Lymph Node Graph Neural Networks for Cancer Metastasis Prediction

Michal Kazmierski and Benjamin Haibe-Kains
Department of Medical Biophysics, University of Toronto &
Princess Margaret Cancer Centre
Toronto, ON, Canada
michal.kazmierski at mail.utoronto.ca, benjamin.haibe.kains at utoronto.ca

Abstract

Predicting outcomes, such as survival or metastasis for individual cancer patients is a crucial component of precision oncology. Machine learning (ML) offers a promising way to exploit rich multi-modal data, including clinical information and imaging to learn predictors of disease trajectory and help inform clinical decision making. In this paper, we present a novel graph-based approach to incorporate imaging characteristics of existing cancer spread to local lymph nodes (LNs) as well as their connectivity patterns in a prognostic ML model. We trained an edge-gated Graph Convolutional Network (Gated-GCN) to accurately predict the risk of distant metastasis (DM) by propagating information across the LN graph with the aid of soft edge attention mechanism. In a cohort of 1570 head and neck cancer patients, the Gated-GCN achieves AUROC of 0.757 for 2-year DM classification and C-index of 0.725 for lifetime DM risk prediction, outperforming current prognostic factors as well as previous approaches based on aggregated LN features. We also explored the importance of graph structure and individual lymph nodes through ablation experiments and interpretability studies, highlighting the importance of considering individual LN characteristics as well as the relationships between regions of cancer spread.

1 Introduction

Accurate prediction of disease trajectory and treatment outcomes is a key element of precision oncology. Nowadays, clinicians have access to an unprecedented amount of data about individual patient, from basic health information to high-resolution medical imaging and molecular biomarkers. Integrating the various signals to extract actionable predictions, however, is becoming increasingly challenging for a human. The growing amount and variety of data generated in complex domains like oncology makes it increasingly difficult for clinicians to integrating these multi-modal data in their decision process. Machine learning (ML) offers a way to make use of all the available data, automatically learning the underlying complex disease patterns from large multi-modal datasets. By delivering accurate, individualised predictions of outcomes and possible course of disease, ML can help doctors make better, data-driven decisions, ultimately translating into a benefit for the patient.

The need for new prognostic tools to guide clinical decision making is particularly apparent in head and neck cancer (HNC), with incidence of 900,000 cases worldwide every year that is projected to increase by 30% by 2030 (Johnson et al. 2020). Despite recent advances in treatment, survival rates still remain suboptimal, largely due to the high heterogeneity in tumour biology and outcomes, making optimal treatment selection challenging. In particular, current prognostic factors are not sufficient to stratify patients by risk of distant metastasis (DM) — the spread of cancer from the original site to other organs (such as lungs or brain) — which is an indicator of particularly poor prognosis (O'Sullivan et al. 2013). Local spread to lymph nodes (LNs) in the neck occurs frequently

Preprint. Under review.

in HNC and often precedes distant metastasis by creating repositories of cancer cells with access to the lymphatic system (Lee et al. 2018) 1, but its current use in prognosis is limited to an aggregate score relying only on simple anatomical and imaging features.

We propose a novel machine learning approach to integrate tumour and LN imaging characteristics and clinical metadata to predict distant metastasis in head and neck cancer. We use a graph-based framework, with vertices representing metastatic LNs and edges the lymphatic connections in an individual patient-level graph. We leverage recent advances in geometric deep learning to develop an edge-gated Graph Convolutional Network (Gated-GCN) which learns the relationships between metastatic LNs in addition to their individual features and multi-task learning to accurately predict distant metastasis-free survival. We investigate the importance of graph structure and imaging features in ablation experiments and compare our approach to previously known prognostic factors.

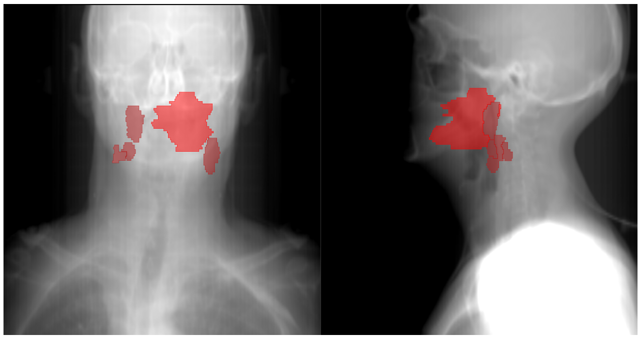


Figure 1: Coronal (left) and saggital (right) projections obtained from a CT scan of a HNC patient showing the primary tumour (lighter red near the centre) and invaded neck lymph nodes (darker red).

