L1 expression, is a standard for identifying individuals with advanced NSCLC that are more likely to respond to IO, used alone or in combination with chemotherapy. In particular, it has been demonstrated a correlation between the level of tissue PD-L1 expression and clinical benefits: when the PD-L1 was higher than one-half of tumor cells (i.e. PD-L1 expression $\geq 50\%$), patients were more likely to respond. Despite this evidence, PD-L1 remains a controversial biomarker for IO response [13].
Another potential biomarker is Tumor Mutational Burden (TMB), which is defined as the total number of mutations per coding area of a tumor genome. It was demonstrated by Rizvi et al [14] that high number of somatic mutations is thought to result in a higher response to checkpoint inhibition. Also this kind of biomarker for IO in NSCLC remains uncertain since, despite these initial positive findings, subsequent data have revealed a statistically nonsignificant benefit in patients with high TMB [15].
Due to these limitations, there is an urgent need to find improved and efficient biomarkers for IO response. To reach this aim Artificial Intelligence (AI) techniques can be used, where radiomics solutions, quantitative approach to medical imaging, are exploited.

The concept of radiomics, which has most broadly been applied in oncology, refers to the extraction of mineable data from medical imaging [16]. It is based on the concept that biomedical images contain information

about disease that are imperceptible by the human eye [17]. Radiomic analysis can be computed on medical images coming from different modalities: magnetic resonance imaging (MRI), computed tomography (CT), and positron-emission-tomography (PET). Traditionally radiomic features are extracted from the region of interest (ROI), usually segmented by expert radiologists, and they regard shape, grey-level intensities, texture, size or volume of the ROI [18]. Both ML and DL algorithms may be used in order to process medical images and extract information.
Multiple studies have explored predicting IO outcomes using radiomics from CT scans. Gonga et al. [19] conducted a study on 224 advanced NSCLC patients using a CT-based radiomics approach. They aimed to predict IO response, examine radiomics' prognostic power for predicting progression-free survival (PFS) and overall survival (OS). Two CT scans per patient were collected: pre-treatment and post-treatment. After tumor segmentation and image resampling, 1118 CT-radiomic features were extracted. Delta radiomic features were calculated by subtracting pre-treatment from post-treatment radiomics. Normalized features were ranked using recursive feature elimination (RFE) to select optimal features. A support vector machine (SVM) classifier was implemented to predict IO response using pre-treatment and delta radiomics features. The study showed that delta-radiomics improved treatment response prediction and enhanced PFS and OS outcomes compared to pre-treatment radiomics. Another interesting work is by He et al. [20], where deep learning solution was exploited to estimate the target tumor area, to distinguish High-TMB from Low-TMB patients and establish a tumor mutational burden radiomic biomarker (TMBRB). CT images from a total of 327 patients with NSCLC, were randomly divided into a training (n=236), validation (n=26), and test cohort (n=65). TMBRB was evaluated for its predictive capability in terms of OS and PFS. The biomarker successfully stratified patients in the IO-treated cohort into high- and low-risk groups, showing superior results in terms of PFS outcome.
Recently, the need of high level of accountability and thus transparency is required in AI algorithms, especially in the medical sector, where decisions derived from such systems affect humans' lives. Explanations for machine decisions and predictions are needed to justify their reliability [21]. For both ML and DL models, explanation can be achieved by using post-hoc methods that approximate the behavior of a model by extracting relationships between feature values and the predictions [22] [23]. Local Interpretable Model-Agnostic Explanations (LIME) [24] and SHapley Additive exPlanations (SHAP) [25] are two examples of post-hoc explanation.

In the present study, baseline CT images and clinical data were retrospectively collected from a cohort of 375 patients diagnosed with advanced NSCLC at Fondazione IRCCS Istituto Nazionale dei Tumori di Milano. These patients received any-line of IO, either alone or in combination with chemotherapy. The primary aim of this study is to contribute to the research on predicting IO response in advanced NSCLC. This is achieved by addressing a binary classification problem, which categorizes each patient if has or not a clinical benefit from IO. The study assesses the predictive power of features extracted from CT scans. To achieve this objective, two distinct pipelines are proposed. The first pipeline (ML solution) involves extracting radiomic features from CT scans and feeding them into ML classifiers to predict IO response. The second pipeline (DL solution) focuses on utilizing an end-to-end neural network to directly extract features from the images (DL features) and predict therapy response. In both the ML and DL solutions, the evaluation is conducted using two distinct data modalities: real-world data (RWD) and features extracted from CT scans. This aims at investigating whether combining these two modalities or using them individually can enhance the classification performance. Significant effort is dedicated to ensure the explainability and interpretability of both ML and DL models by using SHAP technique. This is done to ensure that the results are comprehensible and can potentially be translated into clinical practice.

2. Materials and Methods

2.1. Dataset

The population involved in this retrospective study consisted of data collected at National Cancer Institute of Milan (Fondazione IRCCS Istituto Nazionale dei Tumori) between April 2013 and May 2022. The cohort included 375 patients with advanced NSCLC who received any-line anti-PD(L)1 therapy either alone or in combination with chemotherapy. Specifically, 305 patients were treated with IO, while 70 with the combination of IO and chemotherapy.

2.2. Tumor segmentation

Baseline CT scans were acquired using a third generation dual-source CT scanner, Somatom Force provided by Siemens Helthineers [26]. Baseline non-contrast-enhanced (NCE) and contrast-enhanced (CE) CT scans were analyzed by four experienced radiologists, who identified primary tumors and involved lymph nodes. The

segmentation was performed semiautomatically using syngo.via, an integrated imaging software provided by Siemens Healthineers [27]. If neither the tumor and lymph nodes were present, the distant metastasis were not segmented. In accordance with the guidance of clinicians, the present study included only patients who had a primary tumor (lung tumor). A radiological assessment was performed for each patient, evaluating the follow-up total body (TB-CT) scan. The TB-CT scan was conducted every 9-12 weeks or when signs of disease progression were observed. The evaluation of the response was conducted based on the Response Evaluation Criteria in Solid Tumors (RECIST1.1) criteria released in February 2000 and then updated to current 1.1 version in 2009 [28].

2.3. Data collection and curation

Two types of data were utilized: RWD and features extracted from CT scans. The two types of data were utilized in two distinct pipelines: a Machine Learning (ML) and a Deep Learning (DL) approach.
The RWD were obtained during regular clinical examinations conducted prior to the initiation of treatment. From the extensive range of clinical values available, 16 specific baseline RWD were selected for inclusion in this study based on clinicians hypothesis-driven. The 16 used clinical data are collected and defined in Table 1.

RWD	Definition
Therapy	Therapy administered to the patient: immunotherapy alone (1) or in combination with chemotherapy (0)
Age	Age of the patient at IO baseline
Sex	Patient sex: Male (1), Female (0)
Surgery y/n	Binary variable identifying patients which underwent surgery to reduce tumor mass
Histology	Indicates if the tumor type is squamous (1) or non squamous (0)
Line of therapy	Line of treatment that patient received
Smoking status	Binary variable that identifies if the patient is a smoker or ex-smoker (1) or non-smoker (0)
PDL1 group	Value of Programmed death ligand 1 (PD-L1): < 1% (1), 1-49% (2), >50% (3)
ECOG PS	ECOG Performance Status at IO baseline
Tumor stage	Tumor characterization according to TNM evaluation
Node stage	Node characterization according to TNM evaluation
Metastases stage	Metastasis characterization according to TNM evaluation
Number of metastatic sites	Indicates the number of the metastases
Metastases Brain	Binary variable that indicates brain metastasis
Metastases Bone	Binary variable that indicates bone metastasis
BMI	Body Mass Index at IO baseline

Table 1: RWD; TNM is a system for classification of malignancies: T (Tumor) describes the size of the primary tumor and its' invasion into adjacent tissues, N (Node) describes regional lymph node involvement of the tumor, M (Metastasis) identifies the presence of distant metastases of the primary tumor [29].

An extensive data curation procedure was performed. Data originating from multiple sources were integrated into a coherent dataset, duplicate patients and inconsistent values were eliminated, and missing data were identified and filled with the assistance of clinicians, whenever possible. Finally, textual data were converted into numerical and categorical values, and imputation techniques were employed to address any remaining

missing data. Specifically, a multivariate imputer was utilized to fill the missing values by modeling them as a function of other available data [30]. The other data used were features extracted from the primary tumor volume of patients' baseline CT scans. In ML pipeline, they were calculated with pyradiomics package and encompassed shape characteristics, grey level properties, grey tone differences, and statistical attributes [31]. Conversely, in DL approach, features were directly extracted from the neural network.

2.4. Outcome

The target value is represented by the best overall response, i.e. the best response recorded from the first radiological evaluation until disease progression according to the RECIST1.1 criteria [28].
The outcome analyzed for this study is Clinical Benefit Rate (CBR), which is defined as the percentage of patients who have achieved complete response (CR), partial response (PR), or at least four months of stable disease (SD) as a result of therapy [32].
The outcome, additionally to the at least four months SD, takes into account also the patients that had a progression but with still a clinical benefit after at least 9 months. Two classes were defined, as shown in Table 2:

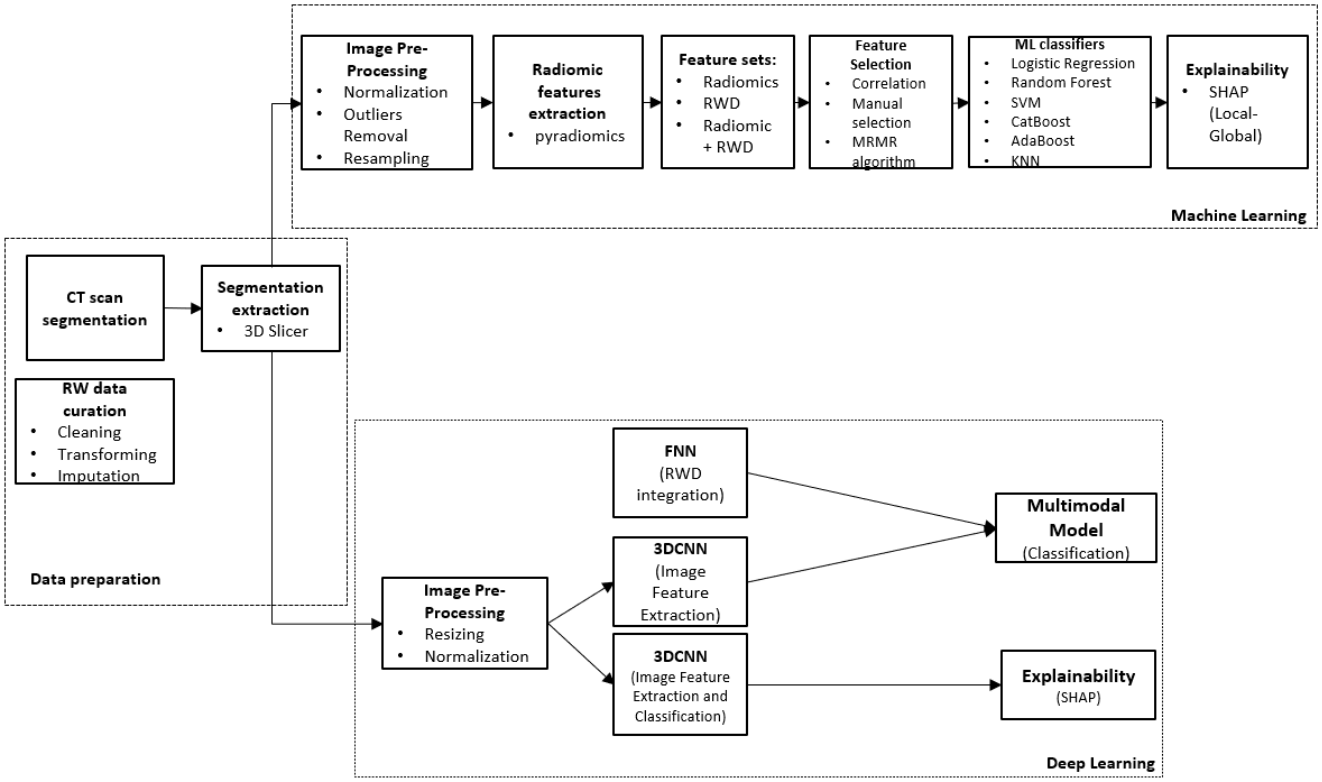
Class	Description
Class 0	Progressive Disease (PD) if TTF < 9 months
	Stable Disease (SD) if PFS < 4 months
Class 1	Complete Response (CR)
	Partial Response (PR)
	Stable Disease (SD) if PFS ≥ 4 months
	Progressive Disease (PD) if TTF ≥ 9 months

Table 2: Clinical Benefit Rate: definition of classes

The choice of using a clinical endpoint rather than the radiologic alone was done after discussions with clinicians, as this endpoint can better distinguish patients who benefit from IO from refractory patients.

2.5. Model development

Two different approaches were used: one ML-based and the other DL-based, as shown in Figure 1.



(a)

Data modality	Machine Learning	Deep Learning
CT scan features	Yes	Yes
RWD	Yes	No
CT scan features + RWD	Yes	Yes

(b)

Figure 1: Two approaches: (a) Machine Learning and Deep Learning Illustration of methodological workflow

2.5.1 Machine Learning

The first approach is a classification using ML pipeline (Fig. 1). The ML pipeline consisted of the segmentation extraction from the CT scans by using 3D slicer, an open source software for visualization, processing, segmentation, registration, and analysis of medical images [33]. Then, both the CT scan and the corresponding segmentation of each patient were converted to NRRD (Nearly Raw Raster Data) format in order to be processed with Pyradiomics package [34]. Subsequently, several steps were undertaken to extract and process the features [18]. Three different feature sets were provided as input to the ML classifiers: radiomic features, a combination of radiomics and RWD, and RWD alone. Finally, ML techniques were employed to predict the outcome and provide corresponding explanations.

