PET-based lesion graphs meet clinical data: An interpretable cross-attention framework for DLBCL treatment response prediction

Oriane Thiery, Mira Rizkallah, Clément Bailly, Caroline Bodet-Milin, Emmanuel Itti, René-Olivier Casasnovas, Steven Le Gouill, Thomas Carlier, Diana Mateus

PET-based lesion graphs meet clinical data: An interpretable cross-attention framework for DLBCL treatment response prediction

Oriane Thiery[a,*], Mira Rizkallah[a], Clément Bailly[b,c], Caroline Bodet-Milin[b,c], Emmanuel Itti[d], René-Olivier Casasnovas[e], Steven Le Gouill[f,g], Thomas Carlier[b,c] and Diana Mateus[a]

[a]*Nantes Université, Centrale Nantes, CNRS, LS2N, UMR 6004, France*
[b]*Nuclear Medicine Department, University Hospital, Nantes, France*
[c]*Nantes Université, INSERM, CNRS, Université d'Angers, CRCI2NA, Nantes, France*
[d]*Nuclear Medicine, CHU Henry Mondor, Paris-Est University, Créteil, France*
[e]*Hematology, CHU Dijon Bourgogne, Dijon, France*
[f]*Institut Curie, Paris, France*
[g]*Université Versailles Saint-Quentin UVSQ, Saint-Quentin, France*

ARTICLE INFO

Keywords:
Multimodal data fusion
Cross-attention
Lesion graphs
Interpretability
Treatment response
DLBCL
PET

ABSTRACT

Diffuse Large B-cell Lymphoma (DLBCL) is a lymphatic cancer of steadily growing incidence. Its diagnostic and follow-up rely on the analysis of clinical biomarkers and 18F-Fluorodeoxyglucose (FDG)-PET/CT images. In this context, we target the problem of assisting the early identification of high-risk DLBCL patients from both images and tabular clinical data. We propose a solution based on a graph neural network model, capable of simultaneously modeling the variable number of lesions across patients, and fusing information from both data modalities and over lesions. Given the distributed nature of the DLBCL lesions, we represent the PET image of each patient as an attributed lesion graph. Such lesion-graphs keep all relevant image information, while offering a compact tradeoff between the characterization of full images and single lesions. We also design a cross-attention module to fuse the image attributes with clinical indicators, which is particularly challenging given the large difference in dimensionality and prognostic strength of each modality. To this end, we propose several cross-attention configurations, discuss the implications of each design and experimentally compare their performances. The last module fuses the updated attributes across lesions and makes a probabilistic prediction of the patient's 2-year progression-free survival (PFS). We carry out the experimental validation of our proposed framework on a prospective multicentric dataset of 545 patients. Experimental results show our framework effectively integrates the multi-lesion image information improving over a model relying only on the most prognostic clinical data. The analysis further shows the interpretable properties inherent to our graph-based design, which enables tracing the decision back to the most important lesions and features.

1. Introduction

Diffuse Large B-cell Lymphoma (DLBCL) is a lymphatic cancer whose incidence is steadily increasing and is expected to account for 30-40% of the 77240 new cases of Non-Hodgkin Lymphoma cases in the US in 2020 [Susanibar-Adaniya and Barta (2021)]. Its diagnosis and follow-up rely on the analysis of clinical biomarkers and semi-quantitative interpretation of 18F-Fluorodeoxyglucose (FDG)-PET/CT images. The tabular clinical data, including information such as blood test results or disease staging, is typically used to support this diagnosis and subsequent follow-up in clinical studies, using classical but interpretable methods [Le Gouill et al. (2021)]. In the field of PET image analysis, trends are to automatically extract quantitative information (radiomic features) to train predictive machine learning methods [Carlier et al. (2024); Jiang et al. (2022); Eertink et al. (2022); Yousefirizi et al. (2024)], or to directly use deep learning methods on the image [Yuan et al. (2022); Liu et al. (2022)]. Despite being prognostic for the

*Corresponding author
oriane.thiery@ls2n.fr (O. Thiery); mira.rizkallah@ec-nantes.fr (M. Rizkallah)
ORCID(s): 0009-0003-1314-1988 (O. Thiery); 0000-0001-7724-9304 (M. Rizkallah); 0000-0001-8313-3287 (C. Bailly); 0000-0002-8219-3592 (C. Bodet-Milin); 0000-0003-1578-4058 (E. Itti); 0000-0002-1156-8983 (R. Casasnovas); 0000-0001-9840-2128 (S.L. Gouill); 0000-0002-6932-7322 (T. Carlier); 0000-0002-2252-8717 (D. Mateus)

pathology [Cottereau et al. (2019)], the above approaches do not explicitly account for multiple lesions and their spatial distribution, or only through simplified measures such as the largest distance between lesions. Furthermore, only a few methods incorporate both imaging and tabular clinical information, and they do so through basic fusion schemes [Carlier et al. (2024); Jiang et al. (2022); Eertink et al. (2022)]. In this context, our aim is to develop a computer-assisted learning-based method to automatically identify high-risk DLBCL patients at diagnosis (*i.e.* those more likely to progress), by efficiently fusing all available information, including imaging information and spatial distribution of multiple lesions, and tabular clinical data. The clinical end-point of this study is the 2-year Progression-Free Survival (PFS) classification, which is designed to predict whether patient's disease will progress within two years after the beginning of the treatment.

While developing a risk classification approach from real-world, prognostic, multicentric, unbalanced and multimodal data is already a complex task, we face additional challenges related to the disease under consideration and the available data. First, the information in the PET volumes of DLBLC patients is spread over multiple, typically small lesions, making feature extraction difficult. Second, both the number of lesions and image resolution can vary significantly between patients, making the generalization difficult. Third, the efficient integration of multiple data modalities is still an open question in the field [Baltrušaitis et al. (2019)]; in our case, the fusion further implies combining an already identified set of highly-predictive clinical features with image information for which there is no consensus (with few exceptions) on their informative value for our task; moreover, there is a consequent gap between the dimensionality of the two modalities, with very few clinical features and many imaging features. Finally, while our work relies on a relatively large DLBCL database for the field, it is not comparable to the huge databases required to train complex fusion models in computer vision [Wang et al. (2022)].

The computer-assisted method developed in this paper first extracts image information at all relevant locations, then relies on a lesion graph and a Graph Neural Network (GNN) to handle the varying number of lesions in each patient, and on a cross-attention fusion module to integrate the two learned modalities. Our contributions are as follows:

- We present a study to determine which type of imaging features are most relevant for a 2-year PFS prediction task in DLBCL;
- We highlight the difficulty of choosing the right model structure to work with complex real-world data, and present an extensive study of our own model, including how best to use a cross-attention module to fuse multimodal data;
- We show that a max pooling operation, when included in a specific configuration of our model, selects a limited number of lesions for prediction, thus supporting the interpretability of our model;
- And we propose a comprehensive study on the interpretation of the results of our best model, highlighting its potential to help clinicians understand both its predictions and which features (both imaging and clinical) are most relevant for identifying high-risk DLBCL patients.

More specifically, we follow and extend the work developed in [Thiery et al. (2023)]. To improve the quality of multimodal fusion and the performance of high-risk patient identification, we conducted an extensive study and explored several variants of the image features retrieved from the lesions and the inputs to the cross-attention module, a key design element that we focus on. We also perform an ablation study on the structural modifications of the best model found, and validate its performance on a complex set of imaging features. Using the best-performing model, we analyze the attention patterns to highlight which lesions are most relevant to the task and provide an interpretation for each patient. We also take advantage of the interpretability of our novel multi-lesion multimodal cross-attention design to rank the most salient clinical features.

2. Related work

First, we briefly review previous work on multi-modal fusion, a problem related to the core of this paper. Beyond the scope of DLBCL PET image analysis, there is a wide variety of methods to fuse different data modalities, ranging from classical early, late and hybrid fusion approaches to kernel-based methods, graphical models and neural networks [Baltrušaitis et al. (2019)].

Deep learning is now widely used in medical image analysis for both decision making and feature extraction, with the majority of methods dedicated to image-to-image fusion problems [Zhou et al. (2023)]. Recently, there has been

considerable interest in the fusion of medical images with text, e.g. from clinical reports. For instance, Wang et al. (2022) achieved the contrastive pre-training of a large text-image CLIP model from [Radford et al. (2021)] exploiting unpaired medical data. Less work has focused on the deep learning fusion of tabular and medical image data. For instance, Pölsterl et al. (2021) addresses the problem by learning affine transformations that dynamically modify CNN-based image features according to tabular data. Also, Hager et al. (2023) propose a contrastive learning framework with a CLIP loss to align image and tabular embeddings in the context of cardiac classification tasks on the large UK Biobank dataset [Sudlow et al. (2015)]. After pretraining, the prediction tasks rely on a concatenation of the latent vectors of unimodal autoencoders. Similarly, Huang (2023) creates a joint embedding of labels and images, followed by a contrastive alignment of tabular data with the image embedding considered as the main modality. During inference, the embeddings of different modalities are concatenated and directly compared with the label encoding for zero shot predictions. Despite these recent advances, there is still no widely accepted method for dealing with image-tabular data fusion for medical data. In particular, contrastive multi-modal fusion frameworks are difficult to implement for our targeted problem, even as pre-training, due to the lack of large publicly available PET-tabular datasets. Moreover, the relevant information in PET images of DLBCL patients is concentrated in a varying number of multiple small and sparsely distributed lesions. It is therefore inefficient to process the whole image directly as input, and the approaches mentioned above fail to process the varying number of lesions considered independently, as would be the case with concatenation operations. Instead, we aim at a unique dynamic encoding that is able to produce an adaptive multi-modal embedding of each lesion, and at the same time is able to integrate the information across the variable number of lesions.

Unlike classical contrastive methods that rely on static encoders (for all [Hager et al. (2023)] or at least one modality [Huang (2023)]), we opt for a dynamic adaptation of the multimodal embedding, similar in idea to [Pölsterl et al. (2021)] but relying on a cross-and-self-attention mechanism instead of an affine transformation. Such transformer-inspired cross-attention modules have been investigated for multi-modal fusion for non-medical [Nagrani et al. (2021)], as well as medical data, as in [Gao et al. (2023); Wang et al. (2024)] and reviewed in [Andrade-Miranda et al. (2023)]. As with contrastive methods, most of existing approaches fuse modalities with similar dimensions, with some exceptions [Golovanevsky et al. (2022); Gao et al. (2023)]. In our case, we are faced with the fusion of modalities with heterogeneous dimensions. More importantly, none of the methods reviewed considers multiple lesions of a patient individually. For the above reasons, we propose to combine our cross-modal fusion with a lesion graph representation.