1.1 Related work

Vallières et al. (2017) first proposed using computed tomography (CT) imaging characteristics of the primary tumour to predict distant metastasis in HNC, with encouraging results. Later, Bogowicz et al. (2019) and Wu et al. (2019) introduced the idea of incorporating metastatic LN characteristics in predicting cancer recurrence and metastasis, respectively. The main limitation of these approaches is the reliance on aggregated features over all metastatic lymph nodes, ignoring the relationships between metastatic LNs as well as their potentially variable individual importance for prognosis. Apart from prognosis, Chao et al. (2020) used a graph neural network to reduce the false positive rate in mediastinal LN detection.

2 Dataset

We used the RADCURE dataset (Kazmierski et al. 2021), which includes pre-treatment CT scans and additional clinical metadata of HNC patients treated with radio(chemo)therapy at a single institution. Primary tumours and metastatic LNs were manually delineated by a trained radiation oncologist as part of the treatment planning process. We included patients with positive lymph node stage according to the TNM criteria (Brierley et al. 2017) and where at least one LN contour could be retrieved, resulting in 1570 patients with median of 4 LN per patient (range 1–30). There were 259 DM events (16%), with a median time to event of 1.1 years.

From each tumour and LN region of interest (ROI), we extracted a set of 11 image features previously identified as prognostic (Wu et al. 2019), as well as the coordinates of the region centroid, signed distance to the body mid-line and LN level (if applicable), which describes the position of a node in the lymphatic system of the neck (Grégoire et al. 2014). We also used the available standard prognostic factors, such as demographic information and staging.

3 Graph-based encoding of lymph node relationships

For each patient, we encoded the relationships between primary tumour and invaded LNs as an undirected graph $G = (V, E)$. Each vertex, representing the tumour or LN, is associated with a vector of imaging features $\mathbf{x}_i$, that is $V = \{\mathbf{x}_i\}_{i=0}^{N_V}$ for an image with $N_V - 1$ metastatic LNs and one primary tumour. The edges represent the underlying connectivity pattern between the tumour and LNs and are initialized with the distance between the 3D centroids $\mathbf{p}$ of the connected regions, $e_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|$ for all $e_{ij} \in E$. We note, however, that the graph convolutional network we used learns a more general edge update than just the distance which depends on the features of the connected vertices as well as previous edge features.

3.1 Edge-gated GCN

To learn prognostic representations from the input graphs we used the edge-gated graph convolutional network (Gated-GCN) (Bresson & Laurent 2018). Gated-GCN learns anisotropic graph convolution operators by explicitly maintaining edge features at each layer and using them to compute dense soft attention coefficients which modulate vertex feature updates. Specifically, the vertex features at $(l+1)$th layer are computed as

$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \mathrm{ReLU}(\mathbf{W}_1^l \mathbf{h}_i^l + \sum_{j \in \mathcal{N}_i} \boldsymbol{\alpha}_{ij}^l \odot \mathbf{W}_2^l \mathbf{h}_j^l), \tag{1}$$

where

$$\boldsymbol{\alpha}_{ij}^l = \frac{\sigma(\mathbf{e}_{ij}^l)}{\sum_{k \in \mathcal{N}_i} \sigma(\mathbf{e}_{ik}^l)}, \quad \sigma(z) = (1 + \exp(-z))^{-1} \tag{2}$$

are the edge attention coefficients, $\mathbf{W}_{1:5} \in \mathbb{R}^{d \times d}$ are learnable weight matrices for the hidden dimension d and $\odot$ denotes elementwise product. The d-dimensional edge features are updated using

$$\mathbf{e}_{ij}^{l+1} = \mathbf{e}_{ij}^l + \mathrm{ReLU}(\mathbf{W}_3^l \mathbf{e}_{ij}^l + \mathbf{W}_4^l \mathbf{h}_i^l + \mathbf{W}_5^l \mathbf{h}_j^l). \tag{3}$$

Since the input vertex and edge feature dimensions do not necessarily match, we embed both in $\mathcal{R}^d$ using a linear projection before the first graph convolutional layer. We also use batch normalization (Ioffe & Szegedy 2015) on both node and edge features, as well as graph size normalization (Dwivedi et al. 2020) to reduce the impact of variable input graph sizes. To help mitigate overfitting, dropout was applied to both the vertex and edge updates, as well as in the final graph-level fully-connected layers.

The graph structure and learned edge features enable the network to exploit the relationships between the existing regions of tumour spread in addition to their individual characteristics, which we hypothesise is crucial for accurate prediction of future cancer spread.

