Lesion graph neural networks for 2-year progression free survival classification of Diffuse Large B-Cell Lymphoma patients

Aswathi Aswathi, Mira Rizkallah, Gauthier Frecon, Clément Bailly, Caroline M Bodet-Milin, Olivier Casasnovas, Steven Le Gouill, Françoise Kraeber-Bodéré, Thomas Carlier, Diana Mateus

LESION GRAPH NEURAL NETWORKS FOR 2-YEAR PROGRESSION FREE SURVIVAL CLASSIFICATION OF DIFFUSE LARGE B-CELL LYMPHOMA PATIENTS

Aswathi[†]*, Mira Rizkallah*[†]*, Gauthier Frecon*[⋆]*, Clément Bailly*[⋆]*,Caroline Bodet-Milin*[⋆]*,*
Olivier Casasnovas[⋆]*,Steven Le Gouill*[⋆]*,Françoise Kraeber-Bodéré*[⋆]*,*
Thomas Carlier[⋆]*, Diana Mateus*[†]

[†] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[⋆] CHU de Nantes, France

ABSTRACT

Survival analysis of DLBCL patients requires the interpretation of PET images characterised by multiple small lesions. Current machine-learning approaches addressing similar problems consider as input the cropped image of a single lesion or the whole volume. In this paper, we incorporate the information of all lesions by modeling their joint survival analysis with a graph learning approach. We propose a compact graph representation of the segmented lesions enriched by radiomics features and edge weights. The representation is fed to a graph attention network to predict the 2-year Progression-Free Survival of a DLBCL patient, formalised as a graph classification problem. Experimental results on a clinical prospective database with 583 patients show that our method improves over three baseline fusion approaches.

***Index Terms*—** DLBCL, Survival Analysis, Graph Attention Networks, Multiple lesion fusion

1. INTRODUCTION

Diffuse Large B-cell Lymphoma (DLBCL) is a haematological disease characterised by the clonal proliferation of lymphocytes. Following current guidelines, the diagnosis and follow-up of DLBCL patients involves the interpretation of full-body 3-D Positron Emission Tomography (PET) images by a nuclear physician. DLBCL PET images are characterised by multiple small lesions (see Figure 1. (a)-(b)). Recently, there has been increased interest in machine learning approaches predicting the survival of DLBCL patients from their PET image at diagnosis. In particular, nuclear physicians target the 2-years progression-free survival (PFS) as a risk indicator. In this paper, we propose a computer-aided decision approach to predict the 2-year PFS based on a graph neural network, with the final aim to early identify patients with a high-risk profile.

In the context of survival analysis from PET images, most existing machine learning methods fall among three categories: combining radiomics features extracted from the PET volumes with classical machine learning methods [1] or deploying a convolutional neural network (CNN) whose input is either a single lesion ([2], [3]), or the full volume [4]. However, resuming a full-body PET image to a single ROI may not fully represent the patient's disease. At the same time, in full-image predictions, the lesions' information is diluted within the large background and thus may be difficult to capture. As an intermediate alternative solution, we propose to model the joint survival analysis of multiple lesions relying on graph-learning approaches. Although graph-based approaches have been explored for histopathological image survival analysis[5], their use on PET images is only very recent[6]. In fact, PET images from DBLCL patients pose several challenges: the lesions are small and sparsely distributed over the body; patients may have from a single to a large number of lesions; and healthy organs like the brain or bladder may have similar intensities to lesions.