More precisely, to deal with the variable number of lesions per patient, we build a lesion graph where the image features of each lesion as well as their structural relationships are encoded in a single graph. Similar lesion graphs have been explored in [Kazmierski and Haibe-Kains (2021)] and [Lv et al. (2023)] in combination with a Graph Neural Network (GNN) to predict the probability of distant metastasis over time and the PFS of head and neck cancer, from CT scans and both CT and PET scans, respectively. Prabhakar et al. (2023) also rely on a lesion-graph and a GNN to predict inflammatory activity in multiple sclerosis, while Rist et al. (2023) presents a comparison of multiple lesion graphs and learning processes to estimate the degree of metastatic progression in liver cancer.

Despite the similar representation, the integration of multimodal data with such lesion graphs has not been fully explored. While Prabhakar et al. (2023); Rist et al. (2023) do not address multi-modal fusion, Kazmierski and Haibe-Kains (2021); Lv et al. (2023) only use a simple concatenation of a clinical data vector with the output of the GNN before the prediction layer (naive late fusion). In general, few graph-based methods have addressed the fusion of multiple data modalities. One exception is the work of Liu et al. (2023) who also rely on a cross-attention module to fuse a patient graph representing the brain of a post-stroke patient with a set of tabular clinical data. However, their approach is based on a fixed graph structure and is therefore incompatible with a variable number of lesions.

Unlike the methods discussed above, our work addresses a two-sided fusion problem, considering multiple lesions per patient and the integration of two data modalities, *i.e.* PET images and clinical tabular data. To this end, we propose a combination of a lesion-graph representation with a cross-attention module, where the latter is designed to deal with the variable number of lesions and the heterogeneous modality dimensions. We investigate different instantiations of such a cross-attention module to optimize the transfer of the most relevant information. In addition to addressing the above challenges, our joint explicit graph and cross-attention model favors intepretability, as it allows tracking the contributions of each lesion to the clinical data (or vice versa). Finally, and interestingly, the output layer of our model acts as a Multiple Instance Learning (MIL) aggregation of the learned multimodal embedding of multiple lesions, retaining lesions that provide both positive (PFS>2 years) and negative (PFS<2 years) evidence, *i.e.* lesions that provide counter-evidence for the presence of PFS>2 years. While explicit designs previously introduced in the literature [Durand et al. (2015, 2019)] have shown a performance interest in explicitly enforcing the search for negative

evidence, a similar behavior emerges from our fusion configuration when relying on the clinical data alone as the value matrix of the cross-attention module.

3. Proposed Graph-based Multi-modal Multi-Lesion Fusion Learning Framework

Consider a DLBCL clinical examination of a patient n at baseline before treatment, consisting of a set of tabular clinical indicators $\mathbf{c}^{(n)} \in \mathbb{R}^{D_{\text{clin}}}$ and a whole-body PET image of the patient $\mathcal{I}^{(n)}$. The proposed framework, presented in Fig. 1, takes this whole-body 3D PET image with 3D lesion segmentation $\mathcal{I}^{(n)}_{seg}$ and the corresponding clinical tabular data as input, and is trained to predict the probability of 2-year Progression-Free Survival (PFS). Patients with a low probability are said to be in the high-risk group.

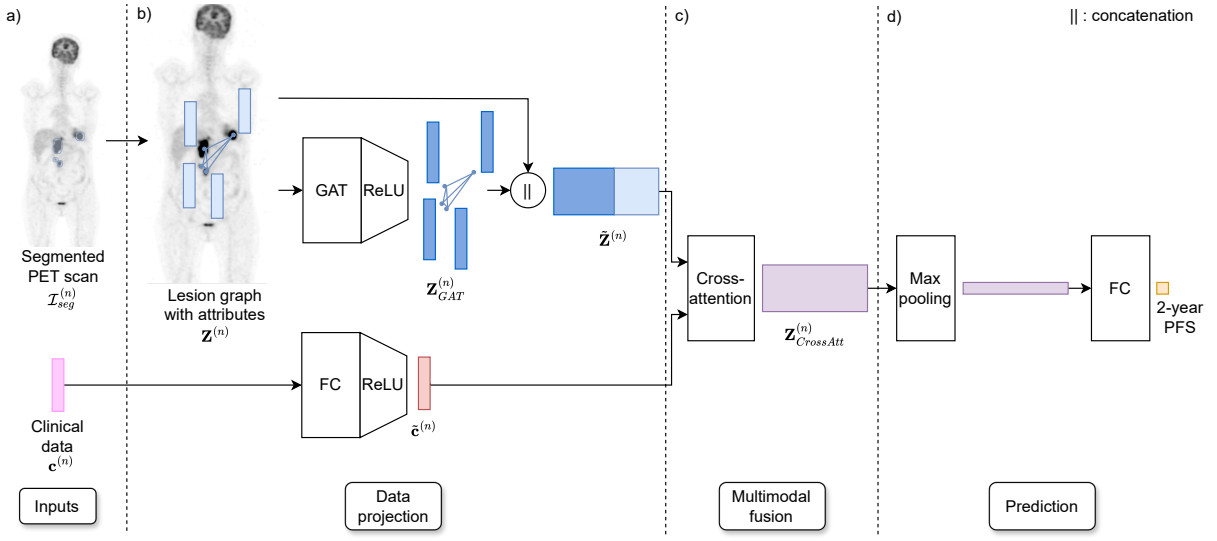


Figure 1: Overview of the proposed framework: a) For a patient n, this framework takes as input the whole-body PET image with 3D lesion segmentation $\mathcal{I}^{(n)}_{seg}$ and the corresponding clinical tabular data $\mathbf{c}^{(n)}$; b) each data modality is projected into a more informative embedding space: a linear projection is applied to the clinical data with a fully connected (FC) layer. The imaging information is first modelled as a lesion graph and the features of all the lesions are merged by a graph attention module (GAT) and concatenated with their initial features; c) the two projected data modalities are fused thanks to a cross-attention module; finally, d) a max pooling operation allows us to obtain a single vector. The latter is fed into an FC layer leading to the 2-year PFS prediction.

Our proposed end-to-end learning framework consists of several blocks. First, we create a lesion graph to simultaneously represent the image features of individual lesions and their spatial distribution. Then, a Graph Neural Network(GNN) coupled with a cross attention module learns to predict 2-year PFS from this lesion graph and the clinical data. To do this, the information from all lesions is first integrated thanks to a Graph Attention Network (GATv2 model [Brody et al. (2022)]), which has the ability to adaptively learn the attentions between lesions for each patient. In parallel, the clinical data is projected with a Fully Connected (FC) layer. The outputs of these operations are then fused by a cross-attention module, which can bring together multiple modalities of data of different dimensions with interpretable behavior. These steps are followed by a pooling operation to reduce the output of the cross-attention layer to a single vector of fixed size, which is sent to an FC layer for prediction.

3.1. Lesion graph creation

The first step of this framework is to create a fully connected graph $\mathcal{G}^{(n)} = \{\mathcal{V}^{(n)}, \mathcal{E}^{(n)}\}$ to group the information of the $L^{(n)}$ lesions appearing on the PET scan of the n^{th} patient. Each of its nodes $v_i^{(n)} \in \mathcal{V}^{(n)}$ corresponds to a single lesion i, from which the image data are extracted and represented in the feature vector $\mathbf{z}_i^{(n)} \in \mathbb{R}^{D_{\text{img}}}$. This image data is composed of both classical intensity-based and radiomics features[1] (see Table 1 in the Supp. material A). From now

[1]Here, classical features are quantitative measurements on the segmented lesion describing the intensity distribution of the voxels. Radiomics features, on the other hand, describe the 3D structure of the lesion, such as shape, or second-order features that characterize the texture of the lesion.

on, we will denote by $\mathbf{Z}^{(n)} \in \mathbb{R}^{L^{(n)} \times D_{\text{img}}}$ the matrix concatenating all the features of the nodes $\mathbf{z}_i^{(n)}$, where D_{img} is the dimension of the vector including classical and radiomics features.

In terms of graph connectivity, we define edges $e_{ij}^{(n)} \in \mathcal{E}^{(n)}$ between each pair of nodes $v_i^{(n)}$ and $v_j^{(n)}$ and include self-loops. To favor message passing between closer and more similar lesions, weights $w_{ij}^{(n)}$ are assigned to each edge. Their values are thus defined based on the distances between both the lesion centroids ($\mathbf{p}_i^{(n)}$, $\mathbf{p}_j^{(n)}$) and the feature vectors ($\mathbf{z}_i^{(n)}$,$\mathbf{z}_j^{(n)}$):

$$w_{ij}^{(n)} = \exp\left(-\frac{||\mathbf{p}_i^{(n)} - \mathbf{p}_j^{(n)}||_2}{\gamma\sigma_1^2}\right) \cdot \exp\left(-\frac{||\mathbf{z}_i^{(n)} - \mathbf{z}_j^{(n)}||_2}{\gamma\sigma_2^2}\right), \tag{1}$$

where $||.||_2$ stands for the L2 norm; γ is a hyper-parameter tuned to find the best edge weight distribution for our task; and σ_1, σ_2 denote the population-level standard deviations of the centroid and feature distances, respectively.

3.2. Multi-Lesion and Clinical data Representation Learning

3.2.1. Multi-lesion

In order to incorporate the information from all the patient's lesions while taking into account their relationships, we design a GNN to be applied over our lesion graph. We define a GATv2 layer with two heads [Brody et al. (2022)] to update each node based only on the information propagated from the most relevant neighboring lesions. To identify the latter, this module learns to compute attention weights $\alpha_{i,j}^{(n)}$ between each pair of nodes as follows [2]:

$$\alpha_{i,j}^{(n)} = \frac{\exp(\mathbf{a}^T \,\text{LeakyReLU}(\mathbf{\Theta}_s\mathbf{z}_i^{(n)} + \mathbf{\Theta}_t\mathbf{z}_j^{(n)} + \mathbf{\Theta}_e w_{ij}^{(n)}))}{\sum_{k\in\mathcal{N}(i)\cup\{i\}} \exp(\mathbf{a}^T \,\text{LeakyReLU}(\mathbf{\Theta}_s\mathbf{z}_i^{(n)} + \mathbf{\Theta}_t\mathbf{z}_k^{(n)} + \mathbf{\Theta}_e w_{ik}^{(n)}))}, \tag{2}$$

with $\mathbf{a}$, $\mathbf{\Theta}_s$, $\mathbf{\Theta}_t$ and $\mathbf{\Theta}_e$ learned parameter matrices and $\mathcal{N}(i)$ the neighboring nodes of $v_i^{(n)}$. The updated feature vector of a node $\mathbf{z}_{i_{\text{GAT}}}^{(n)} \in \mathbb{R}^{D_{\text{GAT}}}$ is then computed as:

$$\mathbf{z}_{i_{\text{GAT}}}^{(n)} = \alpha_{i,i}^{(n)}\mathbf{\Theta}_s\mathbf{z}_i^{(n)} + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}^{(n)}\mathbf{\Theta}_t\mathbf{z}_j^{(n)}. \tag{3}$$

Following this GATv2 operation, a ReLU activation layer is applied on the updated lesion representations. The resulting feature matrix of dimension ($L^{(n)} \times D_{\text{GAT}}$) is a concatenation of the lesion feature vectors: $\mathbf{Z}_{\text{GAT}}^{(n)} = \text{ReLU}([\mathbf{z}_{1_{\text{GAT}}}^{(n)} || \dots || \mathbf{z}_{L^{(n)}{}_{\text{GAT}}}^{(n)}]^\top)$. Finally, a skip connection is added to preserve the original information of the lesions with a simple concatenation operation resulting in $\tilde{\mathbf{Z}}^{(n)} \in \mathbb{R}^{L^{(n)} \times (D_{GAT} + D_{img})}$.