3.2 Predictions and loss function

To predict the probability of distant metastasis over time, we used the multi-task logistic regression framework (Yu et al. 2011). By predicting the probability of DM occurrence across multiple discrete time intervals in a multi-task way, it can better exploit the survival-type training data (also for patients with censored, i.e. partially observed survival times) and learn a flexible, time-varying risk function. Formally, we divide the time axis into K consecutive intervals $[t_{k-1}, t_k)$ for $k = 1, \ldots, K$ and define a sequence of binary targets $y_k = \mathbf{1}\{t_{k-1} \leq T < t_k\}$, where T denotes the time of DM occurrence and $\mathbf{1}\{\}$ is the indicator function. MTLR uses a separate set of parameters $\{\mathbf{w}_k, b_k\}$ for each time interval, giving the predicted logits for event in a given interval as $\hat{y}_k^{(j)} = (\mathbf{w}_k^T \mathbf{z}^{(j)} + b_k)$. The model is trained end-to-end using gradient descent by optimizing the following objective:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{j \text{ uncensored}} \sum_{k=1}^{K-1} y_k^{(j)} \times \hat{y}_k^{(j)} - \sum_{j \text{ censored}} \log(\sum_{i=1}^{K-1} \mathbf{1}\{t_i \geq T_c^{(j)}\} \exp(\sum_{k=i}^{K-1} y_k^{(j)} \times \hat{y}_k^{(j)})) + \sum_j Z^{(j)}, \tag{4}$$

where $T_c^{(j)}$ denotes the censoring time of the jth patient and $Z^{(j)} = \log(\sum_{i=1}^{K} \exp(\sum_{k=i}^{K-1} \hat{y}_k^{(j)}))$ is a normalizing constant. In the Gated-GCN case, we set $\mathbf{z} = \mathrm{MLP}(\mathbf{h}_{\mathrm{out}})$, where $\mathbf{h}_{\mathrm{out}} = \frac{1}{N_V} \sum_{i=1}^{N_V} \mathbf{h}_i^L$ are the pooled vertex features from the final graph convolutional layer, potentially concatenated with the additional patient-level clinical covariates and MLP is a graph-level multi-layer perceptron (fig. 2).

We derived two kinds of predictions from the MTLR output: the probability of event before a specified time point (here 2 years) and the lifetime risk of DM, which takes into account the predictions at all timepoints.

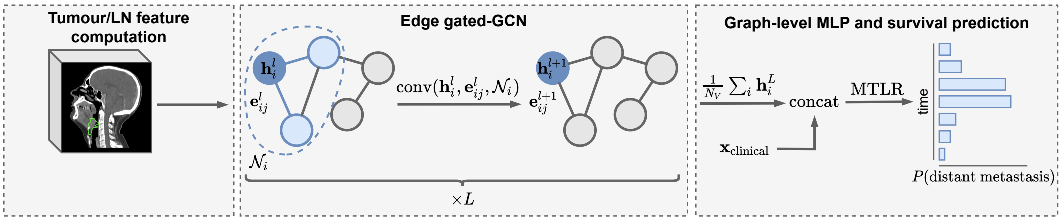

Tumour/LN feature computation
Edge gated-GCN
$\mathbf{h}_i^l$
$\mathbf{e}_{ij}^l$
$\mathcal{N}_i$
$\text{conv}(\mathbf{h}_i^l, \mathbf{e}_{ij}^l, \mathcal{N}_i)$
$\mathbf{h}_i^{l+1}$
$\mathbf{e}_{ij}^{l+1}$
$\times L$
Graph-level MLP and survival prediction
$\frac{1}{N_V}\sum_i \mathbf{h}_i^L$
concat
$\mathbf{x}_{\text{clinical}}$
MTLR
time
$P(\text{distant metastasis})$

Figure 2: Overview of the proposed graph-based framework for DM prediction. Vertex features learned by L Gated-GCN layers (conv) are pooled and concatenated with additional clinical metadata, $\mathbf{x}_{\text{clinical}}$. A graph-level MLP with an MTLR head is used to predict probability of DM jointly at multiple discrete timepoints.

4 Experiments

We evaluated the performance of our approach on 2 tasks: predicting distant metastasis before 2 years (a binary classification task, with target $y_{\text{bin}} = \mathbf{1}\{\text{patient develops DM before 2 years post-treatment}\}$) using area under the ROC curve (AUROC) as the performance metric and predicting the lifetime metastasis risk (a survival prediction task), using the concordance (C) index (Harrell et al. 1996). The inclusion of the binary classification task enabled us to evaluate the performance of the proposed approach at identifying high-risk patients that might require additional interventions.

