Cross-Modality Calibration in Multi-Input Network for Axillary Lymph Node Metastasis Evaluation

Michela Gravina, Domiziana Santucci, Ermanno Cordelli, Paolo Soda, and Carlo Sansone *Senior Member, IEEE*

***Abstract*—The use of deep neural networks (DNNs) in medical images has enabled the development of solutions characterized by the need of leveraging information coming from multiple sources, raising the Multimodal Deep Learning. DNNs are known for their ability to provide hierarchical and high-level representations of input data. This capability has led to the introduction of methods performing data fusion at an intermediate level, preserving the distinctiveness of the heterogeneous sources in modality-specific paths, while learning the way to define an effective combination in a shared representation. However, modeling the intricate relationships between different data remains an open issue. In this paper, we aim to improve the integration of data coming from multiple sources. We introduce between layers belonging to different modality-specific paths a Transfer Module (TM) able to perform the cross-modality calibration of the extracted features, reducing the effects of the less discriminative ones. As case of study, we focus on the axillary lymph nodes metastasis evaluation in malignant breast cancer, a crucial prognostic factor, affecting patient's survival. We propose a Multi-Input Single-Output 3D Convolutional Neural Network (CNN) that considers both images acquired with multiparametric Magnetic Resonance and clinical information. In particular, we assess the proposed methodology using four architectures, namely BasicNet and three ResNet variants, showing the improvement of the performance obtained by including the TM in the network configuration. Our results achieve up to 90% and 87% of accuracy and Area under ROC curve, respectively when the ResNet10 is considered, surpassing various fusion strategies proposed in the literature.**

***Impact Statement*—In the context of breast cancer, the metastatic involvement of axillary lymph nodes (ALN) stands out as a crucial prognostic factor, reflecting the intrinsic behavior of the primary tumor. Multiparametric Magnetic Resonance Imaging (MRI) enables a comprehensive examination, providing both physiological and morphological characteristics through sequences involving pre-contrast and post-contrast agent administration. This highlights the necessity of integrating diverse information, particularly when considering histological data in conjunction with images. However, current state-of-art solutions typically exploit features extracted from the post-contrast series, neglecting the use of the others. The methodology presented in this paper harnesses Multimodal Deep Learning (MDL) to overcome this limitation, efficiently integrating clinical information and features from multiple image modalities. Demonstrating an accuracy of 90% in the metastatic evaluation of ALN, our algorithm has the potential to support radiologists in their daily analysis of breast MRI.**

***Index Terms*—Axillary lymph node, Brest Cancer, Convolutional Neural Networks, Cross-modality Calibration, Medical imaging analysis, Multimodal Deep Leaning**

I. Introduction

Artificial Intelligence (AI) has been applied in medical image analysis with very promising results. Its use provides an efficient way of finding non-invasive and quantitative assessments of diseases, highlighting pattern changes or intrinsic characteristics hidden from the human eye, and offering the opportunity to better understand disease processes [1]. Radiomics is one of the most advanced applications for AI in medical imaging. Initially, it extracts a large amount of quantitative, reproducible information, called features from medical images [2] which may reflect the pathophysiology of the analyzed tissues. Then, after a careful features selection step, it uses Machine Learning (ML) models to provide tools to predict different outcomes all with predictive horizons. Recently, Deep Learning (DL) approaches have improved the handcrafted pipeline by automatically learning from images the set of features that well fits the specific task to solve. A key role is played by Convolutional Neural Networks (CNNs), a set of Deep Neural Networks (DNNs) commonly applied in image processing, for their ability to capture spatial dependencies. The use of DNNs has also enabled the development of DL-based solutions in medical applications characterized by the need of leveraging information coming from multimodal data sources [3], raising the Multimodal Deep Learning (MDL) [4], [5], [6]. Despite the presence of techniques to perform the integration at data-level (early fusion) or decision level (late fusion), the characteristic of the DNNs to transform raw inputs into hierarchical and higher-level representations has encouraged the implementation of methods that integrate information at an intermediate stage. These methods aim to preserve the distinctiveness characteristics of the heterogeneous sources, while providing an effective way for their combination. To this end, the intermediate fusion (IF) technique represents a very flexible approach merging in shared representations units coming from multiple modality-specific paths [4], namely, modality-specific DNNs. Since in DL approaches it is possible to implement end-to-end training, the resulting architecture autonomously learns the shared representation well-suited for

Manuscript received 23 May 2023.

Michela Gravina and Carlo Sansone are with the Department of Electrical Engineering and Information Technology (DIETI) of the University of Naples Federico II, Via Claudio 21, 80125, Naples, Italy (email: michela.gravina@unina.it, carlo.sansone@unina.it).

Domiziana Santucci is with the Department of Radiology, University of Rome "Campus Bio-Medico", Via Alvaro del Portillo, 21, 00128 Rome, Italy (e-mail: d.santucci@policlinicocampus.it).

Ermanno Cordelli is with the Unit of Computer Systems and Bioinformatics, Dept. of Engineering, University of Rome Campus Bio-Medico, via Alvaro del Portillo 21, 00128, Roma, Italy (email: e.cordelli@unicampus.it).

Paolo Soda is with the Department of Diagnostics and Intervention, Radiation Physics, Biomedical Engineering, Umea University, Universitetstorget 4, 90187, Umea, Sweden, and with the Unit of Computer Systems and Bioinformatics, Dept. of Engineering, University of Rome Campus Bio-Medico, via Alvaro del Portillo 21, 00128, Roma, Italy (email: p.soda@unicampus.it)

This paragraph will include the Associate Editor who handled your paper.

the given task. However, the effectiveness of the integration of multiple modalities depends on the ability to accurately account for the intricate relationships between data obtained from diverse sources. Different data-acquisition methods may highlight features that are not equally important, which requires reducing the impact of unnecessary information and reinforcing relevant patterns.

In this paper, we leverage the flexibility of the IF approach, proposing an innovative Transfer Module (TM) to model the complex interaction of multiple sources. When inserted between layers belonging to the different modality-specific DNNs, the TM acts on the extracted features maps and performs the cross-modality calibration taking into account the complementary nature of the inputs. The implemented gating mechanism autonomously highlights the effects of the most discriminative characteristics while reducing the less useful ones. The result is a module that improves the IF strategy considering the inherent relationships between heterogeneous data sources before their combination.

We consider the axillary lymph node status (ALNS) assessment in breast cancer (BC) as a case of study, since it represents one of the most important independent prognostic factors, affecting patients' survival. Currently, the radiologists who are faced with the evaluation of the axilla are led to consider some characteristics of the primary tumor itself, whose intrinsic behavior, morphology, and angioinvasivity reflect the metastatic involvement of the axillary lymph node. They rely on the multiparametric Magnetic Resonance Imaging (MRI), always performed for primary BC stage definition [7], [8], [9], together with histological examination resulting from core-biopsy or surgery. Pre-contrast, such as T2-weighted (T2) and Diffusion weighted (DWI), and post-contrast agent administration image sequences, such as the Dynamic-Contrast Enhanced (DCE), highlight useful and complementary information for tumor evaluation [7], [10], [11], [12], [13], such as morphology and associated edema, tissue organization at the microscopic level and perfusional behavior. This makes the ALNS multimodal by its nature, especially in addition with the histological data. The majority of work in the literature have exploited handcrafted features [14] computed from DCE images and shallow ML algorithms, neglecting the use of the other MRI sequences. Conversely, in this paper, we aim to include the complementary information provided by multiparametric MRI and histological examination. In particular, we consider image acquisitions and histological data as specific input modalities, since each source provides a different perspective of the same disease. As a consequence, the multimodal evaluation of the primary BC for ALNS assessment becomes an interesting case of study for our proposal, including three MRI sequences (DCE, T2, and DWI), and clinical and histological characteristics of the primary tumor. It is worth noting that, although we assess the proposed methodology in a case of study involving a multiparametric MRI, there is nothing preventing the use of the TM in scenarios exploiting heterogeneous acquisition tools.

The rest of the paper is organized as follows: Section II introduces the main concepts of multimodal learning while Section III describes the radiomic-based approaches proposed in the literature for ALNs evaluation; Section IV reports the proposed methodology with the Transfer Module; Section V details the involved dataset; Section VI describes the experimental setup; Section VII discusses the obtained results; finally Section VIII provides some conclusions.

II. Multimodal Deep Learning

The presence of multiple sources allows a deeper understanding of the system under analysis, improving the decision-making process, and identifying the existent relations between data modalities [15]. Despite the potential benefits, how to exploit diverse information is still a challenging task [16], [17], [18]. Data modalities reflect the inherent characteristics of the heterogeneous and complex acquisition tools, making the diversity of the sources the key aspect for complete knowledge, but, at the same time, one of the main complexities to manage [16], due to conflicts and inconsistencies that may occur [18]. The presence of multiple sources requires the introduction of approaches able to preserve the distinctiveness of each modality while providing efficient fusion methods [16].

Multimodal Deep Learning (MDL) exploits DL techniques to implement methods allowing the fusion of complementary information coming from heterogeneous sources. Deep neural networks aim to learn high-level representations of the input data, from simple to complex abstractions, making MDL able to model nonlinear relationships between modalities [17]. Among all deep neural networks, CNNs are widely used in medical imaging with surprising results [19], [20]. A typical CNN consists of stacked relatively complex layers [21], with each of them usually having a convolutional stage, a non-linearity function (i.e., ReLU), and a pooling operation. The set of complex layers constitutes the Convolutional Core (Conv-C), responsible for the features extraction step, while the classification is performed by a Classification Core (CC), typically including fully connected layers.

In recent years, several data fusion techniques have been investigated in the research community [6], [22], [23], [5] resulting in three main categories: early fusion or data-level, late fusion or decision-level, intermediate or joint fusion.

Early fusion consists of the integration of different and heterogeneous sources of data in a single structure that is then used as input to a learning model. In the case of medical image analysis using CNNs, the simplest strategy involves concatenating the acquisitions in a single volume [24]. However, the inherent characteristics of each imaging modality, such as different resolutions or sampling times, may make the creation of a single structure very complex [4], [16], causing the early fusion not fully take into account the complementary nature of the images, generating vectors with redundancy.

Late fusion integrates the decision from different models, each trained on a specific image modality. In other words, this technique combines the decision of independent "experts", exploiting the fact that errors from multiple models should be uncorrelated. There are different combining strategies such as majority voting, averaged fusion, Bayes'rule, or those exploiting the use of a meta-model [4].