Furthermore, we explore two other options to replace the GATv2 module to fuse the lesion information. Thus, a node embedding $\mathbf{z}_i^{(n)}$ can either be updated by a GATv2 layer as in Eq. (3) ; or by a GraphConv module [Morris et al. (2019)] as follows :

$$\mathbf{z}_{i_{\text{GraphConv}}}^{(n)} = \mathbf{W}_1\mathbf{z}_i^{(n)} + \mathbf{W}_2\left(\sum_{j\in\mathcal{N}(i)} w_{ij}^{(n)}\mathbf{z}_j^{(n)}\right), \tag{4}$$

with $\mathbf{W}_1$ and $\mathbf{W}_2$ being learnt projection matrices.

Unlike the previous two message passing schemes, where the structure of the graph and the edge weights are imposed, in the last option the graph (*i.e.* the strength of the connections between lesions) is dynamically learned by a self-attention module [Vaswani et al. (2017)]. In this case the features are updated as:

[2] This implementation corresponds to the `torch_geometric` version of the GATv2 operator.

$$\mathbf{Z}^{(n)}_{\text{SelfAtt}} = \text{softmax}\left(\frac{\mathbf{Z}^{(n)}\mathbf{\Theta}_Q(\mathbf{Z}^{(n)}\mathbf{\Theta}_K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}^{(n)}\mathbf{\Theta}_V, \tag{5}$$

with three learnable matrices $\mathbf{\Theta}_Q \in \mathbb{R}^{D_{\text{img}}\times d_q}$, $\mathbf{\Theta}_K \in \mathbb{R}^{D_{\text{img}}\times d_k}$ and $\mathbf{\Theta}_V \in \mathbb{R}^{D_{\text{img}}\times d_v}$, with the role of projecting the image features to improve the quality of their combination. This last option, based on the self-attention module, directly returns the projected feature matrix for all lesions, $\mathbf{Z}^{(n)}_{\text{SelfAtt}}$, of the same dimension as $\mathbf{Z}^{(n)}$.

3.2.2. Clinical data

In parallel with the multi-lesion representation learning step, a fully connected layer followed by a ReLU activation projects the patient's clinical data $\mathbf{c}^{(n)}$ into a more informative space. We will refer to the output of this step as $\tilde{\mathbf{c}}^{(n)}$, with dimension D_{proj}.

3.3. Multimodal Multi-lesion Cross-Attention

After the construction of a lesion graph and the design of the GNN to efficiently model the relevant image information, a second core element of our model is a cross-attention module for multi-modal fusion. Inspired by self-attention blocks in NLP [Vaswani et al. (2017)], the cross-attention module fuses the information from the lesion graph with attributes $\tilde{\mathbf{Z}}^{(n)}$ and the projected clinical data $\tilde{\mathbf{c}}^{(n)}$. This module takes one input for a query $\mathbf{Q} \in \mathbb{R}^{\mathcal{H}_Q\times\mathcal{W}_Q}$ and another for a key/value pair $\mathbf{K} \in \mathbb{R}^{\mathcal{H}_K\times\mathcal{W}_K}$ and $\mathbf{V} \in \mathbb{R}^{\mathcal{H}_V\times\mathcal{W}_V}$ (with $\mathcal{H}_K = \mathcal{H}_V$) and returns a weighted sum of the values, where the weight assigned to each value is computed from a compatibility function, *i.e.* a scalar product, of the query with the corresponding key (normalized by the key dimension). Formally,

$$\begin{aligned}\mathbf{Z}^{(n)}_{\text{CrossAtt}} &= \text{CrossAtt}(\mathbf{Q},\mathbf{K},\mathbf{V}) \\ &= \text{softmax}\left(\frac{\tilde{\mathbf{Q}}\tilde{\mathbf{K}}^\top}{\sqrt{\mathcal{W}_{\tilde{K}}}}\right)\tilde{\mathbf{V}},\end{aligned} \tag{6}$$

where $\tilde{\mathbf{Q}} = \mathbf{Q}\mathbf{W}_Q$, $\tilde{\mathbf{K}} = \mathbf{K}\mathbf{W}_K$ and $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{W}_V$ are the learned latent representations via three learnable matrices $\mathbf{W}_Q \in \mathbb{R}^{\mathcal{W}_Q\times\mathcal{W}_{\tilde{Q}}}$, $\mathbf{W}_K \in \mathbb{R}^{\mathcal{W}_K\times\mathcal{W}_{\tilde{K}}}$ and $\mathbf{W}_V \in \mathbb{R}^{\mathcal{W}_V\times\mathcal{W}_{\tilde{V}}}$. The output of the cross-attention operation is a matrix $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ of size $\mathcal{H}_Q\times\mathcal{W}_{\tilde{V}}$. Intuitively, the matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ project the query and key information into a common space, before computing their compatibility. The softmax operation then provides the attention scores that should be given to each element of the value matrix. Finally, these attention scores are multiplied by the value matrix which is also projected into a relevant space by $\mathbf{W}_V$.

In this work, we propose different instantiations of this cross-attention module to investigate how such implementations affect the fusion performance and address the specific multi-modal and multi-lesion challenges we have encountered. Not only do we have to deal with different predictive capabilities and dimensions between the two modalities, but we also have to deal with different image data dimensions between patients due to the inconsistent number of their lesions. Therefore, we take advantage of the adaptability of the cross-attention module to let each modality contribute asymmetrically (one modality influencing the weighting of the other) or in a more classical way with a self-attention process (Mult2Mult implementation). Five configurations, presented in Fig. 2, are described below:

- **Img2Clin**: First, we consider a model where $\mathbf{Q} = \tilde{\mathbf{Z}}^{(n)}$, and $\mathbf{K} = \mathbf{V} = \tilde{\mathbf{c}}^{(n)}$. As the clinical features are already known to be predictive for the 2-year PFS [Carlier et al. (2024)], using it as the value of the module should improve this prediction.

- **Mult2Clin**: Similar to Img2Clin, but includes self-attention to the clinical data in addition to cross-attention between the two modalities. If the clinical data supports the imaging information to decide which clinical elements are most relevant for prediction, this could allow us to make the most of its predictive capabilities. So, we still have $\mathbf{K} = \mathbf{V} = \tilde{\mathbf{c}}^{(n)}$, but $\mathbf{Q}$ is equal to $\tilde{\mathbf{Z}}^{(n)}||\tilde{\mathbf{c}}^{(n)\top}_{concat}$, where $\tilde{\mathbf{c}}^{(n)}_{concat}$ is the matrix resulting from the column-wise concatenation of the vector $\tilde{\mathbf{c}}^{(n)}$, $L^{(n)}$ times. Thereby, the shape of $\mathbf{Q}$ is $(L^{(n)} \times D_{concat})$ with $D_{concat} = (D_{GAT} + D_{img}) + D_{proj}$.

- **Mult2Mult**: It uses both self- and cross-attention on each type of data, so $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \tilde{\mathbf{Z}}^{(n)}||\tilde{\mathbf{c}}_{concat}^{(n)\top}$. This corresponds to a common way of fusing multimodal data with a cross-attention module, where the input key, query and value are all a concatenation of both modalities.

- **Clin2Img**: This model is the opposite of Img2Clin, as we invert $\mathbf{Q}$ and $\mathbf{K} = \mathbf{V}$ to get $\mathbf{Q} = \tilde{\mathbf{c}}^{(n)\top}$, and $\mathbf{K} = \mathbf{V} = \tilde{\mathbf{Z}}^{(n)}$. With this configuration, we investigate whether our method can extract information from the lesion data that is relevant to predicting 2-year PFS, based on the clinical data.

- **Mult2Img**: Our final model is an extension of Clin2Img where, in addition to simple cross-attention between the two modalities, we include self-attention to the image data. We expect this configuration, as in our GNN module, to use the relationship between the lesions in addition to the clinical data to select which lesion's updated information is most predictive for the learning task. Thus, we choose $\mathbf{Q} = \tilde{\mathbf{Z}}^{(n)}||\tilde{\mathbf{c}}_{concat}^{(n)\top}$ and $\mathbf{K} = \mathbf{V} = \tilde{\mathbf{Z}}^{(n)}$.

3.4. Prediction module

Finally, a max pooling operation, taking the channel-wise maximum across the node dimension, reduces the dimensionality of $\mathbf{Z}_{\text{CrossAtt}}^{(n)}$ into a single $\mathcal{H}_Q$-dimensional vector (depending on the shape of $\mathbf{Q}$), and a fully connected layer with sigmoid activation predicts the probability of the patient's 2-year PFS from this vector. Learning is controlled by a weighted binary cross-entropy loss function, with weights compensating for class imbalance (ratio of positive to negative samples $\sim 1:5$). The max pooling operation for the Img2clin and Mult2Clin cross-attention configurations forces the model to concentrate all information from the patient on two lesions that positively and negatively influence its 2-year PFS prediction, respectively. This behavior, which is very useful in terms of interpretability, is explained in section 4.5. In general, the whole fusion framework presented is similar to a MIL (Multiple Instance Learning) or late fusion process in the sense that the prediction is made immediately after the cross-attention module responsible for the fusion of the two modalities. However, it has two important differences with these types of models:

- For the image information, we do not consider the features of each lesion separately, as a MIL would do, but rely on all of them and on the underlying topology to update the representation of each of these lesions.

