IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 April 2023, accepted 8 May 2023, date of publication 10 May 2023, date of current version 18 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3275126

RESEARCH ARTICLE

RadioPathomics: Multimodal Learning in Non-Small Cell Lung Cancer for Adaptive Radiotherapy

MATTEO TORTORA[1], (Graduate Student Member, IEEE), ERMANNO CORDELLI[1], ROSA SICILIA[1], LORENZO NIBID[2], EDY IPPOLITO[3], GIUSEPPE PERRONE[2], SARA RAMELLA[3], AND PAOLO SODA[1,4], (Member, IEEE)

[1]Unit of Computer Systems and Bioinformatics, Department of Engineering, Campus Bio-Medico University of Rome, 00128 Rome, Italy
[2]Unit of Anatomical Pathology, Department of Medicine, Campus Bio-Medico University of Rome, 00128 Rome, Italy
[3]Unit of Radiation Oncology, Department of Medicine, Campus Bio-Medico University of Rome, 00128 Rome, Italy
[4]Department of Radiation Sciences, Radiation Physics, Biomedical Engineering, Umeå University, 901 87 Umeå, Sweden

Corresponding author: Matteo Tortora (m.tortora@unicampus.it)

This work was supported in part by University Campus Bio-Medico di Roma under the program "University Strategic Projects," within the project "A CoLlAborative multi-sources Radiopathomics approach for personalized Oncology in non-small cell lung cancer (CLARO)"; in part by Programma Operativo Nazionale (PON) "Ricerca e Innovazione 2014–2020, Azioni IV.4–Dottorati e Contratti di Ricerca su Tematiche dell'Innovazione"; in part by Regione Lazio under the program "PO FSE 2014–2020 Azione Cardine 21"; in part by the project PE0000013-FAIR supported by Piano Nazionale di Ripresa e Resilienza (Recovery and Resilience Plan) of the Ministry of University and Research; and in part by Fondo per la Crescita Sostenibile (FCS) Bando Accordo Innovazione DM 24/5/2017 (Ministero delle Imprese e del Made in Italy), under the project "Piattaforma per la Medicina di Precisione, Intelligenza Artificiale e Diagnostica Clinica Integrata" (CUP B89J23000580005).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee Campus Bio-Medico University under Application No. 60/12 PAR ComEt CBM and Registered at ClinicalTrials.gov under Identifier No. NCT03583723.

ABSTRACT Current practice in cancer treatment collects multimodal data, such as radiology images, histopathology slides, genomics and clinical data. The importance of these data sources taken individually has fostered the recent rise of radiomics and pathomics, i.e., the extraction of quantitative features from radiology and histopathology images collected to predict clinical outcomes or guide clinical decisions using artificial intelligence algorithms. Nevertheless, how to combine them into a single multimodal framework is still an open issue. In this work, we develop a multimodal late fusion approach that combines hand-crafted features computed from radiomics, pathomics and clinical data to predict radiotherapy treatment outcomes for non-small-cell lung cancer patients. Within this context, we investigate eight different late fusion rules and two patient-wise aggregation rules leveraging the richness of information given by CT images, whole-slide scans and clinical data. The experiments in leave-one-patient-out cross-validation on an in-house cohort of 33 patients show that the proposed fusion-based multimodal paradigm, with an AUC equal to 90.9%, outperforms each unimodal approach, suggesting that data integration can advance precision medicine. The results also show that late fusion favourably compares against early fusion, another commonly used multimodal approach. As a further contribution, we explore the chance to use a deep learning framework against hand-crafted features. In our scenario characterised by different modalities and a limited amount of data, as it may happen in different areas of cancer research, the results show that the latter is still a viable and effective option for extracting relevant information with respect to the former.

INDEX TERMS Late fusion, machine learning, multimodal learning, non-small-cell lung cancer, radiomics, pathomics.

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

I. INTRODUCTION

Nowadays, lung cancer is recognised worldwide as one of the most common types of cancer and the leading cause of

tumour death, despite the recent increase in the number of treatment options [1]. There are two main types of lung cancer: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter, accounting for approximately 80-85% of new lung cancer cases [2], is the focus of our study. The current clinical decision-making process relies on multiple data sources to improve the detection and the classification as well as the prognosis of the tumours, such as radiology-based data (e.g. X-ray, CT scan, ultrasound, MRI and metabolic imaging), digital pathology slides, genome profiling and clinical data [3]. Such a variety of modalities catch different clinical aspects of the cancer disease and can help clinicians to pursue the paradigm of precision medicine, i.e., tailoring the treatment to the specific patient. Indeed, the wide variety of complementary quantitative bio-markers extracted from the various modalities can lead to more accurate diagnosis and more efficient treatment plans.

In the last decades, the artificial intelligence (AI) community has directed large efforts towards the detection and the classification of tumours using one or more modalities. However, only in recent years we have assisted to a growing interest directed to the disease outcome prediction using different data modalities. In this respect, the emerging areas of research are:

- **Genomics:** it is an interdisciplinary field of science that focuses on genomes, highlighting the role of human genetic variation in disease diagnosis, prognosis, and treatment response. However, genomics biomarkers still have limitations that hinder the possibility to collect such data in clinical routine due to their complexity and still high cost [4].
- **Radiomics:** it is based on the extraction of quantitative features from radiology images routinely collected in order to predict clinical outcomes or guide clinical decisions using AI algorithms [5], [6].
- **Pathomics:** it refers to the combination of digital pathology, omic science and AI to extract embedded information in digitised high-resolution whole-slide images of tissue biopsy sections to obtain quantitative biomarkers [7].

Given the growing availability of public oncological datasets containing paired samples from different modalities, in the last few years researchers started to take into account the multimodal learning paradigm. Multimodal learning relies on the integration of heterogeneous data from multiple sources into a single machine learning framework. Although several works use genomics, radiomics or pathomics data alone, few works still aim to fuse these modalities together [8], [9], [10], [11], [12], reporting performance improvement. Despite the importance of radiomics and pathomics taken individually, to the best of our knowledge, only one work to date has combined them together into a single machine learning framework [12]. Hence, we present here another investigation that combines radiomics, pathomics and clinical data together into a single multimodal late fusion scheme to predict radiation therapy treatment outcomes for NSCLC patients. In this work, we do not take into account genomics data, because in clinical practice, pending the results of ongoing studies, in patients with locally advanced NSCLC considered for chemoradiation treatment, knowledge of oncogene-dependent characteristics does not change the therapeutic strategy. The late fusion scheme permits us to combine uncorrelated data flows that vary significantly in terms of dimensionality and sampling rates, as in our case.

To summarise, the contributions of this work are:

- We proposed a multimodal late fusion scheme taking into account features extracted from radiomics, pathomics and clinical data.
- We show that the integration of heterogeneous data into a multimodal learning paradigm permits to predict the radiation therapy treatment outcomes in lung cancer.

Furthermore, to offer a deeper analysis of multimodal learning in this context, this work provides other two contributions:

- We compare the proposed multimodal late fusion scheme with the early fusion approach, showing that the latter has lower performance.
- Since deep learning has shown its potential in several healthcare applications, both in unimodal and multimodal learning, here we compare the hand-crafted features against the use of deep neural networks, thus offering a complete analysis of the main different methodologies to process our data.

The rest of this manuscript is organised as follows: section II presents a short overview of the multimodal learning framework and its applications to oncology. Section III introduces the materials, overviewing the multimodal data sources available. Section IV presents the proposed multimodal learning framework, whilst section V and section VI describe the experimental setup and the results respectively. Finally, section VII provides concluding remarks.

II. BACKGROUND

In this section, we first overview the various architectures in the multimodal learning framework, and then we summarise the current state-of-the-art on multimodal-based learning on oncology (section II-B).

A. MULTIMODAL LEARNING

Multimodal learning involves the integration of heterogeneous data from multiple sources extracted from the observation of the same phenomena or problem. Hence, the use of multimodal data sources allows the extraction of a complementary, more robust and richer data representation, with the aim of improving performance compared to the use of a standalone modality. Although there is not any formal proof, this intuition has brought interesting results in many applications, medical imaging included [13].

Multimodal data integration can be performed at different levels using three types of fusion: early, joint, and late fusion [14], [15], respectively (Figure 1), as now described.

Early fusion, also known as data-level fusion or representation learning, concerns the integration of raw inputs

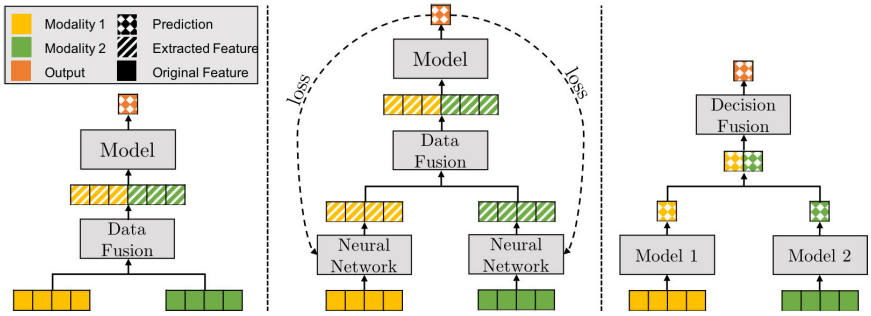


FIGURE 1. Model architectures for different multimodal learning frameworks. From left to right: early or data-level fusion, joint or intermediate fusion, late or decision-level fusion.

from multiple data source modalities into a single feature vector before passing it into a single machine learning model (Figure 1, left panel). In the early fusion, raw input data can be merged into an embedded space according to different policies, such as simple concatenation, addition, pooling, or applying a gated unit [16]. Although the promising results achieved in several applications, how to manage one or more missing modalities, how to handle different sampling rates and/or the time-synchronicity between multiple data sources, and the possible redundancies occurring while generating very large embedded spaces are the main issues of this multimodal approach.