To effectively represent images of multiple and sparsely distributed small lesions, it is essential to characterise the individual lesion's heterogeneity while reducing the background's influence. The question we address here is how to fuse image information from multiple lesions to make better predictions. Indeed, the varying number of lesions makes late fusion through concatenation inadequate. As we later show, early (e.g. averaging of the lesions' features) or late fusion (e.g. through multiple instance learning) are ineffective. Instead, we propose i) building a graph representation on top of the lesions' segmentation ii) characterising the image region covering each lesion with conventional and radiomics

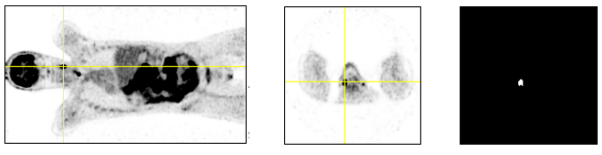

Fig. 1: Cross-sectional views of a 3D PET image of a DLBCL patient (a) and (b). (c) binary mask of one of the lesions.

features, and finally, iii) learning to fuse their information and make predictions with attentional graph. For the choice of GNN, we opt for attention layers to model the relative importance of each node, while optimizing the prediction. In particular, we rely on the recent dynamic Graph Attentional Network (GATv2) [7], which considers different lesions may rank differently the importance of their neighbouring nodes.

From an application perspective, this is the first DBLCL 2-year PFS prediction model to consider multiple lesions explicitly. While some existing methods characterise the lesion spread through dissemination features [8] we tackle the problem from a "learning to fuse" perspective through an effective graph attention model.

We perform experiments on the prospective GAINED phase 3 trial (NCT 01659099) which enrolled 583 newly diagnosed and untreated DLBCL patients. A comparison to three baselines, shows our approach's feasibility and the pertinence of the graph attentional fusion for DBCL prognosis.

2. PROBLEM STATEMENT

We formulate the prediction of the DLBCL Progression-Free Survival (PFS) as a binary classification problem: estimate whether or not an event (death or progression) has occurred between 0 and 2 years after the beginning of the treatment (2-years PFS). We consider as input a PET volume taken at diagnosis and a prior segmentation of the lesions. Our method compactly models the relevant information from the image with a graph connecting the regions surrounding the lesions. With this representation in mind and a supervised learning approach, the problem becomes that of graph classification which we address with a Graph Neural Network (GNN).

3. PROPOSED METHOD

Starting from a dataset of PET images from N patients and their associated 2-year PFS outcome, i.e. $\{I_i, y_i\}_{i=1}^{N}$, we propose a framework that first, builds a graph $\mathcal{G}_i$ from every image, and then, trains the weights θ of a GNN $f_\theta(\cdot)$ to make predictions for a new patient k: $\hat{y}_k = f_\theta(\mathcal{G}_k)$. Fig. 2 illustrates the proposed PFS classification framework with two stages: graph construction and survival learning with a GNN.

3.1. Graph construction

Starting from the image and the segmented lesions of a patient, we build a fully-connected patient-level graph $\mathcal{G}_i = \{\mathcal{N}_i, \mathcal{E}_i\}$, where $\mathcal{N}_i$ represents the set of nodes corresponding to the different lesions of patient i. Each node $m \in \mathcal{N}_i$, identified with its index m over the M_i lesions, has an associated feature vector $\mathbf{x}_{i,m}$. Edges $e_{m,n}$ are drawn between every pair of nodes $m, n \in \mathcal{N}_i$, including a self loop. Finally, weights $w_{m,n}$ are computed for every edge, built based on spatial proximity and feature similarity.

3.1.1. Node features extraction

Feature extraction is performed to compactly characterise the information of the raw 3D image region around lesion $l_{i,m}$, and form the *feature vector* $\mathbf{x}_{i,m}$. In this work, we focus on two types of features: *classical* and *radiomics* features. After the extraction step, we pile the vectors of the different lesions as rows of the node features matrix $\mathbf{X}_i \in \mathbb{R}^{M_i \times D_{in}}$, where D_{in} is the dimensionality of the feature vector of each node ($D_{in} = 11$ in our case).

Classical features are conventional quantitative measurements extracted from the segmented lesions, describing the intensity distribution of individual voxels without taking into account their spatial relationships. In this work, we extract three classical features: the standard uptake value of the maximum intensity voxel within the lesion (SUV_{max}), the *Metabolic Tumour Volume* (MTV) of the lesion and the *Total Lesion Glycolysis* (TLG) from each of the segmented lesions. MTV refers to the volume of the metabolically active region of the lesion and TLG is the product of the metabolic volume with the mean standard uptake value of the lesion.