Intermediate fusion (IF) leverages the ability of DNNs to transform raw inputs into higher-level representations, aiming to create a shared representation [4], [16]. In a basic IF approach, the features maps extracted from Conv-Cs related to different image modalities (modality-specific paths) are merged into a single structure before feeding the CC, as shown in Figure 1, where the symbols ◯ represents a generic fusion operation.

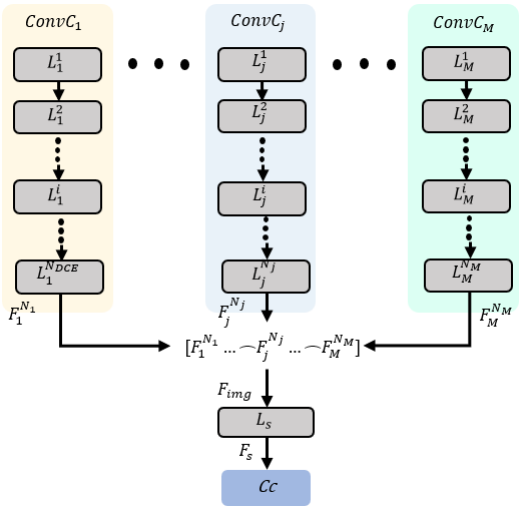

Fig. 1: Basic schema for Intermediate Fusion approach: j is a generic image modality, with j from 1 to M, representing the number of modalities; $ConvC_j$ is the Conv-C of the modality j; and L_j^i indicates the $i-th$ layer in the $ConvC_j$ with i from 1 to N_j that is the number of layers in $ConvC_j$; $F_j^{N_j}$ is the resulting image-specific features map for each Conv-C

Let j be one of the M modalities, $ConvC_j$ be the Conv-C of the modality j, and L_j^i indicate the ith layer in $ConvC_j$, with i ranging in $[1, N_j]$, where N_j that is the number of layers in $ConvC_j$. Denoting with x_j the input image, the resulting image-specific features map $F_j^{N_j}$ for each modality is formalized as follows:

$$F_j^{N_j} = ConvC_j(x_j) = L_j^{N_j}(...L_j^i(...L_j^2(L_j^1(x_j))...)...) \quad (1)$$

The fusion operation ◯ processes the extracted features maps $F_j^{N_j}$, resulting in

$$F_{img} = [F_1^{N_1} \bigcirc F_2^{N_2} ... \bigcirc F_j^{N_j} ... \bigcirc F_M^{N_M}] \quad (2)$$

where F_{img} is the *shared features map*, that represents the input for the *shared path*, specified by the L_s block in Figure 1, and the classification core. In the literature, common strategies to implement ◯ include the concatenation ($\frown$), the element-wise summation ($+$) and product ($\times$), combinations of these operations [17], other functions such as Kronecker product [25], [26], or more complex methods consisting in tensor-based operations or attention mechanisms [27]. Equation 1 highlights that each image-specific features map is computed by only considering a single modality. As a consequence, the complementary information coming from different sources is exploited after the fusion operation (◯) reported in Equation 2, without affecting the features extraction process. During the training phase, the loss is back-propagated to all the convolutional cores so that the CNNs can provide a shared representation that is well-suited to the task to be solved.

It is not easy to understand when the modality-specific representation should be merged into a shared representation. In the literature, while different solutions proposed a single fusion layer, several approaches [28], [29] implement a gradual fusion strategy. The choice of which modality to fuse at which depth of representation can be very challenging, especially in cases where more than two sources are present. In [30] authors proposed a search algorithm called Multimodal Fusion Architecture Search spanning different fusion architectures. The search time depends on the size of the space to be analyzed together with the dimension of the dataset and the complexity of the involved networks. This makes hardware with multiple Graphics Processing Units necessary and limits the applicability of the strategy in diverse scenarios. Moreover, different works in the literature [31], [32] create rich representations concatenating the features maps extracted from the involved CNNs, without taking into account the dependency and the correlation of the extracted features.

The effectiveness of the integration of multiple modalities is affected by the ability to model the complex interactions of data coming from heterogeneous sources [18]. Indeed, different acquisition tools may reveal characteristics that are not equally useful for the specific task to be solved, resulting in the need to reduce the contribution of superfluous or redundant information while enhancing the important patterns [18]. The methodology reported in [33] aims to further intensify the dependency of the modality-specific paths, with the proposal of a Multimodal Transfer Module (MMTM). When inserted between layers belonging to the sub-networks of different sources, the MMTM improves the integration of the features maps, emphasizing the most important features while suppressing the contribution of the less important ones through an excitation process inspired by [34]. In [35], authors proposed a framework for multimodal image synthesis, artificially producing the missing modality for each patient. To exploit the relation between different acquisition tools a Mixed Fusion Block (MFB), an adaptive module designed to integrate heterogeneous information, is proposed. Finally, the methodology described in [36] introduced a Correlation-based Attention Feature Fusion (CAFF) module that is inserted in a neural network to modify the extracted features maps exploiting the correlation among channels.

Taking into account the need of providing an efficient integration of heterogeneous sources, the contributions of our work in the definition of an IF strategy can be summarised as follows:

- We propose an innovative Transfer Module (TM) that can be inserted between layers belonging to different modality-specific paths to model the complex interaction of heterogeneous data.
- We make our TM able to modify the extracted features maps in each modality-specific path taking into account the complementary nature of the inputs. We will refer to this procedure as *cross-modality calibration*.
- We implement in the TM a specific operation, denoted

as *gating mechanism*, that reduces the importance of the least discriminative characteristics in each modality-specific features maps exploiting the descriptors of all the inputs.
- We include the TM in the training process, making the definition of the most relevant features well-suited for the specific task to solve.

III. Radiomic-based Approaches for ALN Evaluation

The application of radiomics for axillary lymph node status prediction from primary tumor is relatively recent, resulting in a small number of papers in the literature [14], which are highly heterogeneous in terms of features extraction/selection and trained classifiers. Most of the proposals extract handcrafted features considering morphological [37], [38], [39], [40], [41], [42], [43], [44], first-order [40], [41], [42], [45], [44], and textural characteristics [37], [40], [41], [43], [46], [39], [45], [47], [44]. To predict ALN metastasis, several solutions consider shallow learners such as Support Vector Machines [40], [41], [42], [37], [48], [44], Logistic Regression [43], [46], [47], Linear Discriminant Analysis [38], [39] and Random Forest [45], [49], [50]. It is interesting to note that there is an almost exclusive use of the post-contrast sequence, that is the DCE, representing the most sensitive imaging modality in comparison to the T2 and DWI acquisitions [51]. However, in the cases where the pre-contrast volumes are also used, the solutions get better performance [52], [13]. As a consequence, recent works started to propose multimodal approaches exploiting multiparametric MRI data. For example, authors in [42], [38] extracted handcrafted features from DCE, T2, and DWI sequences, while the methodology in [44] also added the Positron Emission Tomography (PET) in the analysis.

The application of DL to ALN evaluation is relatively recent, with only a few works using MR images of the primary tumor. In particular, Gao et al. [53] evaluated DCE-MRI of 941 patients and proposed a model containing a 3D Deep residual network (ResNet) architecture and a convolutional block attention module, showing values of area under the ROC curve (AUC) equal to 90.7% and 85.2% in the internal and test cohorts, respectively. Similarly, the solution proposed in [54] involved the fine-tuning of ResNet18 [55], considering the cross-sectional slice with the largest portion of primary lesion as input. Moreover, in our previous work [56], [57], we investigated the role of the tissue surrounding the tumor area, proposing different bounding options. The classification is performed by CNNs, obtaining 74% of sensitivity in a dataset with 153 BC patients. More recently, Zhou et al. [58] proposed an ensemble of three state-of-art networks (ResNet101 [55], DenseNet [59], and ResNetXt101 [60]) for the ALN metastasis prediction in DCE-MRI, obtaining a value of AUC equal to 91.7% in the external test set.

The first attempt to exploit heterogeneous sources of data is proposed in [61] by Nguyen et al. who studied the preoperative DCE MRI of 357 patients from two hospitals. The authors in [61] implement 2D, 3D, and 4D CNNs that integrate histological information of the primary tumor to prevent lymph node metastases, achieving values of sensitivity and AUC equal to 72% and 71% respectively. Then, the presence of different sequences in the multiparametric MRI has prompted recent works in investigating DL-based approaches that consider both pre and post-contrast sequences. In particular, Chen et al. [62] used the ResNet50 [55] pre-trained on ImageNet [63] to extract features from the second and fourth post-contrast volumes, the DWI sequence and the Apparent diffusion coefficient (ADC) map, performing the prediction with the Logistic Regression. Moreover, authors in [64] leveraged the DCE, T2, and DWI acquisitions, fine-tuning a ResNet50 [55] architecture for each data modality and then aggregating the predictions with a weighted voting rule. Similarly, Wang et al. [65] considered the pre-contrast T1, T2, and DWI volumes, adopting the Support Vector Machines model as meta-learner for the aggregation of the results.

It is worth noting that, to the best of our knowledge, the solution in [61] represents the first work in the literature exploiting the IF strategy and integrating in the CNN architecture one image modality, that is the DCE sequence, and the histological information of the primary tumor. Although other approaches [62], [64], [65] consider different MRI acquisitions, they extract features from each modality separately, thus implementing methods based on EF and LF strategies in the case of [62] and [64], [65] respectively.

Taking these aspects into consideration, the contributions of our work for ALN metastasis prediction from primary tumor analysis can be summarised as follows:

- We propose an innovative IF strategy that integrates histological information of the primary tumor, patients' clinical data, and multiparametric MRI (DCE, T2, and DWI sequences) and is able to provide a shared data representation well suited for the specific task to solve.
- We rely on Multi-Input Single-Output 3D CNN architectures to exploit the volumetric and complementary characteristics of the primary tumor.
- We enhance the integration of heterogeneous data by introducing in the IF strategy a Transfer Module that highlights the most discriminative features.

IV. Transfer Module (TM)

Among all MDL fusion techniques, the intermediate fusion is the most flexible solution. It exploits the layers in a deep-based architecture and enables the integration of the learned representations at various levels of abstraction [4], overcoming the limitation of the early fusion method where the shared data structure is determined before the use of a model. Moreover, in contrast to the late fusion, in IF, the heterogeneous sources simultaneously contribute to the decision, avoiding the implementation of a final decision step. However, the combination of the extracted features is still challenging due to the intrinsic characteristics of the different sources [16], [18].

