Bridging Ex-Vivo Training and Intra-operative Deployment for Surgical Margin Assessment with Evidential Graph Transformer

Amoon Jamzad[1(✉)], Fahimeh Fooladgar[4], Laura Connolly[1], Dilakshan Srikanthan[1], Ayesha Syeda[1], Martin Kaufmann[2], Kevin Y.M. Ren[3], Shaila Merchant[2], Jay Engel[2], Sonal Varma[3], Gabor Fichtinger[1], John F. Rudan[2], and Parvin Mousavi[1]

[1] School of Computing, Queen's University, Kingston, Canada
a.jamzad@queensu.ca
[2] Department of Surgery, Queen's University, Kingston, Canada
[3] Department of Pathology and Molecular Medicine, Queen's University, Kingston, Canada
[4] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada

Abstract. PURPOSE: The use of intra-operative mass spectrometry along with Graph Transformer models showed promising results for margin detection on ex-vivo data. Although highly interpretable, these methods lack the ability to handle the uncertainty associated with intra-operative decision making. In this paper for the first time, we propose Evidential Graph Transformer network, a combination of attention mapping and uncertainty estimation to increase the performance and interpretability of surgical margin assessment. METHODS: The Evidential Graph Transformer was formulated to output the uncertainty estimation along with intermediate attentions. The performance of the model was compared with different baselines in an ex-vivo cross-validation scheme, with extensive ablation study. The association of the model with clinical features were explored. The model was further validated for a prospective ex-vivo data, as well as a breast conserving surgery intra-operative data. RESULTS: The purposed model outperformed all baselines, statistically significantly, with average balanced accuracy of 91.6%. When applied to intra-operative data, the purposed model improved the false positive rate of the baselines. The estimated attention distribution for status of different hormone receptors agreed with reported metabolic findings in the literature. CONCLUSION: Deployment of ex-vivo models is challenging due to the tissue heterogeneity of intra-operative data. The proposed Evidential Graph Transformer is a powerful tool that while providing the attention distribution of biochemical subbands, improve the surgical deployment power by providing decision confidence.

Keywords: Intra-operative deployment · Uncertainty estimation · Interpretation · Graph transformer network · Breast cancer margin

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 562–571, 2023.
https://doi.org/10.1007/978-3-031-43990-2_53

1 Introduction

Achieving complete tumor resection in surgical oncology like breast conserving surgery (BCS) is challenging as boundaries of tumors are not always visible/palpable [10]. In BCS the surgeon removes breast cancer while attempting to preserve as much healthy tissue as possible to prevent permanent deformation and to enhance cosmesis. The current standard of care for evaluating surgical success is to investigate the resection margins, which refers to the area surrounding the excised tumor. Up to 30% of surgeries result in incomplete tumor resection and require a revision operation [10]. The intelligent knife (iKnife) is a mass spectrometry device that can address this challenge by analyzing the biochemical signatures of resected tissue using the smoke that is released during tissue incineration [3]. Each spectrum contains the distribution of sampled ions with respect to their mass to charge ratio (m/z). Previously, learning models have been used in combination with iKnife data for ex-vivo tissue characterization and real-time margin detection [16,17].

The success of clinical deployment of learning models heavily relies on approaches that are not only accurate but also interpretable. Therefore, it should be clear how models reach their decisions and the confidence they have in such decision. Studies suggest that one way to improve these factors is through data centric approaches i.e. to focus on appropriate representation of data. Specifically, representation of data as graphs has been shown to be effective for medical diagnosis and analysis [1]. It has also been shown that graph neural networks can accurately capture the biochemical signatures of iKnife and determine the tissue type. Particularly, Graph Transformer Networks (GTN) has have shown to further enhance the transparency of underlying relation between the graph nodes and decision making via attention mechanism [11].

Biological data, specially those acquired intra-opertively, are heterogeneous by nature. While the use of ex-vivo data collected under specific protocols are beneficial to develop baseline models, intra-operative deployment of these models is challenging. For iKnife, the ex-vivo data is usually collected from homogeneous regions of resected specimens under the guidance of a trained pathologist, versus the intra-operative data is recorded continuously while the surgeon cutting through tissues with different heterogeneity and pathology. Therefore, beyond predictive power and explainable decision making, intra-operative models must be able to handle mixed and unseen pathology labels.

Uncertainty-aware models in computer-assisted interventions can provide clinicians with feedback on prediction confidence to increase their reliability during deployment. Deep ensembles [15] and Bayesian networks [9] incur high runtime and computational cost both at training and inference time and thus, less practical for real-time computer-assisted interventions. Evidential Deep Learning [18] is another approach that has been proposed based on the evidence framework of Dempster-Shafer Theory [12]. Since the evidential approach jointly generates the network prediction and uncertainty estimation, it seems more suitable for computationally efficient intra-operative deployment.