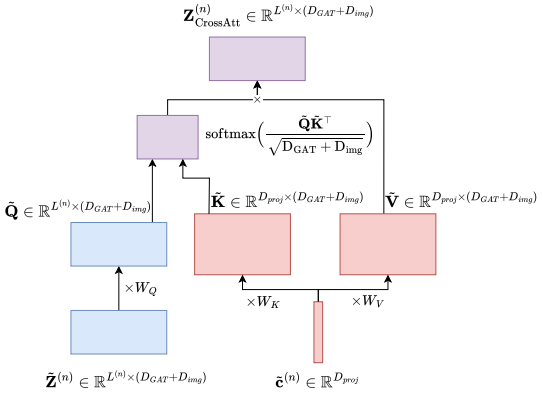
- In the cross-attention module, we do not simply sum or concatenate the features of the two modalities to make the prediction, but we dynamically and adaptively discover from one or both modalities which features of the other are most relevant for a given patient to predict its 2-year PFS.

These differences, as shown in section 4, allow us to outperform conventional and state-of-the-art models by taking advantage of the multiple modalities available.

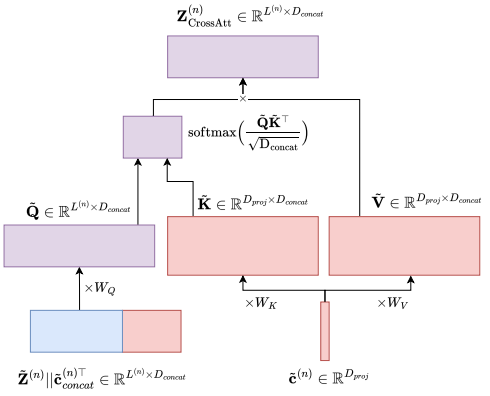
4. Experimental validation

4.1. Dataset

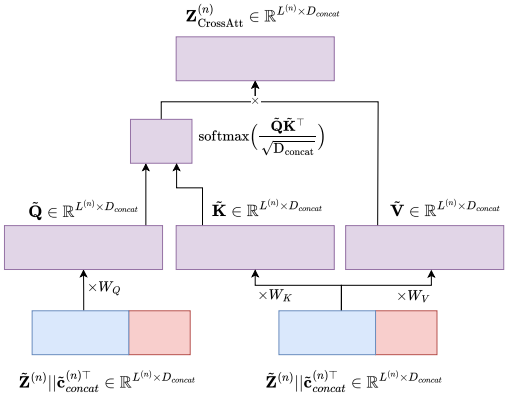
Our validation is based on the prospective GAINED study (NCT 01659099) [Le Gouill et al. (2021)] which enrolled 670 newly diagnosed and untreated DLBCL patients from 99 centers. To obtain a prediction of 2-year PFS, we excluded patients whose data were censored before this time and those whose PET images did not meet the protocol quality criteria. This left 545 patients for our experiments. Each patient is associated with a PET scan at baseline and 10 clinical indicators such as age, Ann Arbor stage or number of extranodal sites (full list in Table 2 in Supp. material A). Lesion detection on the PET images is performed manually by a clinician. The lesions are then segmented within the initial bounding box using different approaches as described in section 4.3. With respect to our preliminary work [Thiery et al. (2023)], the detection and segmentation have been completely revised due to the identification of some inconsistencies that explain the difference in performance for some identical models. Before computing the radiomic imaging features, the segmented lesions were resampled to the same voxel size ($2 \times 2 \times 2$ mm) using a bicubic spline interpolation and then normalized using two approaches: a linear equalization with 64 bins and a fixed SUV bin-width of 0.3. Finally, both the imaging features and the clinical data are standardized by removing the mean and scaling their variance to unity. Normalization of lesion position during the graph creation is performed in the same way, but individually for each patient to account for differences in patient size.



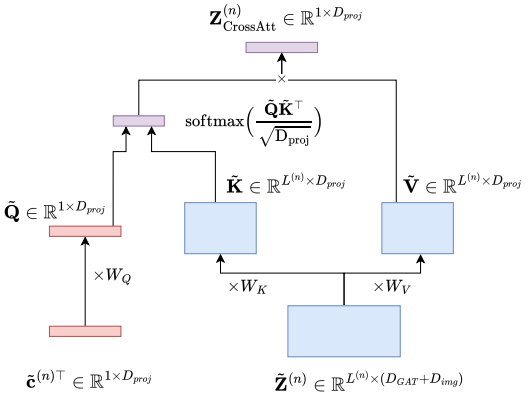
(a) Img2Clin



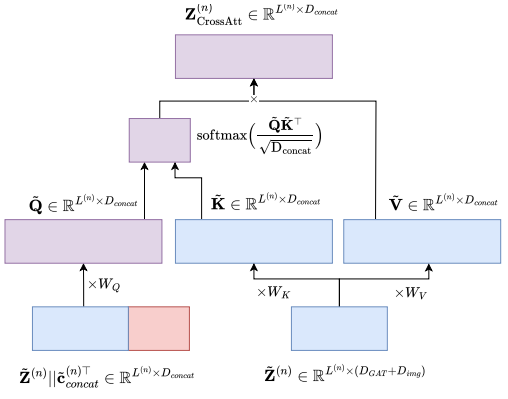
(b) Mult2Clin



(c) Mult2Mult



(d) Clin2Img



(e) Mult2Img

Figure 2: Proposed configurations of the cross-attention module

4.2. Experimental setup

For the different experiments in the following sections, we report the 2-year PFS classification performance of the evaluated methods in terms of AU-ROC, based on a cross-validation scheme and a balanced sampling strategy, with a fixed random seed over all experiments. To account for random uncertainty in the data, we report the average of the results over five random train-validation/test splits. For each of these five sets of test patients, we perform ten

loops of training and validation, with random splits of the patients in the training and validation sets (from those not in the test set) (see Fig. 3.) The proportions of training, validation and test patients in each loop are 80%, 10% and 10% respectively, and we force the proportion of positive patients to be the same in all sets. Furthermore, to ensure that the scores are computed on balanced sets, we repeat the validation and test phases five times: for each run, we build a balanced validation set (resp. test set) with the available positive data and a 1/5 randomly sampled subset of the negative data in the validation set (resp. test set), where this ratio corresponds to the ratio of positive to negative samples in the dataset. We then average the metrics of these five balanced runs to produce the final validation or test results. Thus, the results presented in this paper correspond to the test performance, averaged over the five train-validation/test splits and the ten balancing loops, of the model with the best validation AU-ROC of the loop.

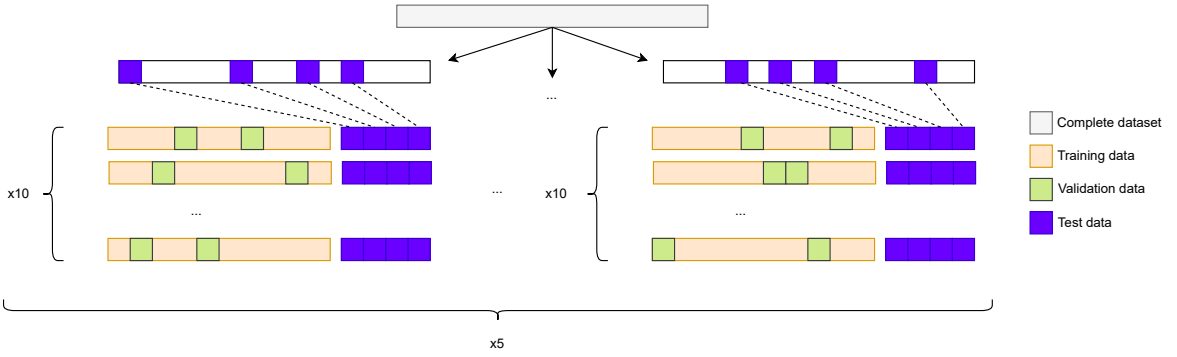


Figure 3: Cross-validation process. A total of five random train-validation/test splits are created. For each of the five sets of test patients, ten loops of training and validation are performed.

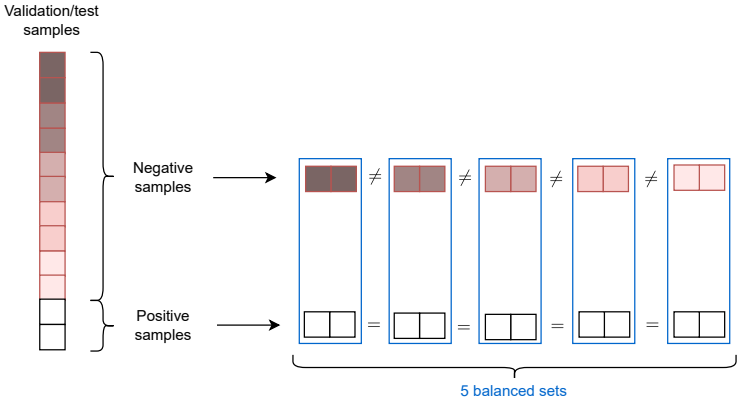


Figure 4: Creation of a balanced validation/test set. We divide the negative samples into five sets, each of which is combined with all the positive samples to create balanced validation/test sets for testing.

A grid search is performed on all evaluated models, including the baseline models and the different configurations studied, by selecting the configuration with the best validation AUC-ROC. These grid searches explore the space of hyperparameters, including the learning rate, the output dimension of the projection layers (FC, GraphConv or GATv2 layers) or, for the GNN methods, the parameter γ (used in the construction of the lesion graphs). The full set of hyperparameters explored is presented in Table 3 in Supp. material A. In addition, we use a McNemar test to validate the statistical significance of our results.

The whole framework was coded in Python using the `PyTorch` and `torch_geometric` modules, and the radiomics were extracted from the lesions using `pyradiomics` [van Griethuysen et al. (2017)].

4.3. Choice of clinical and image input features and comparison to baseline methods

First, we conduct a study to determine the most relevant features for the risk stratification task. A first option would be to feed the whole body images directly into a CNN [Yuan et al. (2022); Liu et al. (2022)], but there is a risk of introducing irrelevant information as lesions occupy only a small part of the image. Furthermore, such CNN features are difficult to interpret. Therefore, we pursue a second option based on the extraction of radiomic features from the segmented lesions to preserve only relevant information while improving interpretability [Carlier et al. (2024); Jiang et al. (2022); Eertink et al. (2022); Yousefirizi et al. (2024)]. However, there is currently no consensus in the community as to which ensemble of radiomic features is the most predictive for the risk stratification task. Furthermore, the extracted features will also depend on the lesion segmentation strategy. To this end, in this experiment we evaluate three common types of segmentation strategies as well as three sets of radiomic features. To broaden the conclusions of this evaluation, we compare these features and segmentation strategies with several baseline models.