With respect to joint and late fusion, these two combination schemes work by aggregating different classification models. Joint fusion, also known as intermediate-level fusion or hybrid fusion, concerns the combination of the extracted intermediate feature vectors from trained neural networks, one per modality, into an abstract fusion layer, also known as a shared representation layer (Figure 1, central panel). Then, this combined feature vector feeds a final classification model whose loss is back-propagated to the feature extracting neural networks during training. Since the loss is back-propagated during the training process, this fusion scheme improves the feature representation at each iteration leading to better multimodal embedded feature spaces. Although joint learning is a very flexible framework, its main issue concerns the design of the architecture in terms of how, when, and which modalities can be fused [15].

Let us now delve into late fusion approaches, as our proposed approach works at this level. In the early- to mid-2000s, late- or decision-level fusion has received considerable interest from the machine learning community due to its potential to improve the performance of stand-alone classifiers. Late fusion concerns the training of independent systems, one per modality, which are then combined by an aggregation function to reduce individual error rates (Figure 1, right panel). This aggregation function takes as input the unimodal decision values provided by the different classifiers that are combined according to a fusion rule (e.g. minimum, maximum, mean, majority vote, etc.). There is a consensus that the key to the success of late fusion is that it builds a mixture of diverse classifiers [17], providing different and complementary points of view to the ensemble. Definitely, the late fusion approach is a well-suited multimodal strategy when input modalities are significantly uncorrelated and they vary significantly in terms of data dimensionality and sampling rates [15]. These are the major reasons that led us to explore this multimodal framework, which will be also experimentally compared against early fusion in section VI.

B. MULTIMODAL ONCOLOGY

Nowadays, the current clinical practice for cancer treatment requires collecting multimodal data for each patient, such as radiological images (e.g. X-ray, CT scan, ultrasound, MRI and metabolic imaging), histopathology slides, genomics and clinical data. Such a variety of modalities describes different clinical aspects of cancer disease and can provide a wide range of complementary bio-markers leading to more accurate diagnosis and more efficient treatment plans. Although there are several works in the current state-of-the-art dealing with the detection, classification and prognostic task taking the aforementioned single modalities individually [18], [19], [20], [21], [22], there are still few works in oncology that aim to fuse these modalities together. Hence, in recent years, researchers focused their efforts on the fusion of these modalities into a single machine learning framework [8], [9], [10], [11], [12].

In [8] the authors proposed a novel multimodal radiomics model for preoperative prediction of lymphatic vascular infiltration (LVI) in rectal cancer based on hand-crafted features extracted from magnetic resonance (MR) and computed tomography (CT) modalities. The authors validated their method on a retrospective cohort of 94 patients with histologically confirmed rectal cancer. The results show as the multimodal (MR/CT) radiomics models can serve as an effective visual prognostic tool for predicting LVI in rectal cancer. It demonstrated the great potential of preoperative prediction to improve treatment decisions over the stand-alone modalities.

In [9] the authors proposed a multimodal deep learning method for NSCLC survival analysis leveraging CT images in combination with clinical data. The authors validated their framework using data from The Cancer Imaging

Archive (TCIA), which contains paired samples of CT scans and clinical data for 422 NSCLC patients [23]. The results show that there is a relationship between prognostic information and radiomics images. In addition, the proposed multimodal model improves the analysis of survival in NSCLC patients compared to the current state-of-the-art which only works with clinical data.

In [10] the authors proposed a deep multimodal fusion framework for the end-to-end multimodal fusion of histopathological images and genomics features (mutations, CNV, mRNAseq) for survival outcome prediction. This work implements the Kronecker product to model pairwise feature interactions across modalities and controls the expressiveness of each modality through a gating-based attention mechanism. The authors validated their framework using glioma and clear cell renal cell carcinoma datasets from The Cancer Genome Atlas (TCGA), which contains paired samples of whole-slide images of hematoxylin-and-eosin-stained specimens, genotype, and transcriptome data for 769 patients [24]. Based on a 15-fold cross-validation, results show that the proposed multimodal fusion paradigm leads to an improvement over the current state-of-the-art in predicting survival outcomes when using each modality independently.

In [11], the authors proposed a hybrid deep multimodal fusion model that merges patients' gene modality data with pathological images to predict breast cancer sub-types. To extract features from the different forms and states of the data, the authors set up separate feature extraction networks and then fused the output of the two networks using a weighted linear aggregation method. The authors used Principal Component Analysis to reduce the dimensionality of the gene modality data and filter the image modality data. The fused features were then used to predict breast cancer subtypes. The authors validated their framework TCGA-BRCA dataset as a sample set for their molecular sub-type prediction of breast cancer. It contains gene expression data, CNVs, and histopathological images for 1098 breast cancer patients. Results show that the model achieved 88.07% accuracy in sub-type prediction, outperforming traditional DL models by 7.45%.

In [12] the authors proposed a deep model merging together radiology scans, molecular profiling, histopathology slides and clinical factors to predict the overall survival of glioma patients. The authors validated their framework by collecting data from the TCIA repository, which contains paired samples of whole-slide images of hematoxylin-and-eosin-stained specimens, MRI scans, DNA sequencing data, and clinical variables for 176 glioma patients. Results show that their model, with a median C-index of 0.788 ± 0.067, significantly outperforms the best performing unimodal model, which has a median C-index equal to 0.718 ± 0.064. Furthermore, the proposed model successfully stratifies patients into clinical subgroups based on overall survival, adding further granularity to clinical prognostic classification and molecular subtyping.

Despite the importance of radiomics and pathomics data taken individually, to the best of our knowledge, at the time this work was written, only one study has combined them into a single machine learning framework for the outcome prediction of radiation therapy treatment [12]. Hence, this paper proposes the second attempt to combine radiomics, pathomics and clinical data into a single model to predict outcomes for NSCLC patients. As mentioned in section I, since in the current clinical practice the knowledge of oncogene-dependent characteristics does not change the therapeutic strategy, in this work we do not consider genomics data

III. MATERIALS

In this work we used an in-house cohort of 33 patients with Locally-Advanced stage III NSCLC, who were enrolled from November 2012 to July 2014 and treated with concurrent chemoradiation at a radical dose with an adaptive approach. The adaptive protocol was approved by Ethical Committee Campus Bio-Medico University on 30 October 2012 and registered at ClinicalTrials.gov on 12 July 2018 with Identifier NCT03583723 after an initial exploratory phase. Enrolled patients underwent a clinical evaluation after chemoradiation treatment and were classified into two groups according to target reduction: (i) adaptive, i.e., patients who achieved a reduction in tumour volume, assessed by two radiation oncologists on weekly chest CT simulations, leading to the implementation of a new treatment plan with which the patient would continue radiation therapy (adaptive approach); (ii) not-adaptive, i.e., patients who did not achieve target shrinkage and continued the chemoradiation with standard treatment. The a priori probability of this patients' cohort consists of 11 and 22 adaptive and not-adaptive patients, respectively.

For this patient cohort we collected heterogeneous data including histological slides, CT scans, as well as clinical data, therefore forming the following unimodal data streams (i.e., pathomics, radiomics and semantics) in the multimodal learning framework investigated in this study:

- **Pathomics modality**: This modality includes samples generated from biopsy slides of lung cancer tissue, stained with haematoxylin and eosin (HE). HE (haematoxylin/eosin) tumour tissue slides were reviewed by a pathologist to confirm sample adequacy. Slides were digitised (APERIO CS2 Leica Biosystems or NanoZoomer 2.0 RT Hamamatsu) at 20x magnification. The digitised slides were loaded and segmented on QuPath. Regions of interest (ROIs), also called crops in the following, were manually defined by lung pathology experts to identify tumour areas avoiding histological artefacts, macrophage clusters and inflammations, fibrosis and necrosis. A total of 1113 tumour areas were manually segmented for the 33 tissue samples, one per patient.
- **Radiomics modality**: This modality includes initial CT scans collected prior to the start of concomitant chemoradiation therapy treatment. The CT scans

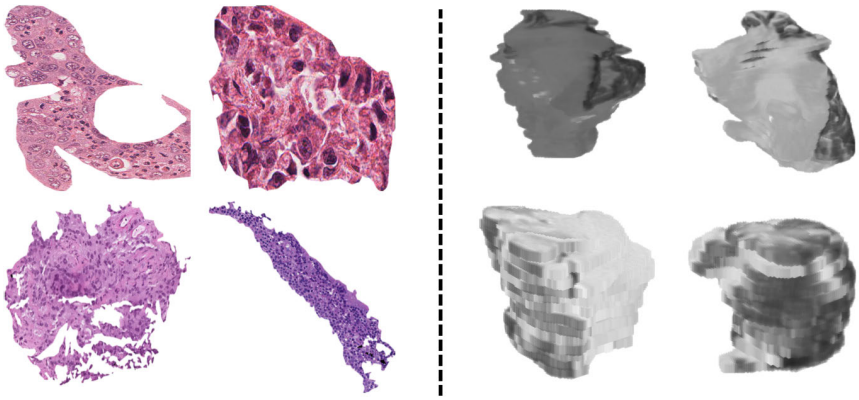


FIGURE 2. Example images for both pathomics and radiomics modality. To the left: examples of crops countered by pathologists. To the right: examples of CTVs manually defined by expert lung pathology on CT scans. For the sake of presentation, we show both the crops and the CTVs scaled to the same size.

consisted of single layer spiral computerised tomography - Siemens Somatom Emotion. Acquisition parameters were 140 Kv, 80 mAs, and 3 mm for slice thick. The scans were pre-processed applying a lung filter (kernel B70) and a mediastinum filter (kernel B31). The characteristics investigated in this work and presented in section IV-B were extracted from 3D ROIs given by the Clinical Target Volume (CTV), manually countered by expert radiation oncologists. The CTV is the volume containing the Gross Tumour Volume (GTV), i.e., the macroscopically demonstrable disease, and therefore, with a probability considered relevant for therapy, the microscopic disease at the subclinical level. It is worth noting that in [25] the authors showed that CTV should be preferred to GTV when computing radiomics features. In total, this modality contains 39 manually contoured CTVs for the 33 patients. It is worth noting that the number of CTVs exceeds the number of patients as multiple tumours can occur in a patient.