Radiomics features While classical features provide limited tumour characteristics, radiomics features based on intra-tumoural heterogeneity can more comprehensively assess the 3D landscape of the lesion[9]. For each lesion $l_{i,m}$, we extract *first-order*, *second-order* and *shape* radiomics features. First-order features do not take into account the spatial relations between the voxels in the lesions. Second-order textural features take into account the inter-relationships among voxels and are extracted from different matrices such as the *Gray Level Co-occurrence Matrix* (GLCM) and the *Gray-Level Run-Length Matrix* (GLRLM) [10]. The shape based features are relevant for the shape characteristics of the lesions. Inspired by [11], we select the following set of radiomics features:

Radiomics features	
Shape	Sphericity
First order	Mean
	Standard deviation
	Entropy
Second order	Contrast
	Correlation
	Inverse difference normalized
	Joint energy

3.1.2. Edge weights

To define the edge weights, we consider the product of two negative exponential terms: the first one computed on the node feature distance, and the second on the Euclidean distance between the 3D centroid coordinates of each lesion:

$$w_{m,n} = e^{-\frac{||\mathbf{c}_{i,m} - \mathbf{c}_{i,n}||}{a\sigma_1^2}} \times e^{-\frac{||\mathbf{x}_{i,m} - \mathbf{x}_{i,n}||}{a\sigma_2^2}}, \qquad (1)$$

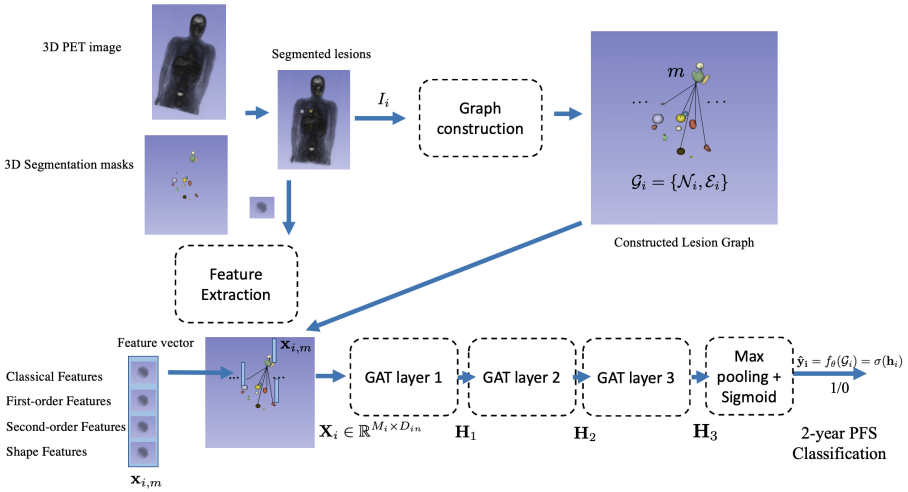

Fig. 2: Overview of the proposed 2-year PFS graph classification framework - The input are 3D PET images and segmented lesions. Radiomics and other features are then extracted from each lesion to be the nodes attributes of a fully connected graphs. The graph is fed as input to a graph attentional neural network whose hyper-parameters are optimized with K-cross validation. At the output of two Graph attention layers, a global max pooling and an activation function are applied to obtain the 2-year PFS. During inference, the GNN receives a graph and makes a 2-year PFS prediciton.

where $\mathbf{C}_{i,m}$, $\mathbf{C}_{i,n}$ are the 3D centroid coordinates vectors of lesions $l_{i,m}$ and $l_{i,n}$ respectively, and $\mathbf{x}_{i,m}$, $\mathbf{x}_{i,n}$ their corresponding feature vectors. σ_1 is the population-level standard deviation of distances between the centroid coordinates, and σ_2 is the population-level standard deviation of distances between the feature vectors. a is a hyper-parameter that can be tuned to select the best edge weights relevant to the PFS classification task. A sample graph constructed from the PET image of a patient is represented in Figure 2, where the individual lesions are the nodes, and the size of each node is equal to the tumour volume.