Image pre-processing

Image processing was located between the image segmentation and feature extraction step. It was used to homogenize and process images from which radiomic features would be extracted with respect to pixel spacing and grey-level intensities [18].
The image pre-processing was composed of three steps: normalization, outliers removal, sampling.
First of all, after computing the region of interest (ROI), normalization was applied in order to rescale grey-levels to a specific range. The normalized intensity ($f(x)$) was calculated by centering the original intensity (x) at the mean (μ_x) with standard deviation (σ_x), where s was the scaling factor (set to default value 1), as shown in Equation (1.3.1):

$$f(x) = \frac{s(x - \mu_x)}{\sigma_x} \tag{1.3.1}$$

After this initial step, outliers were eliminated, and both the CT scan and the segmentation mask were resampled to ensure complete matching resolutions and voxel sizes across all the CT scans and segmentations. The pixel spacing was set to 1, a linear interpolator was applied, and no necessary additional padding was added.

Radiomic Feature Extraction

After the pre-processing process, the extraction of radiomic features was performed. To achieve this, the *featureextractor* module from the Pyradiomics package [34] was utilized with the default parameters (no filter applied, and all the features enabled to extract).
A total of 107 features were extracted, which were categorized into seven different classes: 18 features of the first-order class, 14 shape descriptors, 75 texture features of Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighbouring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM) [31]. The calculated features were stored and returned in an Excel file, where each feature was assigned a unique name comprising the applied filter (no filter applied in this study), the feature class, and the feature name.

Feature selection

Feature selection was performed separately on the three feature sets.
Starting from radiomics, due to the high correlation often observed among radiomic features [35], a three-step feature selection process was implemented to eliminate redundant information in the dataset. The steps involved: (1) correlation analysis, (2) manual selection, and (3) the application of the Maximum Relevance-Minimum Redundancy (MRMR) analysis technique [36]. For the correlation analysis, the Pearson correlation coefficient was computed, and features with correlation coefficients exceeding thresholds of 0.8 (indicating positive correlation) and -0.8 (indicating negative correlation) were identified and removed. Following this initial filtering process, the remaining features were further evaluated to identify and exclude variables that displayed similar behavior across the two classes. This step aimed to eliminate poorly significant variables from the analysis and resulted in 21 final radiomic features. As final step, the MRMR algorithm was applied. MRMR aims at selecting the features that had maximum relevance with respect to the target variable and minimum redundancy with respect to the features that have been selected at previous iterations. In practice, at each iteration i, a score is computed for each feature to be evaluated (f). The feature that has the highest score at each iteration is added to the set of selected features. Once a feature goes into the bucket, it cannot come out. The score is computed by dividing F-statistic between the feature and the target variable by the Pearson correlation between the feature and all the features that have been selected at previous iterations.
Regarding the RWD alone, the Pearson correlation coefficient was also calculated and the MRMR feature selector was directly applied, without any manual selection in this case.
In case of combination of features, the 21 radiomic features were merged with 16 RWD and subsequently MRMR feature selection was performed.

Machine Learning classifiers and evaluation

The study goal was a binary classification carried out with six Machine Learning classifiers: Logistic Regression (LR) [37], Random Forest (RM) [38], Support Vector Machine (SVM) [39], CatBoost [40], AdaBoost [41] and K-nearest neighbors (KNN) [42]. The CatBoost classifier utilized the CatBoost package [43], while all other models were implemented using scikit-learn [30] in Python 3.7.0 [44]. ML models were trained using the three feature sets and were evaluated both on an independent test set and on an external validation set.
In order to evaluate which model reached the best results in terms of prediction, several evaluation metrics were computed: Accuracy, Precision, Recall, F1-score, ROC curve and AUC and Confusion Matrix. Additionally, cross-validation with 5 k-folds was applied.

SHAP values computation

Once the best ML classifier was found for each feature set, SHAP values [45] were used to explain model predictions, identifying the features that had the greatest impact on the outcome and understanding how they influenced it. Shapley values are based on the idea that the outcome of each possible combination of f features (f going from 0 to F, where F is the number of all the possible features available) should be considered to

determine the importance of a single feature [45]. SHAP requires to train a predictive model for each distinct combination of features $S \subseteq F$. The models are equivalent in terms of hyperparameters and dataset, the only thing that changes is the set of features included. The idea is that the gap between the predictions of two combinations of features (that differ from the presence of a specific feature) can be imputed to the effect of this additional feature. This is called "marginal contribution" of a feature and it is the difference between model trained with that feature present and model trained without that feature. Therefore, to obtain the overall effect of a feature the differences are computed for all subsets $S \subseteq F \setminus \{i\}$. All these marginal contributions are then joined in a weighted average. The formula for calculating the SHAP value of a feature is reported below:

$$\phi_i(f) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{2.5.5.1}$$

where F is the total number of features. f(S) is the prediction given the subset S and $f(S \cup \{i\})$ is the prediction given S including feature i and their difference is the so called marginal contribution.
Two approaches were used to provide the explainability of the model: global, for understanding the overall structure of how a model makes decisions, and local, for understanding how the model made decisions for a single prediction.
For the global solution, a SHAP summary plot was exploited [46]. It combines feature importance with feature effects and it shows the positive and negative relationships between the features and the target variable (CBR outcome). Each data point on the plot corresponds to an observation (patient). The features are arranged on the y-axis based on their importance for the model, with the most important feature positioned at the top. On the x-axis, the plot indicates whether the effect of the feature value is associated with a higher (class 1) or lower (class 0) target value. Additionally, a color map is used to represent the feature values, where red indicates high values and blue represents low values.
For the local SHAP analysis, waterfall plots [46] were generated for four types of predictions: one True Positive (TP), one True Negative (TN), one False Positive (FP), and one False Negative (FN). In these plots, the features are ordered from top to bottom based on their importance, and the contribution of each feature to the individual prediction is displayed. Features that move the prediction towards class 1 are represented by red bars, while features that contribute to predict class 0 are represented by blue bars.

2.5.2 Deep Learning

The second approach included the implementation of an end-to-end solution of DL pipeline. End-to-end learning in deep learning means that a single neural network model is trained to directly process raw input data and produce the desired prediction without relying on explicit intermediate representations or manual feature engineering. The model learns to automatically extract relevant features and make classification in a single integrated process, encapsulating multiple stages of a traditional pipeline within a single network architecture [47]. Two different feature sets were used: deep learning features (DL features) coming from the CT scans and a combination of DL features and RWD. For the feature combination, the same RWD used in ML pipeline were introduced to the DL model to obtain complete comparable results between DL and ML techniques.
To perform the classification with the two different feature sets, two models were utilized (see Fig. 1). The first model, a 3D Convolutional Neural Network (3DCNN), solely processed input images. The second model, a bimodal model, received a combination of both DL features and RWD as input. Prior to feeding them into the network, the CT scans and segmentations underwent preprocessing steps. In the two models, the images were resized differently. In the 3DCNN model, the images were resized from their original size of 512x512 pixels to a smaller size of 18x18 pixels. In the bimodal model, the images were resized to the size of 64x64 pixels. Only 10 slices were retained by selecting the slices containing the segmentation. Subsequently, the region of interest (ROI) was computed.
After that, the neural models were trained and tested with the same training, test and external validation sets used for ML.

3D Convolutional Neural Network

The model fed with images only was a 3D Convolutional Neural Network (3DCNN) (Fig. 2).
The 3DCNN took as input both the CT image and the segmentation, along with the labels provided in an Excel file containing the CBR outcomes for each patient. The neural network architecture consisted of three convolutional layers, each followed by a rectified linear unit (ReLu), a max pooling layer, and a dropout layer. Additionally, there were two linear layers, with the final layer dedicated to binary classification. The corresponding architecture is shown in Figure 2.

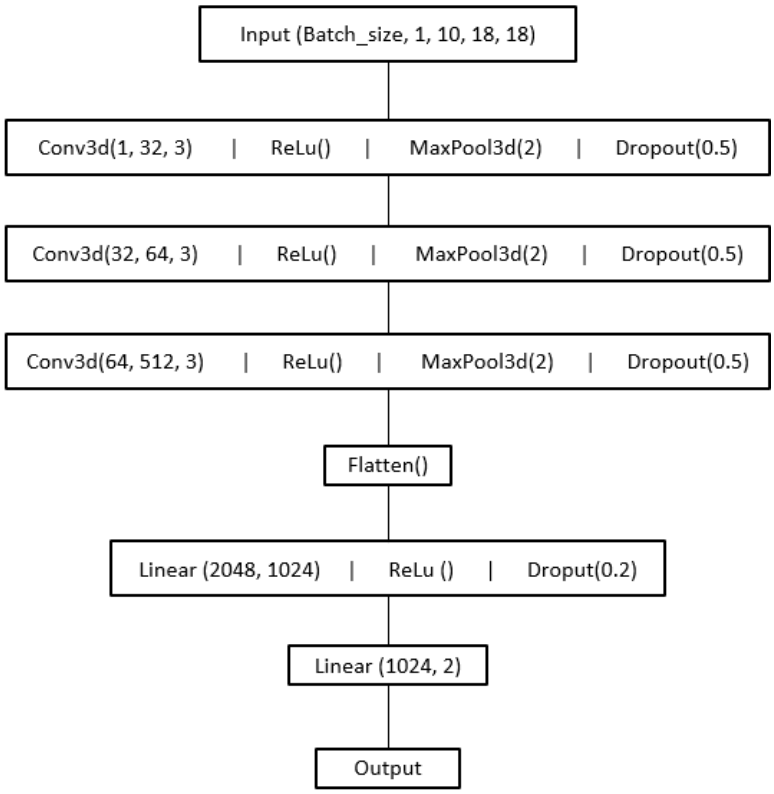


Figure 2: 3D Convolutional Neural Network (3DCNN): to process CT scans and segmentations

Bimodal model

In the bimodal model, two data modalities (RWD and DL features) were processed. Two neural networks were implemented: a 3DCNN (Figure 3a), slightly different from the one used in Section 2.5.2, for processing the 3D images and extract relative features, and a Feed-Forward neural network (FNN) for handling the RWD (Figure 3b). The FNN consisted of two linear layers, each employing a rectified linear unit (ReLU) activation function and a dropout layer.
The features pertaining to CT scans and RWD were extracted from these two neural networks and concatenated within the Bimodal model, where the classification process was carried out.
This is an implementation of intermediate fusion [48], wherein the data corresponding to each modality are concatenated prior to classification. Intermediate fusion involves the transformation of raw inputs into a higher-level representation by mapping them through a stack of layers. By unifying the feature representation, bimodal feature maps are obtained, later used for classification.

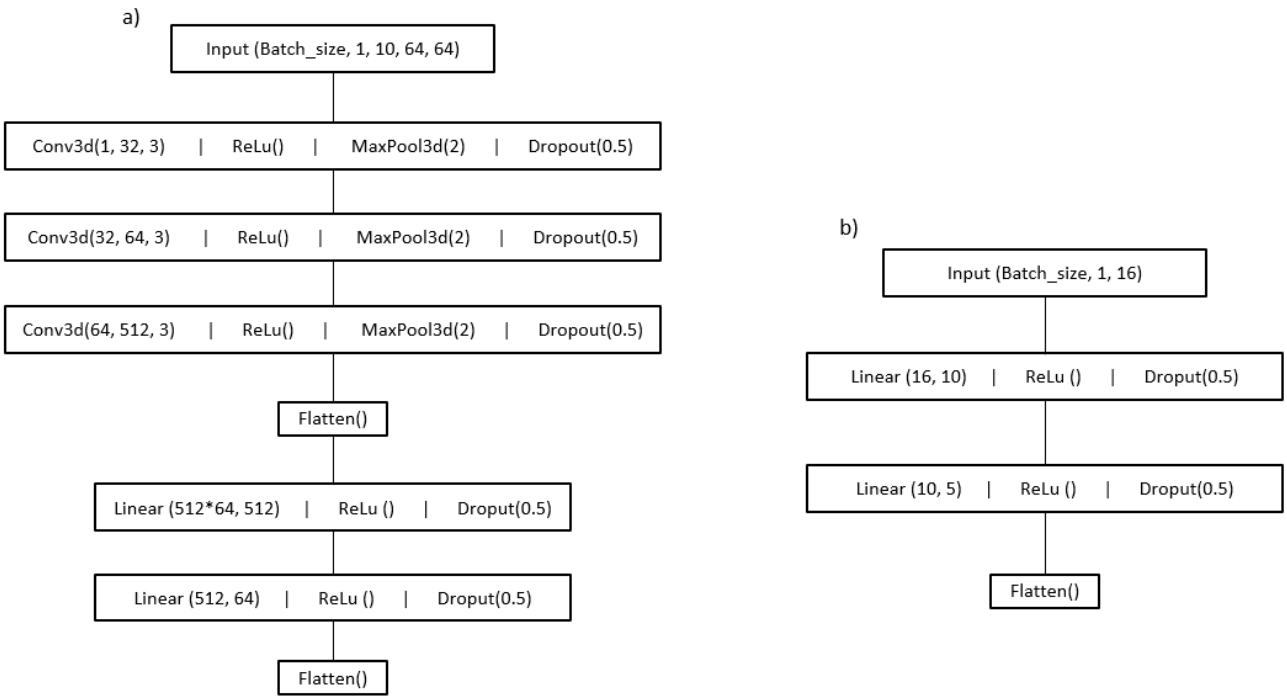


Figure 3: (a) 3D Convolutional Neural Network (3DCNN): to process CT scans and segmentations (b) Feed-Forward Neural Network (FNN): to handle RWD

Model evaluation

Both the deep learning models were trained and evaluated on an independent test set and on an external validation set. 50 epochs were used to train the 3DCNN and 100 epochs for the bimodal model. In order to evaluate the models' performances, Accuracy and Loss function were computed for all the epochs and the best model parameters were saved.

SHAP Explainability

At the end, in case of model with images only, local SHAP values were used to explain the model predictions. As regards deep learning, classification tasks in particular, since features are essentially pixels, model explainability helps to identify pixels which contribute negatively or positively to the predicted class [49]. After computing the SHAP values, the important pixels for the prediction are assigned colors: red pixels represent positive SHAP values that contributed to classify that image in class 1, while blue pixels represent negative SHAP values that contributed to classify that image in class 0.