We experimented with 2 types of graph connectivity: (1) complete (with edges between all pairs of vertices) and (2) K-nearest neighbours (where every vertex is connected to K closest vertices by Euclidean distance) for $K = 2,\ 3$.

To investigate the role of graph structure, we trained a graph-agnostic neural network which shares weights across the vertices, similarly to the Gated-GCN, but ablates the graph connectivity (Luzhnica et al. 2019). The graph-level feature vector is obtained by mean-pooling the final layer features across the vertices. We also compared the performance of the proposed approach to a set of simple benchmark models, which use known prognostic clinical features alone or combined with either aggregated tumour and LN features, tumour features only or total tumour + LN volume. We used the neural MTLR architecture (Fotso 2018) with one or more hidden layers for all of the simple baselines.

In all experiments, we used the same 5-fold stratified cross-validation training setup. We trained each model for 100 epochs using the Adam optimizer (Kingma & Ba 2014). Hyperparameters, including the learning rate, batch size, number and width of hidden layers and dropout rate were tuned using nested random search. The proposed approach and all benchmark models were implemented using PyTorch (Paszke et al. 2019), Lightning (Falcon 2019) `pytorch-geometric` (Fey & Lenssen 2019) and `torchmtlr`[1]. Details of the training and evaluation protocol can be found in Appendix B.

5 Results

The best performance in both 2-year and lifetime risk prediction was achieved by the GCN trained on fully-connected graphs (table 1). All models using imaging features achieved better prognostic performance than baseline relying on clinical metadata only, with further increase in performance when using LN features in addition to the primary tumour. Notably, both the GNN and graph-agnostic NN performed better than the clinical-only baseline, indicating that LN imaging features can be strong, independent predictors of distant metastasis. Furthermore, considering the LN characteristics separately was necessary for higher predictive accuracy (which is further supported by the variable importance of different metastatic LNs in the model, as described in the Interpretability section below).

[1] `https://github.com/mkazmier/torchmtlr`

Table 1: Performance of the Gated-GCN with varying graph connectivity, graph-agnostic neural networks and simple baselines on the classification and lifetime risk tasks. Metrics are reported as mean and standard deviation over the 5 folds.

	2-year AUROC		**C-index**	
	mean	std	mean	std
Gated-GCN + clinical (complete graph)	**0.757**	**0.058**	**0.725**	**0.016**
Gated-GCN + clinical ($K = 3$)	0.750	0.026	0.719	0.021
Gated-GCN + clinical ($K = 2$)	0.747	0.021	0.717	0.024
Graph-agnostic NN + clinical	0.726	0.029	0.696	0.017
Clinical + tumour/node features	0.708	0.039	0.683	0.024
Clinical + tumour/node volume	0.693	0.053	0.665	0.035
Clinical + tumour features only	0.682	0.067	0.653	0.046
Clinical only	0.668	0.053	0.648	0.037

5.1 Role of graph structure

The best graph-based model achieved higher AUROC (.757) and C-index (.725) than the graph-agnostic baseline ($\text{AUROC} = .726,\ C = .696$), indicating that explicitly modelling relationships between the existing metastases is important for predicting further cancer spread. The performance was similar across different choices of graph topology, with slight bias towards stronger connectivity. This suggests that the network might be able to make use of greater number of connections per vertex, with the edge attention process modulating the importance of neighbours during updates (with 3-nearest neighbour connectivity many of the smaller graphs become complete).