In this section, we now introduce our Transfer Module (TM). It improves the integration of heterogeneous modalities, making the specific-modality paths influence each other while extracting the features maps. In particular, TM is able to perform a *cross-modality calibration* by taking into account

the complementary nature of the inputs. In particular, the *gating mechanism* reduces the importance of the least discriminative characteristics in each modality-specific features maps, exploiting the descriptors of all the inputs, and thus allowing them to influence each other during the training.

TM can be inserted between layers belonging to different convolutional cores to take into account the complementary characteristics of the images. Formally, for each modality j, we denote the output of the L_j^i as $F_j^i \in \mathbb{R}^{X_j^i \times Y_j^i \times Z_j^i \times C_j^i}$, where $X_j^i \times Y_j^i \times Z_j^i$ is the spatial dimension, while ${C_j}^i$ is the number of channels. The TM inserted in the layer i is a multi-input multi-output module that considers as input M features maps F_j^i and provides M outputs $\tilde{F}_j^i$, corresponding to the calibrated versions of the features maps, which are obtained by applying the gating procedure. TM is organized in two different steps, namely the *shared vector computation* and the *multimodal calibration*, as shown in Figure 2.

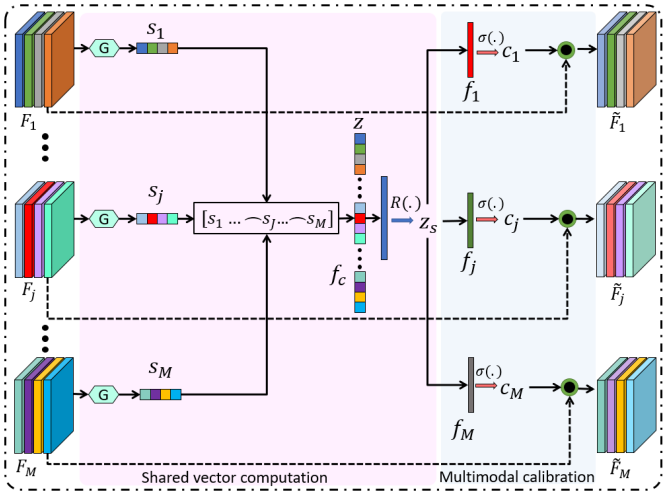

Fig. 2: Architecture of the proposed TM module for three different image modalities. The input consists of the features maps coming from the modality-specific paths, that are further processed by considering two steps acting in a pipeline, namely the *Shared vector computation* and the *Multimodal calibarion*. It is worth noting that the index i of the layer in which the module is inserted is omitted to avoid overly complex notation in the image.

The first computes the shared representation z_s^i of the $i-th$ layer considering the vectors computed from the features maps. The channel descriptor vector $s_j^i \in \mathbb{R}^{1 \times C_j^i}$ for each features map is obtained by using a Global Average Pooling operation [66] $G(\cdot)$, which provides a channel-wise descriptor by computing the average value. This operation is formalized as follows:

$$s_j^i = G(F_j^i) \in \mathbb{R}^{1 \times C_j^i} \tag{3}$$

In other words, the presence of the average operator makes all elements belonging to the same channel equally contribute to its characterization. The concatenation of all the s_j^i make up $z^i \in \mathbb{R}^{1 \times C^i}$ as follows:

$$z^i = [s_1^i \frown s_2^i \dots \frown s_j^i \frown s_M^i] \tag{4}$$

where $C^i = \sum_{j=1}^{M} C_j^i$. Then, z^i is further processed by considering it as input for the fully connected layer f_c, followed by the ReLU function $R(\cdot)$:

$$z_s^i = R(f_c(z^i)) \tag{5}$$

where $z_s^i \in \mathbb{R}^{1 \times C^t}$ is the shared representation, $C^t = C^i/r$, with r representing the reduction ratio, defined according to the work proposed in [33]. The ReLU function makes the TM able to model the complex and nonlinear map between the elements of the input vector z^i, consisting of descriptors relating to the different image modalities.

The *multimodal calibration* uses the shared representation z_s^i to calibrate the features maps F_j^i, thus exploiting information coming from different data sources. z_s^i is considered as input for M fully connected layers f_j, each for a specific modality j. The sigmoid activation function $\sigma(\cdot)$ is used to constraint the output in [0,1], resulting in M calibration vectors $c_j \in \mathbb{R}^{1 \times C_j^i}$, that aim to reduce the contribution of the selected channels. This step is formalized as follows:

$$c_j = \sigma(f_j(z_s^i)) \tag{6}$$

It is worth noting that the number of elements in c_j corresponds to the number of channels C_j^i of the features map of the modality j (F_j^i) and each value of the calibration vector represents the significance of the corresponding channel in F_j^i. As a consequence, the calibrated features maps $\tilde{F}_j^i$ are obtained by implementing a channel-wise product ($\odot$) between the input F_j^i and the corresponding c_j, weighting each channel of F_j^i by its significance level:

$$\tilde{F}_j^i = c_j \odot F_j^i \tag{7}$$

where $\tilde{F}_j^i \in \mathbb{R}^{X_j^i \times Y_j^i \times Z_j^i \times C_j^i}$. This creates a gating mechanism, where the contribution of the filters selected by c_j is reduced. Moreover, the dependence of the M calibration vectors on the shared representation z_s^i, as reported in Equation 6, allows each F_j^i to be influenced (calibrated) by the others during the features extraction step.

In the Multimodal calibration step, we obtain a self-gating mechanism, emphasizing the most informative characteristics while suppressing the less useful ones. The implemented gating mechanism acts on the channels of the extracted features maps, leveraging the shared representation to highlight in each data source the characteristics that contribute the most to the task under consideration. When it is inserted between the $i-th$ layers of the convolutional cores, each $\tilde{F}_j^i$ is affected by the other $M-1$ modalities, allowing the integration of information from the first levels of the network and enhancing the modality-specific paths dependency.

V. Materials

We retrospectively evaluated all the MRI examinations of the breast performed at the Central Radiology Department of Policlinico Umberto I, from January 2017 until January 2020 performed for tumor staging. A total of 153 subjects (average age 55 years; range 30–85), with 155 malignant BC lesions, were included in the study. For each patient/exam, we collected personal anamnestic information, histological and molecular

characteristics from the main tumor, MRI data, and definitive lymph node status.

A. Patient anamnestic-clinical data

Patients were divided into groups, based on collected information, as follows: age; menopausal status (pre- or post-menopausal); hormonal therapy (patients who have performed at least 3 continuous months of hormone therapy of any type, namely contraceptive, replacement therapy, or therapeutic); familiarity (patients with at least 2 female or male family members affected by breast cancer at any age).

B. Tumor histological data

The histological examination was performed on tumor material obtained through core-biopsy or surgery and analyzed by a pathologist with more than 15 years of experience. The tumor histological grade value was assigned in accordance with the Next-generation Sequencing (NGS) for which a score from 1 to 3 was given. The immunohistochemical analysis was conducted, evaluating the estradiol (ER), the progesterone (PgR), the herceptin2 receptor (HER2), and the proliferation index Ki67.

C. MRI imaging

The MRI investigations were performed with a 3 Tesla magnetic field using a Discovery 750 machine (by GE Healthcare, Milwaukee, WI, USA), using an 8- channel coil (8 US TORSOPA) dedicated for breast study, with the patient in prone position. T2-weighted, diffusion-weighted (DWI), and axial T1-weighted 3D dynamic contrast-enhanced (DCE) sequences are performed for each subject. The images were analyzed by two radiologists with 10 and 3 years of experience and the following characteristics were recorded: tumor localization based on breast quadrant; tumor distribution (unifocal, multifocal, multicentric); diameter of the target lesion; tumor margins (regular, irregular, lobulated, or spiculated); lesion intensity signal timing curve (IS/T curve) on DCE sequence; visibility on T2 of the lesion; visibility on DWI of the lesion; ADC values.

Bilateral tumors were considered as two different cases. For each case, the subtracted post-contrast T1-MRI was selected. The second phase (60–120 s) was selected for ROI segmentation, due to its higher contrast resolution. The lesions were manually drawn through manual and assisted thresholding segmentation techniques on the axial projection, reproducing the same technique of our previous works [56], [50]. When present, necrosis was avoided by segmentation. For multifocal or multicentric tumors, all lesions, even the smallest, were segmented.

D. Lymph Node Status (LNS)

The state of the axillary cavity was histologically determined after the diagnosis of breast cancer. The patients in our study were classified as positive (LN+) or negative (LN-), depending on whether there was, in the former case, at least one lymph node involved, or, in the latter case, no positive lymph nodes. On this basis, the dataset accounts for 27 positive and 128 negative samples.

E. Pre-processing

In MRI examination, the DCE sequence consists of the intravenous injection of a contrast agent (CA), whose absorption and release determine the specific wash-in and wash-out times respectively. Indeed, the DCE involves the acquisition of 3D volumes at different times, considering MRI images taken before (pre-contrast) and after (post-contrast) CA injection. The result is a 4D data with three spatial and one temporal dimension, that can be interpreted as a 3D image with several channels. Following the methodology proposed in [56], [57], we select four subtractive volumes: the first, the second, the last volume, and the median index volume between the third and the second-to-last volume, with the aim of preserving information about the wash-in and wash-out of CA flowing.

Differently from DCE, T_2 and DWI scans consist of the acquisition of a 3D volume without the temporal information, which can be considered as a 3D image with 1 channel. Since the breast lesion is segmented considering the second subtractive volume, the T_2 and DWI acquisitions are aligned to the segmentation mask generated from the $T1$ volume. Moreover, before applying the information about the lesion localization, the DWI scan is co-registered to the DCE volume considering the second post-contrast acquisition as a reference and using mutual information as a similarity metric [67], [68].

Based on the studies proposed in [56], [57], which evaluate how the amount of the included non-tumor tissue impacts the ALN assessment, we select the *Single Isotropic-size Box* (SIB) as tumor bounding option. In particular, all the DCE, T_2, and DWI volumes are re-sampled to obtain isotropic voxels with dimension $1 \times 1 \times 1mm^3$, before selecting the smallest box surrounding the tumor area.

Finally, information coming from images needs to be merged with the clinical features (CL) that include age, familiarity, hormone therapy, menopausal status, dimensions, ER, PgR, HER2, ki-67, and grading, resulting in a set of 10 features.

VI. Experimental Set-up

This paper aims to propose a TM to improve the integration of heterogeneous sources of data, modeling the complex relations between them and implementing the calibration of the extracted features maps. As already mentioned in the previous section, we focus on ALN status assessment considering multiparametric breast MRI of the primary tumor with different sequences, namely DCE, T_2, DWI, and patients' clinical features. Consequently, the number of image modalities M is 3, with $j \in \{DCE, DWI, T_2\}$.