3. Results

3.1. Dataset description

The patient cohort used in this study was derived from a larger database comprising a total of 556 patients with NSCLC treated with IO as any-line of therapy for advanced disease. Figure 4 illustrates the workflow of patients throughout the study. The initial training and test cohorts comprised 426 patients, while the initial external validation cohort consisted of 130 patients. The external validation set consisted of patients collected at the same institution (Fondazione IRCCS Istituto Nazionale dei Tumori), but at a later time.
Within these two main groups, certain patients had to be excluded from the analysis due to either the inability to evaluate their response to therapy or the absence of a target lesion.
In consultation with the clinicians, it was decided that only patients with a primary lung tumor would be included in the study, resulting in a final cohort size of 375 patients.
In particular, a total of 236 patients were used as the training set, while 59 patients were allocated to the test set. Additionally, 80 patients were reserved as an external validation set exclusively for the best-performing model.

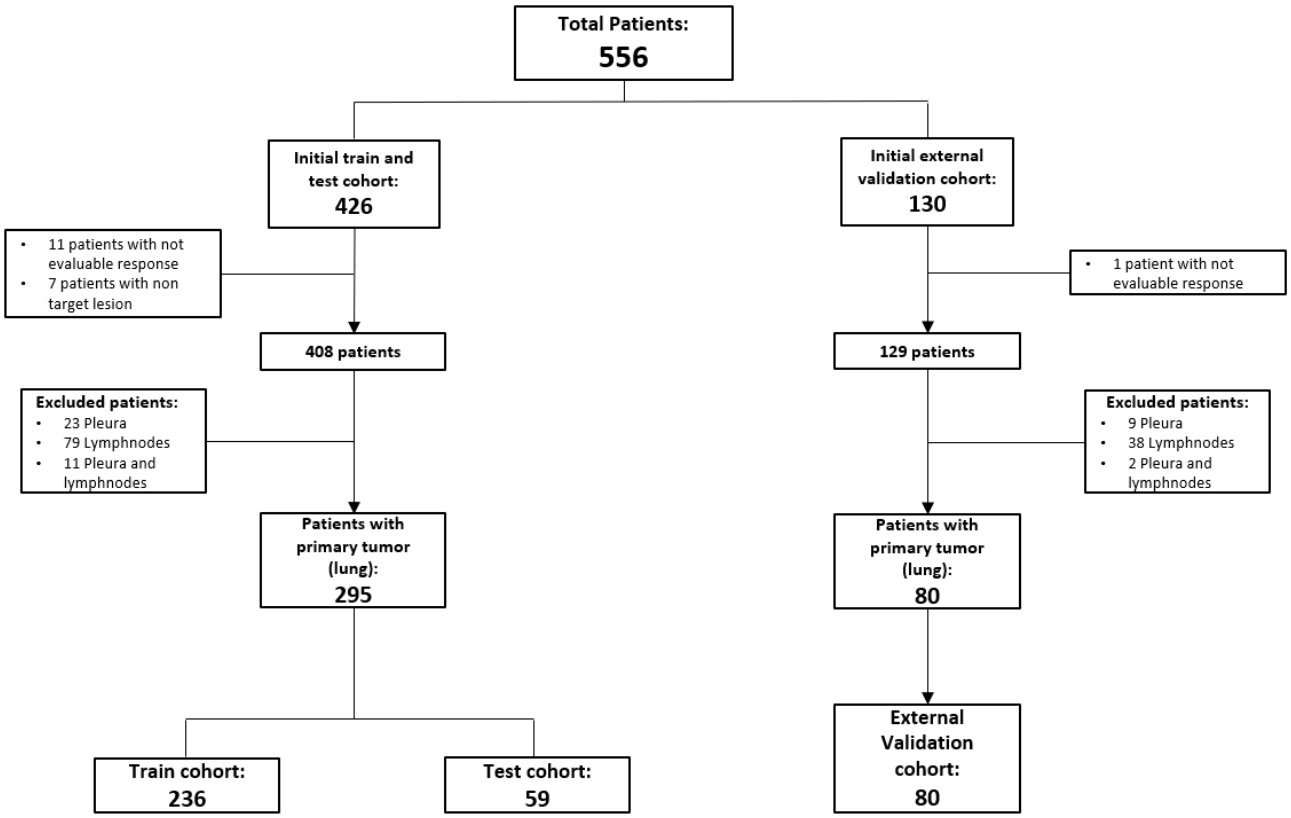


Figure 4: Patients subdivision: train, test and external validation cohorts

Patients included in the database exploited in this study shown a very high heterogeneity, including patients with different stages (IIIA, IIIB, IVA, IVB). Concerning the RWD, some values were missing. Specifically, 1% of smoking history, 1% of histology type, 11% of surgery, 4.5% of line of therapy, 33% of PDL1 group, 3% of ECOG PS, 8.5 % of tumor stage, 10% of node stage, 9% of metastases stage and 11% of number of metastatic sites were not possible to recover. Data imputation was applied to fill these missing values by modeling them as a function of other available data. The patients characteristics are summarized in Table 3. It should be noted that the subgroups of each characteristic may not provide the total count due to the presence of missing data.

Characteristic	Total patients (n = 375)	Training set (n = 236)	Test set (n = 59)	External Validation set (n = 80)
Age (mean ± SD)	67 ± 9,6	67,65 ± 9,6	65,1 ± 10,3	67 ± 9,6
Gender				
Male	228 (61%)	148 (63%)	33 (56%)	47 (59%)
Female	147 (39%)	88 (37%)	26 (44%)	33 (41%)
Outcome				
Class 1	197 (53%)	119 (50.5%)	30 (51%)	48 (60%)
Class 0	178 (47%)	117 (49.5%)	29 (49%)	32 (40%)
Therapy				
IO	289 (77%)	206 (87%)	51 (86%)	32 (40%)
IO-CT	86 (23%)	30 (13%)	8 (14%)	48 (60%)
Smoking History				
Yes	319 (85%)	208 (88%)	50 (85%)	61 (76%)
No	51 (14%)	27 (11%)	8 (14%)	16 (20%)
HistoType				
Squamous	73 (19%)	50 (21%)	9 (15%)	14 (18%)
Non Squamous	299 (80%)	187 (79%)	49 (83%)	63 (79%)
Surgery				
Yes	58 (15%)	45 (19%)	9 (15%)	4 (5%)
No	276 (74%)	190 (81%)	49 (83%)	37 (46%)
Lines of therapy				
0	21 (6%)	12 (5%)	4 (7%)	5 (6%)
1	195 (52%)	108 (46%)	23 (39%)	64 (80%)
2	90 (24%)	72 (31%)	15 (25%)	3 (4%)
3	34 (9%)	28 (12%)	5 (8%)	1 (1%)
4	9 (2%)	7 (3%)	1 (2%)	1 (1%)
5	4 (1%)	3 (1%)	1 (2%)	0 (0%)
6	2 (0.5%)	2 (1%)	0 (0%)	1 (1%)
7	2 (0.5%)	1 0.4%)	1 (2%)	0 (0%)
8	2 (0.5%)	2 (1%)	0 (0%)	0 (0%)
BMI (mean ± SD)	24,65 ± 4,32	24,69 ± 4,2	25,20 ± 4,31	24 ± 4,7
ECOG PS				
0	122 (33%)	87 (37%)	19 (32%)	16 (20%)
1	200 (53%)	122 (52%)	29 (49%)	49 (61%)
2	42 (11%)	27 (11%)	2 (3%)	13 (16%)
PDL1 group				
1	93 (25%)	60 (25%)	15 (25%)	18 (23%)
2	102 (27%)	72 (31%)	14 (24%)	16 (20%)
3	95 (25%)	56 (24%)	13 (22%)	26 (33%)

Table 3: Patients characteristics

3.2. Classification with Machine Learning pipeline

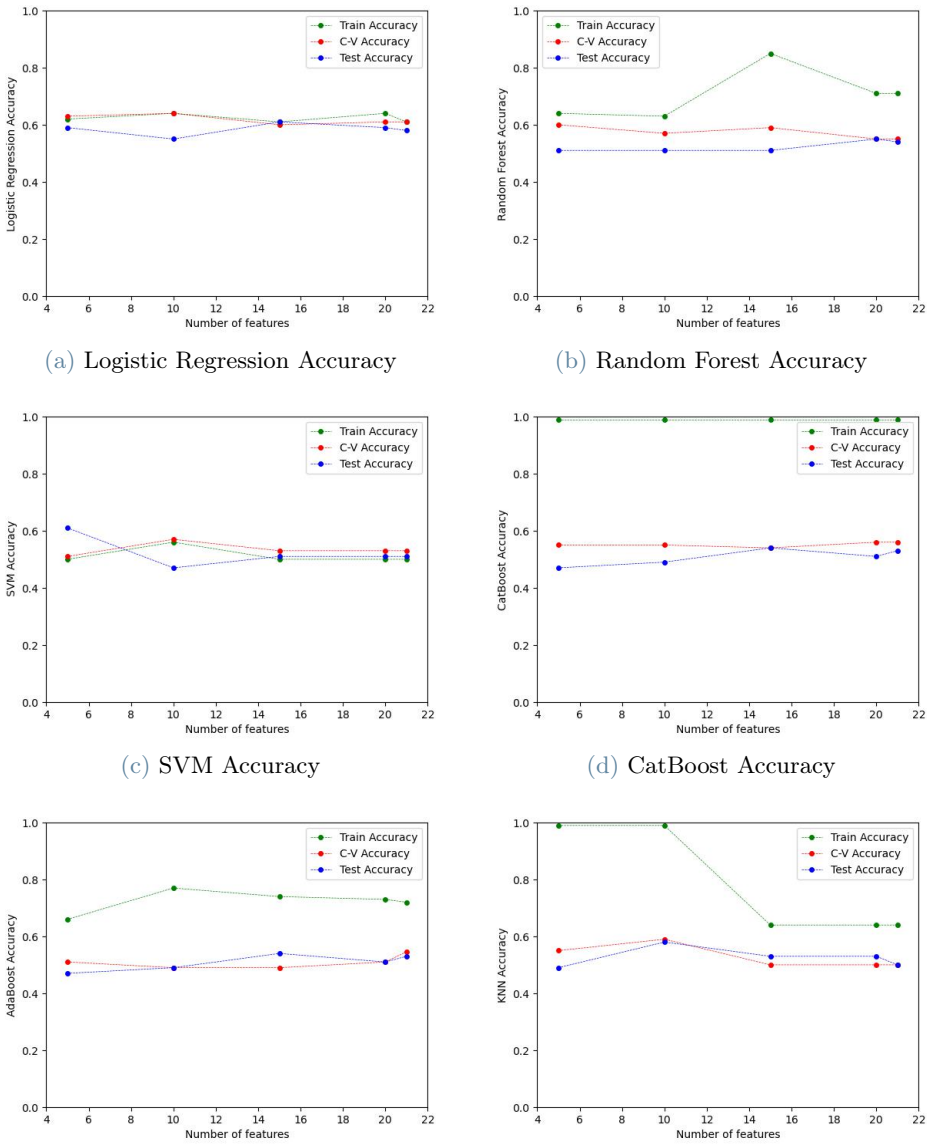
In this section, the results of each ML model are reported and divided according to the three different feature sets used in the study.

3.2.1 Radiomics

Feature Selection

To select the best set of features to feed the ML models, the first step involved checking for highly correlated features by displaying the correlation matrix (see Figure 20 in Appendix A.1). Highly correlated features were removed as they carry nearly identical information, rendering it redundant to include all of them in the model. Out of the initial 107 radiomic features, 77 were found to be highly correlated, leaving 30 remaining features. Subsequently, the remaining features were plotted based on the outcome, and those that did not differentiate between class 1 and class 0, as well as those with many outliers considered as noisy, were eliminated. After consultation with clinicians, a total of 9 radiomic features were removed, resulting in a set of 21 remaining radiomic features.

To determine the optimal number of features for each ML classifier, various values were tested by the MRMR feature selector, and their corresponding performances were calculated.
Specifically, in Fig. 5, accuracy performances were plotted for 5, 10, 15, 20, and 21 features, allowing for the selection of the values that yielded the best results. The performances were evaluated using training, cross-validation and testing, where training was mainly used in order to check if the model was overfitting. For each model, the number of features that yielded the best performances in terms of train accuracy, cross-validation (C-V) accuracy and test accuracy was selected.



(a) Logistic Regression Accuracy

(b) Random Forest Accuracy

(c) SVM Accuracy

(d) CatBoost Accuracy

(e) AdaBoost Accuracy

(f) KNN Accuracy

Figure 5: Radiomic features. MRMR feature selection: Accuracies on train, cross-validation and test set for different numbers of selected features: (a) Logistic Regression, (b) Random Forest (c) SVM (d) CatBoost (e) AdaBoost (f) KNN

In particular, by referring to Figure 5, 15 radiomic features were selected for LR (Fig. 5a), 20 for Random Forest (Fig. 5b), 10 for SVM (Fig. 5c), 21 for CatBoost (Fig. 5d) and AdaBoost (Fig. 5e), and 10 features for KNN (Fig. 5f).

Performances

Performances of all classifiers with optimal number of features are shown in Table 4. The LR classifier demonstrated the most promising results on radiomics dataset with an accuracy = 0.61 and AUC = 0.58.

Outcome	Model	Features	Class	N. class	Precision	Recall	F1	ACC	AUC
DCR	Logistic Regression	15	0	29	0.60	0.62	0.61	0.61	0.58
			1	30	0.62	0.60	0.61		
Class 0 (PD) 146 patients	Random Forest	20	0	29	0.57	0.69	0.62	0.59	0.57
			1	30	0.60	0.50	0.56		
Class 1 (SD+PR+CR) 149 patients	CatBoost	21	0	29	0.52	0.59	0.55	0.53	0.57
			1	30	0.54	0.47	0.50		
	KNN	10	0	29	0.57	0.55	0.56	0.57	0.58
			1	30	0.58	0.60	0.59		
	AdaBoost	21	0	29	0.53	0.59	0.56	0.54	0.50
			1	30	0.56	0.50	0.53		
	Support Vector Machine	10	0	29	0.46	0.38	0.42	0.47	0.49
			1	30	0.49	0.57	0.52		

Table 4: Radiomics: Performances of classification models on the test set

The corresponding confusion matrix, which is used to display the results according to predicted and actual values, is shown in Figure 23 in Appendix A.2. To assess the robustness and performance of this model, the external validation set was utilized. The performance metrics of the LR model on the external validation set are presented in Table 5.