5.2 Interpretability

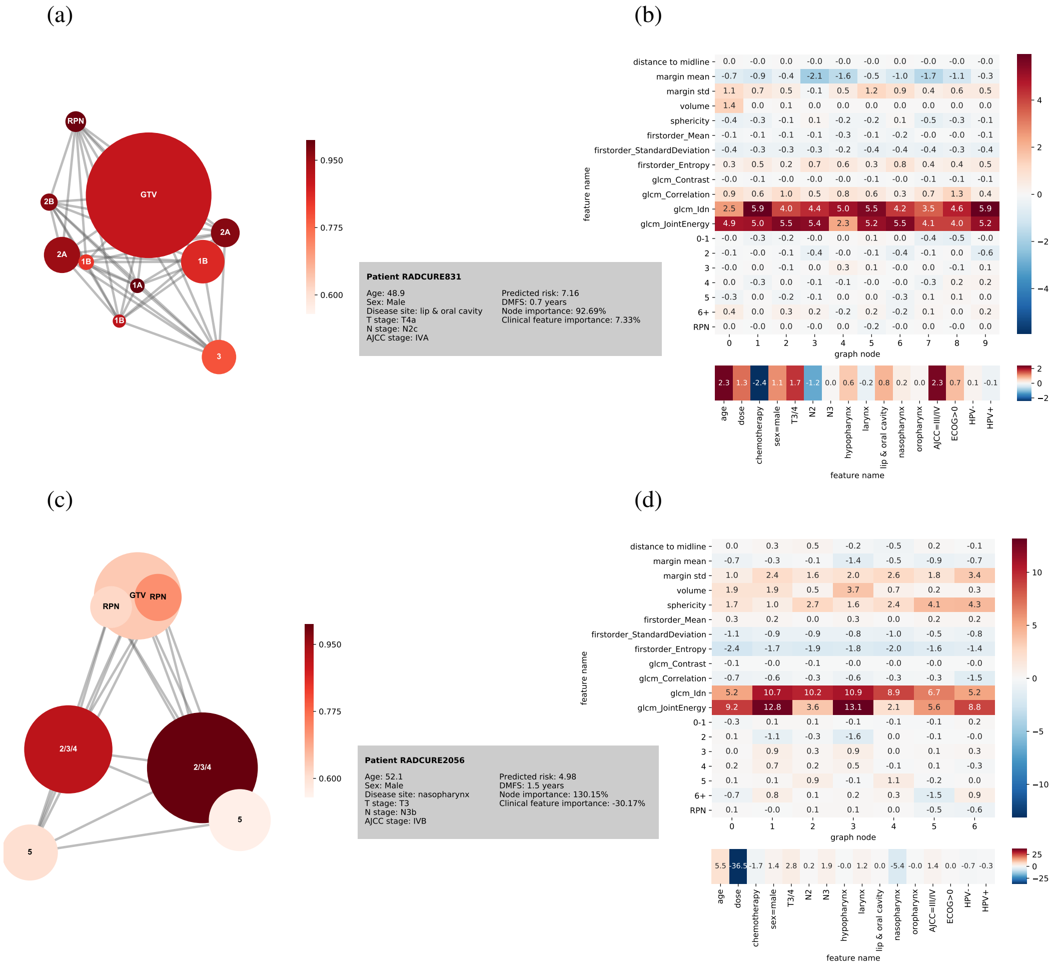
Understating the factors contributing to the model's prediction can be just as important as high discriminative performance for both doctors and patients, particularly when the predictions factor into clinical decisions (Tonekaboni et al. 2019). We used integrated gradients (Sundararajan et al. 2017) to evaluate the contributions of the graph vertices, as well as individual vertex and patient-level features to the predictions of the best model (fig. 3). Vertex-level attribution scores were computed as sums of absolute values of feature attributions for each vertex, normalized by sum of attributions over all vertices.

The attributions can be used to develop a qualitative intuition of how the model operates. Inspecting vertex integrated gradients for a random sample of test patients revealed variable relevance of individual vertices for the predictions. Notably, in some cases a LN received over 60% higher attributions than the primary tumour. Additionally, we found that several features had consistently high importance for model predictions, including 2 image heterogeneity features and a region margin measure.

6 Conclusions

We presented a method for graph-based distant metastasis prediction in head and neck cancer. Our approach using edge-gated graph convolutional network achieved better performance than the graph-agnostic baselines on both 2-year classification and lifetime risk prediction, and is a promising new way of incorporating data from multiple regions of interests and their relationships in a prognostic model. Additionally, our results highlight the relevance of lymph node imaging biomarkers for DM prediction. Although we used HNC as a case study, due to high incidence of LN metastases, our approach is straightforward to apply in other cancer sites where local LN spread commonly occurs, e.g. breast.

The study has several potential limitations. The dataset we used was relatively small (although still substantial in comparison with other publicly-available HNC datasets) and single-institution. We



Patient RADCURE831

Age: 48.9	Predicted risk: 7.16
Sex: Male	DMFS: 0.7 years
Disease site: lip & oral cavity	Node importance: 92.69%
T stage: T4a	Clinical feature importance: 7.33%
N stage: N2c	
AJCC stage: IVA	