Our experimental set-up not only deals with IF technique, where we introduce our TM, but we also compare our proposal against several competitors, i.e the unimodal (U) approach and the other multimodal fusion strategies, namely Early Fusion (EF) and Late Fusion (LF).

A. CNN Architectures

To deal with the presence of multiple data modalities, we propose a Multi Input - Single output network architecture, implementing the fusion at an intermediate level. In particular,

we exploit two different CNNs, that in turn reflect in two different convolutional (Conv-C) and classification (CC) parts. Such two networks are named in the following as BasicNet and ResNet.

The Conv-C of the BasicNet, responsible for features maps extraction, consists of five reduction layers, including blocks with a 3D-convolutional operation, followed by batch normalization and ReLU function. Each convolutional layer consists of a 3D operation with $4 \times 4 \times 4$ kernels and a stride set to 2 in order to extract features from the input volume while having a gradual dimensionality reduction. The padding is set to 1 in each layer, excluding the last one where it is set to 0. Moreover, each layer doubles the number of channels, while the convolutional operation in the first block presents 8 output channels. Finally, the classification core consists of two fully connected layers.

The ResNet is inspired by the state-of-art network proposed in [55], considering the 3D ResNet variants described in [69] which also offers a set of architectures pre-trained with medical images for segmentation tasks. The architecture consists of a Conv-C, which is an encoder with residual layers, and a set of decoders for the generation of the segmentation masks. In particular, the Conv-C consists of a first convolutional layer, followed by batch normalization, ReLU and Max Pooling layers, and a chain on four blocks containing the layers implementing the residual network as proposed in [55]. Hence, To adapt the ResNet we retain the encoder while the set of decoders is replaced with a CC consisting of a global average pooling and a fully connected layer.

B. IF approach for ALN status assessment

The network proposed for the specific task to solve is a Multi Input - Single output CNN, whose architecture is presented in Figure 3 and consists of 3 convolutional cores, the Multilayer Perceptron (MLP) for the clinical features, and two concatenating operations ($\frown$). The architecture implemented in this paper exploits for each $ConvC_j$ only the convolutional part of BasicNet or ResNet. The number of input channels of the first convolutional layer in the ConvC depends on the imaging modality. In particular, the first layer in the $ConvC_{DCE}$ presents 4 channels, while $ConvC_{T_2}$ and $ConvC_{DWI}$ consider 1 channel volumes as input. The MLP used to process the CL set presents a fully connected layer with 10 input neurons and 4 output features, followed by the ReLU function. The aim is to project the tabular data in a space that allows it to be combined with the features extracted from the images. As described in Figure 3, the first concatenation operation integrates $F_{DCE}^{N_{DCE}}$, $F_{T_2}^{N_{T_2}}$, $F_{DWI}^{N_{DWI}}$, obtaining the F_{img} features map, with C channels ($C = C_{DCE} + C_{T_2} + C_{DWI}$). Then, the *shared image features map* $\dot{F}_{img}$ is generated by the block L_s that includes a convolutional operation with a number of input and output channels set to C, a $1 \times 1 \times 1$ kernel, values of stride and padding set to 1 and 0 respectively, followed by batch normalization and ReLU function. In particular, this set of operations aims to further process the integration between the features maps modeling a nonlinear relation. Moreover, when the ResNet architecture is exploited for the ConvC, L_s also includes a Global Average Pooling layer to obtain a features vector. The second concatenation operation merges the $\dot{F}_{img}$ with features vector F_{cl} coming from the MLP that processes the clinical information. The resulting representation $F_s \in \mathbb{R}^{1\times(C+4)}$ is considered as input for the CC, which consists of two fully connected layers spaced by ReLU function. In particular, the first layer in the classification core considers an input features vector of $C + 4$ elements, with $C/3$ output features, implementing a map between the elements of F_s while performing a dimensionality reduction. It is worth noting that if the convolutional cores are implemented exploiting the same architecture, C_{DCE} is equal to C_{T_2}, and C_{DWI}, making the quantity $C/3$ integer. However, we argue that this aspect does not introduce a limitation in the proposed methodology since when the assumption is not assumed a ceiling or floor function can be applied. The second fully connected layer is responsible for the prediction, presenting two output neurons.

In the definition of TM, as suggested in [33], the reduction ratio r is set to 4, even if, as reported by the experiments conducted in [34], different values do not seem to influence the results. Although the TM can be inserted at any level of the convolutional cores, two simple rules are applied to decide the number of modules to be included, avoiding the implementation of a large number of experiments. As suggested in [33], in the case of BasicNet the TM is applied in the second half of the network, after the third, fourth, and fifth reduction layers. When the ResNet architecture is exploited, the TM is inserted after each block containing the layers implementing the residual network, resulting in the use of 4 transfer modules. It is possible to note that the implemented rules prevent the calibration of features maps that strongly depend on the specific image modality. Indeed, the first convolutional layers usually learn features such as edges and shapes that are inherent to input volumes, increasing the generalization in the following levels. For the sake of completeness, we also conducted additional experiments to evaluate the effects of varying the number of transfer modules. Moreover, the influence of the TM is studied by considering solutions in which that module is excluded. In IF experiments, fine-tuning is exploited by initializing the convolutional cores with the weights determined in the U approach.

C. Unimodal Approach

In the Unimodal (U) approach, the heterogeneous image modalities and the clinical features are exploited to build different models that do not cooperate for the determination of a single prediction. Indeed, the result is a set of 4 classifiers, each of them trained on a specific source of data. The aim is to separately evaluate each sub-component of the proposed Multi Input network. In the case of MRI sequences, the BasicNet and the ResNet are trained considering the DCE, the T2, and the DWI volumes separately. In particular, in the case of BasicNet, the CNN is trained from scratch, while when the ResNet is exploited, transfer learning is used, considering the pre-trained Conv-C [69] as a starting point and implementing the fine-tuning for adapting the network for the specific task

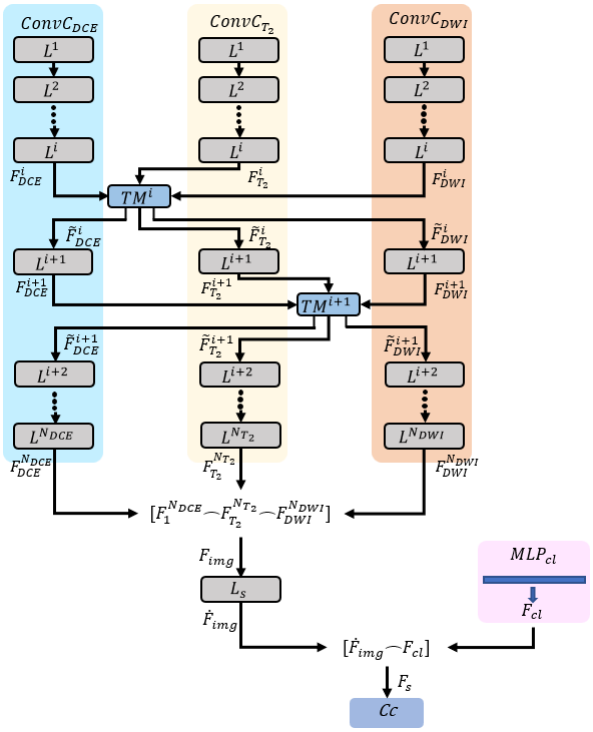

Fig. 3: Architecture of the IF approach including TM between layers belonging to different convolutional cores.

to solve. On the other hand, the classification core is trained from scratch. As described in Section V, clinical information is reported in the form of tabular data representing the set CL with 10 features, that are considered as input for a MLP, consisting of two fully connected layers, spaced by ReLU function. The first hidden layer presents 4 output features, while the output consists of two neurons, responsible for the classification.

D. Early Fusion approach

In the Early Fusion (EF) the different image modalities are organized in a single structure before being considered as an input for the single classifier. The simplest strategy involves concatenating the acquisitions in a multi-channels volume. However, the described fusing approach can not be applied in this work for two main reasons: i) it is not possible to integrate clinical features, ii) the characteristic of the DCE to be an image in which the temporal information is concatenated on the channels makes the idea of concatenating all the images along the channels unfeasible. As a consequence, a higher-level representation is extracted from each imaging modality by using the networks trained in the unimodal approach as features extractors. In particular, the features maps $F^{N_{DCE}}_{DCE}$, $F^{N_{T_2}}_{T_2}$, $F^{N_{DWI}}_{DWI}$ coming from the convolutional cores of the networks trained for each imaging modality are considered, resulting in the architecture represented in Figure 4 where the classification is performed by a neural network (NN), exploiting the block L_s and the MLP introduced for the set CL.

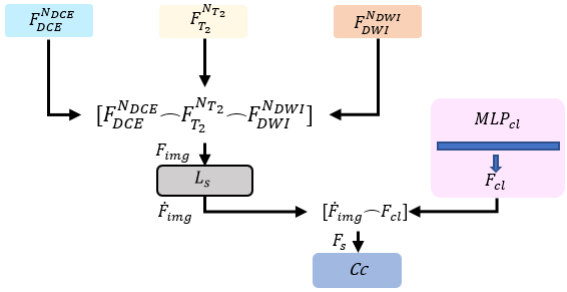

Fig. 4: Architecture of the EF approach, where the classification is performed by a neural network.

E. Late Fusion approach

The late fusion (LF) completely relies on the unimodal approach since it aggregates the prediction coming from the classifiers implemented for each data source. To this goal, the predictions coming from the three networks and the MLP trained with the CL set are combined using Weighted Majority Voting (WMV), in which a weight is assigned to each prediction according to the model output probability. However, in LF each classifier acts independently, not taking advantage of the complementary characteristics of the image modalities and clinical data that do not influence each other during the prediction.

F. Implementation Details

As reported in Section V-E, the selected tumor bounding option (SIB) considers a box whose size varies according to each patient's region of interest. Moreover, the presence of multiple image modalities with different resolutions causes the creation of volumes with different dimensions. As a consequence, a resize stage is used to give the volumes a common size of $64 \times 64 \times 64$, before feeding them to the involved CNNs. In experiments involving ResNet, the ResNet10, the ResNet18, the ResNet34, and the ResNet50 architectures are used, exploiting transfer learning [69]. Data augmentation is used in the training stage by applying random rotations and flips, while the dataset is balanced by replicating some randomly chosen volumes belonging to the minority class. Moreover, the greyscale intensity in each extracted volume is normalized in $[0; 1]$ to ensure that, in the classification step, the CNNs operate with volumes having the same scale across different patients. During the experiments, the maximum number of epochs is set to 500, the batch size is set to 32 for BasicNet and all the ResNet architectures. The learning rate for the cross-entropy loss was set to 10^{-6}. Adam optimizer is used with a weight decay set to 10^{-4}. To find the appropriate hyper-parameters, a grid search is implemented by varying the batch size in [8, 64], the learning rate in $[10^{-7}, 10^{-3}]$ and the weight decay in $[0, 10^{-4}]$.