Model	Class	N. class	Precision	Recall	F1-score	Accuracy	AUC
Logistic Regression	0	32	0.44	0.53	0.48	0.54	0.54
	1	48	0.63	0.54	0.58		

Table 5: Radiomics: Performance of LR on external validation set

Explainability analysis

In this section are outlined the results obtained applying the SHAP algorithm on the best performing model LR classifier by using 15 radiomic features for the classification of the CBR outcome. SHAP values are calculated on test set. Both the global and the local Explainability results were reported in the following plots, where Figure 6 shows how the features impact globally the predictions for all the patients, while in Figure 7 is shown how the features influenced four single instances: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

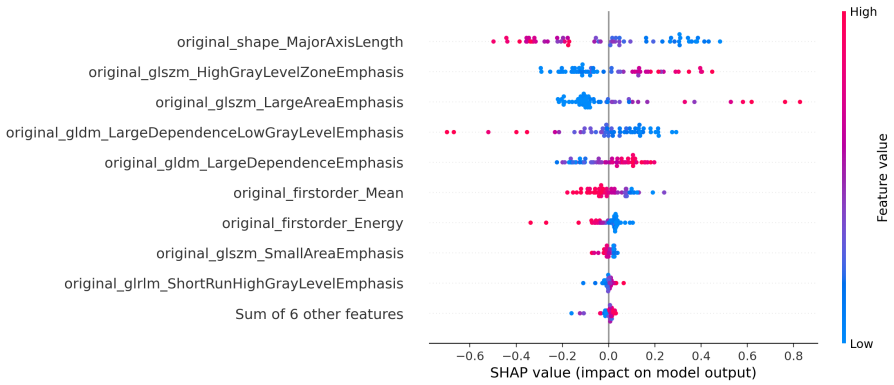
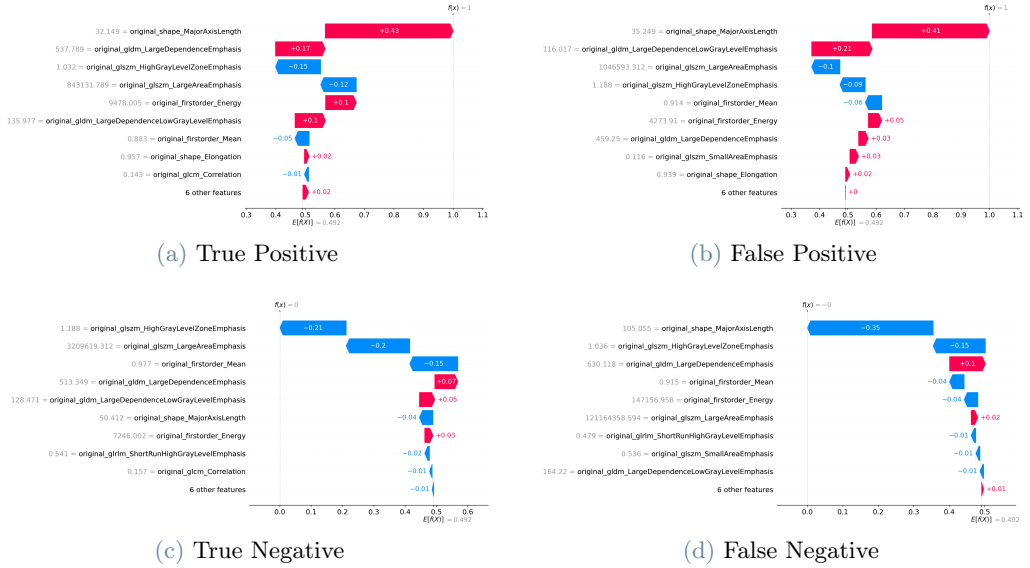


Figure 6: Radiomics: Global Explainability of LR on test set

Figure 6 represents the global explainability performed on test set predictions. It's evident that the feature that has highest influence in the model's outcome is Major Axis Length. It belongs to the shape class, where features are descriptors of the three-dimensional size and shape of the ROI and are independent from the gray level intensities distribution in the ROI. This kind of feature measures the largest axis length of the ROI-enclosing ellipsoid [34]. Specifically, based on the plot, lower values of this feature move the prediction towards class 1, meaning that they are associated with class 1, while higher values are more likely to be present in patients belonging to class 0.



(a) True Positive

(b) False Positive

(c) True Negative

(d) False Negative

Figure 7: Local Explainability with radiomics on the test set: (a) Class 1 patient correctly classified as such; (b) Class 0 patient incorrectly classified as class 1; (c) Class 0 patient correctly classified; (d) Class 1 patient incorrectly classified as class 0

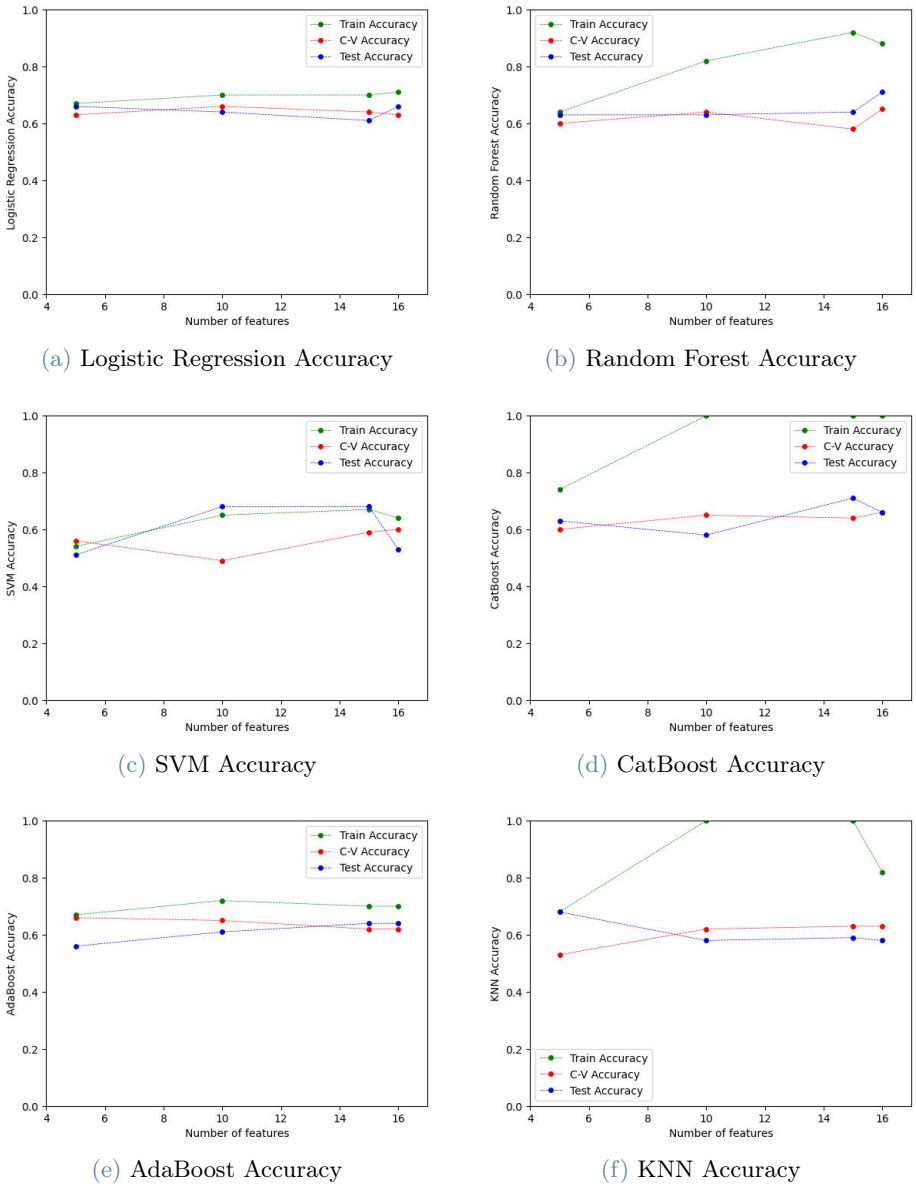
In Figure 7, four waterfall plots are depicted. Taking as examples True Positive (Fig. 7a) and False Positive (Fig. 7b) E[*f(x)*] = 0.492 indicates the average of the predicted outcomes, while *f(x)* is the predicted outcome (1 in both cases). Numbers on the bars represent the SHAP values. The sum of all the SHAP values is equal to the quantity E[*f(x)*] - *f(x)*. Also in this case the feature that has a highest influence on both the predictions is the Major Axis Length. More precisely, if the value it's low (mean is 60,38), which happens in both the predictions (32.149 in TP and 35.249 in FP) it is associated with class 1. This could be clarified by referring to Figure 7d, which explains a false negative prediction. In this case, the value of the same radiomic feature is high (105.055), and it influences the prediction towards class 0.

3.2.2 RWD

In this section, results obtained with RWD alone are presented.

Feature Selection

The correlation matrix demonstrated that none of the RWD were correlated with each other (see Fig. 22 in Appendix A.1). Therefore, all of them were utilized as input for the MRMR algorithm, without any additional manual selection before.
The optimal number of features among 5,10,15 and 16 for each model was determined by considering train, cross-validation, and test accuracies, as outlined in Figure 8:



(a) Logistic Regression Accuracy

(b) Random Forest Accuracy

(c) SVM Accuracy

(d) CatBoost Accuracy

(e) AdaBoost Accuracy

(f) KNN Accuracy

Figure 8: RWD. MRMR feature selection: Accuracies on train, cross-validation and test set for different numbers of selected features: (a) Logistic Regression, (b) Random Forest (c) SVM (d) CatBoost (e) AdaBoost (f) KNN

In case of RWD, 10 features were selected for LR (Fig. 8a), 16 for Random Forest (Fig. 8b), 15 for SVM (Fig. 8c), CatBoost (Fig. 8d) and AdaBoost (Fig. 8e), and 16 features for KNN (Fig. 8f).

Performances

Subsequently, each ML model was trained using the determined optimal number of features, and the corresponding performance metrics are presented in the table below:

Outcome	Model	Features	Class	N. class	Precision	Recall	F1	ACC	AUC
DCR	Logistic Regression	10	0	29	0.67	0.55	0.60	0.64	0.67
			1	30	0.63	0.73	0.68		
Class 0 (PD) 146 patients	Random Forest	16	0	29	0.64	0.62	0.63	0.64	0.71
			1	30	0.65	0.67	0.66		
Class 1 (SD+PR+CR) 149 patients	CatBoost	15	0	29	0.73	0.60	0.66	0.68	0.70
			1	30	0.67	0.77	0.72		
	KNN	16	0	29	0.58	0.48	0.53	0.58	0.65
			1	30	0.57	0.67	0.62		
	AdaBoost	15	0	29	0.61	0.79	0.69	0.64	0.69
			1	30	0.71	0.50	0.59		
	Support Vector Machine	15	0	29	0.65	0.76	0.70	0.68	0.71
			1	30	0.72	0.60	0.65		

Table 6: RWD: Performance of classification models on the test dataset

SVM fed with 15 features, resulted as the best performing model with an accuracy = 0.68 and AUC = 0.71. The SVM confusion matrix is represented in Figure 24 in Appendix A.2. Then, it was tested on the external validation set, bringing to the results presented in Table 7:

Model	Class	N. class	Precision	Recall	F1-score	Accuracy	AUC
Support Vector Machine	0	32	0.67	0.50	0.57	0.7	0.71
	1	48	0.71	0.83	0.77		

Table 7: RWD: Performance of SVM on the external validation set

Explainability analysis

As done before, SHAP values were calculated to compute the local and global explanation of best performing model SVM on the test set.

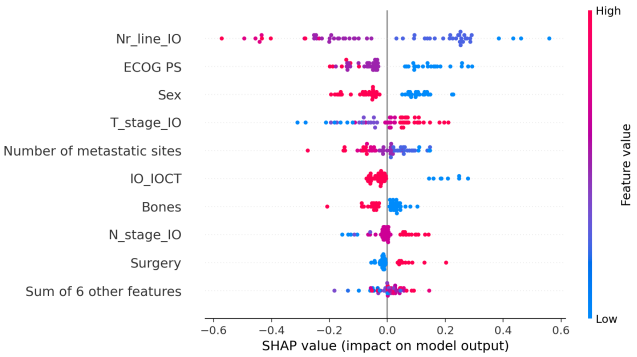
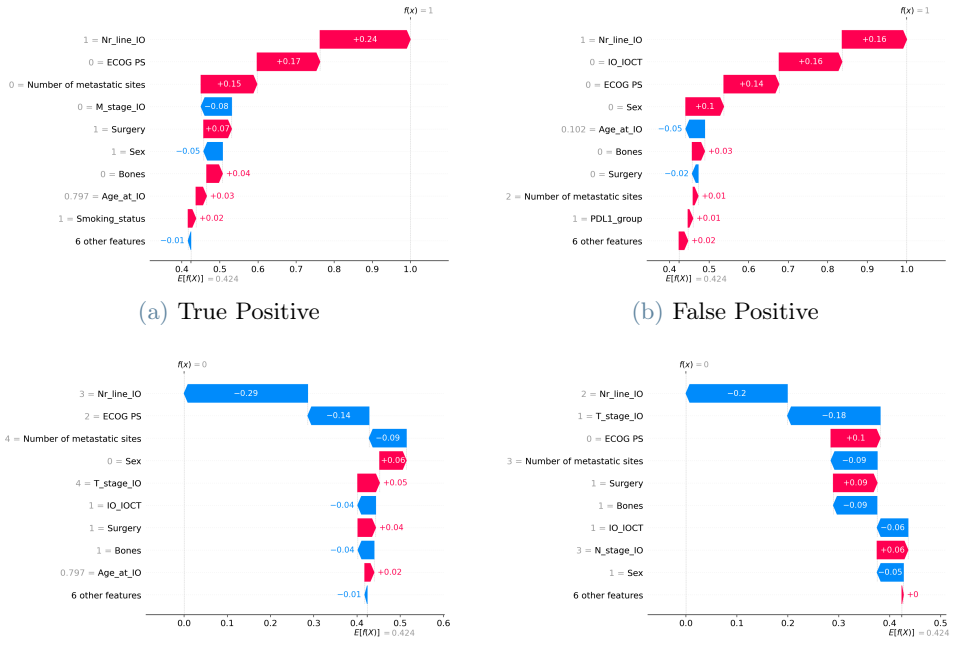


Figure 9: RWD: Global Explainability of SVM on test set

Figure 9, in which is shown the global explanation, the RWD that has the highest influence in the model's outcome is Line of therapy. Specifically, lower values of this data move the prediction towards class 1, while higher values are more likely to impact "negatively" on the model, meaning that they are associated with class 0.