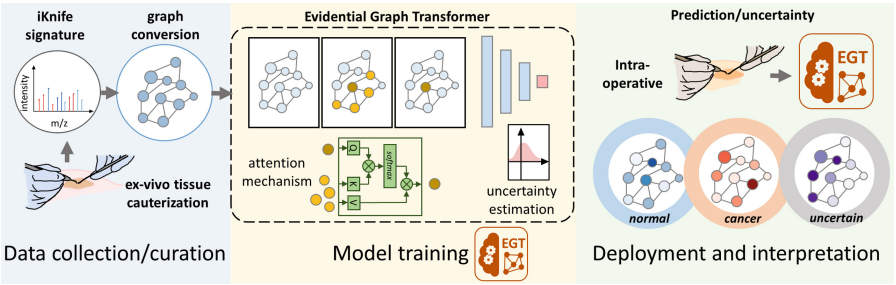

Fig. 1. An overview of the proposed approach including data collection and preprocessing, graph conversion, and interpretation of uncertainty and attentions.

In this paper, we propose Evidential Graph Transformer (EGT), a combination of graph-based feature-level attention mechanism with sample-level uncertainty estimation, to increase the performance and interpretability of surgical margin assessment. This is done by implementing the evidential loss and prediction functions within a graph transformer model to output the uncertainty, intermediate attention, and model prediction. To demonstrate the state-of-the-art performance of the proposed approach on mass spectrometry data, the model is compared with different baselines in both cross-validation and prospective schemes on ex-vivo data. Furthermore, the performance of model is also investigated intraoperatively. In addition to the proposed model, we present a new visualization approach to better correlate the graph nodes with the spectral content of the data, which improves interpretability. In addition to the ablation study on the network and graph strictures, we also investigate the metabolic association of breast cancer hormone receptor status.

2 Materials and Methods

Figure 1 presents the overview of the proposed approach. Following data collection and curation, each burn (spectrum) is converted to a single graph structure. The proposed graph model learns from the biochemical signatures of the tissue to classify cancer versus normal tissue. The uncertainty and intermediate attentions generated by the model are visualized and explored for their association with the biochemical mechanisms of cancer.

2.1 Data Curation

Ex-vivo: Data is collected from fresh breast tissue samples from the patients referred to BCS at Kingston Health Sciences Center over two years. The study is approved by the institutional research ethics board and patients consent to be included. Peri-operatively, a pathologist guides and annotates the ex-vivo point-burns, referred to as spectra, from normal or cancerous breast tissue immediately after excision. In addition to spectral data, clinicopathological details such as the

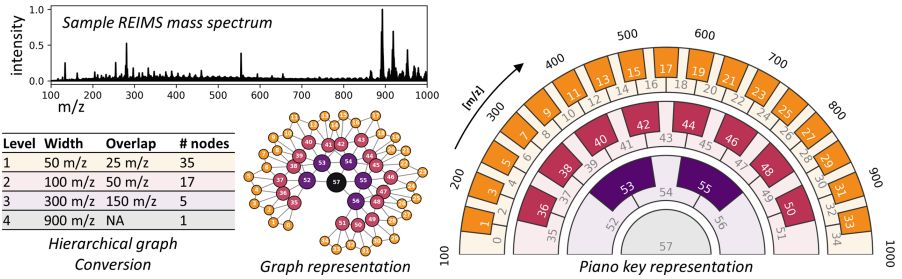

Fig. 2. Graph conversion: each spectrum is converted to a multi-level hierarchical graph. For intuitive interpretation of the process, Piano-key visualization is introduced. Each key represent a node with m/z range equal to the angular extent of the key.

status of hormone receptors is also provided post-surgically. In total 51 cancer and 149 normal spectra are collected and stratified into five folds (4 for cross validation and 1 prospectively) with each patient restricted to one fold only.

Intra-operative: A stream of iKnife data is collected during a BCS case (27 min) at Kingston Health Sciences Center. At the sampling rate of 1 Hz, a total of 1616 spectra are recorded. Each spectrum is then labeled based both on surgeons comments during the operation and post-operative pathology report.

Preprocessing: Each spectrum is converted to a hierarchical graph as illustrated in Fig. 2. The nodes are generated from a specific subband in each spectrum. Different subband widths (50, 100, 300, and 900 m/z) are used to create different levels of hierarchy (Fig. 2). The edges connect nodes with overlapping subbands within and between levels. As a result, each graph (spectrum) consists of 58 nodes and 135 edges. For details on graph conversion please refer to [2] and [11]. For easier interpretation of nodes with respect to their corresponding subbands, we visualize the graph as a Piano-key plot in Fig. 2, where each key represents a node with m/z range equal to the angular extent of the key. The dark keys show the subband overlaps between adjacent nodes.