- **Semantic modality**: Two experienced radiation oncologists independently reviewed all CT scans and scored each tumour for four semantic imaging features, divided into tumour staging scores (T, N and tumour stage), and histological evaluation. They also added the age and sex of the patients. Each radiation oncologist blindly assigned staging scores, and, in case of disagreement, they reviewed the CT scans together and any discrepancies were resolved through discussion until consensus was reached.

As can be seen from these descriptions, the data sources used are highly heterogeneous and are uncorrelated unimodal flows. Thus, as we previously mentioned in section I, this motivated the choice of using the late fusion approach as a multimodal approach.

Figure 2 shows four examples of both crops extracted by the pathologists from the high-resolution whole-slide images which contain the selected tumour area of interest, and CTVs extracted by expert oncological radiotherapist by CT scans weekly collected during the radiation therapy treatment. Moreover, the first three rows of Table 1 summarise the a priori sample distribution for each of the three different modalities.

TABLE 1. A priori distribution of samples for each unimodal flow.

	Modality	Adaptive	Not-Adaptive	Total
Raw Data	Pathomics Crops	303	810	1113
	Radiomics CTVs	13	26	39
	Semantic	11	22	33
Pre-Processed Data	Pathomics Patches	10869	42681	53550
	Radiomics Slices	301	627	928

IV. METHODS

This section introduces the proposed fusion framework to handle the binary classification task introduced before. It is composed of four main blocks shown in Figure 3 identified by the bars at its bottom and presented in the following. First, a pre-processing phase is applied to the different unimodal flows (section IV-A). This stage uniforms the data, increases the dimensionality of unimodal flows, and encodes categorical features into numerical ones. The second step, presented in section IV-B, extracts the features from both images belonging to the pathomics and the radiomics flows. The third step consists in patient aggregation, i.e., we merge each instance of the same patient of a single modality to get a single label for each patient (section IV-C). This is necessary so that the following data fusion step can work on consistent samples, i.e., one sample per patient, and not on single histologicals (patches or CTs' slices). Section IV-D presents the eight fusion rules we investigated that belong to three different paradigms, this offering a view of how different fusion techniques can fuse three sources of information. On the one side, they are the *product*, *maximum*, *minimum*, *mean*, *decision template* and *Dempster-Shafer*, and all of them are based on the *decision profile*, i.e., a matrix organising the

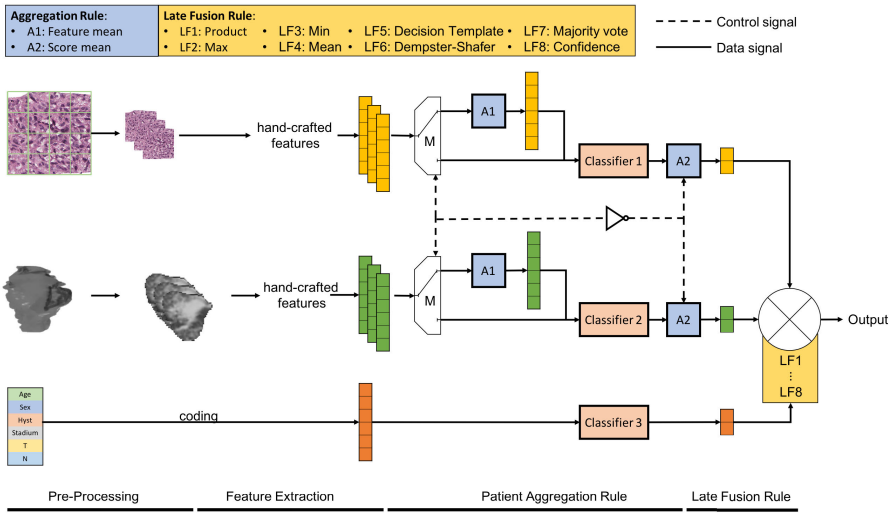


FIGURE 3. Proposed late fusion classification framework consisting of four blocks identified by the bars at the bottom of the figure. Let M be a multiplexer that allows us switching between two patient aggregation modes, A_1 and A_2, respectively. If mode A_1 is active, mode A_2 is deactivated by a control signal passing through the logic NOT port, and vice versa.

output of l different soft classifiers in a multiple classier framework, with $l = 3$ in our case. On the other side, the other two are the *majority voting rule* that works with the crisp labels and the *confidence rule*, which is a classifier selection technique.

A. PRE-PROCESSING

This section presents the pre-processing applied to each unimodal flow which differs for each modality due to the heterogeneity of the data.

In the case of pathomics images, we applied a patch extraction operation to the original crops manually contoured by expert pathologists on the high-resolution whole-slide images. The patch extraction phase was performed by a sliding window with a size equal to 100×100 and a stride equal to 60, both chosen empirically. This step permits us to increase the cardinality of the available images by exploiting the variability typical of different regions of the same image. Furthermore, it also provides homogeneous images, as the original ones are characterised by a wide size variability. After this operation, we empirically removed the extracted patches with more than 20% of pixels belonging to the background to keep only the most informative images. In the end, the new repository of pathomics patches is composed of 53550 instances.

Let us now focus on the pre-processing for radiomics. As already reported in section III, the radiomics modality consists of CTVs manually contoured by expert radiation oncologists on CT scans. To increase the dimensionality of this modality, we decomposed the CTVs, i.e., the volumes containing the macroscopically demonstrable tumour mass and the microscopic disease at the sub-clinical level, into their component slices. Thus we passed from having few 3D CTVs per patient to multiple 2D slices per patient, for a total of 928 slices.

Finally, in the case of semantic information, in order to have numerical features, we applied an ordinal encoding for T, N and stadium features, and one-hot encoding for sex and diagnosis features as no ordinal relation exists for these latter categorical variables.

The second part of Table 1 shows the a priori distribution of instances obtained after the pre-processing.

B. FEATURES EXTRACTION

This section describes the features extraction stage implemented for both the pathomics and the radiomics unimodal flows. It is straightforward that this step is not applied to the semantic unimodal flow, since it includes patients' medical data already processed. Note also that the features computed for each modality were selected to optimise each unimodal flow, but this is out of the scope of this work and, for the sake of brevity, we do not present this phase here. Nevertheless, the starting feature set consists of 2D intensity and texture features well established in the medical image processing scenario [26] and, specifically, in radiomics [27] and digital pathology [28]. They are statistical features extracted from the first-order image histogram, and several descriptors extracted from the results provided by both the Grey Level Co-Occurrence Matrix (GLCM) and the Local Binary Pattern (LBP) operators. Please note that we also investigated the use of deep learning, as we will discuss in section V.

1) PATHOMICS

Among the descriptors mentioned before, histopathological images are represented by measures derived from the GLCM, computed from each patch.[1]

[1]Note that the GLCM operator is applied on the S-channel of the HSV colour model.

The GLCM is a texture feature used to analyse the spatial distribution of grey levels within a 2D image at the microscale [29]. This operator can be parameterised in terms of δ and θ. The former represents the relative distance in pixels between two points in I, while the latter is the relative orientation between two points in I. Here, taking into account preliminary experiments and findings in related fields [30], [31], [32], [33], for pathomics we used $\delta = 1$ and $\theta \in [0°, 45°, 90°, 135°]$, so that each patch has four GLCMs.

From each of these matrices we extracted six Haralick descriptors [29], i.e., contrast, dissimilarity, homogeneity, energy, correlation and the angular second moment, listed in depth in Appendix B. Their concatenation provides 24 textural descriptors per patch.

2) RADIOMICS

For the radiomics modality, we used the same features as presented in our previous work [34]. Hence, for each slice that makes up the 3D ROI extracted from the CT scans by the radiologists, we computed 12 statistics features and 104 textural features.

Statistical features consist of the moments up to the fourth-order of the first-order image histogram, i.e., the mean, the standard deviation, the skewness and the kurtosis. Furthermore, the picture of grey-level distribution is also grasped by the histogram width, the energy, the entropy, the value of the histogram absolute maximum and the corresponding grey-level value, the energy around such maximum, the number of relative maxima in the histogram and their energy. These statistics are listed in depth in Appendix B.

Texture features are derived from the GLCM and from the LBP. The former is parameterised by a unit distance δ between pixels and an orientation $\theta \in [0°, 45°, 90°, 135°]$ and we extracted six second-order statistical features as the pathomics flow.