(a) True Positive

(b) False Positive

(c) True Negative

(d) False Negative

Figure 10: Local Explainability with RWD on the test set: (a) Class 1 patient correctly classified as such; (b) Class 0 patient incorrectly classified as class 1; (c) Class 0 patient correctly classified; (d) Class 1 patient incorrectly classified as class 0

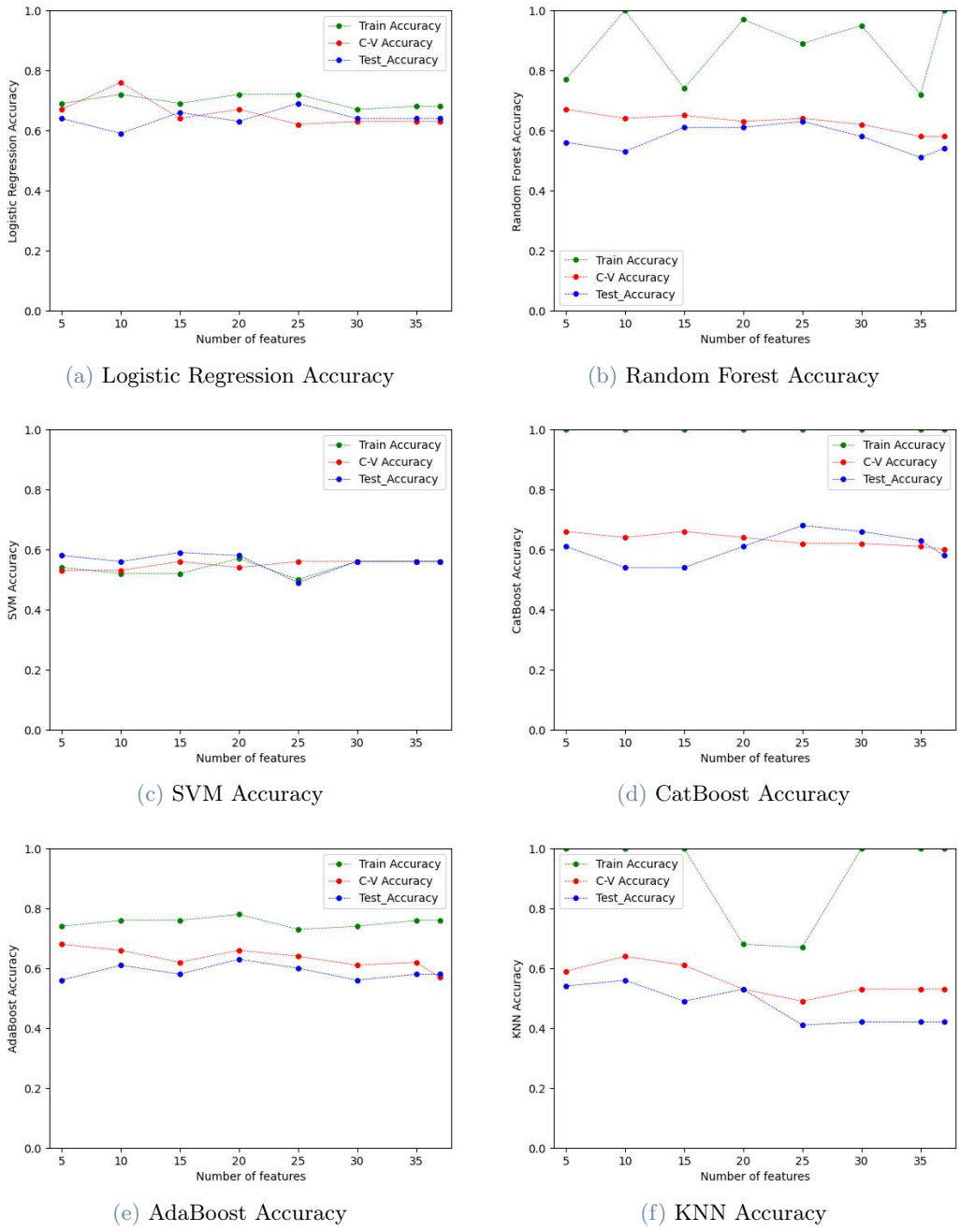
The four waterfall plots presented in Figure 10 depict predictions for True Positive, True Negative, False Positive, and False Negative. Taking TN (Fig. 10c) and FN (Fig. 10d) as examples, the Line of therapy is equal to 3 and 2, respectively. This implies that the TN patient received therapy in the third line, while the FN patient received therapy in the second line. In both cases, this feature contributes to predict class 0. However, for TN, a higher value of Line of Therapy (3) results in a larger negative SHAP effect (-0.29), whereas for FN, with a lower line of therapy (2), the negative SHAP effect diminishes (-0.2).

3.2.3 Radiomics and RWD combination

In this section, results achieved from the combination of radiomics and RWD are shown.

Feature Selection

Correlation Matrix containing both radiomics and RWD is shown in Figure 21 in Appendix A.1. The matrix revealed no correlation between the RWD either between each other or with the radiomics.
Same 77 correlated features were found, and the same set of 21 radiomic features, obtained after applying the procedure described in Section 3.2.1 to remove highly correlated, poor informative and noisy features, was utilized.
The 16 baseline RWD were included in this dataset, resulting in a dataset composed of 37 features. The MRMR feature selector was then applied to select the optimal number of features for each model. Specifically, the accuracy of all six ML models on cross-validation, training, and testing datasets was plotted for 5, 10, 15, 20, 25, 30, 35, and 37 features. This allowed to determine the optimal number of features that achieved the highest performances for each model.



(a) Logistic Regression Accuracy

(b) Random Forest Accuracy

(c) SVM Accuracy

(d) CatBoost Accuracy

(e) AdaBoost Accuracy

(f) KNN Accuracy

Figure 11: Radiomics and RWD. MRMR feature selection: Accuracies on train, cross-validation and test set for different numbers of selected features: (a) Logistic Regression, (b) Random Forest (c) SVM (d) CatBoost (e) AdaBoost (f) KNN

By referring to accuracies plotted in Figure 11, 25 features were selected for LR (Fig. 11a) and Random Forest (Fig. 11b), 10 for SVM (Fig. 11c), 25 for CatBoost (Fig. 11d), 20 for AdaBoost (Fig. 11e) and 10 features for KNN (Fig. 11f).

Performances

The evaluation metrics on test set for the six ML models with optimal number of radiomics and RWD are shown in the Table 8.

Outcome	Model	Features	Class	N. class	Precision	Recall	F1	ACC	AUC
DCR	Logistic Regression	25	0	29	0.72	0.62	0.67	0.69	0.73
			1	30	0.68	0.77	0.72		
Class 0									
(PD)	Random Forest	25	0	29	0.61	0.66	0.63	0.63	0.66
146 patients			1	30	0.64	0.60	0.62		
	CatBoost	25	0	29	0.67	0.69	0.68	0.68	0.72
Class 1									
			1	30	0.69	0.67	0.68		
(SD+PR+CR)									
149	KNN	10	0	29	0.55	0.55	0.55	0.56	0.58
patients			1	30	0.57	0.57	0.57		
	AdaBoost	20	0	29	0.61	0.69	0.65	0.63	0.70
			1	30	0.65	0.57	0.61		
	Support Vector Machine	10	0	29	0.57	0.41	0.48	0.56	0.55
			1	30	0.55	0.70	0.62		

Table 8: Radiomics and RWD: Performance of classification models on the test dataset

As for radiomic features dataset, the LR classifier performed better than the other models, achieving accuracy = 0.69 and AUC = 0.73 with 25 features. The confusion matrix is shown in Figure 25 in Appendix A.2. LR was tested on the external validation set and the results are presented in Table 9.

Model	Class	N. class	Precision	Recall	F1-score	Accuracy	AUC
Logistic Regression	0	32	0.62	0.56	0.59	0.69	0.71
	1	48	0.73	0.77	0.75		

Table 9: Radiomics and RWD: Performance of LR on external validation set

Explainability analysis

The features coming from the MRMR feature selection process are listed in Table 10.

	N	Feature Name
Radiomic Features	**11**	• original_gldm_LargeDependenceLowGrayLevelEmphasis • original_firstorder_Mean • original_glszm_GrayLevelNonUniformityNormalized • original_glszm_SmallAreaEmphasis • original_shape_MajorAxisLength • original_gldm_LargeDependenceEmphasis • original_shape_Elongation • original_glrlm_ShortRunHighGrayLevelEmphasis • original_firstorder_Energy • original_shape_Sphericity • original_firstorder_90Percentile
RW Features	**14**	• ECOG PS • IO/IOCT • Number of metastatic sites • M_stage_IO • Sex • Nr_line_IO • BMI_IO_Baseline • T_stage_IO • Bones • N_stage_IO • Brain • PDL1_group • Age_at_IO • Smokin_status

Table 10: List of radiomics and RWD selected for LR

The SHAP values for LR fed with 25 features were computed on the test set. Global SHAP (see Fig. 12) revealed that the two features that most strongly influenced the predictions were first ECOG PS (RWD) and Large Dependence Emphasis (LDE), a radiomic feature. In the case of ECOG PS, higher values of this feature shifted the predictions towards class 1, which is confirmation of what is assessed by clinical practice, as a lower value on the ECOG PS scale indicates better patient clinical condition [50].
Large Dependence Emphasis belongs to the Gray Level Dependence Matrix (GLDM) class, wherein the features quantify the number of connected voxels within a distance δ that depend on the center voxel. A neighbouring voxel with a gray level j is considered dependent on center voxel with gray level i if $|i - j| \leq \alpha$ [34].
LDE can be translated into a measure of the texture of the lesion, where higher values indicate a more homogeneous texture. The plot in Figure 12, reveals that high values for LDE are correlated with class 1, suggesting that a more homogeneous texture is most likely to correspond to a class 1 patient.

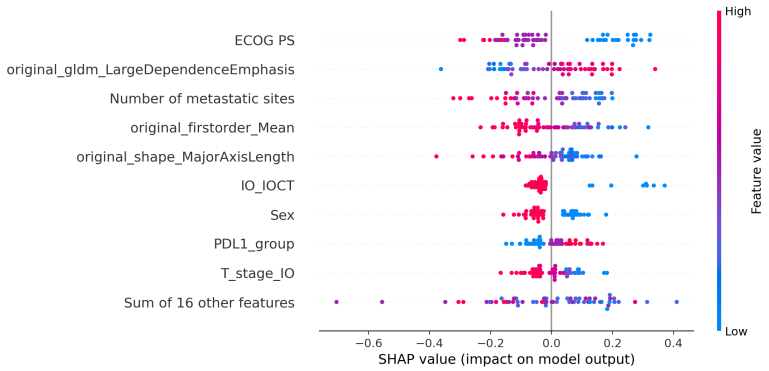


Figure 12: Radiomics and RWD: Global explainability of LR on test set

In order to understand how the features impact single predictions, waterfall plots (Fig. 13) for the local explainability were computed, by choosing again one TP, one TN, one FP and one FN.
Taking Figure 13c as an example, the feature that has the most significant impact on the prediction is ECOG PS, which has a value of 2 and it brings the prediction towards class 0. In this case, the patient was correctly classified as belonging to class 0. The same trend could be seen for the number of metastatic sites, which is also

(a) True Positive

(b) False Positive

(c) True Negative

(d) False Negative

Figure 13: Local Explainability with radiomics and RWD on test set: (a) Class 1 patient correctly classified as such; (b) Class 0 patient incorrectly classified as class 1; (c) Class 0 patient correctly classified; (d) Class 1 patient incorrectly classified as class 0

high in this case and it moves prediction towards class 0.
Moving on Figure 13d, it demonstrates that the two primary features influencing the prediction towards class 0 are ECOG PS and Large Dependence Emphasis (LDE). The incorrect prediction is reasonable considering that ECOG PS (1) is relatively high (ranging from 0 to 2 in the present study), while LDE is low (below the mean value of 426.3). As clarified by the plot in Figure 12, high values of ECOG PS and low values of LDE are correlated with class 0, which further justifies the erroneous prediction.