Performance is evaluated in terms of Accuracy (ACC), Sensitivity (SENS), Specificity (SPE), and Area under ROC curve (AUC). All the experiments were run in a 10-folds Cross Validation (CV) setting, to better assess the generalization ability of each approach. In particular, patient-based cross-validation is performed, to reliably compare the performance

of different models by avoiding the use of volumes from the same patient during the training and evaluation phase. All the DL experiments were carried out using Pytorch (version 1.10), while the pre-processing step was implemented in MATLAB 2020b. A Linux workstation equipped with Intel(R) Core(TM) i7-10700KF CPU, 64 GB of DDR4 RAM and a Nvidia RTX 3090 GPU is used. The source code of the implemented experiments is available at this link. [1]

VII. Results and Discussions

This section reports in Table I the results of the implemented experiments, considering both the proposed solution and different fusion techniques. In particular, each configuration is detailed according to the fusion approach, the used data modalities, and the model involved for the classification in columns *App.*, *Mod.*, and *Model* respectively. Moreover, the performance is evaluated in terms of ACC, Sens, Spe, and AUC, reporting the average rate computed adopting a 10-fold CV and the standard deviation after the ± symbol. For readability, we denote the involved networks, namely BasicNet, ResNet10, ResNet18, ResNet34, and ResNet50 as Ba and Re10, Re18, Re34, and Re50 respectively.

Table I is organized into four main sections where different fusion approaches are exploited. In the configurations considering the IF, the involvement of the proposed transfer module is highlighted by the presence of TM in the name of the network architecture (i.e. Ba-TM, in the case of BasicNet). In the EF, all data modalities are considered, exploiting the CNN as a features extractor, the MLP for the clinical features, and performing the classification using NN. The solution with the LF merges the predictions coming from the CNNs trained with the unimodal approach and the experiment considering the CL set and the MLP. In the U approach, the network is trained for each imaging modality, leveraging the fine-tuning in the case of ResNet [69].

The first section focuses on the IF approach, showing the results obtained by varying the CNNs and reporting the effect of the transfer module. When TM is considered, ResNet10 presents the best performance in terms of accuracy, and specificity, achieving 90.91% and 92.91% respectively while obtaining 81.48% in sensitivity together with ResNet18-TM, and BasicNet-TM that also reports the highest value in terms of AUC (90.14%). In the configuration without the transfer module, each network reports a slight decrease in performance. Indeed ResNet10 achieves 90.26% in accuracy, 92.13% in specificity, 87.43% in AUC, and the same sensitivity as BasicNet, that is 81.48%. To assess the statistical significance of the comparison between the experiments with and without the TM for each CNN, we perform a Wilcoxon rank-sum test with a significance level of 0.05. This analysis considers the probability distributions of predictions from different networks as paired observations. In particular, in the case of BasicNet, we compare the configurations involving Ba and Ba-TM (i.e. the first two rows of Table I), that report a significant difference with a p-value equal to 0.0023. Similarly, ResNet10, ResNet18, and ResNet50 differ significantly from ResNet10-TM, ResNet18-TM, and ResNet50-TM respectively, reporting p-values equal to 0.0038, 0.0048 and 0.0113, while the comparison between the experiments with ResNet34 and ResNet34-TM do not show a statistical difference with a p-value equal to 0.0518.

TABLE I: Performance of the implemented experiments evaluated in 10-fold CV setting considering different models and multimodal data fusion techniques. In each section, the best values are reported in bold.

App.	Mod.	Model	Acc	Spe	Sens	AUC
IF	ALL	Ba-TM	87.01±0.08	88.19±0.08	**81.48±0.17**	**90.14±0.10**
	ALL	Ba	84.42±0.11	85.04±0.12	**81.48±0.17**	81.28±0.15
	ALL	Re10-TM	**90.91±0.08**	**92.91±0.07**	**81.48±0.17**	87.17±0.14
	ALL	Re10	90.26±0.08	92.13±0.08	**81.48±0.17**	87.43±0.12
	ALL	Re18-TM	89.61±0.06	91.34±0.06	**81.48±0.17**	85.04±0.14
	ALL	Re18	86.36±0.10	88.19±0.10	77.78±0.16	84.54±0.14
	ALL	Re34-TM	87.66±0.10	89.76±0.09	77.78±0.16	83.79±0.15
	ALL	Re34	87.01±0.07	88.98±0.05	77.78±0.16	84.49±0.12
	ALL	Re50-TM	84.42±0.11	85.83±0.11	77.78±0.16	83.96±0.16
	ALL	Re50	81.82±0.11	82.68±0.12	77.78±0.16	81.31±0.13
	DCE+CL-4	CNN[61]	**68.18±0.15**	**70.87±0.21**	**55.56±0.33**	**67.07±0.20**
	ALL	Ba-MMTM[33]	82.47±0.09	82.68±0.10	**81.48±0.17**	86.56±0.13
	ALL	Re10-MMTM[33]	89.62±0.10	91.34±0.07	**81.48±0.17**	85.42±0.13
	ALL	Re18-MMTM[33]	85.71±0.08	87.40±0.09	77.78±0.15	80.78±0.12
	ALL	Re34-MMTM[33]	87.02±0.09	88.98±0.08	77.78±0.15	80.84±0.13
	ALL	Re50-MMTM[33]	83.12±0.10	84.25±0.09	77.78±0.16	81.69±0.14
EF	ALL	Ba+MLP+NN	79.87±0.13	81.10±0.14	**74.07±0.15**	80.87±0.15
	ALL	Re10+MLP+NN	**84.42±0.09**	**86.61±0.09**	**74.07±0.15**	**85.62±0.11**
	ALL	Re18+MLP+NN	83.12±0.10	85.04±0.10	**74.07±0.15**	84.60±0.12
	ALL	Re34+MLP+NN	83.77±0.08	**86.61±0.09**	70.37±0.15	77.72±0.14
	ALL	Re50+MLP+NN	79.87±0.09	81.89±0.10	70.37±0.15	81.16±0.12
LF	ALL	Ba+MLP	83.23±0.11	87.50±0.11	62.96±0.14	82.29±0.15
	ALL	Re10+MLP	**89.03±0.07**	92.19±0.07	**74.07±0.05**	**93.14±0.10**
	ALL	Re18+MLP	86.45±0.08	90.63±0.08	66.67±0.10	90.16±0.10
	ALL	Re34+MLP	83.23±0.10	85.94±0.11	70.37±0.14	84.69±0.19
	ALL	Re50+MLP	82.58±0.11	85.16±0.11	70.37±0.11	87.18±0.15
	IMG	Re50+WV [64]	80.52±0.13	84.25±0.09	62.96±0.17	83.03±0.12
	IMG	Re50+SVM [65]	78.57±0.12	**94.49±0.10**	0.04±0.45	49.10±0.33
U	CL	MLP	75.97±0.11	78.74±0.10	62.96±0.22	75.61±0.17
	DCE	Ba	78.06±0.11	78.91±0.12	**74.07±0.15**	78.53±0.09
		Re10	**85.16±0.07**	87.50±0.07	**74.07±0.17**	81.34±0.12
		Re18	84.52±0.08	86.72±0.08	**74.07±0.15**	80.93±0.10
		Re34	78.71±0.10	80.47±0.12	70.37±0.15	74.13±0.12
		Re50	78.06±0.08	79.69±0.08	70.37±0.20	**82.26±0.11**
		CNN[61]	63.76±0.16	69.67±0.19	41.67±0.38	65.21±0.22
		RF[49]	84.52±0.08	**96.10±0.06**	29.63±0.42	62.86±0.20
	T2	Ba	74.84±0.11	78.13±0.11	59.26±0.28	62.36±0.21
		Re10	**85.16±0.08**	**89.84±0.09**	**62.96±0.08**	76.71±0.14
		Re18	78.71±0.13	82.03±0.16	**62.96±0.08**	72.60±0.16
		Re34	76.77±0.10	80.47±0.13	59.26±0.19	67.85±0.16
		Re50	67.10±0.10	68.75±0.13	59.26±0.11	64.06±0.09
	DWI	Ba	79.87±0.13	85.04±0.12	55.56±0.25	66.00±0.23
		Re10	83.77±0.08	88.98±0.09	59.26±0.11	71.60±0.15
		Re18	81.17±0.11	85.83±0.12	59.26±0.11	71.30±0.13
		Re34	74.03±0.08	77.95±0.10	55.56±0.20	64.36±0.13
		Re50	72.08±0.12	75.59±0.13	55.56±0.25	64.01±0.17

As reported in Section VI-B, the reduction ratio r is defined as 4 in the transfer module. However, we investigate the impact of different values of r on performance. To do so, we choose the experiment with the BasicNet, representing the configuration with the lowest computational cost, and repeat executions with reduction ratios set to 8 and 16. The results show AUC values of 90.03% and 90.11% for r set to 8 and 16, respectively. There is no statistically significant difference compared to the case where r is equal to 4, as confirmed by the Wilcoxon rank-sum test with significance levels of 0.05 (p-values equal to 0.5574 and 0.4892 for r set to 8 and 16, respectively). It is worth noting that the obtained results are in line with the work proposed in [34], according to which different values of r do not seem to influence the performance of the model.

We compare our methodology with the solution proposed in the literature by Nguyen et al. [61] that represents the first attempt to apply an approach based on IF for ALN metastasis prediction. The authors proposed a solution that exploits a CNN, DCE sequence of the primary tumor and four clinical

[1] https://github.com/Michela94CE/Cross-Modality-Calibration-with-Transfer-Module

features, namely age, ER, ki-67, and HER2, that in this paper represent the CL-4 set. The authors implemented a 3D CNN to process DCE-MRI images using a subtractive approach that works with the third, fourth, and fifth post-contrast volumes. A 3D cuboidal bounding box of size $50 \times 50 \times 50$, encompassing the tumor region, is used to crop each DCE-MRI data. The CL-4 set is inserted in the first fully connected layer of the classification core of the implemented network. It is possible to note that our methodology with the Transfer Module outperforms by a wide margin the approach described in [61]. We argue that different aspects contribute to explaining this result. First, in [61] the authors used a dataset consisting of 357 patients, and hence the performance can be reasonably affected by the size of the dataset involved in our work (153 patients). Moreover, the fixed-size bounding box used in [61] could reduce the generalization capability of the implemented model as reported in [56], [57]. In this paper we also exploit different complementary imaging modalities that contribute to the characterization of the primary tumor, leveraging the transfer module to improve the integration of data, while authors in [61] only consider the DCE sequence and four clinical information.