The latter is an operator that describes the local texture by assigning each pixel in an image a binary code according to its local circular neighbourhoods of P points located on the circumference of radius R centred on the pixel itself [35]. In this work, we used an extension of the original operator making it invariant to both local monotonic greyscale variations and rotation [36]. Once we have applied this operator to each pixel in the image, we can compute a histogram of the LBP decimal codes' occurrences.

In this work, we empirically parameterised R with a unit distance and we set P equal to 8. Finally, the same 12 statistical features reported above on the top of this section (i.e., mean, standard deviation, skewness, kurtosis, etc.) are then computed from the histogram of LBP distribution.

C. PATIENT AGGREGATION RULE

As mentioned above, each patient is composed of several samples both for the pathomics and the radiomics flows. For the former, the samples correspond to the patches extracted from the crops contoured by the pathologists, whilst for the latter, the samples correspond to the various slices included in the segmented CT VOIs.

For this reason, in order to have a single classification per patient and consistent sample fusion, a samples' patient-wise aggregation is necessary. In this work, we used two different patient-wise aggregation rules, denoted as A_1 and A_2 in Figure 3, respectively.

The former is applied before the classification step, and it averages out each component of the feature vector $\boldsymbol{x} \in \Re^n$ belonging to the same patient p:

$$\boldsymbol{x}^p = \frac{1}{N^p} \sum_{\boldsymbol{x} \in \mathcal{X}^p} \boldsymbol{x}$$

where $\mathcal{X}^p$ is the set of feature vectors computed from all the samples of the same patient p (i.e., histopathology patches or CT slices), and N^p is its cardinality.

The latter works after the classification process, and it averages the soft labels of all the instances of a patient. Formally, given a classification problem with C class labels and L unimodal flows,[2] and assuming one classifier per modality, let $\mathcal{D} = \{D_i\}_{i=1}^{L}$ denotes the set of classifiers. Hence, given $\boldsymbol{x}$, a soft classifier outputs a C-dimensional vector given by

$$D_i(\boldsymbol{x}) = \left[d_{i,1}(\boldsymbol{x}), \ldots, d_{i,C}(\boldsymbol{x})\right]^T,$$

where $d_{i,j}(\boldsymbol{x}) \in [0, 1]$ is the soft label and it represents the degree of support provided by classifier D_i for the hypothesis that $\boldsymbol{x}$ comes from the class ω_j. On this premise, the $A2$ patient-wise aggregation rule is defined by:

$$d^p{}_{i,j} = \frac{1}{N_p} \sum_{\boldsymbol{x} \in \mathcal{X}^p} d_{i,j}(\boldsymbol{x})$$

where, thus, $d^p_{i,j}$ represents the average soft label per class computed over all the instances of the i-th modality of the same patient (i.e., $\boldsymbol{x} \in \mathcal{X}^p$).

D. LATE FUSION RULES

This section introduces the late fusion rules that merge the multimodal information extracted from the different unimodal flows.

Using the notation already introduced, in a multimodality framework we organise the outputs returned by the L unimodal classifiers into a patient-wise *decision profile* DP^p, defined by the following matrix:

$$DP^p = \begin{bmatrix} \mu_{1,1} & \cdots & \mu_{1,j} & \cdots & \mu_{1,C} \\ \vdots & \ddots & \vdots & & \vdots \\ \mu_{i,1} & \cdots & \mu_{i,j} & \cdots & \mu_{i,C} \\ \vdots & & \vdots & \ddots & \vdots \\ \mu_{L,1} & \cdots & \mu_{L,j} & \cdots & \mu_{L,C} \end{bmatrix}$$

where $\mu_{i,j}$ is computed according to $A1$ or $A2$ aggregation rule. This implies that, using $A1$, $\mu_{i,j} = d_{i,j}(\boldsymbol{x}^p)$, whereas using $A2$ we have $\mu_{i,j} = d^p_{i,j}$. Thus, the patient-wise data is projected into a new feature space with dimension $L \times C$ and

[2]Note that in our case $C = 2$ and $L = 3$.

this new representation combining the unimodal classification stages is depicted by the symbol $\bigotimes$ in Figure 3.

The fusion methods calculate the support χ_j for the class ω_j by applying some mathematical procedure described below on the DP^p representation and, using the maximum membership rule, we then assign the patient p to the class ω_s if:

$$\chi_s \geq \chi_z, \quad \forall\ z = 1, \ldots, C$$

In this work, to compute χ_j we apply eight late fusion techniques, represented with the tag LF_t, with $t = 1, \ldots, 8$, in Figure 3. They include four fusion rules computing the support for the j-th class independently of the support of the other classes:

1) **Product rule (LF_1)**: it computes the support χ_j for the class ω_j as:

$$\chi_j = \prod_{i=1}^{L} \mu_{i,j}$$

2) **Max rule (LF_2)**: it computes the support χ_j for the class ω_j as:

$$\chi_j = \max_i \mu_{i,j}$$

3) **Min rule (LF_3)**: it computes the support χ_j for the class ω_j as:

$$\chi_j = \min_i \mu_{i,j}$$

4) **Mean rule (LF_4)**: it computes the support χ_j for the class ω_j as:

$$\chi_j = \frac{1}{L} \sum_{i=1}^{L} \mu_{i,j}$$

We also investigated others two rules computing the class supports comparing the entire DP^p feature space with the decision templates (DTs) of each class. DTs-based methods have been found to be among the best combination techniques and show stable performance over a range of experimental settings [37]. They are:

1) **Decision Templates, DTs (LF_5)**: its use was proposed in [37] and consists of calculating C DTs, one per class, that capture the pattern of each. The decision template DT_i for class ω_i is the centroid of class ω_i in the training $L \times C$ feature space DP^p and it is calculated as follows:

$$DT_i = \frac{1}{N_i} \sum_{p=1}^{N_i} DP^p,$$

where N_i is the number of patients belonging to the class ω_i.
Finally, the p-th patient's support degree χ_i for the class ω_i is computed by measuring the similarity between the current DP^p and DT_i:

$$\chi_i = 1 - \frac{1}{L \cdot C} \sum_{j=1}^{C} \sum_{i=1}^{L} \left(\mu_{k,j} - dt^i_{k,j}\right)^2,$$

where $dt^i_{k,j}$ is the k, j-th entry in the i-th decision template DT_i.

2) **Dempster-Shafer rule (LF_6)**: it is still based on the use of DTs. The p-th patient's proximity Φ^p between the output of the i-th classifier D^p_i and DT^i_j is defined as [38]:

$$\Phi^p_{j,i} = \frac{\left(1 + \left\|DT^i_j - D^p_i\right\|^2\right)^{-1}}{\sum_{k=1}^{C} \left(1 + \left\|DT^i_j - D^p_i\right\|^2\right)^{-1}},$$

where DT^i_j denotes the i-th row of decision template for the class ω_j, D^p_i denotes the output of the i-th classifier on the p-th patient and $\|\cdot\|$ is any matrix norm. Then, the final support degree for the j-th class is:

$$\chi_j = K \prod_{i=1}^{L} \frac{\Phi^p_{j,i} \prod_{k \neq j} \left(1 - \Phi^p_{k,i}\right)}{1 - \Phi^p_{j,i} \left[1 - \prod_{k \neq j} \left(1 - \Phi^p_{k,i}\right)\right]}$$

where K is a scaling factor.

For the sake of completeness, we also investigated other two rules working with different paradigms.

On the one side, we use the *majority voting rule* (LF_7) that works with crisp label outputs of each modality by assigning the patient p the class label ω_s that is most represented among those returned by the L unimodal classifiers. Formally:

$$s = \arg\max_j \sum_{i=1}^{L} \mu^{crisp}_{i,j}, \quad \text{for } j = 1, \ldots, C$$

where

$$\mu^{crisp}_{i,j} = \begin{cases} 1, & \text{if } j = \arg\max_w \mu_{i,w} \\ 0, & \text{Otherwise} \end{cases}, \quad \forall\ i = 1, \ldots, L \ \wedge\ j = 1, \ldots, C$$

On the other side, we also applied the *confidence rule* (LF_8), which assigns patient p the class label ω_s:

$$s = \arg\max_j DP^p_q, \quad \text{for } j = 1, \ldots, C$$

which corresponds to the q-th unimodal classifier output with the largest degree of support:

$$q = \arg\max_i \left(\max DP^p_i\right), \quad \text{for } i = 1, \ldots, L$$

where DP^p_i denotes the i-th modality whose classifier output is represented by row the i-th of DP^p.

V. EXPERIMENTAL SETUP

Here we introduce the experimental setup adopted, presenting in section V-A the set of experiments carried out, and in section V-B the validation adapted as well as the evaluation metrics used.

TABLE 2. Summary of the 16 rule combinations performed in this work.

	Product	Max.	Min.	Mean	Decision Template	Dempster Shafer	Majority Vote	Confidence
Features Mean	$A_1 + LF_1$	$A_1 + LF_2$	$A_1 + LF_3$	$A_1 + LF_4$	$A_1 + LF_5$	$A_1 + LF_6$	$A_1 + LF_7$	$A_1 + LF_8$
Score Mean	$A_2 + LF_1$	$A_2 + LF_2$	$A_2 + LF_3$	$A_2 + LF_4$	$A_2 + LF_5$	$A_2 + LF_6$	$A_2 + LF_7$	$A_2 + LF_8$

A. SET OF EXPERIMENTS

The first set of experiments consists of evaluating the late fusion paradigm through all the different combinations of the fusion and aggregation rules, LF_x and A_y, respectively, for a total of 16 combinations since $x \in \{1, \ldots, 8\}$ and $y \in \{1, 2\}$, which are summarised in Table 2. Furthermore, these 16 experiments were performed for all combinations of modalities, i.e. *Pathomics*+*Radiomics*+*Semantic* (*P*+*R*+*S*), *Pathomics*+*Semantic* (*P*+*S*), *Radiomics*+*Semantic* (*R*+*S*), and, finally, *Pathomics* + *Radiomics* (*P* + *R*), for a total of 64 experiments.