3.2. Graph neural network

Once the extracted features are assigned to the nodes and the weights to the edges, the graph is built. We then aim at exploiting the relationships between lesions along with their features to classify the 3D PET images and provide an output class (event or no-event). To do so, we rely on a GAT Network [12] fed with the graph $\mathcal{G}_i$ as shown in Fig. 2.

From the architecture standpoint, the network is composed of 3 graph attentional convolutional layers. Each GAT_p layer (with $p \in \{1, 2, 3\}$) is composed of a graph dynamic attention operator [7] followed by an activation function (ReLU) for both the first and second layer. The third has only the graph attention operator. The final layer consists of a global max pooling operation and an activation function (a sigmoid) to get a value equal to 0 or 1.

Each graph attention layer gets as input the feature matrix from the previous layer $\mathbf{H}_{p-1}$ ($\mathbf{X}_i$ for the first layer) and provides as output the transformed matrix $\mathbf{H}_p$. The final output node feature matrix $\mathbf{H}_{out} = \mathbf{H}_3 \in \mathbb{R}^{N \times D_{out}}$ contains the node representation encoding both graph structural properties and node features. In our case, we fix D_{out} to 1 for binary event probability. The feature matrix $\mathbf{H}_{out}$ is aggregated with a global maximum graph pooling layer, retaining the maximum feature values across the nodes to obtain a fixed-size vector $\mathbf{h}_i \in \mathbb{R}^{D_{out}}$. A sigmoid activation $\sigma(\cdot)$ is then applied on $\mathbf{h}_i$ to obtain the probability of an event. The prediction for the entire PET image is computed as $\hat{\mathbf{y}}_\mathbf{i} = f_\theta(\mathcal{G}_i) = \sigma(\mathbf{h}_i)$ where the parameters θ of the GNN are trained by minimizing an image-wise weighted cross-entropy loss.

4. EXPERIMENTAL VALIDATION

4.1. Dataset description and preprocessing

The dataset comes from the prospective multicentric GAINED phase III study including imaging data of 583 patients diagnosed with DLBCL. The patients are divided into *two classes* based on the PFS two-year categorical value: positive class if an event has occurred for the patient (PFS=1) and negative class if there is no event occurrence after 2 years (PFS=0). 494 patients belong to the negative category, whereas only 89 patients belong to the positive category.

The imaging data provided for each patient includes 3D 18-FDG PET scans, acquired with different machines during diagnosis. First, volume resampling is performed to obtain a fixed voxel size of $2 \times 2 \times 2mm^3$. The PET volumes are then converted to the *Standardized Uptake Value* (SUV). Along with the images, we have access to rough 3D segmentation masks of the lesions performed by expert nuclear

	MLP	MIL	GCN	GAT (ours)
AUC	0.58 ±0.09	0.56 ±0.09	0.59 ±0.06	**0.63** ±0.06
Bal. Acc	0.51 ±0.06	0.58 ±0.08	0.59 ±0.06	**0.60** ±0.07
Sensitivity	0.66 ±0.01	**0.72** ±0.01	0.67 ±0.03	0.61 ±0.02
Specificity	0.36 ±0.12	0.45 ±0.17	0.52 ±0.13	**0.59** ±0.15
Micro F1	0.51 ±0.06	0.58 ±0.09	0.58 ±0.07	**0.60** ±0.07
Macro F1	0.50 ±0.07	0.57 ±0.1	0.58 ±0.07	**0.60** ±0.07
Weig. F1	0.48 ±0.09	0.57 ±0.1	0.58 ±0.07	**0.60** ±0.08

Table 1: Results of the proposed method vs three baselines.

physicians. Considering these masks as initial bounding boxes, and following current practices in PET image segmentation [13], we perform majority voting on the outcome of three lesion segmentation methods: SUV 2.5, 41% SUV$_{max}$ and K-means.The radiomics features were extracted using *PyRadiomics* [14], a Python package for the extraction of radiomics features, which is IBSI compliant [15].