feature name	0	1	2	3	4	5	6	7	8	9
distance to midline	0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0
margin mean	-0.7	-0.9	-0.4	-2.1	-1.6	-0.5	-1.0	-1.7	-1.1	-0.3
margin std	1.1	0.7	0.5	-0.1	0.5	1.2	0.9	0.4	0.6	0.5
volume	1.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
sphericity	-0.4	-0.3	-0.1	0.1	-0.2	-0.2	0.1	-0.5	-0.3	-0.1
firstorder_Mean	-0.1	-0.1	-0.1	-0.1	-0.3	-0.1	-0.2	-0.0	-0.0	-0.1
firstorder_StandardDeviation	-0.4	-0.3	-0.3	-0.3	-0.2	-0.4	-0.4	-0.4	-0.3	-0.4
firstorder_Entropy	0.3	0.5	0.2	0.7	0.6	0.3	0.8	0.4	0.4	0.5
glcm_Contrast	-0.0	-0.1	-0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.1	-0.1
glcm_Correlation	0.9	0.6	1.0	0.5	0.8	0.6	0.3	0.7	1.3	0.4
glcm_Idn	2.5	5.9	4.0	4.4	5.0	5.5	4.2	3.5	4.6	5.9
glcm_JointEnergy	4.9	5.0	5.5	5.4	2.3	5.2	5.5	4.1	4.0	5.2
0-1	-0.0	-0.3	-0.2	-0.0	0.0	0.1	0.0	-0.4	-0.5	-0.0
2	-0.1	-0.0	-0.1	-0.4	-0.0	-0.1	-0.4	-0.1	0.0	-0.6
3	-0.0	0.0	-0.1	0.0	0.3	0.1	-0.0	0.0	-0.1	0.1
4	0.0	-0.3	0.0	-0.1	-0.1	-0.0	-0.0	-0.3	0.2	0.2
5	-0.3	0.0	-0.2	-0.1	0.0	0.0	-0.3	0.1	0.1	0.0
6+	0.4	0.0	0.3	0.2	-0.2	0.2	-0.2	0.1	0.0	0.2
RPN	0.0	-0.0	0.0	-0.0	0.0	-0.2	-0.0	0.0	-0.0	0.0
	graph node									

age	dose	chemotherapy	sex=male	T3/4	N2	N3	hypopharynx	larynx	lip & oral cavity	nasopharynx	oropharynx	AJCC=III/IV	ECOG>0	HPV-	HPV+
2.3	1.3	-2.4	1.1	1.7	-1.2	0.0	0.6	-0.2	0.8	0.2	0.0	2.3	0.7	0.1	-0.1

Patient RADCURE2056

Age: 52.1	Predicted risk: 4.98
Sex: Male	DMFS: 1.5 years
Disease site: nasopharynx	Node importance: 130.15%
T stage: T3	Clinical feature importance: -30.17%
N stage: N3b	
AJCC stage: IVB	

feature name	0	1	2	3	4	5	6
distance to midline	0.0	0.3	0.5	-0.2	-0.5	0.2	-0.1
margin mean	-0.7	-0.3	-0.1	-1.4	-0.5	-0.9	-0.7
margin std	1.0	2.4	1.6	2.0	2.6	1.8	3.4
volume	1.9	1.9	0.5	3.7	0.7	0.2	0.3
sphericity	1.7	1.0	2.7	1.6	2.4	4.1	4.3
firstorder_Mean	0.3	0.2	0.0	0.3	0.0	0.2	0.2
firstorder_StandardDeviation	-1.1	-0.9	-0.9	-0.8	-1.0	-0.5	-0.8
firstorder_Entropy	-2.4	-1.7	-1.9	-1.8	-2.0	-1.6	-1.4
glcm_Contrast	-0.1	-0.0	-0.1	-0.0	-0.0	-0.0	-0.0
glcm_Correlation	-0.7	-0.6	-0.3	-0.6	-0.3	-0.3	-1.5
glcm_Idn	5.2	10.7	10.2	10.9	8.9	6.7	5.2
glcm_JointEnergy	9.2	12.8	3.6	13.1	2.1	5.6	8.8
0-1	-0.3	0.1	0.1	-0.1	-0.1	-0.1	0.2
2	0.1	-1.1	-0.3	-1.6	0.0	-0.1	-0.0
3	0.0	0.9	0.3	0.9	0.0	0.1	0.3
4	0.2	0.7	0.2	0.5	-0.1	0.1	-0.3
5	0.1	0.1	0.9	-0.1	1.1	-0.2	0.0
6+	-0.7	0.8	0.1	0.2	0.3	-1.5	0.9
RPN	0.1	-0.0	0.1	0.1	0.0	-0.5	-0.6
	graph node						

age	dose	chemotherapy	sex=male	T3/4	N2	N3	hypopharynx	larynx	lip & oral cavity	nasopharynx	oropharynx	AJCC=III/IV	ECOG>0	HPV-	HPV+
5.5	-36.5	-1.7	1.4	2.8	0.2	1.9	-0.0	1.2	0.0	-5.4	-0.0	1.4	0.0	-0.7	-0.3