Then, we compared the late fusion approach with an early fusion framework. Concerning this last approach, in this work we considered two early fusion rules: a simple approach in which the different modalities are concatenated without any processing on the feature space, and a concatenation given by the application of the Kronecker product, as presented in [10]. Indeed, the latter rule was chosen to bring out a correlation of the different modalities in the various combinations of them. For the sake of consistency of samples, in the early fusion paradigm we only applied the A1 aggregation rule, i.e. samples' patient-wise aggregation is performed by averaging each component of the feature vectors belonging to the same patient.

In all the experiments, we used the same learning paradigm in the *classifier* blocks of Figure. 3, in which is a Random Forest [39] with entropy as a function to measure the quality of a split, whilst, for all the other parameters, we used the default values provided by the Scikit-learn framework [40], without any fine-tuning. Indeed, it was empirically observed in [41] that in many cases the use of tuned parameters cannot significantly outperform the default values of a classifier suggested in the literature, as also confirmed in other works [42], [43], [44].

Although we focus on the potential of multimodal learning in outcome prediction for NSCLC, in this work we also compare hand-crafted features with a deep learning framework to provide a thorough and complete analysis. For this comparison, the 64 experiments described above were also performed using deep features extracted from both pathomics and radiomics modalities using the ResNet-18 [45] and GoogLeNet [46] networks respectively, pre-trained on ImageNet dataset. The choice of using ImageNet as a pre-training tool is motivated by the fact that this dataset provides enough rich image detail of different objects and targets, and therefore we believe that these pre-trained network feature extraction capabilities can be transferred to both pathomics and radiomics tasks. For each patient we trained the CNNs with a transfer learning process performed with all samples from the other patients for 20 epochs, as suggested by our previous work [47]. Furthermore, given the reduced amount of training samples, during the training, we froze the weights for all the layers except the ones of the new final fully connected layer. Straightforwardly in this last layer, we removed the original 1000 neurons, which are replaced by two softmax neurons with random weights. These experiments were performed using the PyTorch framework [48].

TABLE 3. Results for the unimodal flows in terms of AUC.

Aggregation Rules	Modalities		
	Pathomics	Radiomics	Semantic
-	-	-	.705
A1	.686	**.870**	-
A2	.711	.731	-

B. EVALUATION METHODS

We tested all the proposed approaches with a *leave-one-patient-out* (LOPO) cross-validation paradigm so that we performed a number of runs equal to the number of patients. Therefore, in each run, the test set consisted of all samples belonging to one patient, whereas all the others were allocated to the training set.

The patient-wise performances were computed by averaging the Area under the ROC curve (AUC) of each run, where, as a reminder, the positive and negative classes correspond to adaptive and non-adaptive patients, respectively. It is worth recalling that AUC is a figure of merit widely adopted in the medical community to characterise the performance of a prediction model. Furthermore, to compare the results we also applied some statistical tests that will be introduced hereinafter and formally presented in Appendix A

VI. RESULTS AND DISCUSSION

This section presents and analyses the results in several directions, starting from the raw outputs reported in Table 3 and Table 4. The former reports the scores attained by every single

TABLE 4. Overall results for the 64 experiments performed in terms of AUC, where *P*, *R* and *S* stand for pathomics, radiomics and semantic, respectively.

Rules Combination	Modalities Combinations			
	P+R	R+S	P+S	P+R+S
$A_1 + LF_1$	.853	.888	.752	.866
$A_1 + LF_2$	**.909**	.837	.756	.860
$A_1 + LF_3$	.812	.864	.740	.824
$A_1 + LF_4$	.903	.893	.764	.907
$A_1 + LF_5$	.888	**.909**	.748	.905
$A_1 + LF_6$	.901	.781	.731	**.909**
$A_1 + LF_7$	.903	.893	.764	.907
$A_1 + LF_8$	.853	.876	.748	.857
$A_2 + LF_1$	.756	.752	.756	.793
$A_2 + LF_2$	.760	.688	.684	.709
$A_2 + LF_3$	.715	.754	.760	.777
$A_2 + LF_4$	.793	.752	.748	.798
$A_2 + LF_5$	.789	.756	.748	.810
$A_2 + LF_6$	.769	.740	.731	.773
$A_2 + LF_7$	.793	.752	.748	.798
$A_2 + LF_8$	.715	.740	.740	.764

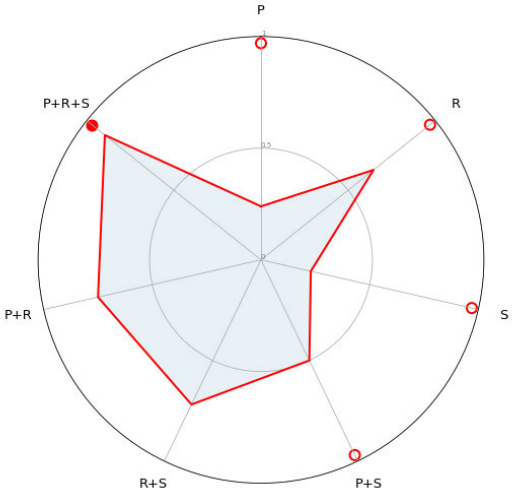


FIGURE 4. Radar chart showing the performance in terms of AUC of the unimodal and multimodal approaches, where P stands for Pathomics, R for Radiomics, and S for Semantic. The filled circle represents the flow with the highest rank, whilst blank circles represent unimodal or multimodal approaches with statistically different performances from the best approach according to Friedman test with the Iman-Davenport amendment followed by the pairwise Bonferroni-Dunn post-hoc test ($p < 0.1$).

modality, eventually using one aggregation rule to combine the features describing the samples. The radiomics modality with the use of the feature mean as patient aggregation rule is the best unimodal flow with an AUC equal to 0.870. Furthermore, this performance score confirms the effectiveness of the radiomics signature identified in our previous work [34], as the AUC obtained there, equal to 0.82, is of comparable magnitude to the one presented in this work. Nevertheless, the cohorts of patients included in the two studies are not directly comparable in terms of the dimensionality of the dataset. Table 4 shows the performance attained by the pairwise fusion approaches and by the trimodal combination, for all the 16 fusion rules. With an AUC equal to 0.909, the best results are achieved by the multimodal triplet $P+R+S$, and the pairwise combinations $R+S$ and $P+R$, with the following fusion rules respectively: $A_1 + LF_6$, $A_1 + LF_5$ and $A_1 + LF_2$. Hence, all of them are given by the use of the feature mean as patient aggregation rule at the feature level followed by the Dempster-Shafer, Decision Template, and Maximum as fusion rule, respectively.

To discuss these results, the rest of this section deepens the results in three directions. First, we present the results provided by the late fusion approaches introduced in section IV-D and schematically depicted in Figure 3 (section VI-A). Second, in section VI-C we compare late fusion and early fusion approaches. Third, in section VI-C we show a comparison of hand-crafted and deep features.

A. LATE FUSION RESULTS

The contribution of this experiment is three-fold, so it permits us to answer the following three questions:

1) What is the best combination of modes?
2) Within the multimodal combination, which is the unimodal mode contributing more to the best performance?
3) What is the best fusion rule?

Let us now explore each of these questions. In all the cases we will introduce figures that offer a high-level synthesis of the huge amount of results provided by all the experiments.

1) BEST MULTIMODAL COMBINATION

Figure. 4 shows a radar chart plotting the performances in terms of AUC of the various unimodal and multimodal approaches. As mentioned above and summarised in Table 2 we have a total of 16 different rules, given by the different combinations of the aggregation rules (A_y) and the late fusion rules (LF_x). So for each of these rule combinations we rank each approach so that the one with the highest performance receives a score of 7, whilst the one with the lowest performance gets a score of 1. At the end of this iterative process the rank of each stream is given by the sum of the ranks received for each of the 16 experiments mentioned above. We then normalise for the maximum rank achievable. In the figure, we adopt a filled circle to mark the flow with the highest rank, while the blank circles denote those approaches with a lower rank, whose performances are statistically different from the best one according to the Friedman test with the Iman-Davenport amendment followed by the Bonferroni-Dunn pairwise post-hoc test ($p < 0.1$). Furthermore, we do not report any circle when the rank of a flow is lower than the best one and the corresponding performance are not statistically different.

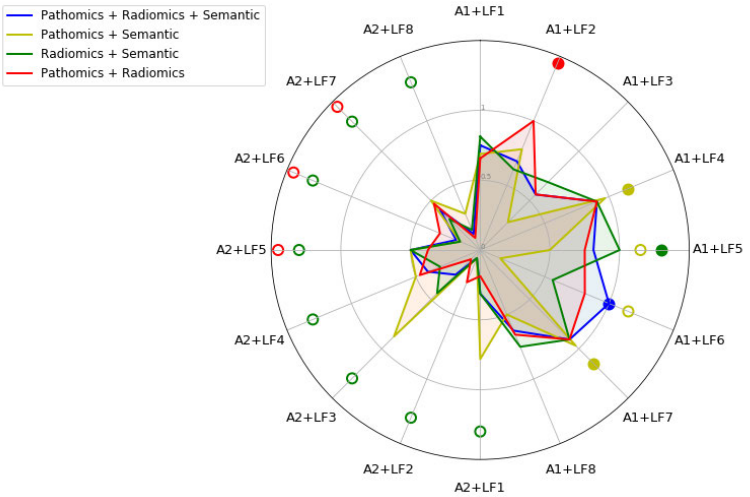


FIGURE 5. Radar chart showing the performance in terms of AUC of the fusion rules given by the combination of LF_x and A_y (Figure. 3) and varying the modalities combination. Filled circles represent the fusion rule combination with the highest rank, whilst blank circles represent models with statistically different performances from the best model according to the Wilcoxon signed-rank test ($p < 0.1$).