3.3. Classification with the Deep Learning pipeline

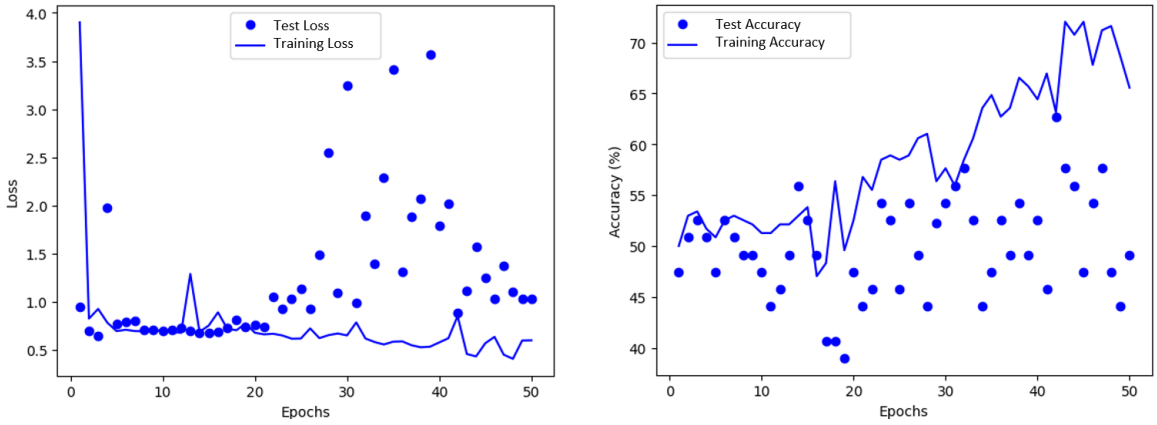
For the training of DL models, Pytorch 1.13.1 library was used [51]. Cross Entropy Loss [52] was used in the pipeline with only CT scans, while Binary Cross-Entropy with logits Loss (BCEWithLogitsLoss) [53] for bimodal model. The loss function is used to quantify the model's error by assessing the disparity between the predicted output and the true target value. For both the solutions, Adaptive Moment Estimation (Adam) optimizer was employed with a learning rate equal to 0.005. Adam optimizer is really used in deep learning problems since it is fast, efficient and it requires little memory [54]. The method computes adaptive learning rates for different parameters from estimates of first and second moments of the gradients and its primary goal is to minimize the loss function.
The results obtained from the end-to-end deep learning solutions are categorized into the two feature sets used: DL features and combination of DL features and RWD.

3.3.1 DL features

Cross Entropy Loss is applied. Cross Entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1. Cross Entropy loss increases as the predicted probability diverges from the actual label.

Performances

CT scans and the corresponding segmentations were given as input to the 3DCNN. The plots showing training and test accuracy and loss function are reported in Figure 14.
Based on both metrics, overfitting is present, which means that the model performs well during training but performs poorly on test data, which represents unseen data. In case of accuracy, the majority of epochs show



(a) Train and Test Loss (b) Train and Test Accuracy

Figure 14: DL features: Performance on training and testing: (a) Loss function (b) Accuracy

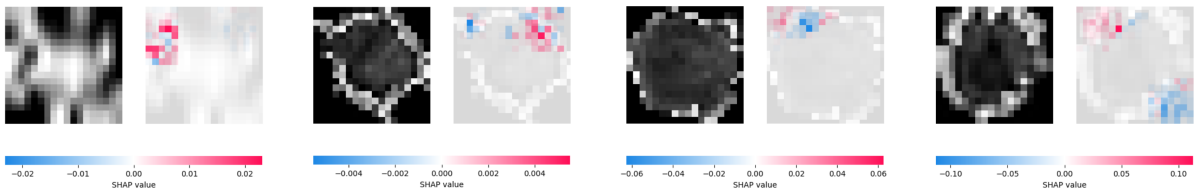
significantly lower test accuracy compared to the training accuracy. The train loss demonstrates a consistent pattern of decreasing as the number of epochs increases. However, the test loss function exhibits an irregular behavior, increasing instead of decreasing from left to right. The best value for test accuracy is 0.63, reached at epoch 42, where the test loss is 0.887. In the same epoch, train accuracy is 0.63 and train loss is 0.852. As done for ML pipeline, the external validation set was used in order to test the robustness and performance of the model. The results on the external validation set are collected in Table 11.

Model	Correct positive	Correct negative	Accuracy	Loss
3DCNN	27/48	17/32	0.55	0.913

Table 11: 3DCNN: Performance on external validation set in the best epoch

Explainability analysis

Local explanation on the test set was carried out using the SHAP algorithm, showcasing four predictions: TP, TN, FP, and FN (Figure 15). Specifically, two images are presented for each prediction. The left image displays one slice of the CT scan. In the slice, only segmentation is present, as the model resized the original CT scan by zooming in on the ROI region. The right image represents the pixel-level explanation of the corresponding slice.



(a) True Positive (b) False Positive (c) True Negative (d) False Negative

Figure 15: Explainability with DL features on test set: (a) Class 1 patient correctly classified as such; (b) Class 0 patient incorrectly classified as class 1; (c) Class 0 patient correctly classified; (d) Class 1 patient incorrectly classified as class 0

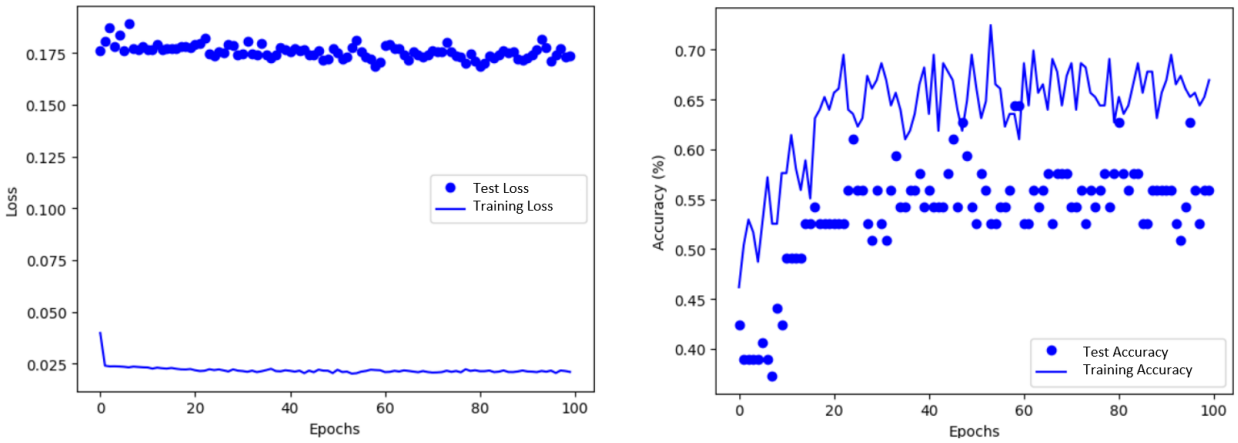
In SHAP explanations, pixels that contribute to a specific prediction are assigned colors. Specifically, red pixels influence the prediction towards class 1, while blue pixels influence the prediction towards class 0. Figure 15a (TP) and Figure 15b (FP) provide examples. In both explanations, a significant number of red pixels can be observed. In the case of TP, the region of the ROI highlighted in red leads to the correct prediction, whereas in the case of FP, the highlighted pixels contribute to an incorrect prediction.

3.3.2 DL features and RWD

Binary Cross-Entropy with logits Loss (BCEWithLogitsLoss), commonly used for binary classification problems, was utilized [53]. This kind of loss combines a Sigmoid layer and the Binary Cross Entropy Loss (BCELoss) in one single class. This version is more numerically stable than using a Sigmoid layer followed by a BCELoss as, by combining the operations into one layer.

Performances

By adding RWD, training and test performances are plotted in Figure 16:



(a) Train and Test Loss

(b) Train and Test Accuracy

Figure 16: DL features and RWD: Performance on training and testing: (a) Loss function (b) Accuracy

In Figure 16a, the loss behavior is illustrated. The training loss, again, follows a typical trend, decreasing from the initial epochs as the number of epochs increases. However, the test loss indicates that the model may not be learning effectively, as the loss value remains relatively stable without decreasing as seen in training.
In the accuracy plot (Fig. 16b), both the training and test accuracies show an increasing trend from left to right. For the test set, the highest accuracy of 0.64 is achieved at epoch 60, where test loss is equal to 0.170. In the same epoch, the train accuracy is equal to 0.64, while train loss is 0.022.
After these results were computed and the best model parameters were saved, the same model was tested on the external validation set and the following results were reached:

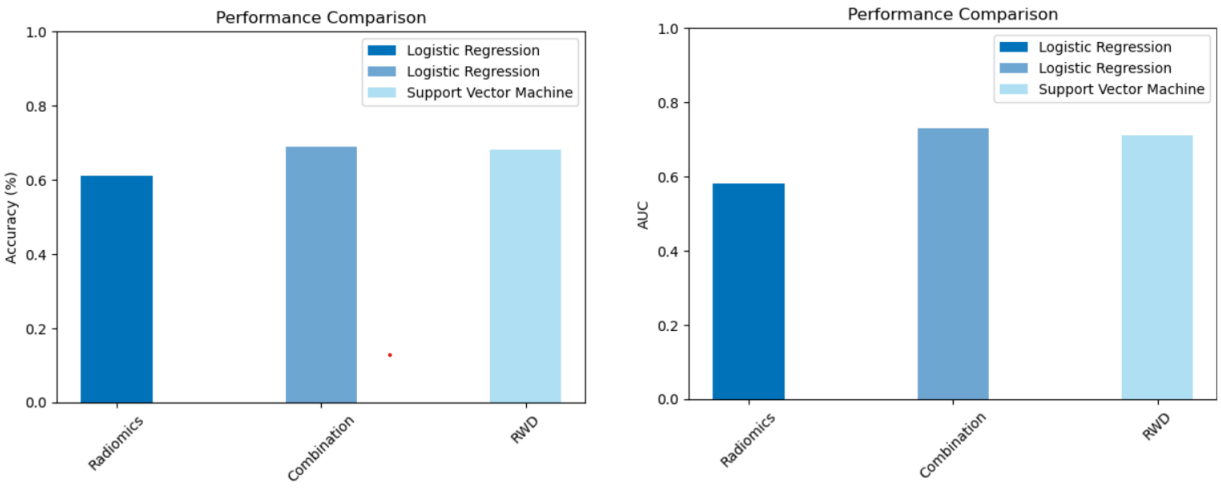
Model	Correct positive	Correct negative	Accuracy	Loss
Multimodal Model	39/48	13/32	0.65	0.119

Table 12: Bimodal model: Performance on external validation set in the best epoch

4. Discussion

4.1. Machine Learning: Response Outcome Results

One of the objectives of this study was to evaluate the predictive capability of radiomic features, either alone or in combination with Real World Data (RWD), for determining the response to IO in patients with advanced NSCLC. The Clinical Benefit Rate (CBR) was chosen as the outcome measure to predict, and three different feature sets were utilized to investigate this purpose: radiomic features, RWD and a combination of radiomics and RWD. Six different machine learning classifiers were trained to predict the CBR outcome for each of the three feature sets. The best performing model was selected for each feature set: LR was found to be the best performing model for both the radiomics and combination feature sets, while SVM was chosen as the best performing model when only RWD were considered.
Figure 17b illustrates the performance of these best performing ML models for each feature set, as measured by accuracy and AUC metrics.



(a) Accuracy (b) AUC

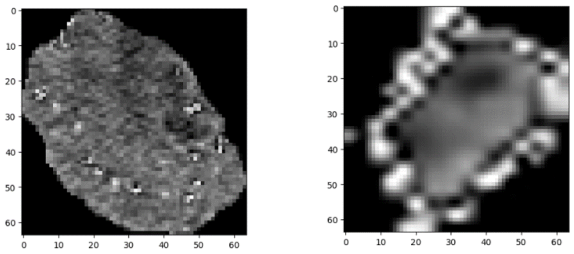
Figure 17: ML performance comparison between three feature sets: radiomics, RWD and radiomics combination, RWD

From figures it evident that in ML, radiomics alone were not efficient in the classification (accuracy = 0.61), while RWD demonstrated greater robustness and efficiency in predicting therapy response (accuracy = 0.68). However, the results of this study revealed that the addition of radiomics to RWD did not add significantly more information compared to a prediction model based solely on RWD. In fact, accuracy reached with the combination was 0.69, while RWD got 0.68. These findings are consistent with other studies, as explored by Peisen et al. [55], where baseline CTs did not add additional information to the prediction of response after three months, compared to a prediction model based on baseline clinical parameters alone. To provide a broader perspective, the AUC metric is plotted (see Fig.17b), confirming the low predictive power of radiomics alone (AUC = 0.58), while a similar performance of RWD and radiomics combination (AUC = 0.73) and RWD (AUC = 0.71). Results in terms of AUC are useful for comparing the present study with other works found in the literature, such as the work by Yu et al. [56], where the combination of RWD and radiomics achieved an AUC of 0.81. The results in the present study are slightly lower but still comparable. Furthermore, similar findings were reported, as the combination of RWD and radiomics exhibited higher predictive power compared to radiomics alone.

4.2. Explainability analysis Results

In order to provide a comprehensive explanation of how both radiomics and RWD are utilized to predict the response in ML, SHAP Summary Plot obtained from the LR model trained on the combination of radiomics and RWD (refer to Fig. 12 in Section 3.2.3) are considered. Firstly, it is important to note that the model achieved an accuracy of 0.69 and an AUC of 0.73. These results indicate that further work is required to improve the model's performance and enhance its reliability. Although the AUC of 0.73 is considered acceptable, it does not reach the level of excellence, particularly for medical applications, where AUC should be higher than 0.8 [57]. Given that the model's performance is not outstanding, the reliability of its interpretability may be compromised. As a result, the interpretability of the model may exhibit bias towards certain features and their corresponding value explanations, which may not be entirely reliable.
As already discussed in Section 3.2.3, the RWD ECOG PS emerged as the most influential in making predictions with the combination of RWD and radiomics. Additionally, the treatment received by the patients also provides valuable information. Notably, the IO_IOCT indicates that patients who received a combination of chemotherapy and IO (coded as IO_IOCT = 0) have a higher probability of positive response outcomes. This is inline with clinical knowledge, since there are many studies that confirm that anti-PD-(L)1 antibody combined with chemotherapy is more efficient than IO monotherapy for advanced NSCLC patients [58] [59]. The SHAP Summary Plot reveals interesting insights into the radiomic features. For instance, the second feature, Large Dependence Emphasis (LDE), is considered as an example. A higher value of LDE feature is associated with a more homogeneous texture in terms of radiomic meaning. Homogeneity of the texture can be further supported by visualizing the regions of interest (ROI) in two CT scans: one corresponding to a high value of LDE and the other corresponding to a low LDE. To illustrate this, ROI images are presented in Figure 18. The left image (Fig. 18a) represents the CT scan with a high LDE value, while the right image (Fig. 18b) corresponds to the

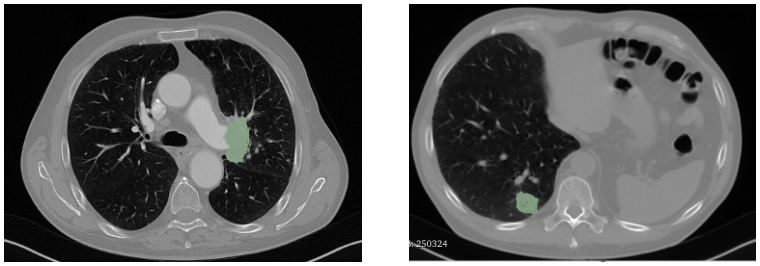
CT scan with a low LDE value. By comparing these images, it becomes evident that the high LDE value is associated with a more homogeneous texture, as indicated by the visually consistent appearance of the ROI. In contrast, the low LDE value is indicative of a more heterogeneous texture, as observed by the presence of pixel variations and irregularities within the ROI. This visual demonstration reinforces the relationship between the LDE feature and the homogeneity of the texture in the CT scans.