4.3.1. Segmentation strategies

The segmentation strategies considered include raw bounding boxes, majority voting and threshold-based segmentation. In more detail, the respective binary segmentation masks are obtained from:

- the raw polygonal bounding boxes drawn by clinicians ('Raw' in Table 1);
- a majority vote ('Majority vote' in Table 1) between three automatic segmentations computed on the raw bounding boxes combining i) a K-means clustering with K = 2, ii) a thresholding to retain only voxels with intensity values greater than 40% of the maximum intensity, and iii) a second thresholding to retain voxels whose normalized SUV (Standard Uptake Value) is greater than 2.5;
- a normalized SUV thresholding on the bounding boxes with a threshold set to 4 ('SUV4' in Table 1).

4.3.2. Sets of radiomic features

Depending on the segmentation strategy, we investigate three sets of radiomic features extracted from each segmented lesion. These radiomic features, after extraction, correspond to the vector $\mathbf{z}_i^{(n)} \in \mathbb{R}^{D_{\text{img}}}$ assigned to the lesion i of a patient n. First, we consider a set of 15 features inspired by the work of Kazmierski and Haibe-Kains (2021) on head and neck cancer and used in [Thiery et al. (2023)] ('HeadNeck' in Table 1). The second option investigated is the 'FirstOrder' set of features, corresponding to those from [Kazmierski and Haibe-Kains (2021); Thiery et al. (2023)] without the second-order features, computed both on images pre-processed as described in section 4.1 and with additional pre-processing to improve the image characteristics: a wavelet transform using coiflet-1 filters to decompose the original image into 8 frequencies and an edge enhancement Laplacian of Gaussian filter (with 2- and 6-mm σ). This makes 34 features. The last feature set considered ('All' in Table 1) is the full set of 650 image features developed in Supp. material B. These radiomics can only be computed on raw bounding boxes to avoid computing texture and statistics on regions that are too small, which is known to be unreliable [Zwanenburg et al. (2016)].

Five combinations of these segmentation strategies and radiomic feature sets were investigated. First, the Featv1 feature set corresponds to 'HeadNeck' featuresextracted from lesions segmented by majority vote. The Featv2, Featv3 and Featv4 feature sets are all composed of first-order features but are extracted from raw bounding box, majority vote and thresholding segmentations respectively. Finally, the Featv5 set consists of the full 650 radiomic features computed directly on the bounding boxes of the lesions.

4.3.3. Baseline models

We then measure the influence of the above image features set selection on the 2-year PFS prediction task. To this end, we compare several baseline models using either clinical data, imaging data or both. These models deal differently with the variable number of lesions in a patient and rely on different fusion approaches when both modalities are used. The methods investigated are:

- **MLP**: This model consists of two linear layers with ReLU activations and a 1-dim linear output layer with sigmoid activation. The two intermediate layers have the same dimension, chosen by a grid search. The MLP model allows us to see how predictive each modality is for the given task with a very simple approach, and to test a simple early fusion scheme. However, as it requires fixed size inputs, the lesion image information must be reduced to a fixed size vector, either by averaging the features from all lesions or by considering only the data from a single lesion.

	clin	***rad***	Featv1	Featv2	Featv3	Featv4	Featv5
Segmentation			Majority vote	Raw	Majority vote	SUV4	Raw
Image features			HeadNeck	FirstOrder	FirstOrder	FirstOrder	All
MLP	1		$\underline{0.65 \pm 0.09}$	$\underline{0.65 \pm 0.09}$	$\underline{0.65 \pm 0.09}$	0.65 ± 0.09	$\mathbf{0.65 \pm 0.09}$
MLP		1	0.53 ± 0.08	0.52 ± 0.12	0.56 ± 0.09	0.56 ± 0.11	0.50 ± 0.09
MLP		2	0.61 ± 0.11	0.52 ± 0.08	0.53 ± 0.10	0.51 ± 0.09	0.56 ± 0.08
MIL		3	0.52 ± 0.10	0.46 ± 0.09	0.51 ± 0.10	0.50 ± 0.08	0.55 ± 0.12
GraphConv		3	0.54 ± 0.11	0.56 ± 0.11	0.54 ± 0.09	0.54 ± 0.09	0.61 ± 0.07
GAT		3	0.55 ± 0.10	0.54 ± 0.10	0.53 ± 0.10	0.51 ± 0.11	0.59 ± 0.10
MLP	1	1	0.61 ± 0.10	0.62 ± 0.12	0.61 ± 0.11	$\underline{0.66 \pm 0.09}$	0.58 ± 0.09
MLP	1	2	0.63 ± 0.05	0.60 ± 0.09	0.62 ± 0.07	0.60 ± 0.08	0.59 ± 0.09
Cross-att.	1	3	0.59 ± 0.11	0.63 ± 0.08	0.64 ± 0.08	0.63 ± 0.08	0.64 ± 0.08
Cross-att. 2	1	3	$\mathbf{0.67 \pm 0.09}$	$\mathbf{0.66 \pm 0.08}$	$\mathbf{0.67 \pm 0.08}$	$\mathbf{0.67 \pm 0.07}$	$\underline{0.64 \pm 0.10}$

Table 1
AU-ROC baseline results: ***clin*** and ***rad*** indicate whether clinical (***clin***) and/or radiomic (***rad***) data are used for the model tested. For ***rad***, this means (1) that we consider only the data from the largest lesion; (2) that we average the imaging data from all lesions; and (3) that we consider the data from each lesion individually. For a set of features, the first is bold and second is underlined.

- **MIL**: A Multiple Instance Learning approach which takes as input the image features from a patient's $L^{(n)}$ lesions, applies a one-layer MLP followed by a ReLU to each lesion's feature vector, aggregates the results by a maximum operation, and projects them linearly (with a sigmoid activation) to obtain the prediction. This approach allows us to use the image information from each lesion individually, but without using the topological information from the graph.
- **GraphConv**: A GNN approach that takes as input the lesion graph presented in Sec. 3.2. It consists of two GraphConv [Morris et al. (2019)] convolutional layers, the first with an output dimension determined by grid search and followed by ReLU activation, and the second with an output size of 1 and followed by max pooling and sigmoid activation to predict PFS. Both this model and the following GAT aim to explore the interest of graph structure over just considering lesion image features, as the MIL does.
- **GAT**: A model with the same structure as the GraphConv model, but with GATv2 layers instead of GraphConv layers.
- **Cross-attention:** The intermediate fusion model presented in section 3, with its original configuration from our preliminary work [Thiery et al. (2023)] based on two Img2Clin cross-attention modules, shows that it is possible to take advantage of both modalities while considering all lesions.
- **Cross-attention 2:** The updated Cross-attention model based on two Img2Clin cross-attention modules, with the addition of two successive fully connected layers followed by ReLU activations to project the clinical data (one before each cross-attention module).

The above models take as input the image information $\mathbf{Z}^{(n)} \in \mathbb{R}^{L^{(n)} \times D_{\text{img}}}$, the clinical data vector $\mathbf{c}^{(n)} \in \mathbb{R}^{D_{\text{clin}}}$ or both. For the models based on imaging features, we explore both single and multiple lesion approaches. For example, for the MLP models that require a fixed-size vector as input, we consider either averaging the imaging data from all lesions or using only the data from the largest lesion. For multiple lesion approaches, we compare a simple MIL strategy with graph-based fusion approaches. For the MLPs where we use both clinical and image data, the corresponding vectors are concatenated before being given as input to the model (early fusion).

4.3.4. Results

Regarding the comparison of **clinical-only vs image-only methods**, the results in Table 1 show a first observation, which is the performance gap between the two modalities. This is an expected result, as the clinical data considered have shown their efficacy for the task at hand [Carlier et al. (2024)]. On the other hand, although there is some evidence that radiomics supports the prediction for DLBCL with simpler prediction models [Carlier et al. (2024)], their combination

with more complex models has not yet been shown to be more effective than clinical data [Carlier et al. (2024)], and to the best of our knowledge no other groups have reported multi-lesion predictions on the DLBCL task.

Secondly, with respect to **image-only models**, the type of feature and the choice of segmentation have a variable impact on the performance depending on the prediction method. However, some trends can be observed. Regarding the choice of the feature set, the HeadNeck features (Featv1) perform best when averaged over all lesions segmented by majority voting (MLP with ***rad***=2), while other image-based methods perform less well on this feature set . On the other hand, reducing the number of features to keep only the first-order set from the HeadNeck list (Featv2-Featv3-Featv4) does not usually improve the results, except when relying only on the features of the largest lesion (MLP with ***rad***=1). Nevertheless, among the image-based methods dealing with the first-order feature sets (Featv2-Featv3-Featv4), the majority vote segmentation (Featv3) ranks higher in all cases except for the graph-based methods (GraphConv and GAT), which perform better without additional segmentation on the raw bounding boxes (Featv2). On the other hand, adding more features (Featv5) degrades the MLP performance compared to the HeadNeck set (Featv1), but improves the scores for methods that process data from each lesion individually (MIL and graph-based methods).

Regarding the performance of the **multimodal models**, the results in Table 1 show that the cross-attention model with projection of clinical data is statistically significantly better (p-value < 0.05) than all models (including unimodal ones) using the same type of features, except for the set with the highest number of imaging features (Featv5) where the model using only clinical data is the best, and the robust features on a SUV4 segmentation of lesions (Featv4)where the difference with the second best model is not significant. This highlights the importance of the cross-attention design to efficiently use both types of data, compared to simple fusion methods, which mostly degrade the results compared to the reference MLP relying only on clinical data. Regarding the Featv5 set, which includes the 650 radiomic features, we observe that its combination with clinical data leads to a significant drop in performance for simple fusion models (less important for our cross-attention models) compared to other feature sets; and this despite the predictive power of the clinical features, as shown by the MLP performance on clinical data only. We attribute this drop in performance to the difference in dimensionality between the two modalities, which may make this feature set more difficult to learn. Finally, we observe a significant difference between the performance of the cross-attention model with and without clinical data projection: this motivates our experiment in section 4.4 to explore a wide range of cross-attention model instantiations to find the most efficient one for this complex task.

Overall, the type of features extracted from the lesions has an impact on the performance, depending on the model considered. Therefore, in the following experiments, we will focus on the Featv3 set because of its good performance with the Cross-attention 2 model..

4.4. Cross-modal attention design

Next, we evaluate the impact of the configuration of the cross-attention module for multimodal fusion. For each configuration presented in section 3.3, we study the impact of several structural elements of our model, including:

- the learning mechanism for multi-lesion fusion between a GATv2, a GraphConv or a self-attention module;
- the presence of a skip connection to keep the raw image data accessible (or not) after the multi-lesion fusion stage;
- the projection of the clinical data with a linear layer vs the implementation of cross-attention directly on the raw clinical data;
- the use or not of a ReLU activation after the projection of the clinical data;
- and the number of cross-modal fusion blocks (1 or 2), taking into account that one block consists of a multi-lesion fusion module and a multi-modal cross-attention layer. In the case of two blocks, the output of the first cross-attention module is sent to another GAT with ReLU activation, then concatenated with the initial image information $\mathbf{Z}^{(n)}$ and fused again with the projected clinical data $\tilde{\mathbf{c}}^{(n)}$ by a second cross-attention module before the max pooling operation and the prediction.