Under these premises, in this figure the lengths of the spokes show how the multimodal approaches generally perform better than the unimodal ones, as the latter always differ from the best combination in a statistically significant way. Furthermore, as you can see from the filled circle, the trimodal combination given by $P+R+S$ is the best approach and it significantly differs from all unimodal approaches (i.e. P, R, S) $P + S$. Although in a different clinical context, similar considerations about the inherent superiority of multimodal approaches over the unimodal flows were obtained from [12] and [49]. Indeed in [12] and [49], the overall survival analyses performed for glioma and lung cancer, respectively, show how the multimodal combination significantly outperforms the best performing unimodal flows. Furthermore, as in our work, in both studies the triplet multimodal combination emerged as the best approach.

2) MOST INFORMATIVE UNIMODAL APPROACH

Let us now rank the unimodal approaches in order to understand which flow is the most informative in terms of AUC. The ranks are computed as described before. Indeed, given a fusion rule, the multimodal approaches are ranked so that the flow with the best performance gets rank 4, as we have four different multimodal approaches. Then, all the unimodal approaches that make up the multimodal flow get the same rank as the multimodal one. So, for instance, if the case $P + R + S$ multimodal flow got rank 3, the unimodal approaches of pathomics, radiomics and semantic all get the rank 3. This operation is repeated for all 16 different rule combinations and the final ranks are updated by accumulation and, finally, normalised for the maximum rank achievable.

Hence, with a rank equal to 0.667, the radiomics approach is the most informative unimodal flow, outperforming both the pathomics and semantic flows which got a rank equal to 0.549 and 0.535, respectively. Similar considerations were obtained from [12] where, although in a different clinical context, the overall survival analysis of glioma dealing with the radiomics unimodal flow emerges as the most informative modality in terms of Cox Loss.

3) BEST FUSION RULE

Figure 5 shows a radar chart plotting the performance in terms of AUC of the various fusion rules varying the way we combine the modalities. Let us remember that, as mentioned above and summarised in Table 2, the rules represented in the figure are the combination of the aggregation and late fusion rules, A_y and LF_x, respectively. For each multimodal combination, we ranked each fusion rule in terms of AUC so that the worst rules receives rank 1, whilst the best receives rank 16. Filled circles in the figure represent the fusion rule combination with the highest rank, whilst blank circles represent models with statistically different performances from the best fusion rule according to the Wilcoxon signed-rank test ($p < 0.1$). Note that here we used such test rather than Friedman's method since, for each late fusion rule, we compared the four different multimodal combinations, a number limiting the application of Friedman's test. Furthermore, we do not report any circle when the rank of a flow is lower than the best one and the corresponding performance are not statistically different.

The chart shows that the patient aggregation rule named as score mean aggregation (A_2) performs generally worst than the feature mean rule (A_1). On the other hand, if we focus on the late fusion rules, we observe that they perform almost equally well on all multimodal combinations.

Moreover, the best combinations of rules depend on the modality combination considered. For the $P + R + S$ combination, the best rule combination is denoted as $A_1 + LF_6$, which is therefore given by the use of the feature mean as patient aggregation rule at feature level followed by the

TABLE 5. Exhaustive comparison of late and early multimodal learning for the various modality combinations in terms of AUC. Each cell shows the amount of win–tie–loss of a combination of the corresponding multimodal combination handled in the late fusion paradigm. For each modalities combination, the cells are highlighted in grey when the late fusion approaches are significantly better than the early fusion rules according to the one-tailed sign test ($p < 0.05$).

Multimodal Combination	Simple Concatenation	Kronecker
Pathomics + Radiomics + Semantic	5-2-1	8-0-0
Pathomics + Semantic	8-0-0	8-0-0
Radiomics + Semantic	4-1-3	7-0-1
Pathomics + Radiomics	5-0-3	7-1-0

Dempster Shafer as fusion rule working on the outputs of each classifier. For the $P + S$ combination, the best rules are denoted as $A_1 + LF_4$ and $A_1 + LF_7$, which are given by the combination of the feature mean as patient aggregation rule and, respectively, the mean and majority vote as fusion rule. For the $R + S$ combination, the best rule combination is denoted as $A_1 + LF_5$, which is therefore given by the use of the feature mean followed by the Decision Template as fusion rule. For the $P + R$ combination, the best rule combination is denoted as $A_1 + LF_2$, which is therefore given by the use of the feature mean followed by the maximum as fusion rule. Globally, the best fusion rules are $A_1 + LF_4$ and $A_1 + LF_7$, since on average they ranked highest across all modality combinations. It suggests that these rules are well adapted to different situations, generalising successfully across different types of datasets, which, in turn, are characterised by different data types, sizes and dimensionalities.

B. LATE VS EARLY FUSION

Table 5 shows an exhaustive comparison of late and early multimodal learning for the various modality combinations in terms of AUC. As we already mentioned, with the early fusion paradigm we only tested the A_1 aggregation rule for the sake of consistency of samples. Each cell reports the amount of win–tie–loss of the corresponding multimodal flow handled in the late fusion paradigm. Since early fusion only handles the A_1 aggregation rule, the comparison was restricted to only the 8 late fusion rules that use this method of aggregation. Given the number of patient data, we statistically validated this comparison with the sign test, a simple but powerful statistical test. In Table 5, for each modalities combination, the grey cells highlight when the late fusion approaches are significantly better than early ones according to the one-tailed sign test ($p < 0.05$). The table shows that the late fusion paradigm almost always outperforms the early fusion paradigm for the performance metric considered. This suggests us that there is a certain difficulty in the data-level fusion process when the data are significantly uncorrelated, and have such a different nature and dimensionality, as in the medical task we are dealing with in this work.

C. HAND-CRAFTED VS DEEP FEATURES

This experiment compares the discriminant capacity of hand-crafted features (as discussed in section IV-B) with that of deep features, using the same multimodal late fusion framework (as depicted in Figure 2). Hence, for the two descriptor groups, we have the same number of experiments, i.e., 64 as discussed in section V-A.

TABLE 6. Exhaustive comparison between the performance of modality combinations expressed in terms of AUC, where P stands for Pathomics, R for Radiomics, and S for Semantic. Each cell shows the amount of win–tie–loss of a combination in a row compared with a combination in a column, respectively performed using hand-crafted and deep features. The cells highlighted in grey represent the modality combination significantly better than another according to the one-tailed sign test ($p < 0.05$).

		Deep Features			
		P+R+S	P+S	R+S	P+R
Hand-crafted Features	P+R+S	16-0-0	16-0-0	16-0-0	16-0-0
	P+S	11-0-5	15-1-0	10-3-3	16-0-0
	R+S	16-0-0	16-0-0	14-2-0	16-0-0
	P+R	16-0-0	16-0-0	13-1-2	16-0-0

We summarised the comparison in Table 6, which offers an exhaustive comparison between the modality combinations in terms of AUC. Each cell shows the amount of win–tie–loss of a pair of modality combinations indexed by row and column, respectively performed using hand-crafted and deep features. For instance, the second cell in row 1, with indexes $(1, 2)$, counts the wins, ties and losses obtained by the modalities triplet $P+R+S$ performed with hand-crafted features against the modalities pair $P + S$ performed with deep features. Since for each combination we have a total of 16 different combinations of the fusion and aggregation rules, LF_x and A_y respectively, the total amount of scores for each cell is equal to 16.

Again, we statistically validated this comparison with the sign test. In Table 6 the grey cells represent the modality combination significantly better than another according to the one-tailed sign test ($p < 0.05$). From the table we can see how the hand-crafted features perform better than the deep features. The reason for this result can be found in the low dimensionality of the dataset, as could be expected. Indeed, this limits the ability of the deep neural networks to fully express their power of abstraction, generalisation and discrimination.

VII. CONCLUSION

In this work, we have presented a multimodal late fusion framework combing radiomics, pathomics and clinical data to predict radiation therapy treatment outcomes for NSCLC patients. We fed the proposed framework with hand-crafted features extracted from the aforementioned data sources. Here, we explored the combinations of eight different late fusion rules (i.e., product, maximum, minimum, mean, decision template, Dempster-Shafer, majority voting, and confidence rule) with two samples' patient-wise aggregation rules (i.e., feature mean and score mean) implemented to

have a single classification per patient and consistent sample fusion, for a total a 64 experiments.

The take-home message emerging from this work is that the multimodal learning framework leads to a significant improvement of a learning system in terms of performance. Indeed, in this work the simultaneous fusion of the three modalities is the best approach and it significantly differs from all the models fed with the stand-alone data flows. Although in a different clinical context, similar considerations about the inherent superiority of multimodal approaches over the unimodal ones were obtained from [12] and [49].

While our work demonstrates the potential of the multimodal framework to predict radiotherapy outcomes, some limitations must be acknowledged. Although handcrafted features still show remarkable performance in low-dimensional datasets, such as the one in our study, they may be prone to human biases, which could adversely affect the accuracy and limit the generalisability of the results.