Figure 3: Attributions for randomly selected patients with time to DM shorter (top) and longer (bottom) than median. 3a and 3c show the input graph, where the size of each node is proportional to the volume of the associated region and the colour intensity corresponds to the total attribution score normalized by the maximum score. The vertices are labeled by the nodal level (number or `RPN` for retropharyngeal nodes) or `GTV` for the primary tumour. 3b and 3d show the normalized integrated gradients for node features (top) and clinical metadata (bottom).

plan to extend our dataset with more patients, as well as test the proposed approach using data from a different hospital. Additionally, our model relies on engineered image features. It is likely that learned representations from a convolutional network could further improve the performance of our method and we are working on an end-to-end approach learning directly from images. We leave this for future work.

References

Bogowicz, M., Tanadini-Lang, S., Guckenberger, M. & Riesterer, O. (2019), 'Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer', *Sci Rep* **9**(1), 15198.

Bresson, X. & Laurent, T. (2018), 'Residual Gated Graph ConvNets', *arXiv:1711.07553 [cs, stat]* .

Brierley, J., Gospodarowicz, M. K. & Wittekind, C., eds (2017), *TNM Classification of Malignant Tumours*, eighth edition edn, John Wiley & Sons, Inc, Chichester, West Sussex, UK ; Hoboken, NJ.

Chao, C.-H., Zhu, Z., Guo, D., Yan, K., Ho, T.-Y., Cai, J., Harrison, A. P., Ye, X., Xiao, J., Yuille, A., Sun, M., Lu, L. & Jin, D. (2020), 'Lymph Node Gross Tumor Volume Detection in Oncology Imaging via Relationship Learning Using Graph Neural Network', *arXiv:2008.13013 [cs]* .

Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y. & Bresson, X. (2020), 'Benchmarking Graph Neural Networks', *arXiv:2003.00982 [cs, stat]* .

Falcon, W. (2019), 'PyTorch Lightning', *GitHub* **3**.

Fey, M. & Lenssen, J. E. (2019), Fast graph representation learning with PyTorch Geometric, *in* 'ICLR Workshop on Representation Learning on Graphs and Manifolds'.

Fotso, S. (2018), 'Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework', *arXiv:1801.05512 [cs, stat]* .

Grégoire, V., Ang, K., Budach, W., Grau, C., Hamoir, M., Langendijk, J. A., Lee, A., Le, Q.-T., Maingon, P., Nutting, C., O'Sullivan, B., Porceddu, S. V. & Lengele, B. (2014), 'Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines', *Radiotherapy and Oncology* **110**(1), 172–181.

Harrell, F. E., Lee, K. L. & Mark, D. B. (1996), 'Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Stat Med* **15**(4), 361–387.

Ioffe, S. & Szegedy, C. (2015), 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift', *arXiv:1502.03167 [cs]* .

Johnson, D. E., Burtness, B., Leemans, C. R., Lui, V. W. Y., Bauman, J. E. & Grandis, J. R. (2020), 'Head and neck squamous cell carcinoma', *Nat Rev Dis Primers* **6**(1), 92.

Kazmierski, M., Welch, M., Kim, S., McIntosh, C., Head, P. M., Group, N. C., Rey-McIntyre, K., Huang, S. H., Patel, T., Tadic, T., Milosevic, M., Liu, F.-F., Hope, A., Bratman, S. & Haibe-Kains, B. (2021), 'A Machine Learning Challenge for Prognostic Modelling in Head and Neck Cancer Using Multi-modal Data', *arXiv:2101.11935 [cs, eess]* .

Kingma, D. P. & Ba, J. (2014), 'Adam: A Method for Stochastic Optimization', *arXiv:1412.6980 [cs]* .

Lee, N. C., Kelly, J. R., Park, H. S., An, Y., Judson, B. L., Burtness, B. A. & Husain, Z. A. (2018), 'Patterns of failure in high-metastatic node number human papillomavirus-positive oropharyngeal carcinoma', *Oral Oncology* **85**, 35–39.

Levman, J. E. & Martel, A. L. (2011), 'A Margin Sharpness Measurement for the Diagnosis of Breast Cancer from Magnetic Resonance Imaging Examinations', *Academic Radiology* **18**(12), 1577–1581.

Luzhnica, E., Day, B. & Liò, P. (2019), 'On Graph Classification Networks, Datasets and Baselines', *arXiv:1905.04682 [cs, stat]* .