We consider and evaluate variable combinations of these elements for each instantiation of the cross-attention module presented in section 3.3, and detail the results in Table 2, including the performance of the best model in terms of test AU-ROC and retained structure for each cross-attention design.

Regardind the design of the **cross-attention**, our results show that for the problem studied it is crucial to use the computed attention scores to modulate the clinical data, *i.e.* to include the clinical data as part of the 'Value' matrix as it

Cross-attention module	Best AU-ROC	GAT	GraphConv	Self-attention	Skip connection	No skip connection	1 multimodal block	2 multimodal blocks	Clinical FC layer	No clinical FC layer	ReLU	No ReLU
Img2Clin	$\underline{0.68 \pm 0.08}$	✓				✓		✓	✓		✓	
Mult2Clin	$\mathbf{0.68 \pm 0.08}$			✓	✓		✓		✓		✓	
Mult2Mult	0.67 ± 0.09			✓		✓		✓	✓		✓	
Clin2Img	0.61 ± 0.07	✓			✓		✓	-	✓			✓
Mult2Img	0.62 ± 0.08		✓			✓	✓		✓		✓	

Table 2
Best structure obtained for each cross-attention configuration and corresponding performance (- means impossible).

is the case for the Img2Clin, Mult2Clin and Mult2Mult modules. Instead, the poorer performance of the Clin2Img and Mult2Img models shows that we fail to extract meaningful information from using image data as the 'Value' matrix of the cross-attention module, even when using the clinical data. We attribute this behavior gap to the higher predictive power of the clinical data. This conclusion suggests that in a cross-attention module used for fusion, it is better to use the more predictive information as the key/value, regardless of the data used as query. Other choices, such as using self-attention in addition to cross-attention within this module (Img2Clin vs Mult2Clin and Clin2Img vs Mult2Img), have less impact on the results.

Furthermore, we observe that all models require the projection of clinical data to perform best. However, when this is excluded, the best configuration is very different depending on the use of cross-attention, highlighting the complexity of constructing an efficient model when working with real-world, multicentric, prognostic and multimodal data.

Finally, we observed a correlation between the number of weights of the configuration retained by the grid search for the different cross-attention instantiations, and their performance. More precisely, the best performing cross-attention configurations were those with the lowest number of parameters, which is consistent with the limited size of our dataset and the associated risk of overfitting.

Overall, the Mult2Clin cross-attention module outperforms the others in its ability to predict 2-year PFS: we therefore retain the Mult2Clin design with its best configuration for the next experiments.

Ablation study To show the influence of each element of the structure of our best model, we perform an ablation study, presented in Table 3. This study highlights the relevance of each structural choice, with the model obtained from our previous experiments performing better than any other on the Featv3 feature set. From this experiment, we can also identify the elements in our model structure that are the most important among those evaluated. The most important in this case is the skip connection, followed by the choice of the multi-lesion fusion block and the clinical projection. However, relying on one or two blocks for multimodal fusion gives similar results.

Translation to another feature set To test the robustness of our best Mult2Clin model, whose structure was optimized on the Featv3 set of imaging features (first-order features on a majority vote segmentation), we also train it on the Featv5 set (full set of image radiomics on raw segmentation), which was the most difficult to learn for models based on both clinical and imaging data (cf Table 1). With an AU-ROC of 0.69 ± 0.06, there is a slight increase in performance compared to all baseline models in Table 1 (with a maximum performance of 0.67 ± 0.08). This is especially true compared to the previous best baseline using the Featv5 set of image features (0.64 ± 0.10), and even compared to the same best proposed model learning on the Featv3 set of features (0.68 ± 0.08). Even if this gain is not consequent, this experiment proves the ability of our model to take advantage of both modalities even when there is an important difference in dimensionality between the two, which is especially helpful in cases when we do not know which features of a modality are the most important and thus cannot restrict manually the set to rely on. For the sake of clarity, we will in the following restrict the analysis to the best performing Mult2Clin model (with the optimized structure) trained on the full set of lesion features (Featv5).

Best AU-ROC	GAT	GraphConv	Self-attention	Skip connection	No skip connection	1 multimodal block	2 multimodal blocks	Clinical FC layer	No clinical FC layer	ReLU	No ReLU
0.64 ± 0.09			✓		✓	✓		✓		✓	
0.65 ± 0.09		✓		✓		✓		✓		✓	
0.65 ± 0.08			✓	✓		✓		✓			✓
0.66 ± 0.09	✓			✓		✓		✓		✓	
0.67 ± 0.09			✓	✓		✓			✓	-	-
0.68 ± 0.09			✓	✓			✓	✓		✓	
$\mathbf{0.68 \pm 0.08}$			✓	✓		✓		✓		✓	

Table 3
Ablation study over the components of our best model (with Mult2Clin cross-attention module). Circled tickmarks highlight the modified component with respect to the best reference configuration (bottom) .

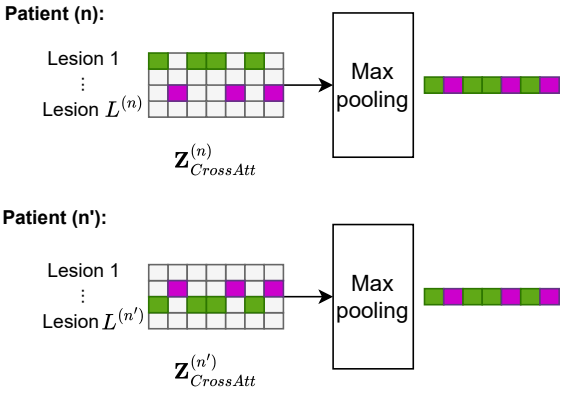
4.5. Model interpretability

In this section, we show how our model is designed to enhance interpretability, potentially helping clinicians to understand its predictions. To this end, we will analyze the learned attention patterns and illustrate how these provide valuable insights at both the lesion and feature level.

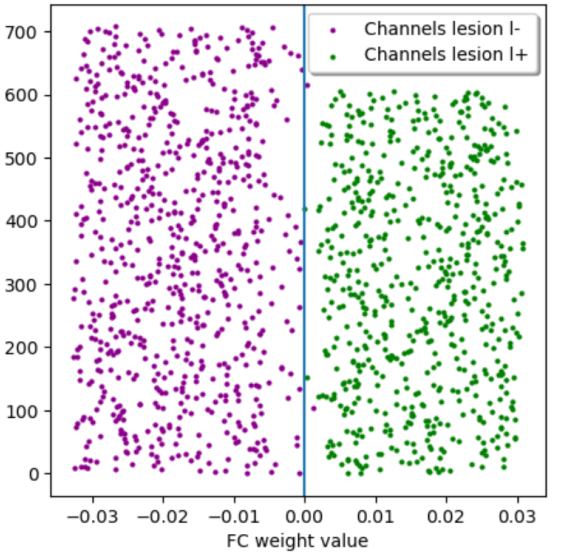
The first element supporting the interpretability of our model comes from the combination of the multi-lesion graph with the proposed cross-attention module. With this combination and the Mult2Clin cross-attention configuration, the computed crossmodal attentions can be interpreted as weights that modulate the clinical data according to the (graph-updated) image features of each lesion. A second design element important for interpretability, which we will focus on below, is the max pooling operation applied to the cross-attention output. While the primary role of this operation is to recover a fixed-sized embedding vector from a variable number of lesions as input, we observe two interesting mechanisms: (1) the identification of exactly two meaningful lesions for each patient; (2) one of these lesions is positively associated with 2-year PFS, while the other is negatively associated with it.

Two lesions retained by the Max Pooling In our design, the max pooling operation is applied to the columns of the cross-attention output $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ (see Fig 1). In practice, the pooling operation only retains features from two rows/lesions of $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ (see Fig 5a). This selection results from our cross-attention design and, more specifically, from the projection matrix $\mathbf{W}_V$ applied to the clinical data (cf. Fig 2). Indeed, given that the clinical vector $\tilde{\mathbf{c}}^{(n)}$ has a single dimension, the effect of the $\mathbf{W}_V$ projection is to replicate $\tilde{\mathbf{c}}^{(n)}$ on the columns of the 'Value' matrix $\tilde{\mathbf{V}}$, each time weighted by a different learned scalar, *i.e.* $\tilde{\mathbf{V}} = \tilde{\mathbf{c}}^{(n)}\mathbf{W}_V = [\beta_1\tilde{\mathbf{c}}^{(n)}, \beta_2\tilde{\mathbf{c}}^{(n)}, \dots, \beta_{D_{concat}}\tilde{\mathbf{c}}^{(n)}]$. Since the $\beta_.$ weights can be either positive or negative, the highest values in each $\tilde{\mathbf{V}}$ column correspond to either the highest or lowest values in $\tilde{\mathbf{c}}^{(n)}$. Also, since the elements of the attention matrix are all positive because they pass through a softmax function, they do not change the sign of the $\tilde{\mathbf{V}}$'s columns when multiplied. Therefore, when the max-pooling operation is reached, the elements of $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ still carry the signs of $\beta_.$ Let $\mathbf{a}_i$ denote the cross-attention row corresponding to lesion i, then the elements of $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ are all scaled versions of $\mathbf{a}_i\tilde{\mathbf{c}}^{(n)}$. Since the elements of a $\mathbf{Z}^{(n)}_{\text{CrossAtt}}$ column share the $\beta_.$'s scale and sign, the only remaining difference that matters in the max-pooling operation in that column is the attention that each lesion $\mathbf{a}_i$ gives to the clinical data $\tilde{\mathbf{c}}^{(n)}$. Therefore, the two lesions that "agree" most with the positive or negative versions of $\tilde{\mathbf{c}}^{(n)}$ get higher values and are retained by the max pooling. Furthermore, since the $\beta_.$ are constant across patients, the two sets of columns where the value of the first or second lesion is selected form two distinct groups that are strictly the same for all patients (cf Fig. 5a).