Future work will focus on the following directions to overcome these limitations. External validation on independent datasets, when will be available, would help us robustly assess the performance of the proposed multimodal framework on new patient cohorts. Furthermore, by increasing the dimensionality of the dataset, deep learning approaches can be reconsidered. These have the potential to automatically learn complex patterns from the data, which could potentially improve the accuracy, reliability, and robustness of the prediction model, as happened in other fields. Next, we deem that incorporating the eXplainable AI paradigm into the proposed multimodal late fusion framework is a direction worth investigating. Indeed, it underlies the mechanisms that drive predictions, which is required to help clinicians to justify and make informed evidence-based decisions [50]. By making the proposed framework interpretable, we can facilitate the adoption of AI techniques into current medical practice and improve patient outcomes.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence (bias) the work reported in this paper.

AUTHOR CONTRIBUTIONS

Matteo Tortora: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing–original draft, writing–review and editing, visualization. Ermanno Cordelli: Conceptualization, resources, data curation, writing–review and editing. Rosa Sicilia: Conceptualization, data curation, writing–review and editing. Lorenzo Nibid: Data curation, writing–review and editing. Edy Ippolito: Data curation, writing–review and editing. Giuseppe Perrone: Conceptualization, resources, data curation, writing–review and editing. Sara Ramella: Conceptualization, resources, data curation, writing–review and editing, project administration, funding acquisition. Paolo Soda: Conceptualization, methodology, validation, resources, formal analysis, investigation, writing–original draft, writing–review and editing, visualization, supervision, project administration, funding acquisition.

APPENDIX A
STATISTICAL TESTS

In this work, we used various statistical tests to validate and compare our approaches, which are outlined in the following sections. We recommend referring to [51] for interested readers to provide further insight into these methods.

A. FRIEDMAN TEST WITH IMAN AND DAVENPORT AMENDMENT

The Friedman test with an amendment by Iman and Davenport is a non-parametric statistical test used to compare multiple models. For each of the 16 different combinations of fusion rules ($A_y + LF_x$), each flow is ranked so that the best receives rank 1, whilst the worst receives rank 7. Tied ranks are shared equally as explained above. The test statistic with the amendment proposed by Iman and Davenport is the following:

$$F_F = \frac{(N-1)\,\chi_F^2}{N\,(M-1) - \chi_F^2}$$

which follows the F-distribution with $(M-1)$ and $(M-1)\,(N-1)$ degrees of freedom and where:

$$\chi_F^2 = \frac{12N}{M\,(M+1)} \left(\sum_{j=1}^{M} R_j^2 - \frac{M(M+1)^2}{4} \right)$$

where $R_j = \frac{1}{N} \sum_{i=1}^{N} r_i^j$ is the average rank of the j-th flow and r_i^j is the rank of the j-th flow when considering the i-th fusion rule, where $i = 1, \ldots, N$ and $j = 1, \ldots, M$ ($N = 16, M = 7$). Once the F-statistic has been computed, we can carry out the test comparing it with the critical value for the chosen level of significance. If it is greater than this value we can reject the null hypothesis H_0 and accept that there is a difference between the flows.

Bonferroni-Dunn Post-Hoc Test: If H_0 is rejected, Bonferroni-Dunn post-hoc test is applied to find exactly where the differences are:

$$z = \frac{R_1 - R_j}{\sqrt{\frac{M(M+1)}{6N}}}$$

where R_1 is the average rank of the best flow and R_j is the average rank of j-th flow. Two flows are statistically different if the obtained p-value from this z-value is smaller than $\frac{\alpha}{M-1}$, where α is the desired level of significance.

B. WILCOXON SIGNED RANK TEST

The Wilcoxon signed-rank test is a non-parametric statistical test that tests if two models are statistically different. Given the error estimates of two models for the N folds of the LOPO validation paradigm, the test computes the differences

of these errors d_i. Then it ranks the absolute values of the differences $|d_i|$ so that the smallest value receives rank 1, whilst the largest one receives rank N. If there is a tie, all the ranks are shared so that the total sum stays $1 + 2 + \cdots + N$. Subsequently, it splits the ranks into positive and negative according to the sign of d_i and calculates the following amounts:

$$R_+ = \sum_{d_i>0} r_i + \frac{1}{2}\sum_{d_i=0} r_i, \qquad R_- = \sum_{d_i<0} r_i + \frac{1}{2}\sum_{d_i=0} r_i$$

Finally, the test computes the following statistic:

$$z = \frac{T - \frac{1}{4}N\,(N+1)}{\sqrt{\frac{1}{24}N\,(N+1)\,(2N+1)}}$$

which is approximately distributed as a normal distribution and where $T = \min(R_+, R_-)$. Once the z-statistic has been computed, we can carry out the test comparing it with the critical value for the chosen level of significance. If it is greater than this value we can reject the null hypothesis and state that the two methods have statistically different performances.

C. SIGN TEST

The sign test is simply performed counting wins, ties and losses, with or without statistical significance, of each method pair. This test is based on the intuition that if two methods are equivalent, each one will perform better than the other one on approximately $N/2$ of the tests. Hence, following the binomial distribution, we can claim that the first method is significantly better than the second one if its amount of wins is greater than $N/2 + 1.96\sqrt{N/2}$, at a level of significance of 0.05.

APPENDIX B
STATISTICAL DESCRIPTORS

The following are the statistical measures extracted from the intensity histogram and used in this work to provide a synthetic description of both the LBP and the grey level histogram:

- **Mean**:

$$m = \sum_{k=0}^{l-1} r_k p\,(r_k)$$

- **Standard deviation**:

$$\sigma = \sqrt{\sum_{k=0}^{l-1} (r_k - m)^2 p\,(r_k)}$$

- **Smoothness**:

$$R = 1 - \frac{1}{\left(1+\sigma^2\right)}$$

- **Skewness**:

$$\text{skewness} = \sum_{k=0}^{l-1} (r_k - m)^3 p\,(r_k)$$

- **Kurtosis**:

$$\text{kurtosis} = \sum_{k=0}^{l-1} (r_k - m)^4 p\,(r_k)$$

- **Energy**:

$$\text{energy} = \sum_{k=0}^{l-1} p\,(r_k)^2$$

- **Entropy**:

$$\text{entropy} = -\sum_{k=0}^{l-1} p\,(r_k)\log_2\left[p\,(r_k)\right]$$

- **Absolute maximum**:

$$\max_{k=0}^{l-1}\left[p\,(r_k)\right]$$

- **Maximum value**:

$$\arg\max_{k=0}^{l-1}\left[p\,(r_k)\right]$$

where l is the number of grey levels in the image I and $p\,(r_k)$ is the number of pixels with a grey level equal to r_k.

Given instead a grey-level co-occurrence matrix G extracted from image I, the following are the formal definition of Haralick features used in this work:

- **Contrast**:

$$\sum_{i,j=0}^{l-1} G_{i,j}\left(g_i - g_j\right)^2$$

- **Dissimilarity**:

$$\sum_{i,j=0}^{l-1} G_{i,j}\left|g_i - g_j\right|$$

- **Homogeneity**:

$$\sum_{i,j=0}^{l-1} \frac{G_{i,j}}{1 + (g_i - g_j)^2}$$

- **Angular Second Moment (ASM)**:

$$\sum_{i,j=0}^{l-1} G_{i,j}^2$$

- **Energy**:

$$\sqrt{ASM}$$

- **Correlation**:

$$\sum_{i,j=0}^{l-1} G_{i,j}\left[\frac{(g_i - \mu_i)\,(g_j - \mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right]$$

where $G_{i,j}$ denotes the i,j-th entry of G, l is the number of grey levels in the image I, g_i and g_j denote two grey level values $\in \left[0, 2^l - 1\right]$, and, finally, μ_i denotes the mean value of the one-dimensional marginal distributions of G.

REFERENCES

[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros. (2020). *Global Cancer Observatory: Cancer Today*. Accessed: Apr. 4, 2023. [Online]. Available: https://gco.iarc.fr/today

[2] American Cancer Society. (2023). *Key Statistics for Lung Cancer*. Accessed: Apr. 4, 2023. [Online]. Available: https://www.cancer.org/cancer/lung-cancer

[3] P. E. Postmus, K. M. Kerr, M. Oudkerk, S. Senan, D. A. Waller, J. Vansteenkiste, C. Escriu, and S. Peters, "Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 28, pp. iv1–iv21, Jul. 2017.

[4] G. Lee, H. Y. Lee, H. Park, M. L. Schiebler, E. J. R. van Beek, Y. Ohno, J. B. Seo, and A. Leung, "Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art," *Eur. J. Radiol.*, vol. 86, pp. 297–307, Jan. 2017.

[5] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. W. L. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012.

[6] V. Kumar, "Radiomics: The process and the challenges," *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012.

[7] R. Gupta, T. Kurc, A. Sharma, J. S. Almeida, and J. Saltz, "The emergence of pathomics," *Current Pathobiol. Rep.*, vol. 7, no. 3, pp. 73–84, Sep. 2019.

[8] Y. Zhang, K. He, Y. Guo, X. Liu, Q. Yang, C. Zhang, Y. Xie, S. Mu, Y. Guo, Y. Fu, and H. Zhang, "A novel multimodal radiomics model for preoperative prediction of lymphovascular invasion in rectal cancer," *Frontiers Oncol.*, vol. 10, p. 457, Apr. 2020.

[9] Y. Wu, J. Ma, X. Huang, S. H. Ling, and S. W. Su, "DeepMMSA: A novel multimodal deep learning method for non-small cell lung cancer survival analysis," 2021, *arXiv:2106.06744*.