Although the implemented experiments show the effectiveness of the proposed TM, we also compare our module with other fusion strategies presented in the literature. In particular, we followed the description reported in [33], adapting the MMTM for a task involving three modalities. We replaced the TM with the MMTM in each CNN, showing that our proposal outperforms the approach described in [33], even if it obtains good results. Indeed, we argue that in contrast with the MMTM, the presence of the ReLU function in the Shared vector computation step makes our TM able to model the complex and nonlinear map between the different images. In addition, another difference can be noted in the Multimodal Calibration stage where we do not change the output range of the sigmoid function with the multiplication by a scalar value, allowing the network automatically to learn how to select the most discriminative features, reducing the contribution of less significant ones. Taking into account other state-of-art methods, it is worth noting that in comparison with the CAAF module proposed in [36], which considers the correlation among features as a measure of redundancy, our TM consists of trainable parameters, making the calibration well suited for the specific task to solve. Furthermore, the presence of the correlation measure in [36] requires two features maps with the same spatial dimensions, while the methodology proposed in this paper does not place any limitations on the number of image modalities and the characteristics of the extracted features maps. Then, similarly to the work proposed in [35], the fusion is defined in the course of the training. However, the MFB defined in [35] only affects the network implemented for the image synthesis, learning the underlying correlation among data, without disturbing the modality-specific paths, thus the extracted features maps.

The second part of Table I focuses on EF, where the CNNs are used as features extractors, the MLP is exploited to process the clinical features, and the classification is performed by a NN, as described in Section VI-D. The configuration with the ResNet10 achieves the highest performance in terms of accuracy, specificity, and AUC, reporting 84.42%, 86.61%, and 85.62% respectively while obtaining the same sensitivity as experiments exploiting BasicNet and ResNet34 (74.07%).

The third section focuses on the LF method, which integrates predictions from four models trained independently on DCE, T_2, DWI sequences, and the CL set, respectively. In particular, the column *Model* differs for the CNN architecture used to process the image modalities. The configuration exploiting the ResNet10 outperforms the others in each metric, achieving a value of accuracy equal to 89.03%, 92.19% in terms of specificity, a value of sensitivity equal to 74.07%, and 93.14% in AUC. In the last two rows of the section, we also report the results obtained implementing the state-of-art approaches presented in [64] and [65] that exploit the fine-tuning a ResNet50 [55] architecture for each data modality. Then, they aggregate the predictions with a weighted voting (WV) and Support Vector Machines (SVM) model in [64] and [65], respectively. In these experiments, the configuration of the involved data modalities is denoted as *IMG*, since the solutions in [65] and [64] do not consider clinical and histological information.

The U approach is reported in the fourth section of Table I, consisting of experiments involving a single data modality. In particular, the CL set is processed with the MLP, achieving values equal to 75.97%, 78.74%, 62.96%, and 75.61% in terms of accuracy, specificity, sensitivity, and AUC respectively. When the image modalities are considered, the configurations with different CNNs are explored. In the case of the DCE sequence, the BasicNet, the ResNet10, and the ResNet18 achieve the highest values of sensitivity (74.07%). Moreover, the ResNet10 presents the best performance in terms of accuracy 85.16%, while the ResNet50 reports the best AUC (82.26%) among all the experiments in the fourth section. When the T_2 sequence is considered, the ResNet10 achieves the best performance in each metric, compared with the configurations exploiting the same modality, and it is confirmed as the best-performing network also in the case of the DWI modality. In line with the current state-of-art, the DCE represents the most discriminating series, thus confirming its use in the majority of the works in the literature, as reported in Section III. For comparison, we consider again the solution proposed in [61], where authors presented a version of the CNN exploiting only the DCE sequence. Moreover, we report the results obtained with the methodology proposed in our previous work [49], where we extracted several characteristics from the second post-contrast volume, including first-order, gray level co-occurrence matrix, three orthogonal planes-Local binary patterns features, and then rely on Randon Forest (RF) to perform the classification.

To assess the positioning of the TM within both the BasicNet and ResNet architectures, we conducted a series of experiments by varying the number of transfer modules. Specifically, these modules are strategically placed after each group of layers – the reduction layer for BasicNet and the residual block for ResNet. The placement follows the reverse direction, starting from the last group. In the case of ResNet, we focused on ResNet50 due to the significant impact observed

on performance metrics in terms of ACC and AUC. Indeed, as highlighted in the initial section of Table I, the configurations with Re50 and Re50-TM present the biggest gaps among all the experiments involving the ResNet architecture ($+2.60$ and $+2.65$ in ACC and AUC, respectively). This choice is also motivated by the observation that, in comparison with smaller architectures (i.e., ResNet10, ResNet18, and ResNet34), each residual block in ResNet50 comprises numerous layers that contribute to a more profound feature extraction stage, making the cross-modality calibration more meaningful.

Table II presents the outcomes of experiments involving the variation of the number of TMs, detailed in the *#TM* column. The table encompasses two sections dedicated to BasicNet and ResNet50, respectively. It's important to note that when *#TM* is set to 0, we present the results of the configuration where the transfer module is not utilized. As emphasized in the initial section of Table II, the optimal performance across all metrics is achieved with the presence of three TMs when the BasicNet architecture is considered. This configuration effectively prevents the calibration of feature maps that might strongly rely on specific image modalities. In the case of ResNet50, the best results emerge when a transfer module is inserted after each residual block, excluding the first convolutional layer. This strategic placement leverages the chain of multiple layers within each residual block, enhancing the extraction of high-level representations from the input data.

TABLE II: Performance of the implemented experiments evaluated in 10-fold CV setting and obtained by varying the number of transfer modules. In each section, the best values are reported in bold.

Model	#TM	Acc	Spe	Sens	AUC
Ba	4	81.17±0.07	81.10±0.10	**81.48±0.15**	82.15±0.12
	3 (Proposed)	**87.01±0.08**	**88.19±0.08**	**81.48±0.17**	**90.14±0.10**
	2	86.36±0.08	87.40±0.12	**81.48±0.17**	81.51±0.14
	1	79.87±0.13	81.10±0.14	74.07±0.18	80.87±0.15
	0	84.42±0.11	85.04±0.12	**81.48±0.17**	81.28±0.15
Re50	4 (Proposed)	**84.42±0.11**	85.83±0.11	**77.78±0.16**	**83.96±0.16**
	3	83.12±0.10	84.25±0.10	**77.78±0.15**	81.69±0.11
	2	83.77±0.13	**86.61±0.09**	70.37±0.11	77.72±0.17
	1	79.87±0.14	84.10±0.12	74.07±0.10	79.24±0.13
	0	81.82±0.11	82.68±0.12	**77.78±0.16**	81.31±0.13

The results reported in Table I highlight that, for each network, the solutions exploiting the TM component achieve the best performance, outperforming by a wide margin, the experiments involving a single data modality (U). Moreover, the LF performs better than the EF approaches, thus supporting the preference of the former over the latter. Indeed, the LF leverages different models, each trained for a specific data modality. As a consequence, the four classifiers learn to extract features that reflect the distinctive characteristics of each modality, delaying the combination of the results in a post-processing step and also exploiting the uncorrelated nature of errors performed by models. In solutions involving the IF configuration, the shared representation is created by concatenating features from the convolutional cores at an intermediate level, thus preserving the distinctiveness of the different image modalities, which is then exploited in the classification core. Moreover, the training strategy exploiting the backpropagation algorithm allows the definition of a representation well suited for the specific task to solve, improving the integration of heterogeneous data compared to the EF and LF experiments. In particular, the presence of the proposed transfer module makes the layers of the convolutional cores influence each other during the features extraction step, enabling the CNN to self-select the characteristics in each modality that best contribute to the problem to be solved. Indeed, the gating mechanism reduces the impact of the selected features according to a shared representation that considers all data sources. This results in a more focused and context-aware representation of the input data. Furthermore, the TM contributes to the mitigation of the domain gap between heterogeneous modalities, calibrating the features at an intermediate level. The inclusion of the transfer module in the training process allows the definition of an integrated set of features specifically suited for the specific task under analysis. This characteristic leads, therefore, to an improvement in performance, as reported in Table I, speeding up the convergence of the network. We focus again on BasicNet and ResNet50, providing in Figure 5 the loss curves obtained in the configuration with and without the proposed transfer module. As aforementioned in Section VI-F, we conducted a 10-fold cross-validation, resulting in ten curves for both the training and validation sets. To provide a more concise representation for each classification task, we calculated the mean value and standard deviation of the loss function for each epoch. The bold lines represent the average training and validation curves, while the shaded areas indicate the region defined by the standard deviation. It is worth noting that for both CNNs, the inclusion of TM improves the convergence, reducing the gap between the training and validation curves.

Finally, Table I also reveals that performance decreases for each approach when the ResNet architectures are exploited: we deem that this happens since the complexity of the network increases, and it can be further explained by the limited amount of data used in the study. Although fine-tuning is exploited, the small number of MR images may affect the convergence of the CNNs, limiting their generalization ability.