4.2. Experimental setup and results

We split the data into a training and a testing sets, with 466 and 117 patients respectively. We then deploy a 5-fold cross-validation with a grid search to automatically obtain the best hyper-parameters - learning rate, number of epochs and the edge weight constant *a*- on the training set. After this step, the GNN model is trained again with the selected hyper-parameters and tested on the test set.

To deal with the imbalance, we implement stratified splits, where the same ratio of positive and negative samples are maintained in the train-test sets and K-folds. We also employ a weighted loss during training, which gives higher weights to the loss of the minority class samples. Finally, we draw 5 class-balanced sub-sets from the test set and report the mean and standard deviation over them.

We compare the results of the proposed method against three baselines. The first, is a *MultiLayer Perceptron (MLP)* with early fusion, which receives as input, a feature vector per patient computed by averaging the feature vectors of his/her M_i lesions. The MLP is composed of three dense layers of dimension 32, the first two followed by a ReLU non linearlity and the last one by a sigmoid. The second baseline is a *Multiple Instance Learning (MIL)* approach, where each patient is represented as a *bag of lesions*. This method passes the feature vector of each lesion through an MLP, with the same architecture as before but with a final aggregation step (late fusion) consisting of a global max pooling operation over the lesion-wise predictions. Finally, the third baseline model is a *Graph Convolutional Network (GCN)* with 3 layers but no attention module (graph convolutional layers).

Table 1 compares the performance of our GAT model against the three baselines. The overall performances reflect the difficulty of the task. When looking at the mean *AUC* for the MLP classifier, we observe it has a low performance, with

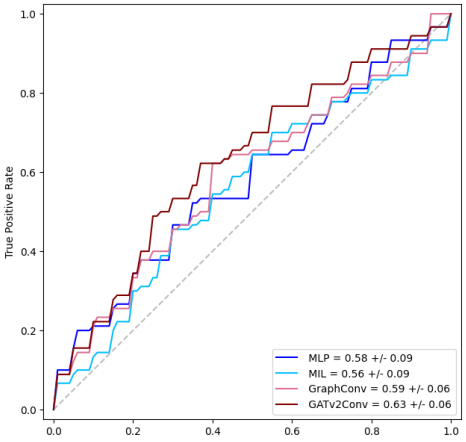

Fig. 3: ROC-AUC plots for MLP model, MIL model, GraphConv (GCN) model and our GATv2Conv (GAT) model.

a mean AUC of 0.58. Its standard deviation is high for all the metrics except for sensitivity, indicating that the model is less reliable and does not perform consistently for different input sets. It can be reasoned that much of the imaging information of the patient is lost by averaging the imaging features of all the lesions.

The MIL approach also performs poorly, with an AUC of 0.56. The standard deviation of this model is very high. The MIL classifier performs better than the MLP model. This can be explained by the fact that all the lesion features are fed as the input and late fusion looks promising. In what concerns the graph-based methods, the GCN model performs slightly better with an AUC of 0.59. This classification model has a lower standard deviation than the baseline models. This result indicates that the model is more reliable and displays lower sensitivity to different input sets. The GAT model performs best, with an AUC of 0.63. The GAT model also displays the best mean performance in terms of mean *balanced accuracy*, *specificity*, *balanced accuracy*, *micro*, *macro* and *weighted F1-scores*. The model is very stable, with a low standard deviation for all the metrics.

5. CONCLUSION

This paper proposes a novel framework for the 2-year survival classification of DLBCL patients from PET images, based on lesion graph attention-based neural networks. By applying a graph neural network coupled with attention, we are able to fusion the features of different lesions in the PET volume in a meaningful manner. The experimental results show that our framework improves over other machine learning models relying on simpler fusion strategies. These results highlight the importance of the information in graph representations. The framework is potentially generalizable to metastatic cancer PET images. Future work will include the combination with an automatic lesion segmentation approach.