O'Sullivan, B., Huang, S. H., Siu, L. L., Waldron, J., Zhao, H., Perez-Ordonez, B., Weinreb, I., Kim, J., Ringash, J., Bayley, A., Dawson, L. A., Hope, A., Cho, J., Irish, J., Gilbert, R., Gullane, P., Hui, A., Liu, F.-F., Chen, E. & Xu, W. (2013), 'Deintensification Candidate Subgroups in Human Papillomavirus–Related Oropharyngeal Cancer According to Minimal Risk of Distant Metastasis', *JCO* **31**(5), 543–550.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019), PyTorch: An imperative style, high-performance deep learning library, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox & R. Garnett, eds, 'Advances in Neural Information Processing Systems 32', Curran Associates, Inc., pp. 8024–8035.

Sundararajan, M., Taly, A. & Yan, Q. (2017), 'Axiomatic Attribution for Deep Networks', *arXiv:1703.01365 [cs]* .

Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. (2019), 'What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use', *arXiv:1905.05134 [cs, stat]* .

Vallières, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J. W. L., Khaouam, N., Nguyen-Tan, P. F., Wang, C.-S., Sultanem, K., Seuntjens, J. & El Naqa, I. (2017), 'Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer', *Sci Rep* **7**(1), 10117.

van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S. & Aerts, H. J. (2017), 'Computational Radiomics System to Decode the Radiographic Phenotype', *Cancer Research* **77**(21), e104–e107.

Wu, J., Gensheimer, M. F., Zhang, N., Han, F., Liang, R., Qian, Y., Zhang, C., Fischbein, N., Pollom, E. L., Beadle, B., Le, Q.-T. & Li, R. (2019), 'Integrating Tumor and Nodal Imaging Characteristics at Baseline and Mid-Treatment Computed Tomography Scans to Predict Distant Metastasis in Oropharyngeal Cancer Treated With Concurrent Chemoradiotherapy', *International Journal of Radiation Oncology*Biology*Physics* **104**(4), 942–952.

Yu, C.-N., Greiner, R., Lin, H.-C. & Baracos, V. (2011), Learning patient-specific cancer survival distributions as a sequence of dependent regressors, *in* J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems 24', Curran Associates, Inc., pp. 1845–1853.

A Dataset details

The detailed description of the clinical and imaging data can be found in Kazmierski et al. (2021). Here we highlight the some of the details relevant to the present study.

We used the following clinical metadata: age, sex, tumour location and stage, overall health status (ECOG criteria), radiation dose, chemotherapy use and human papillomavirus (HPV) infection status. Continuous features were standardized to zero mean, unit variance and discrete features were one-hot encoded, with missing values handled by adding additional missingness indicator level.

CT features were computed using PyRadiomics version 3.0 (van Griethuysen et al. 2017), with the exception of margin features, for which we used an in-house implementation following (Levman & Martel 2011). We used the same feature set and preprocessing protocol as Wu et al. (2019). The following image features were extracted from each region of interest: margin mean and standard deviation, volume, sphericity, firstorder-Mean, firstorder-StandardDeviation, firstorder-Entropy, GLCM-Contrast, GLCM-Correlation, GLCM-Idn (equivalent to GLCM-Homogeneity1), GLCM-JointEnergy, distance to midline and lymph node level. For the lymph node benchmark model, the features were averaged over all lymph nodes (except volume which was summmed) and additional features were added (LN count, maximum tumour-node and node-node distance), following (Wu et al. 2019).

B Training and hyperparameter selection

For each cross-validation split, we initially reserved one of the training folds for validation and tuned the hyperparameters using random search (see table 2 for hyperparameter distributions). We selected the hyperparameter set with the lowest validation loss out of 60 search iterations and re-trained the model on all training folds before evaluating on the test fold.

Table 2: Hyperparameter distributions used in random search. Square brackets indicate discrete uniform distribution over the values. $\text{loguniform}(a, b)$ is the uniform distribution in log domain between $\log(a)$ and $\log(b)$.

Hyperparameter	Distribution
Learning rate	$\text{loguniform}(10^{-4},\ 5 \times 10^{-3})$
Batch size	$[64, 128, 256, 512]$
Weight decay	$\text{loguniform}(10^{-5},\ 10^{-3})$
MTLR ℓ_2 regularization	$[10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$
Dropout probability	$\text{uniform}(0,\ .5)$
Number of GCN/node-level MLP layers	$[1, 2, 3, 4]$
Number of MLP layers	$[1, 2, 3, 4]$
Layer width	$[32, 64, 128, 256]$