Association between the selected lesions and the 2-year PFS In fact, our learning architecture allows us to understand not only which lesions the model relied on, but also how these lesions influenced the prediction. More specifically, by analyzing the behavior of the fully connected layer that follows the max pooling operation and outputs



(a) Behavior of the max pooling operation



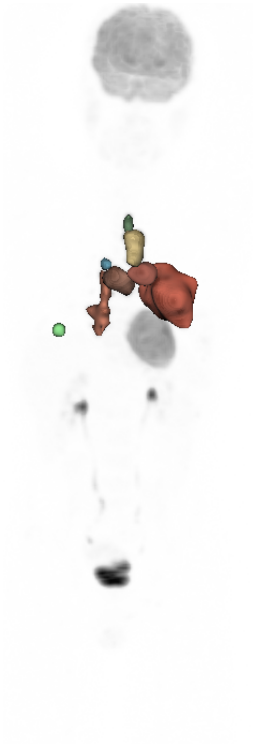
(b) Distribution of final fully connected layer weights as a function of the lesion from which the max pooling operation selected the corresponding feature (the y-axis: weight identifier).

Figure 5: Explanation of the max pooling role for interpretability.

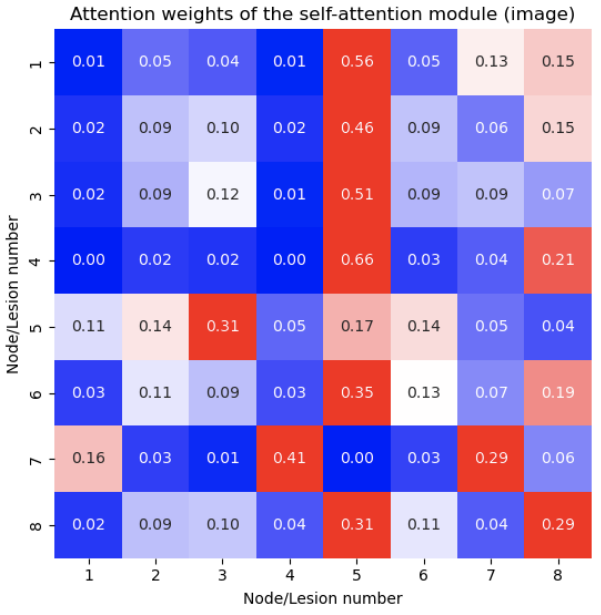
the 2-year PFS prediction, we observe that one of the two selected lesions, called l_+ from here on, is positively associated with the predicted 2-year PFS, while the other l_- is negatively associated with it. Indeed, this fully connected layer learns weights that are consistent with the lesions retained by max pooling: the weights related to the channels retrieved from the feature vector of a given lesion (green or purple in Fig. 5b) are almost exclusively of the same sign (except for 2 weights over 1316), indicating a very strong and complementary influence of each of these lesions on the 2-year PFS prediction.

Relevance of the selected lesions To experimentally demonstrate that our model selects lesions that are relevant for predicting 2-year PFS, we investigate how it performs when we prevent it from using these lesions. To do this, we iteratively remove pairs of lesions from the lesion graphs (from 0 to 16 lesions). At each iteration, the removed pairs are the two lesions selected by the model, plus those already removed in previous iterations. For each number of lesions removed, we examine the performance of the model using AU-ROC only on those patients for whom we still retain at least 80% of their lesions. For the sake of statistical significance, especially when we remove many lesions and thus have fewer patients to test on, we perform this test on our full dataset. For a given number lesions removed, we compare the performance of the model on the same subset of patients (1) when we remove the pairs of lesions selected with the max pooling (blue in Fig. 7), (2) when we remove randomly selected pairs of lesions (orange in Fig. 7) to see if the lesions retained by max pooling are particularly significant, and (3) without any lesions removed (green in Fig. 7) to take into account the effect of changing the subset of patients we are testing on. We observe that for almost any number of lesions removed, the performance of the model does not drop much when we randomly remove lesions compared to keeping all lesions. However, it drops much more when we remove the lesions selected by the max pooling operation. This highlights the prognostic importance of the lesions that the model has learnt to include in its prediction.

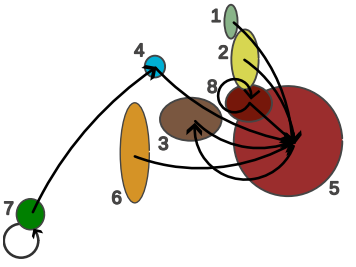
Lesion to lesion Attentions While the two previous mechanisms make it possible to identify the two complementary salient lesions that are mainly used for prediction, the self-attention matrix indicates in a less straightforward way, secondary lesions that are also taken into account for the prediction. They correspond to those to which l_+ and l_- pay attention in the self-attention module (which replaces the GAT in our model following the structural study) (see Fig. 6b). Indeed, the activated output of the self-attention module is concatenated with the initial feature vector of the lesion before the cross-attention module. Thus, the lesions l_+ and l_- used for the prediction contain image data from



(a) Maximum Intensity projection of a patient and his segmented lesions. A total of 8 lesions were segmented.



(b) Lesion self-attention matrix for this patient, where $l\_$ is lesion 3, l_+ is lesion 7, and both the label and the prediction of this patient's 2-year PFS 1 (meaning there was a progression before 2 years after starting treatment).



(c) Schematic representation of the patient's lesions (color and size as in 6a) and the attention they pay to each other (only the most important attentions are represented).

Figure 6: Example of interpretation of the Lesion self-attention matrix.

both their own feature vector and the feature vectors of other lesions after the self-attention module. Therefore, these secondary lesions are also associated with the prediction of 2-year PFS with the same type of influence on it (positive or negative) as the lesion (l_+ or $l\_$) to which it pays attention in the self-attention module. This ability to identify lesions associated with either a positive or negative 2-year PFS could, following appropriate external validation, provide clinicians with valuable diagnostic insight and address the current lack of information on how individual lesions may influence treatment response.

Interpretability at the feature level Another advantage of our model is its ability to extract information about the most important features for predicting 2-year PFS. First, there is a positive bias towards some clinical features that are more strongly represented in the vector of projected clinical data $\mathbf{c}'^{(n)}$, *i.e.* the sum of the weights associated with these values in the fully connected layer projecting the clinical data is higher than the sum of the weights associated with other features (see Fig. 8). More specifically, we observe three salient features: the Ann Arbor stage of the disease, the number of extranodal sites and the binary value indicating whether the patient underwent salvage therapy. Regarding the image-based information, we observe that some features of the lesions l_+ and $l\_$ (among the image features of the Featv5 set) have statistically different distributions, either one lesion against the other or each of them against the distribution of the features of all lesions. This could mean that these features are important for identifying the lesions that are essential for predicting 2-year PFS. This behavior is observed in particular for a subset of 9 textural features presented in Fig. 1 in Supp. material A.

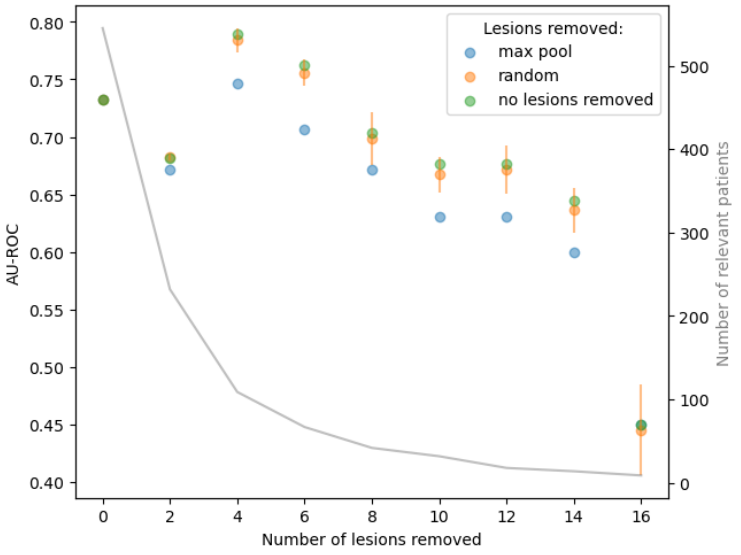


Figure 7: AU-ROC performance of the best model in patients with progressively removed lesions, depending on how these removed lesions are selected (randomly or lesions retained by the max pooling operation).

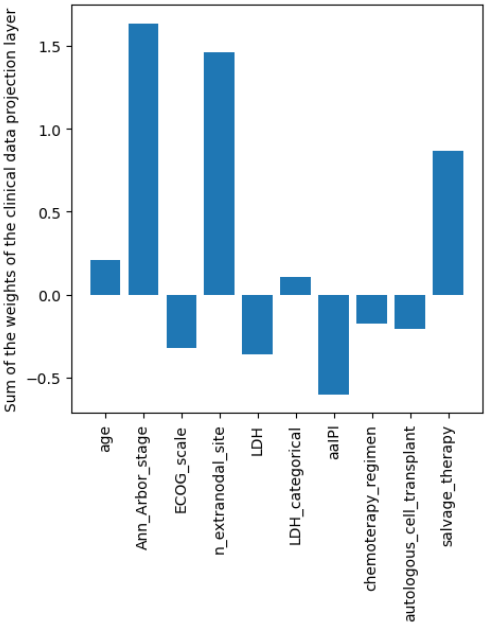


Figure 8: Sum of the clinical Fully Connected layer weights, for each clinical characteristic.

5. Conclusion and perspectives

We address the problem of identifying high-risk DLBCL patients by automatically predicting their 2-year PFS. To this end, we propose a model that is able to exploit imaging information from a variable number of lesions thanks to the creation of a lesion graph. We then fuse this information with tabular clinical data through a cross-attention module. Several studies of this graph-based cross-attention model, exploring both the input features to be used and

the structure of the model, have highlighted the complexity of constructing a model that performs well on complex, real-world, multicentric, unbalanced and multimodal data. Our experiments have shown the interest of using a large number of image features in the case of multimodal fusion, but only if special attention is paid to defining the structure of the model so that this dimensionality does not prevent the exploitation of the features of other modalities. Regarding the choice of cross-attention module inputs, our experiments showed that using the modality with the higher predictive power as the 'Value' matrix outperformed the other strategies for multimodal fusion. Finally, the strength of our model in terms of interpretability was highlighted: we showed how its structure allows clinicians to better understand the prediction and gain clinical insight into which clinical features and lesions are most relevant for predicting 2-year PFS in DLBCL.

However, there are several limitations to our work. First, this framework should be applied to other multimodal/multi-lesion datasets to validate its generalizability. Furthermore, it should be noted that our performance is not yet high enough for clinical practice, and therefore the results regarding the relevance of the features considered should be interpreted with caution.

Some perspectives for this work include an extended study of the graph structure considered for the lesion graph, which could help to take into account dispersion of the lesions more efficiently. Another interesting development would be to consider cost functions specifically adapted to survival tasks, allowing us to obtain a more accurate representation of patients' risk in the short and long term, and to take advantage of patients censored before two years. Finally, we would like to validate the robustness of our best model and its ability to generalize to other pathologies.