6. ACKNOWLEDGMENTS

Approval for the GAINED study was granted by the Ethics Committee of the Nantes CHU.

7. REFERENCES

[1] Jakoba J Eertink, Gerben JC Zwezerijnen, Matthijs CF Cysouw, Sanne E Wiegers, Elisabeth AG Pfaehler, Pieternella J Lugtenburg, Bronno van der Holt, Otto S Hoekstra, Henrica CW de Vet, Josée M Zijlstra, et al., "Comparing lesion and feature selections to predict progression in newly diagnosed dlbcl patients with fdg pet/ct radiomics features," *European Journal of Nuclear Medicine and Molecular Imaging*, pp. 1–10, 2022.

[2] A Amyar, S Ruan, I Gardin, C Chatelain, P Decazes, and R Modzelewski, "3-d rpet-net: development of a 3-d pet imaging convolutional neural network for radiomics analysis and outcome prediction," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 225–231, 2019.

[3] Hongming Li, Pamela Boimel, James Janopaul-Naylor, Haoyu Zhong, Ying Xiao, Edgar Ben-Josef, and Yong Fan, "Deep convolutional neural networks for imaging data based survival analysis of rectal cancer," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 846–849.

[4] Amine Amyar, Romain Modzelewski, Vincent Morard, Pierre Vera, and Su Ruan, "Weakly supervised pet tumor detection using class response," *Journal of Nuclear Medicine*, vol. 62, no. supplement 1, pp. 1432–1432, 2021.

[5] Cheng Lu, Xiangxue Wang, Prateek Prasanna, German Corredor, Geoffrey Sedor, Kaustav Bera, Vamsidhar Velcheti, and Anant Madabhushi, "Feature driven local cell graph (fedeg): predicting overall survival in early stage lung cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 407–416.

[6] Bastien Jamet, Ludivine Morvan, Cristina Nanni, Anne-Victoire Michaud, Clément Bailly, Stéphane Chauvie, Philippe Moreau, Cyrille Touzeau, Elena Zamagni, Caroline Bodet-Milin, et al., "Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including fdg-pet radiomics: a combined analysis of two independent prospective european trials," *European journal of nuclear medicine and molecular imaging*, vol. 48, no. 4, pp. 1005–1015, 2021.

[7] Shaked Brody, Uri Alon, and Eran Yahav, "How attentive are graph attention networks?," in *International Conference on Learning Representations*, 2021.

[8] Anne-Ségolène Cottereau, Christophe Nioche, Anne-Sophie Dirand, Jérôme Clerc, Franck Morschhauser, Olivier Casasnovas, Michel Meignan, and Irène Buvat, "18f-fdg pet dissemination features in diffuse large b-cell lymphoma are predictive of outcome," *Journal of Nuclear Medicine*, vol. 61, no. 1, pp. 40–45, 2020.

[9] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[10] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, , no. 6, pp. 610–621, 1973.

[11] Michal Kazmierski and Benjamin Haibe-Kains, "Lymph node graph neural networks for cancer metastasis prediction," *arXiv preprint arXiv:2106.01711*, 2021.

[12] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *stat*, vol. 1050, pp. 20, 2017.

[13] Sally F Barrington, Ben GJC Zwezerijnen, Henrica CW de Vet, Martijn W Heymans, N George Mikhaeel, Coreline N Burggraaff, Jakoba J Eertink, Lucy C Pike, Otto S Hoekstra, Josée M Zijlstra, et al., "Automated segmentation of baseline metabolic total tumor burden in diffuse large b-cell lymphoma: which method is most successful? a study on behalf of the petra consortium," *Journal of Nuclear Medicine*, vol. 62, no. 3, pp. 332–337, 2021.

[14] Fedorov A. Parmar C. Hosny A. Aucoin N. Narayan V. Beets-Tan R. G. H. Fillon-Robin J. C. Pieper S. Aerts H. J. W. L. an Griethuysen, J. J. M., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. (21), 2017.

[15] Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al., "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.