VIII. Conclusion

In this work, we presented a novel methodology for the integration of heterogeneous sources of data. We proposed an approach based on the intermediate fusion method that involves the use of a transfer module for cross-modality calibration. TM modifies the features extracted in the convolutional cores, emphasizing the most relevant features and minimizing the impact of less informative ones. To demonstrate the effectiveness of our approach, we implemented a Multi Input-Single Out CNN that utilizes both multiparametric breast MRI sequences and patients' clinical information to predict the presence of axillary lymph node metastasis. Despite the role of magnetic resonance imaging of primary tumors in predicting the involvement of the axillary state has been strongly demonstrated in the literature, it still remains unused in international guidelines [70], [71], [72]. The results presented in this work intend to highlight not only, as previously demonstrated in other works [61], [73], the importance of the role of DL approaches in predicting the state

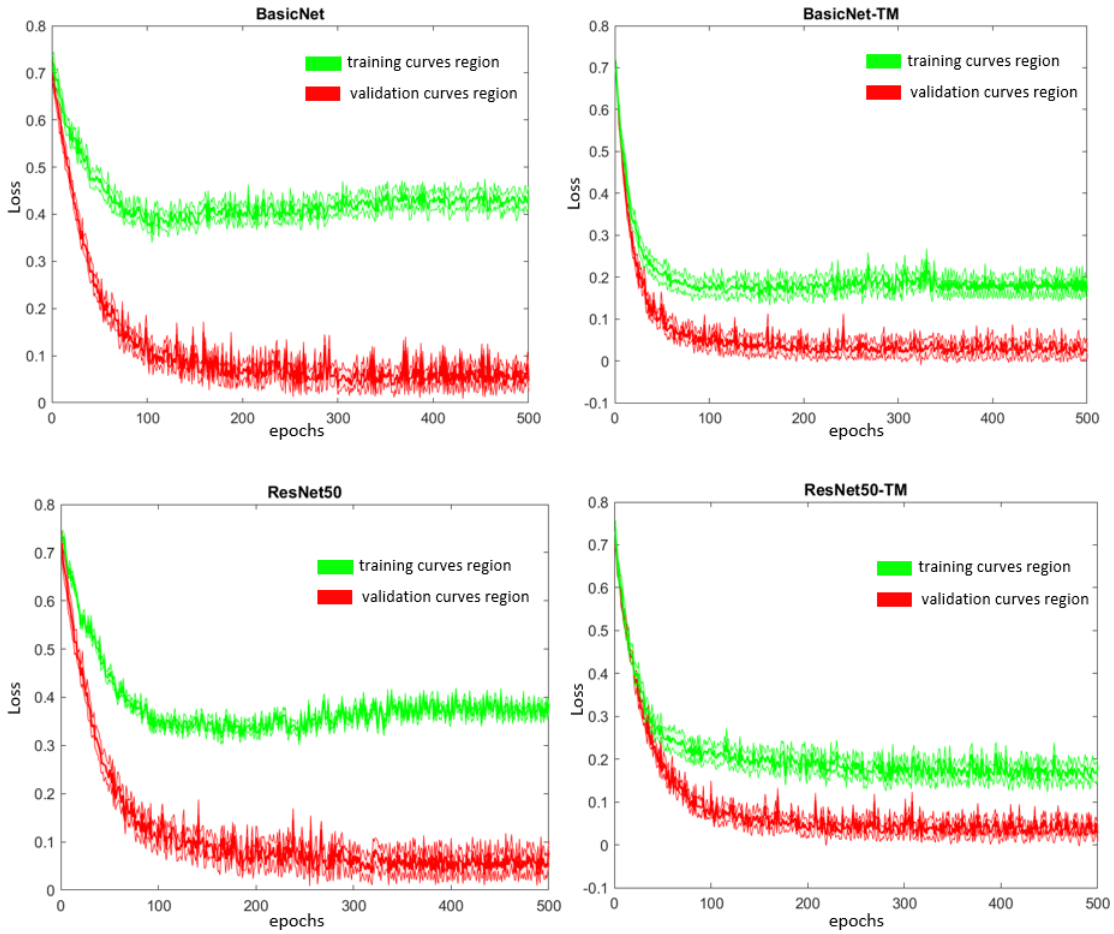

Fig. 5: Training and validation loss curves for the experiments involving the BasicNet and the ResNet50 with and without the presence of TM. The lines in bold represent the average training and validation curves computed considering the 10-fold CV, while the underlying area determines the region identified by the standard deviation

of the axillary lymph node, but also the contribution that pre-contrast sequences can offer. In particular, the T2w sequence allows highlighting the morpho-structural characteristics of the tumor while the DWI describes the intrinsic properties and tumor aggressiveness. As demonstrated in our previous works [13], [74], both the peritumor edema evaluated in the T2w sequences and the ADC value, a quantitative expression of the cellular restriction of the tumor lesion, correlate significantly with the state of the axilla.

Despite the promising results, we argue that our work presents some limitations. In particular, the reduced size of the population and the involvement of a single medical center may affect the evaluation of the generalization ability of the proposed methodology. Additionally, the presence of a multiparametric MRI dataset might be viewed as a constraint when contemplating the applicability of our solution within a truly multimodal context. Although the number of patients is similar to that used in [43], [46], [47], future efforts will concentrate on evaluating the proposed methodology on data collected from different centers, thereby expanding the size of the considered population. Moreover, we plan to assess the proposed approach in diverse tasks involving heterogeneous medical imaging procedures, such as PET-MRI. This will leverage the inherent flexibility of the proposed TM, which, by definition, can manage a variable number M of image modalities, adapting to different applications. Finally, the integration of uncertainty quantification in our solutions stands out as a critical avenue for our future research. Indeed, acknowledging and quantifying uncertainties associated with model predictions are crucial for enhancing the robustness, interpretability, and real-world applicability of DL systems. To this aim, we will explore the development of an uncertainty-aware version of the TM, drawing inspiration from approaches proposed in [75], [76].

ACKNOWLEDGMENT

This work is partially funded by: i) PNRR MUR project PE0000013-FAIR ii) project n. F/130096/01-05/X38 - Fondo per la Crescita Sostenibile - ACCORDI PER L'INNOVAZIONE DI CUI AL D.M. 24 MAGGIO 2017 - Ministero dello Sviluppo Economico (Italy), iii) Programma Operativo Nazionale (PON) "Ricerca e Innovazione" 2014-2020 CCI2014IT16M2OP005 Azione IV.4, iii) FONDO PER LA CRESCITA SOSTENIBILE (F.C.S.), Bando Accordo Innovazione DM 24/5/2017 (Ministero delle Imprese e del Made in Italy), CUP B89J23000580005.

REFERENCES

[1] O. Oren, B. J. Gersh, and D. L. Bhatt, "Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints," *The Lancet Digital Health*, vol. 2, no. 9, pp. e486–e488, 2020.

[2] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020.

[3] M. A. Azam, K. B. Khan, S. Salahuddin, E. Rehman, S. A. Khan, M. A. Khan, S. Kadry, and A. H. Gandomi, "A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics," *Computers in Biology and Medicine*, vol. 144, p. 105253, 2022.

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

[5] H. Hermessi, O. Mourali, and E. Zagrouba, "Multimodal medical image fusion review: Theoretical background and recent advances," *Signal Processing*, vol. 183, p. 108036, 2021.

[6] Y. Xu, "Deep learning in multimodal medical image analysis," in *International Conference on Health Information Science*. Springer, 2019, pp. 193–200.

[7] F. Michelle C. Walters, D.O. and F. Lennard Nadalo M.D. (2015) Mri breast clinical indications: A comprehensive review. [Online]. Available: https://www.jaocr.org/articles\/mri-breast-clinical-indications-a-comprehensive-review

[8] V. Cipolla, D. Santucci, D. Guerrieri, F. M. Drudi, M. L. Meggiorini, and C. de Felice, "Correlation between 3 t apparent diffusion coefficient values and grading of invasive breast carcinoma," *European journal of radiology*, vol. 83, no. 12, pp. 2144–2150, 2014.

[9] C. De Felice, V. Cipolla, D. Guerrieri, D. Santucci, A. Musella, L. Porfiri, and M. Meggiorini, "Apparent diffusion coefficient on 3.0 tesla magnetic resonance imaging and prognostic factors in breast cancer," *Eur J Gynaecol Oncol*, vol. 35, no. 4, pp. 408–414, 2014.

[10] L. W. Turnbull, "Dynamic contrast-enhanced mri in the diagnosis and management of breast cancer," *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, vol. 22, no. 1, pp. 28–39, 2009.

[11] R. M. Mann, N. Cho, and L. Moy, "Breast mri: state of the art," 2019.

[12] C. J. Moran, B. A. Hargreaves, M. Saranathan, J. A. Lipson, J. Kao, D. M. Ikeda, and B. L. Daniel, "3d t2-weighted spin echo imaging in the breast," *Journal of Magnetic Resonance Imaging*, vol. 39, no. 2, pp. 332–338, 2014.

[13] D. Santucci, E. Faiella, E. Cordelli, A. Calabrese, R. Landi, C. de Felice, B. Beomonte Zobel, R. F. Grasso, G. Iannello, and P. Soda, "The impact of tumor edema on t2-weighted 3t-mri invasive breast cancer histological characterization: a pilot radiomics study," *Cancers*, vol. 13, no. 18, p. 4635, 2021.

[14] A. Calabrese, D. Santucci, R. Landi, B. Beomonte Zobel, E. Faiella, and C. de Felice, "Radiomics mri for lymph node status prediction in breast cancer patients: the state of art," *Journal of Cancer Research and Clinical Oncology*, vol. 147, no. 6, pp. 1587–1597, 2021.

[15] T. Zhou, Q. Cheng, H. Lu, Q. Li, X. Zhang, and S. Qiu, "Deep learning methods for medical image fusion: A review," *Computers in Biology and Medicine*, p. 106959, 2023.

[16] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[17] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 2022.

[18] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.

[19] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[20] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[21] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[22] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[23] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[24] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[25] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[26] R. J. Chen, M. Y. Lu, J. Wang, D. F. 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 Transactions on Medical Imaging*, vol. 41, no. 4, pp. 757–770, 2020.

[27] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. Landman, and Y. Huo, "Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review," *Progress in Biomedical Engineering*, 2023.

[28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[29] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[30] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6966–6975.

[31] R. Yan, F. Ren, X. Rao, B. Shi, T. Xiang, L. Zhang, Y. Liu, J. Liang, C. Zheng, and F. Zhang, "Integration of multimodal data for breast cancer classification using a hybrid deep learning method," in *Intelligent Computing Theories and Application: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I 15*. Springer, 2019, pp. 460–469.

[32] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 250–258.

[33] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 289–13 299.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[35] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: hybrid-fusion network for multi-modal mr image synthesis," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.

[36] W. Ma, J. Shen, H. Zhu, J. Zhang, J. Zhao, B. Hou, and L. Jiao, "A novel adaptive hybrid fusion network for multiresolution remote sensing images classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[37] X. Cui, N. Wang, Y. Zhao, S. Chen, S. Li, M. Xu, and R. Chai, "Preoperative prediction of axillary lymph node metastasis in breast cancer using radiomics features of dce-mri," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[38] "Differentiating axillary lymph node metastasis in invasive breast cancer patients: a comparison of radiomic signatures from multiparametric breast MR sequences, author=Chai, Ruimei and Ma, He and Xu, Mingjie and Arefan, Dooman and Cui, Xiaoyu and Liu, Yi and Zhang, Lina and Wu, Shandong and Xu, Ke, journal=Journal of Magnetic Resonance Imaging, volume=50, number=4, pages=1125–1132, year=2019, publisher=Wiley Online Library."