6. Acknowledgements

This work has been funded by the AIby4 project (Centrale Nantes-Project ANR-20-THIA-0011), INCa-DGOS-INSERM-ITMO Cancer 18011 (SIRIC ILIAD) with the support from the Pays de la Loire region (GCS IRECAN 220729), the European Regional Development Fund (FEDER), the Pays de la Loire region on the Connect Talent MILCOM programme and Nantes Métropole (Conv. 2017-10470).

CRediT authorship contribution statement

Oriane Thiery: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Mira Rizkallah:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Clément Bailly:** Data curation, Writing – review & editing. **Caroline Bodet-Milin:** Data curation, Writing – review & editing. **Emmanuel Itti:** Data curation, Writing – review & editing. **René-Olivier Casasnovas:** Data curation, Writing – review & editing. **Steven Le Gouill:** Data curation, Writing – review & editing. **Thomas Carlier:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Diana Mateus:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

References

Andrade-Miranda, G., Jaouen, V., Tankyevych, O., Cheze Le Rest, C., Visvikis, D., Conze, P.H., 2023. Multi-modal medical transformers: A meta-analysis for medical image segmentation in oncology. Computerized Medical Imaging and Graphics 110, 102308. doi:10.1016/j.compmedimag.2023.102308.

Baltrušaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 423–443.

Brody, S., Alon, U., Yahav, E., 2022. How attentive are graph attention networks?, in: International Conference on Learning Representations.

Carlier, T., Frécon, G., Mateus, D., Rizkallah, M., Kraeber-Bodéré, F., Kanoun, S., Blanc-Durand, P., Itti, E., Le Gouill, S., Casasnovas, R.O., Bodet-Milin, C., Bailly, C., 2024. Prognostic value of 18 F-FDG PET radiomics features at baseline in PET-guided consolidation strategy in Diffuse Large B-Cell Lymphoma: A machine-learning analysis from the GAINED study. Journal of Nuclear Medicine 65, 156–162. doi:10.2967/jnumed.123.265872.

Cottereau, A.S., Nioche, C., Dirand, A.S., Clerc, J., Morschhauser, F., Casasnovas, O., Meignan, M., Buvat, I., 2019. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. The Journal of Nuclear Medicine 61, 40 – 45.

Durand, T., Thome, N., Cord, M., 2015. MANTRA: Minimum maximum latent structural SVM for image classification and ranking. IEEE International Conference on Computer Vision , 2713–2721.

Durand, T., Thome, N., Cord, M., 2019. Exploiting negative evidence for deep latent structured models. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 337–351.

Eertink, J.J., Zwezerijnen, G.J.C., Wiegers, S.E., Pieplenbosch, S., Chamuleau, M.E.D., Lugtenburg, P.J., de Jong, D., Ylstra, B., Mendeville, M., Dührsen, U., Hanoun, C., Hüttmann, A., Richter, J., Klapper, W., Jauw, Y.W.S., Hoekstra, O.S., de Vet, H.C.W., Boellaard, R., Zijlstra, J.M.,

2022. Baseline radiomics features and MYC rearrangement status predict progression in aggressive B-cell lymphoma. Blood Advances 7, 214 – 223.

Gao, X., Shi, F., Shen, D., Liu, M., 2023. Multimodal transformer network for incomplete image generation and diagnosis of alzheimer's disease. Computerized Medical Imaging and Graphics 110, 102303. URL: https://www.sciencedirect.com/science/article/pii/S0895611123001210, doi:https://doi.org/10.1016/j.compmedimag.2023.102303.

Golovanevsky, M., Eickhoff, C., Singh, R., 2022. Multimodal attention-based deep learning for Alzheimer's disease diagnosis. Journal of the American Medical Informatics Association .

van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R., Fillion-Robin, J.C., Pieper, S.D., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer research 77 (21), e104–e107.

Hager, P., Menten, M.J., Rueckert, D., 2023. Best of both worlds: Multimodal contrastive learning with tabular and imaging data, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23924–23935.

Huang, W., 2023. Multimodal contrastive learning and tabular attention for automated Alzheimer's disease prediction, in: IEEE/CVF International Conference on Computer Vision Workshops, pp. 2465–2474. doi:10.1109/ICCVW60793.2023.00261.

Jiang, C., Li, A., Teng, Y., Huang, X., Ding, C., Chen, J., Xu, J., Zhou, Z., 2022. Optimal PET-based radiomic signature construction based on the cross-combination method for predicting the survival of patients with diffuse large B-cell lymphoma. European Journal of Nuclear Medicine and Molecular Imaging 49, 2902–2916.

Kazmierski, M., Haibe-Kains, B., 2021. Lymph node graph neural networks for cancer metastasis prediction URL: http://arxiv.org/abs/2106.01711.

Le Gouill, S., Ghesquières, H., Oberic, L., Morschhauser, F., Tilly, H., Ribrag, V., Lamy, T., Thieblemont, C., Maisonneuve, H., Gressin, R., Bouhabdallah, K., Haioun, C., Damaj, G., Fornecker, L., Bouhabdallah, R., Feugier, P., Sibon, D., Cartron, G., Bonnet, C., André, M., Chartier, L., Ruminy, P., Kraeber-Bodéré, F., Bodet-Milin, C., Berriolo-Riedinger, A., Brière, J., Jais, J.P., Molina, T.J., Itti, E., Casasnovas, R.O., 2021. Obinutuzumab vs rituximab for advanced DLBCL: a PET-guided and randomized phase 3 study by LYSA. Blood 137, 2307–2320.

Liu, P., Zhang, M., Gao, X., Li, B., Zheng, G., 2022. Joint lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images via multitask convolutional neural networks. IEEE Access 10, 81612–81623. doi:10.1109/ACCESS.2022.3195906.

Liu, S., Zhang, B., Fang, R., Rueckert, D., Zimmer, V.A., 2023. Dynamic graph neural representation based multi-modal fusion model for cognitive outcome prediction in stroke cases, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, p. 338–347. doi:10.1007/978-3-031-43993-3_33.

Lv, W., Zhou, Z., Peng, J., Peng, L., Lin, G., Wu, H., Xu, H., Lu, L., 2023. Functional-structural sub-region graph convolutional network (FSGCN): Application to the prognosis of head and neck cancer with PET/CT imaging. Computer Methods and Programs in Biomedicine 230, 107341.

Morris, C., Ritzert, M., Fey, M., Hamilton, W., Lenssen, J., Rattan, G., Grohe, M., 2019. Weisfeiler and Leman go neural: Higher-order graph neural networks. AAAI Conference on Artificial Intelligence 33, 4602–4609.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C., 2021. Attention bottlenecks for multimodal fusion. Neural Information Processing Systems .

Pölsterl, S., Wolf, T.N., Wachinger, C., 2021. Combining 3D image and tabular data via the dynamic affine feature map transform, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 688–698.

Prabhakar, C., Li, H.B., Paetzold, J.C., Loehr, T., Niu, C., Mühlau, M., Rueckert, D., Wiestler, B., Menze, B., 2023. Self-pruning graph neural network for predicting inflammatory disease activity in multiple sclerosis from brain MR images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 226–236.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning.

Rist, L., Taubmann, O., Mühlberg, A., Denzinger, F., Thamm, F., Sühling, M., Nörenberg, D., Holch, J.W., Maurus, S., Gebauer, L., Huber, T., Maier, A., 2023. Spatial lesion graphs: Analyzing liver metastases with geometric deep learning for cancer survival regression, in: IEEE International Symposium on Biomedical Imaging, pp. 1–5. doi:10.1109/ISBI53787.2023.10230367.

Sudlow, C.L.M., Gallacher, J.E., Allen, N.E., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M.J., Liu, B., Matthews, P.M., Ong, G.J., Pell, J.P., Silman, A.J., Young, A., Sprosen, T., Peakman, T.C., Collins, R., 2015. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine 12.

Susanibar-Adaniya, S., Barta, S.K., 2021. 2021 update on Diffuse large B cell lymphoma: A review of current data and potential applications on risk stratification and management. American journal of hematology 96, 617–629.

Thiery, O., Rizkallah, M., Bailly, C., Bodet-Milin, C., Itti, E., Casasnovas, R.O., Le Gouill, S., Carlier, T., Mateus, D., 2023. Graph-based multimodal multi-lesion DLBCL treatment response prediction from PET images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 103–112.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Neural Information Processing Systems.

Wang, G., Fan, F., Shi, S., An, S., Cao, X., Ge, W., Yu, F., Wang, Q., Han, X., Tan, S., Tan, Y., Wang, Z., 2024. Multi modality fusion transformer with spatio-temporal feature aggregation module for psychiatric disorder diagnosis. Computerized Medical Imaging and Graphics 114, 102368. doi:https://doi.org/10.1016/j.compmedimag.2024.102368.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. MedCLIP: Contrastive learning from unpaired medical images and text, in: Conference on Empirical Methods in Natural Language Processing.

Yousefirizi, F., Gowdy, C., Klyuzhin, I.S., Sabouri, M., Tonseth, P., Hayden, A.R., Wilson, D., Sehn, L.H., Scott, D.W., Steidl, C., Savage, K.J., Uribe, C.F., Rahmim, A., 2024. Evaluating outcome prediction via baseline, end-of-treatment, and delta radiomics on PET-CT images of primary mediastinal large B-cell lymphoma. Cancers 16.

Yuan, C., Shi, Q., Huang, X., Wang, L., He, Y., Li, B., Zhao, W.L., Qian, D., 2022. Multimodal deep learning model on interim 18F-FDG PET/CT for predicting primary treatment failure in diffuse large B-cell lymphoma. European Radiology 33, 77–88.

Zhou, T., Cheng, Q., Lu, H., Li, Q., Zhang, X., Qiu, S., 2023. Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine 160, 106959. doi:10.1016/j.compbiomed.2023.106959.

Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H.J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G.J.R., Davatzikos, C., Depeursinge, A., Desseroit, M.C., Dinapoli, N., Dinh, C.V., Echegaray, S., Naqa, I.M.E., Fedorov, A., Gatta, R., Gillies, R.J., Goh, V.J., Götz, M., Guckenberger, M., Ha, S.M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R.T.H., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E.A.G., Rahmim, A., Rao, A., Scherer, J., Siddique, M.M., Sijtsema, N.M., Fernandez, J.S., Spezi, E., Steenbakkers, R.J.H.M., Tanadini-Lang, S., Thorwarth, D., Troost, E.G.C., Upadhaya, T., Valentini, V., van Dijk, L.V., van Griethuysen, J.J.M., van Velden, F.H.P., Whybra, P., Richter, C., Löck, S., 2016. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology , 191145.