(a) LDE = 581.914 (b) LDE = 257.381

Figure 18: Examples of two CT scans: (a) High value of Large Dependence Emphasis (b) Low value of Large Dependence Emphasis

Another interesting feature to analyze is the Major Axis Length, which measures the largest axis of the ellipsoid containing the ROI. By comparing two CT scans, the change in value can be easily assessed. In Figure 19a, the image shows a larger tumor area (green zone), and as expected, the Major Axis Length value is higher. Conversely, the Figure 19b depicts a smaller tumor region, which is confirmed by the lower value of this radiomic feature.



(a) MajorAxisLength = 80.663 (b) MajorAxisLength = 30.327

Figure 19: Examples of two CT scans: (a) High value of Major Axis Length (b) Low value of Major Axis Length

In the DL pipeline, SHAP values provided initial insights into the functioning of neural networks. However, further analysis is necessary to fully comprehend which tumor characteristics contribute to specific predictions. From figure 15 in Section 3.3.1, it seems that the network primarily focuses on the edges of the ROI. In all four explanations, the colored pixels are concentrated along the edges, while the internal region of the ROI does not significantly influence the prediction.

4.3. Comparison between ML and DL pipelines

One of the objectives of this study was to evaluate the performance of ML and DL techniques (Table 13) and determine which approach could be superior in predicting IO response in this clinical context. Since the dataset was balanced with respect to classes in the test set, the performances can be compared based on test accuracy results.
When considering models that uses features derived from CT scans, DL (acc = 0.63) on the test set demonstrates slightly higher efficiency compared to ML (acc = 0.61). However, when incorporating both RWD and features derived from CT scans, ML technique (acc = 0.69) outperforms DL approach (acc = 0.64). Furthermore, when combining CT scans features with RWD, ML techniques exhibit an increase in their predictive performance. Indeed acc = 0.61 is achieved with radiomics only, while acc = 0.69 with the combination. However, this does not happen for DL, where performance achieved using CT scan features (acc = 0.63) and performance with the combination (acc = 0.64) can be considered comparable. Considering external validation, more significant improvement when adding RWD to CT scan features is observed.

Data modality	Machine Learning Accuracy		Deep Learning Accuracy	
	Test	Validation	Test	Validation
CT scan features	0.61	0.54	0.63	0.55
RWD	0.68	0.7	No	
CT scan features + RWD	0.69	0.69	0.64	0.65

Table 13: Machine Learning and Deep Learning performance comparison

4.4. Limitations and Future research

Finally, none of the ML and DL approaches with the present data types yielded satisfactory results that would allow the model to be applied in possible clinical practice. There could be several reasons for this. Firstly the population included an heterogeneous cohort of patients treated with IO in a wide range of time (2013 – 2023), when different CT image acquisition protocol were applied. In addition, CT scan exams were performed at different Institutions. These two considerations could have produced some intrinsic noise during the feature extraction. In both ML and DL pipelines, including wider and more homogeneous cohort of patients to the current dataset would likely improve performance, particularly for DL methods, which benefit from large volumes of data to effectively learn patterns. Secondly, the number of features utilized could be expanded. The current image pre-processing and feature extraction and selection methods may not be the optimal solutions for this type of problem. In the ML approach, it is important to note that no specific filter was applied to extract the features. However, a broader range of features could be extracted by employing various types of filters, such as Wavelet filters, Laplacian of Gaussian filters, and logarithmic filters. These filters are commonly used in the field and have the potential to capture different aspects of the underlying data [31] [60]. his approach would enable the inclusion of a larger quantity of radiomic features that could be explored in terms of their predictive capabilities. Additionally, conducting a more precise analysis of highly correlated features would be crucial to avoid excluding important features that are associated with the outcome. For both ML and DL approaches, the implementation of additional image preprocessing techniques can enhance the quality of input images and contribute to more accurate feature extraction. The third main limitation of the present study is that it is not sufficient to apply ML and DL solely for predicting response outcomes. Survival outcomes, such as progression-free survival (PFS) and overall survival (OS), should also be considered to gain a deeper understanding of the problem. Additionally, the combination of RWD and CT scans with other data types, since they are more relevant clinical outcomes. This is supported by numerous studies in the literature, including the work by Bohem et al. **??**, which demonstrated the improvements achieved by integrating features from different data modalities (histopathological, radiomic, genomic, and clinical data) into multimodal models. Another potential solution, considering the data types used in the present study, is the utilization of delta-radiomic features. These features represent the differences between radiomics extracted from two CT scans: the baseline CT scan and the post-treatment CT scan **??**. Incorporating delta-radiomic features may provide valuable information on the changes in radiomic characteristics over the course of treatment, potentially improving the predictive capabilities of the models. Last main problem concerns DL approach, where preliminary were found. To improve performance, transfer learning techniques could be employed. Transfer learning is a commonly used technique to improve generalization in the current task by leveraging knowledge gained from previous tasks and datasets. In practice, models are initially trained on a different dataset that is unrelated to the current task. The model's weights are saved, and then are used in the current model to avoid starting from random initializations. This approach allows the model to benefit from the learned previous task, potentially enhancing performance on the current task. Another potential approach for improvement is the exploration of different neural network architectures.

5. Conclusions

The objective of this study was to identify radiomic and clinical biomarkers associated with IO benefit in a cohort of patients with NSCLC. With clinical and radiological data collected at Istituto Nazione dei Tumori di Milano, both ML and DL pipelines were developed. The findings suggest that incorporating clinical evidence with quantitative information extracted from medical images can improve the predictive performance of the outcome. However, this improvement is observed only in the ML solution. It is worth noting that in the

DL pipeline, the addition of RWD does not lead to a significant enhancement in performance in the test set. Considering external validation, in DL more significant improvement when adding RWD to CT scan features is observed. Furthermore, the main findings of the present study indicate that the ML solution outperforms the DL approach when both data modalities are used. On the other hand, the DL model exhibits a slightly superior performance compared to the ML model with CT scans features only. Medical applications require very high reliability and performances in order to be applied in clinical practice. These initial achievements could be the base of the ultimate and ambitious goal of developing novel tools for selection of ideal candidates for IO. So by investigating and incorporating future perspectives, this research may have the potential to contribute to the development of innovative approaches that can be applied in clinical practice.