[10] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[11] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng, "A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data," *IRBM*, vol. 43, no. 1, pp. 62–74, Feb. 2022.

[12] N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe, and J. Venkataraman, "Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 667–677.

[13] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–9, Oct. 2020.

[14] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[15] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[16] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5198–5204.

[17] G. D. C. Cavalcanti, L. S. Oliveira, T. J. M. Moura, and G. V. Carvalho, "Combining diversity measures for ensemble pruning," *Pattern Recognit. Lett.*, vol. 74, pp. 38–45, Apr. 2016.

[18] H. Abdel-Nabi, M. Ali, A. Awajan, M. Daoud, R. Alazrai, P. N. Suganthan, and T. Ali, "A comprehensive review of the deep learning-based tumor analysis approaches in histopathological images: Segmentation, classification and multi-learning tasks," *Cluster Comput.*, pp. 1–41, Jan. 2023.

[19] M. Ferro, "Radiomics in prostate cancer: An up-to-date review," *Therapeutic Adv. Urol.*, vol. 14, Jul. 2022, Art. no. 17562872221109020.

[20] L. Balkenende, J. Teuwen, and R. M. Mann, "Application of deep learning in breast cancer imaging," *Seminars Nucl. Med.*, vol. 52, no. 5, pp. 584–596, Sep. 2022.

[21] P. Monkam, S. Qi, H. Ma, W. Gao, Y. Yao, and W. Qian, "Detection and classification of pulmonary nodules using convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 78075–78091, 2019.

[22] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.

[23] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[24] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[25] N. C. D'Amico, R. Sicilia, E. Cordelli, L. Tronchin, C. Greco, M. Fiore, A. Carnevale, G. Iannello, S. Ramella, and P. Soda, "Radiomics-based prediction of overall survival in lung cancer using different volumes-of-interest," *Appl. Sci.*, vol. 10, no. 18, p. 6425, Sep. 2020.

[26] A. Chaddad, P. O. Zinn, and R. R. Colen, "Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 84–87.

[27] N. Aggarwal and R. K. Agrawal, "First and second order statistics features for classification of magnetic resonance brain images," *J. Signal Inf. Process.*, vol. 3, no. 2, pp. 146–153, 2012.

[28] R. Rashmi, K. Prasad, C. B. K. Udupa, and V. Shwetha, "A comparative evaluation of texture features for semantic segmentation of breast histopathological images," *IEEE Access*, vol. 8, pp. 64331–64346, 2020.

[29] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[30] P. Kral and L. Lenc, "LBP features for breast cancer detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2643–2647.

[31] E. Cordelli, G. Maulucci, M. D. Spirito, A. Rizzi, D. Pitocco, and P. Soda, "A decision support system for type 1 diabetes mellitus diagnostics based on dual channel analysis of red blood cell membrane fluidity," *Comput. Methods Programs Biomed.*, vol. 162, pp. 263–271, Aug. 2018.

[32] R. Sicilia, E. Cordelli, M. Merone, E. Luperto, R. Papalia, G. Iannello, and P. Soda, "Early radiomic experiences in classifying prostate cancer aggressiveness using 3D local binary patterns," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 355–360.

[33] L. Joo, S. C. Jung, H. Lee, S. Y. Park, M. Kim, J. E. Park, and K. M. Choi, "Stability of MRI radiomic features according to various imaging parameters in fast scanned T2-FLAIR for acute ischemic stroke patients," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, Aug. 2021.

[34] S. Ramella, M. Fiore, C. Greco, E. Cordelli, R. Sicilia, M. Merone, E. Molfese, M. Miele, P. Cornacchione, E. Ippolito, G. Iannello, R. M. D'Angelillo, and P. Soda, "A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0207455.

[35] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[36] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[37] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognit.*, vol. 34, no. 2, pp. 299–314, Feb. 2001.

[38] L. Kuncheva, *Combining Pattern Classifiers Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2004.

[39] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, pp. 278–282.

[40] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: Experiences from the scikit-learn project," in *Proc. ECML PKDD Workshop, Lang. Data Mining Mach. Learn.*, 2013, pp. 108–122.

[41] A. Arcuri and G. Fraser, "Parameter tuning or default values? An empirical investigation in search-based software engineering," *Empirical Softw. Eng.*, vol. 18, no. 3, pp. 594–623, Jun. 2013.

[42] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, Oct. 2017.

[43] N. C. D'Amico, "Tackling imbalance radiomics in acoustic neuroma," *Int. J. Data Mining Bioinf.*, vol. 22, no. 4, pp. 365–388, 2019.

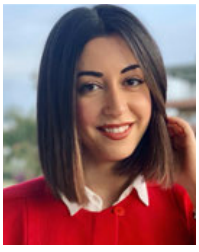
[44] P. Soda, "AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102216.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[47] C. Z. Liu, R. Sicilia, M. Tortora, E. Cordelli, L. Nibid, G. Sabarese, G. Perrone, M. Fiore, S. Ramella, and P. Soda, "Exploring deep pathomics in lung cancer," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2021, pp. 407–412.
[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017, pp. 1–4.
[49] P. Grossmann, O. Stringfield, N. El-Hachem, M. M. Bui, E. R. Velazquez, C. Parmar, R. T. Leijenaar, B. Haibe-Kains, P. Lambin, R. J. Gillies, and H. J. Aerts, "Defining the biological basis of radiomic phenotypes in lung cancer," *eLife*, vol. 6, Jul. 2017, Art. no. e23421.
[50] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107161.
[51] L. I. Kuncheva, *Combining Pattern Classifiers: Methods Algorithms*. Hoboken, NJ, USA: Wiley, 2014.

MATTEO TORTORA (Graduate Student Member, IEEE) received the bachelor's degree in industrial engineering and the master's degree (Hons.) in biomedical engineering from the Campus Bio-Medico University of Rome, Rome, Italy, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree in biomedical engineering (artificial intelligence area) with the Computer Systems and Bioinformatics (CoSBi) Research Laboratory, Faculty of Engineering. His current research interests include reinforcement learning, machine learning, computer vision, and multimodal learning.

ERMANNO CORDELLI received the master's degree in biomedical engineering and the Ph.D. degree in biomedical engineering (computer science area) from the Campus Bio-Medico University of Rome (UCBM), in 2014 and 2017, respectively. He was on a collaboration contract with the Computer Systems and Bioinformatics (CoSBi) Research Laboratory, Faculty of Engineering, UCBM. He is currently an Assistant Professor with CoSBi Laboratory, UCBM. His main research interests include artificial intelligence and its applications in the health sector, federated learning, computer vision, radiomics, and the IoT, and working on a project on the creation of an intelligent pen for diabetes treatment.



ROSA SICILIA was born in Cosenza, Italy, in 1993. She received the degree (Hons.) in biomedical engineering and the Ph.D. degree in biomedical engineering (computer science area) from the Campus Bio-Medico University of Rome (UCBM), Rome, in 2016 and 2020, respectively. After one year as a Postdoctoral Researcher, she is currently an Assistant professor (RTDA) with UCBM, a position co-funded by Regione Lazio to work on the project "Ease-it: A smart and intelligent outpatient clinic for Hospital 4.0." Her main research interests include machine learning and multimodal data mining.

LORENZO NIBID received the Medical Doctor degree from the Campus Bio-Medico University of Rome (UCBM), Rome, in 2020. He is currently a Resident Physician in pathological anatomy with the University Hospital, Department of Medicine, UCBM.

EDY IPPOLITO received the Medical Doctor and Specialist degrees in radiation oncology from the Catholic University of the Sacred Heart, Rome, in 2004 and 2008, respectively. She is currently a Researcher in diagnostic imaging and radiotherapy with the Campus Bio-Medico University of Rome (UCBM), a member of the Breast Unit, and a Doctor in radiation oncology with the University Hospital, UCBM.

GIUSEPPE PERRONE received the Medical Doctor and Specialist (Hons.) degrees in anatomical pathology from the Campus Bio-Medico University of Rome (UCBM), Rome, in 2001 and 2006, respectively, and the Ph.D. degree from the Sapienza University of Rome, in 2009. He is currently a Full Professor of pathology with the Medical School of Medicine, UCBM, and the Director of the Pathology Department, University Hospital, UCBM. His research interests include tissue biomarkers research characterization, lung, liver, biliary, and pancreatic pathology.

SARA RAMELLA received the Medical Doctor and Specialist (Hons.) degrees in radiation oncology from the Catholic University of the Sacred Heart, Rome, in 2000 and 2004, respectively. She is currently the Director of radiation oncology with the University Hospital, Campus Bio-Medico University of Rome (UCBM), a Delegate Prorector for integration and social impact with UCBM, the President of the degree course in radiology, diagnostic imaging, and radiotherapy techniques with UCBM, a Full Professor of diagnostic imaging and radiation oncology with UCBM, and the Director of the Residency Program in Radiation Oncology, UCBM.

PAOLO SODA (Member, IEEE) received the degree (Hons.) in biomedical engineering and the Ph.D. degree in biomedical engineering (computer science area) from the Campus Bio-Medico University of Rome (UCBM), Rome, in 2004 and 2008, respectively. He is currently a Full Professor of computer science and computer engineering with UCBM. His research interests include artificial intelligence, pattern recognition, machine learning, big data analytics, and data mining applied to data, signals, 2D and 3D image and video processing and analysis.

Open Access funding provided by 'Università ''Campus Bio-Medico'' di Roma' within the CRUI CARE Agreement