[39] D. Arefan, R. Chai, M. Sun, M. L. Zuley, and S. Wu, "Machine learning prediction of axillary lymph node metastasis in breast cancer: 2d versus 3d radiomic features," *Medical physics*, vol. 47, no. 12, pp. 6334–6342, 2020.

[40] J. Liu, D. Sun, L. Chen, Z. Fang, W. Song, D. Guo, T. Ni, C. Liu, L. Feng, Y. Xia *et al.*, "Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer," *Frontiers in oncology*, vol. 9, p. 980, 2019.

[41] L. Han, Y. Zhu, Z. Liu, T. Yu, C. He, W. Jiang, Y. Kan, D. Dong, J. Tian, and Y. Luo, "Radiomic nomogram for prediction of axillary lymph node metastasis in breast cancer," *European radiology*, vol. 29, no. 7, pp. 3820–3829, 2019.

[42] Y. Yu, Z. He, J. Ouyang, Y. Tan, Y. Chen, Y. Gu, L. Mao, W. Ren, J. Wang, L. Lin *et al.*, "Magnetic resonance imaging radiomics predicts preoperative axillary lymph node metastasis to support surgical decisions and is associated with tumor microenvironment in invasive breast cancer: A machine learning, multicenter study," *EBioMedicine*, vol. 69, p. 103460, 2021.

[43] C. Liu, J. Ding, K. Spuhler, Y. Gao, M. Serrano Sosa, M. Moriarty, S. Hussain, X. He, C. Liang, and C. Huang, "Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced mri," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 1, pp. 131–140, 2019.

[44] V. Romeo, P. Kapetas, P. Clauser, S. Rasul, R. Cuocolo, M. Caruso, T. H. Helbich, P. A. Baltzer, and K. Pinker, "Simultaneous 18f-fdg pet/mri radiomics and machine learning analysis of the primary breast tumor for the preoperative prediction of axillary lymph node status in breast cancer," *Cancers*, vol. 15, no. 20, p. 5088, 2023.

[45] S. Samiei, R. W. Granzier, A. Ibrahim, S. Primakov, M. B. Lobbes, R. G. Beets-Tan, T. J. van Nijnatten, S. M. Engelen, H. C. Woodruff, and M. L. Smidt, "Dedicated axillary mri-based radiomics analysis for the prediction of axillary lymph node metastasis in breast cancer," *Cancers*, vol. 13, no. 4, p. 757, 2021.

[46] Y. Dong, Q. Feng, W. Yang, Z. Lu, C. Deng, L. Zhang, Z. Lian, J. Liu, X. Luo, S. Pei *et al.*, "Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of t2-weighted fat-suppression and diffusion-weighted mri," *European radiology*, vol. 28, no. 2, pp. 582–591, 2018.

[47] M. Liu, N. Mao, H. Ma, J. Dong, K. Zhang, K. Che, S. Duan, X. Zhang, Y. Shi, and H. Xie, "Pharmacokinetic parameters and radiomics model based on dynamic contrast enhanced mri for the preoperative prediction of sentinel lymph node metastasis in breast cancer," *Cancer Imaging*, vol. 20, no. 1, pp. 1–8, 2020.

[48] Y. Liu, X. Li, L. Zhu, Z. Zhao, T. Wang, X. Zhang, B. Cai, L. Li, M. Ma, X. Ma *et al.*, "Preoperative prediction of axillary lymph node metastasis in breast cancer based on intratumoral and peritumoral dce-mri radiomics nomogram," *Contrast Media & Molecular Imaging*, vol. 2022, 2022.

[49] D. Santucci, E. Faiella, E. Cordelli, R. Sicilia, C. de Felice, B. B. Zobel, G. Iannello, and P. Soda, "3t mri-radiomic approach to predict for lymph node status in breast cancer patients," *Cancers*, vol. 13, no. 9, p. 2228, 2021.

[50] E. Cordelli, R. Sicilia, D. Santucci, C. de Felice, C. C. Quattrocchi, B. B. Zobel, G. Iannello, and P. Soda, "Radiomics-based non-invasive lymph node metastases prediction in breast cancer," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 486–491.

[51] C. Chen, Y. Qin, H. Chen, D. Zhu, F. Gao, and X. Zhou, "A meta-analysis of the diagnostic performance of machine learning-based mri in the prediction of axillary lymph node metastasis in breast cancer patients," *Insights into Imaging*, vol. 12, pp. 1–12, 2021.

[52] Y. Kayadibi, B. Kocak, N. Ucar, Y. N. Akan, E. Yildirim, and S. Bektas, "Mri radiomics of breast cancer: Machine learning-based prediction of lymphovascular invasion status," *Academic Radiology*, vol. 29, pp. S126–S134, 2022.

[53] J. Gao, X. Zhong, W. Li, Q. Li, H. Shao, Z. Wang, Y. Dai, H. Ma, Y. Shi, H. Zhang *et al.*, "Attention-based deep learning for the preoperative differentiation of axillary lymph node metastasis in breast cancer on dce-mri," *Journal of Magnetic Resonance Imaging*, 2022.

[54] "Development and validation of convolutional neural network-based model to predict the risk of sentinel or non-sentinel lymph node metastasis in patients with breast cancer: a machine learning study, author=Chen, Mingzhen and Kong, Chunli and Lin, Guihan and Chen, Weiyue and Guo, Xinyu and Chen, Yaning and Cheng, Xue and Chen, Minjiang and Shi, Changsheng and Xu, Min and others, journal=EClinicalMedicine, volume=63, year=2023, publisher=Elsevier."

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] D. Santucci, E. Faiella, M. Gravina, E. Cordelli, C. de Felice, B. Beomonte Zobel, G. Iannello, C. Sansone, and P. Soda, "Cnn-based approaches with different tumor bounding options for lymph node status prediction in breast dce-mri," *Cancers*, vol. 14, no. 19, p. 4574, 2022.

[57] M. Gravina, E. Cordelli, D. Santucci, P. Soda, and C. Sansone, "Evaluating tumour bounding options for deep learning-based axillary lymph node metastasis prediction in breast cancer," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 4335–4342.

[58] H. Zhou, Z. Hua, J. Gao, F. Lin, Y. Chen, S. Zhang, T. Zheng, Z. Wang, H. Shao, W. Li *et al.*, "Multitask deep learning-based whole-process system for automatic diagnosis of breast lesions and axillary lymph node metastasis discrimination from dynamic contrast-enhanced-mri: A multicenter study," *Journal of Magnetic Resonance Imaging*, 2023.

[59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[60] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[61] S. Nguyen, D. Polat, P. Karbasi, D. Moser, L. Wang, K. Hulsey, M. C. Çobanoğlu, B. Dogan, and A. Montillo, "Preoperative prediction of lymph node metastasis from clinical dce mri of the primary breast tumor using a 4d cnn," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 326–334.

[62] Y. Chen, L. Wang, X. Dong, R. Luo, Y. Ge, H. Liu, Y. Zhang, and D. Wang, "Deep learning radiomics of preoperative breast mri for prediction of axillary lymph node metastasis in breast cancer," *Journal of Digital Imaging*, pp. 1–9, 2023.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR, 2009*. IEEE, 2009, pp. 248–255.

[64] X. Zhang, M. Liu, W. Ren, J. Sun, K. Wang, X. Xi, and G. Zhang, "Predicting of axillary lymph node metastasis in invasive breast cancer using multiparametric mri dataset based on cnn model," *Frontiers in Oncology*, vol. 12, p. 1069733, 2022.

[65] Z. Wang, H. Sun, J. Li, J. Chen, F. Meng, H. Li, L. Han, S. Zhou, and T. Yu, "Preoperative prediction of axillary lymph node metastasis in breast cancer using cnn based on multiparametric mri," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 3, pp. 700–709, 2022.

[66] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[67] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank, "Nonrigid multimodality image registration," in *Medical imaging 2001: image processing*, vol. 4322. Spie, 2001, pp. 1609–1620.

[68] S. Rahunathan, D. Stredney, P. Schmalbrock, and B. D. Clymer, "Image registration using rigid registration and maximization of mutual information," in *13th Annu. Med. Meets Virtual Reality Conf*, 2005.

[69] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[70] M. Zhao, Q. Wu, L. Guo, L. Zhou, and K. Fu, "Magnetic resonance imaging features for predicting axillary lymph node metastasis in patients with breast cancer," *European Journal of Radiology*, vol. 129, p. 109093, 2020.

[71] M. Dietzel, P. A. Baltzer, T. Vag, T. Gröschel, M. Gajda, O. Camara, and W. A. Kaiser, "Application of breast mri for prediction of lymph node metastases–systematic approach using 17 individual descriptors and a dedicated decision tree," *Acta Radiologica*, vol. 51, no. 8, pp. 885–894, 2010.

[72] E. H. Jeong, E. J. Choi, H. Choi, E. H. Park, and J. S. Song, "Prediction of axillary lymph node metastasis in early breast cancer using dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted imaging," *Investigative Magnetic Resonance Imaging*, vol. 23, no. 2, pp. 125–135, 2019.

[73] W. Guo, H. Li, Y. Zhu, L. Lan, S. Yang, K. Drukker, E. A. Morris, E. S. Burnside, G. J. Whitman, M. L. Giger *et al.*, "Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data," *Journal of medical imaging*, vol. 2, no. 4, p. 041007, 2015.

[74] D. Santucci, E. Faiella, A. Calabrese, B. Beomonte Zobel, A. Ascione, B. Cerbelli, G. Iannello, P. Soda, and C. de Felice, "On the additional information provided by 3t-mri adc in predicting tumor cellularity and microscopic behavior," *Cancers*, vol. 13, no. 20, p. 5167, 2021.

[75] M. Abdar, M. A. Fahami, S. Chakrabarti, A. Khosravi, P. Pławiak, U. R. Acharya, R. Tadeusiewicz, and S. Nahavandi, "Barf: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification," *Information Sciences*, vol. 577, pp. 353–378, 2021.

[76] M. Abdar, M. A. Fahami, L. Rundo, P. Radeva, A. F. Frangi, U. R. Acharya, A. Khosravi, H.-K. Lam, A. Jung, and S. Nahavandi, "Hercules: Deep hierarchical attentive multilevel fusion model with uncertainty quantification for medical image classification," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 274–285, 2022.