A. Appendix

A.1. Correlation Matrices

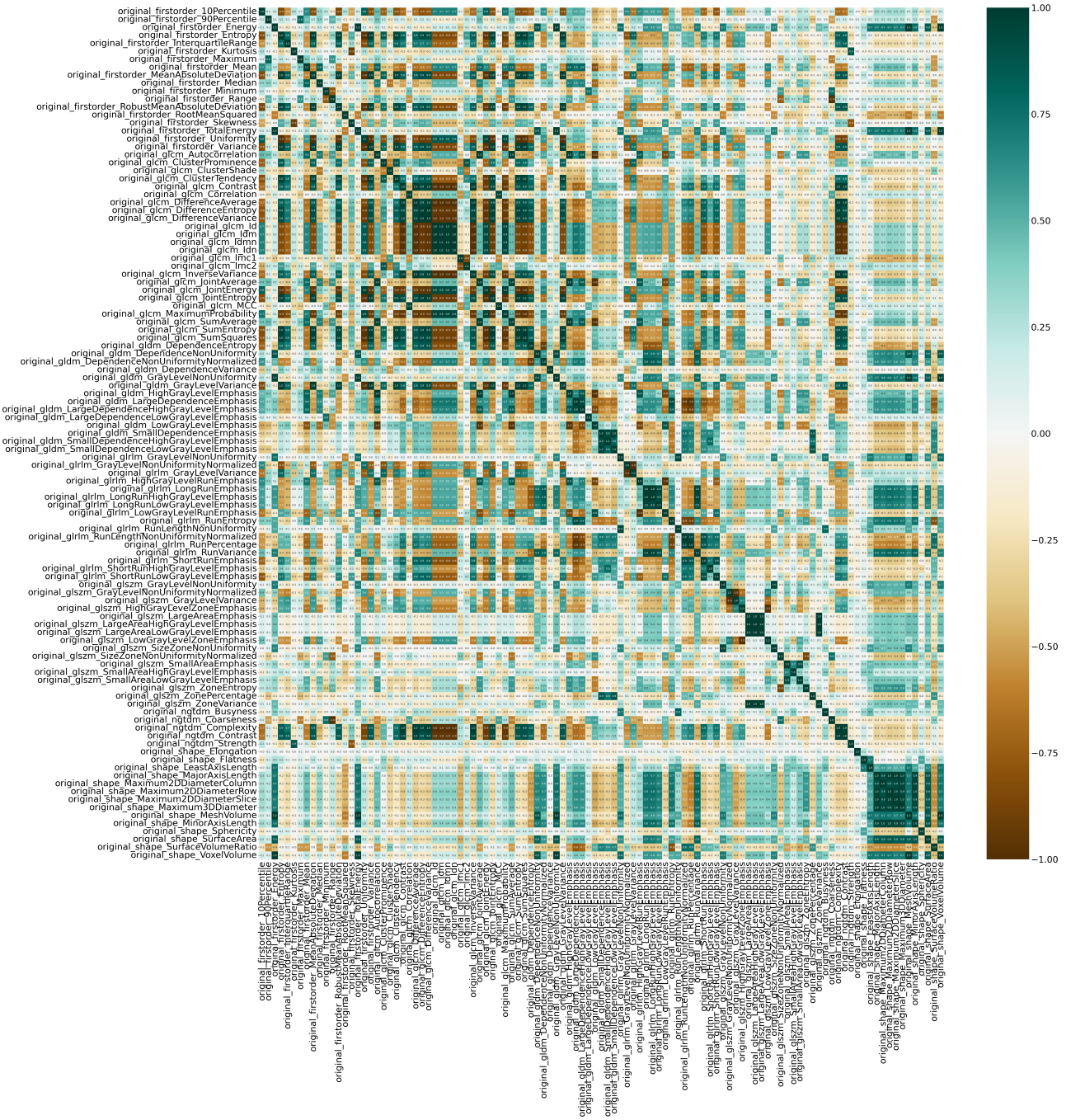


Figure 20: Radiomic features Correlation Matrix

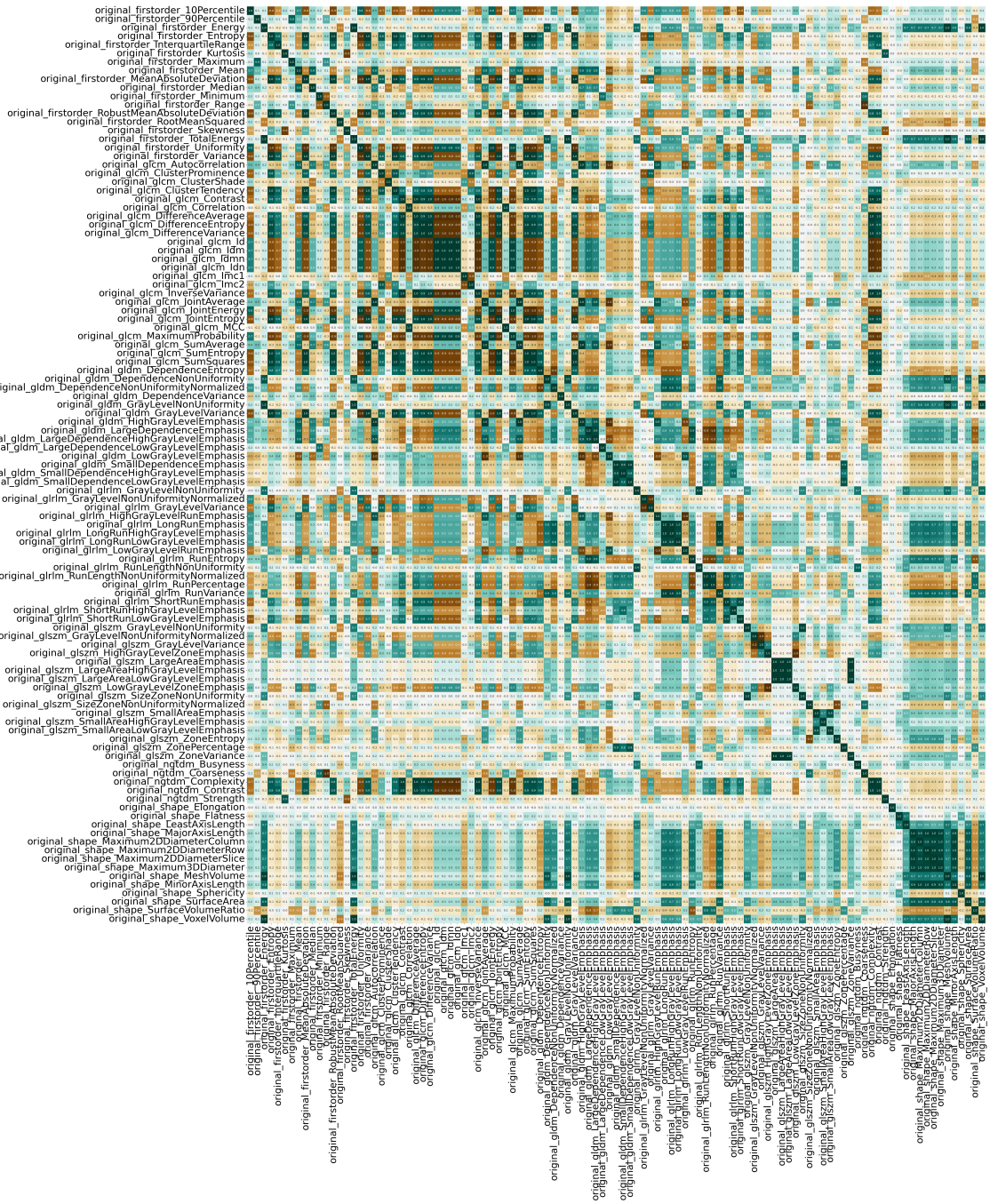


Figure 21: Radiomics and RWD Correlation Matrix

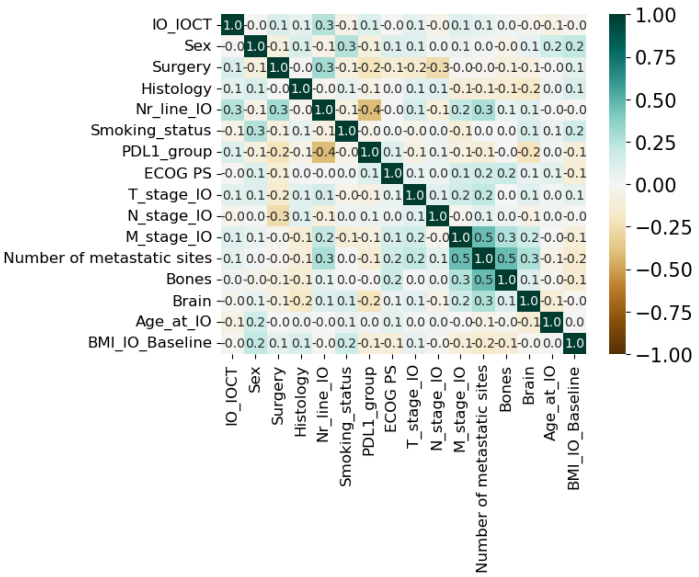


Figure 22: RWD Correlation Matrix

A.2. Confusion Matrices

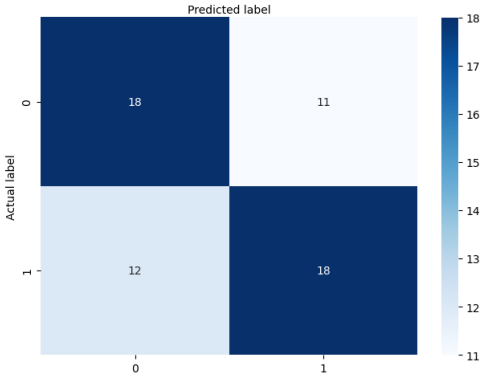


Figure 23: Radiomic features: LR Confusion Matrix on the test set

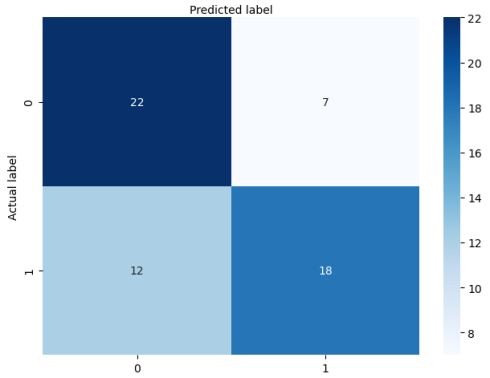


Figure 24: RWD: SVM Confusion Matrix on the test set

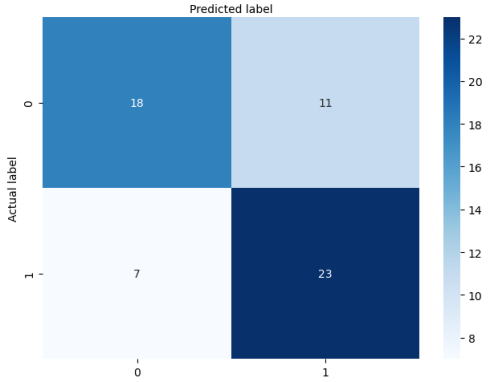


Figure 25: Radiomics and RWD: LR Confusion Matrix on the test set

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023.

[3] M Malvezzi, C Santucci, P Boffetta, G Collatuzzo, F Levi, C La Vecchia, and E Negri. European cancer mortality predictions for the year 2023 with focus on lung cancer. *Annals of Oncology*, 34(4):410–419, 2023.

[4] Global cancer observatory: Cancer today. `https://gco.iarc.fr/today`, . Accessed: 15/06/2023.

[5] PDQ Adult Treatment Editorial Board. Non-small cell lung cancer treatment (pdq®). In *PDQ Cancer Information Summaries [Internet]*. National Cancer Institute (US), 2022.

[6] Christina Fitzmaurice, Tomi F Akinyemiju, Faris Hasan Al Lami, Tahiya Alam, Reza Alizadeh-Navaei, Christine Allen, Ubai Alsharif, Nelson Alvis-Guzman, Erfan Amini, Benjamin O Anderson, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA oncology*, 4(11):1553–1568, 2018.

[7] Luis G Paz-Ares, Filippo de Marinis, Mircea Dediu, Michael Thomas, Jean-Louis Pujol, Paolo Bidoli, Olivier Molinier, Tarini Prasad Sahoo, Eckart Laack, Martin Reck, et al. Paramount: final overall survival results of the phase iii study of maintenance pemetrexed versus placebo immediately after induction treatment with pemetrexed plus cisplatin for advanced nonsquamous non–small-cell lung cancer. *Journal of clinical oncology*, 31(23):2895–2902, 2013.

[8] Marina C Garassino, Shirish Gadgeel, Giovanna Speranza, Enriqueta Felip, Emilio Esteban, Manuel Dómine, Maximilian J Hochmair, Steven F Powell, Helge G Bischoff, Nir Peled, et al. Pembrolizumab

plus pemetrexed and platinum in nonsquamous non–small-cell lung cancer: 5-year outcomes from the phase 3 keynote-189 study. *Journal of Clinical Oncology*, 41(11):1992, 2023.

[9] Marek Z Wojtukiewicz, Magdalena M Rek, Kamil Karpowicz, Maria Górska, Barbara Polityńska, Anna M Wojtukiewicz, Marcin Moniuszko, Piotr Radziwon, Stephanie C Tucker, and Kenneth V Honn. Inhibitors of immune checkpoints—pd-1, pd-l1, ctla-4—new opportunities for cancer patients and a new challenge for internists and general practitioners. *Cancer and Metastasis Reviews*, 40:949–982, 2021.

[10] Satya Das and Douglas B Johnson. Immune-related adverse events and anti-tumor efficacy of immune checkpoint inhibitors. *Journal for immunotherapy of cancer*, 7(1):1–11, 2019.

[11] Martin Reck, Delvys Rodríguez-Abreu, Andrew G Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi, Sinead Cuffe, et al. Five-year outcomes with pembrolizumab versus chemotherapy for metastatic non–small-cell lung cancer with pd-l1 tumor proportion score 50%. *Journal of Clinical Oncology*, 39(21):2339, 2021.

[12] Ye Wang, Zhuang Tong, Wenhua Zhang, Weizhen Zhang, Anton Buzdin, Xiaofeng Mu, Qing Yan, Xiaowen Zhao, Hui-Hua Chang, Mark Duhon, et al. Fda-approved and emerging next generation predictive biomarkers for immune checkpoint inhibitors in cancer patients. *Frontiers in oncology*, 11:683419, 2021.

[13] Joseph E Grossman, Divya Vasudevan, Cailin E Joyce, and Manuel Hildago. Is pd-l1 a consistent biomarker for anti-pd-1 therapy? the model of balstilimab in a virally-driven tumor. *Oncogene*, 40(8):1393–1395, 2021.

[14] Dan Sha, Zhaohui Jin, Jan Budczies, Klaus Kluck, Albrecht Stenzinger, and Frank A Sinicrope. Tumor mutational burden as a predictive biomarker in solid tumors. *Cancer discovery*, 10(12):1808–1825, 2020.

[15] J Nicholas Bodor, Yanis Boumber, and Hossein Borghaei. Biomarkers for immune checkpoint inhibition in non–small cell lung cancer (nsclc). *Cancer*, 126(2):260–270, 2020.

[16] Joshua D Shur, Simon J Doran, Santosh Kumar, Derfel Ap Dafydd, Kate Downey, James PB O'Connor, Nikolaos Papanikolaou, Christina Messiou, Dow-Mu Koh, and Matthew R Orton. Radiomics in oncology: a practical guide. *Radiographics*, 41(6):1717–1732, 2021.

[17] Manoj Mannil, Jochen von Spiczak, Robert Manka, and Hatem Alkadhi. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: unveiling the invisible. *Investigative radiology*, 53(6):338–343, 2018.

[18] Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):1–16, 2020.

[19] Jing Gong, Xiao Bao, Ting Wang, Jiyu Liu, Weijun Peng, Jingyun Shi, Fengying Wu, and Yajia Gu. A short-term follow-up ct based radiomics approach to predict response to immunotherapy in advanced non-small-cell lung cancer. *Oncoimmunology*, 11(1):2028962, 2022.

[20] Bingxi He, Di Dong, Yunlang She, Caicun Zhou, Mengjie Fang, Yongbei Zhu, Henghui Zhang, Zhipei Huang, Tao Jiang, Jie Tian, et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. *Journal for immunotherapy of cancer*, 8(2), 2020.

[21] 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.

[22] Milad Moradi and Matthias Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, 165:113941, 2021.

[23] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[26] Somatom force. `https://www.siemens-healthineers.com/it/computed-tomography/dual-source-ct/somatom-force`, . Accessed: 10/06/2023.

[27] syngo.via. `https://www.siemens-healthineers.com/digital-health-solutions/syngovia`, . Accessed: 10/06/2023.

[28] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[29] Ryan D Rosen and Amit Sapra. Tnm classification. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Pyradiomics. `https://pyradiomics.readthedocs.io/en/latest/`, 2016. Accessed: 09/06/2023.

[32] Amanda Delgado and Achuta Kumar Guddati. Clinical endpoints in oncology-a primer. *American journal of cancer research*, 11(4):1121, 2021.

[33] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9): 1323–1341, 2012.

[34] 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.

[35] Ke Wang, Ying An, Jiancun Zhou, Yuehong Long, and Xianlai Chen. A novel multi-level feature selection method for radiomics. *Alexandria Engineering Journal*, 66:993–999, 2023.

[36] Rihab Laajili, Mourad Said, and Moncef Tagina. Application of radiomics features selection and classification algorithms for medical imaging decision: Mri radiomics breast cancer cases study. *Informatics in Medicine Unlocked*, 27:100801, 2021.

[37] David G Kleinbaum, Mitchel Klein, David G Kleinbaum, and Mitchel Klein. Introduction to logistic regression. *Logistic regression: a self-learning text*, pages 1–39, 2010.

[38] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[39] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[40] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[41] Robert E Schapire. Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 37–52, 2013.

[42] Antonio Mucherino, Petraq J Papajorgji, Panos M Pardalos, Antonio Mucherino, Petraq J Papajorgji, and Panos M Pardalos. K-nearest neighbor classification. *Data mining in agriculture*, pages 83–106, 2009.

[43] Catboost. `https://catboost.ai/en/docs/concepts/python-installation`, . Accessed: 05/06/2023.

[44] Python 3.7.0. `https://www.python.org/downloads/release/python-370/`, . Accessed: 09/06/2023.

[45] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[46] Welcome to the shap documentation. https://shap.readthedocs.io/en/latest/, . Accessed: 17/06/2023.

[47] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[48] Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

[49] Deep learning example with deepexplainer. https://shap.readthedocs.io/en/latest/, . Accessed: 20/06/2023.

[50] Ecog-acrin cancer research group. https://ecog-acrin.org/resources/ecog-performance-status/, . Accessed: 14/06/2023.

[51] Pytorch. https://pytorch.org/, . Accessed: 14/06/2023.

[52] Crossentropyloss. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html, . Accessed: 20/06/2023.

[53] Bcewithlogitsloss. https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html, . Accessed: 20/06/2023.

[54] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[55] Felix Peisen, Annika Hänsch, Alessa Hering, Andreas S Brendlin, Saif Afat, Konstantin Nikolaou, Sergios Gatidis, Thomas Eigentler, Teresa Amaral, Jan H Moltz, et al. Combination of whole-body baseline ct radiomics and clinical parameters to predict response and survival in a stage-iv melanoma cohort undergoing immunotherapy. *Cancers*, 14(12):2992, 2022.

[56] Yang Yu, Yuping Bai, Peng Zheng, Na Wang, Xiaobo Deng, Huanhuan Ma, Rong Yu, Chenhui Ma, Peng Liu, Yijing Xie, et al. Radiomics-based prediction of response to immune checkpoint inhibitor treatment for solid cancers using computed tomography: a real-world study of two centers. *BMC cancer*, 22(1):1241, 2022.

[57] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[58] Xue Wang, Xiaomin Niu, Na An, Yile Sun, and Zhiwei Chen. Comparative efficacy and safety of immunotherapy alone and in combination with chemotherapy for advanced non-small cell lung cancer. *Frontiers in oncology*, 11:611012, 2021.

[59] Yimin Wang, Hedong Han, Fang Zhang, Tangfeng Lv, Ping Zhan, Mingxiang Ye, Yong Song, and Hongbing Liu. Immune checkpoint inhibitors alone vs immune checkpoint inhibitors—combined chemotherapy for nsclc patients with high pd-l1 expression: a network meta-analysis. *British Journal of Cancer*, 127(5): 948–956, 2022.