2.2 Network Architecture and Training

Graph Transformer Network: The GTN consists of a node embedding layer, L Graph Transformer Layers (GTL), a node aggregation layer, multiple dense layers, and a prediction layer [8]. Assume a graph G with N nodes and $h_i \in R^{d\times 1}$ as node features of node i. In each GTL, the H headed attention mechanism updates the features of node i based on all neighboring node features h_j that are directly connected to node i via e_{ij} edges. The attention mechanism for node update at layer $l+1$ is formulated as:

$$w_{ij}^{k,l} = softmax_j\left(\frac{Q^{k,l}h_i^l \cdot K^{k,l}h_j^l}{\sqrt{d}}\right), \quad \hat{h}_i^{l+1} = O^l \Big\|_{k=1}^{H}\left(\sum_{j\in N_i} w_{ij}^{k,l} V^{k,l} h_j^l\right) \tag{1}$$

where $Q^{k,l}$, $K^{k,l}$, and $V^{k,l}$ are trainable linear weights. The weights w_{ij}^{kl} defines the k-th attention that is paid by node j to update node i at layer l. The concatenation of all H attention heads multiplied by trainable parameters O^l generates final attention $\hat{h}_i^{l+1}$, which is passed through batch normalization and residual layers to update the node features for the next layer. After the last GTL, features from all nodes are aggregated, then passed to the dense layers to construct a final prediction output.

Evidential Graph Transformer: Evidential deep learning provides a well-defined theoretical framework to jointly quantify classification prediction and uncertainty modeling by assuming the class probability follows a Dirichlet distribution [18]. We propose to modify the loss and prediction layer of GTN, considering the same assumption, to formulate the Evidential Graph Transformer model. Therefore, there are two mechanisms embedded in EGT: i) node-level attention calculation - via aggregation of neighboring nodes according to their relevance to the predictions, and ii) graph-level uncertainty estimation - via fitting the Dirichlet distribution to the predictions.

In the context of surgical margin assessment, the attentions reveal the relevant metabolic ranges to cancerous tissue, while uncertainty helps identify and filter data with unseen pathology. Specifically, the attentions affect the predictions by selectively emphasizing the contributions of relevant nodes, enabling the model to make more accurate predictions. On the other hand, the spread of the outcome probabilities as modeled by the Dirichlet distribution represents the confidence in the final predictions. Combining the two provides interpretable predictions along with the uncertainty estimation.

Mathematically, the Dirichlet distribution is characterized by $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_C]$ where C is the number of classes in the classification task. The parameters can be estimates as $\boldsymbol{\alpha} = f(x_i|\Theta) + 1$ where $f(x_i|\Theta)$ is the output of the Evidential Graph Transformer parameterized by Θ for each sample(x_i). Then, the expected probability for the c-th class p_c and the total uncertainty u for each sample (x_i) can be calculated as $p_c = \frac{\alpha_c}{S}$, and $u = \frac{C}{S}$, respectively, where $S = \sum_{c=1}^{C} \alpha_c$. To fit the Dirichlet distribution to the output layer of our network, we use a loss function consisting of the prediction error $\mathcal{L}_i^p$ and the evidence adjustment $\mathcal{L}_i^e$

$$\mathcal{L}_i(\Theta) = \mathcal{L}_i^p(\Theta) + \lambda \mathcal{L}_i^e(\Theta) \tag{2}$$

where λ is the annealing coefficient to balance the two terms. $\mathcal{L}_i^p$ can be cross-entropy, negative log-likelihood, or mean square error , while $\mathcal{L}_i^e(\Theta)$ is KL divergence to the uniform Dirichlet distribution [18].

2.3 Experiments

Network/Graph Ablation: We explore the hyper-parameters of the proposed model in an extensive ablation study. The attention parameters include the number of attention heads (1–15 with step size of 2) and the number of hidden features (7–14). For the evidential loss, we evaluate the choice of loss function (the 3 previously mentioned), and the annealing coefficient (5–50 with step size

of 5). The number of GTLs and dense layers are both fixed at 3. Additionally, we run ablation studies on the graph structure themselves to show the importance of presenting the data as graphs. We try randomizing the edge connections and dropping the nodes with overlapping m/z subbands.

Ex-vivo Evaluation: The performance of the proposed network is compared with 3 baseline models including GTN, graph convolution network [14], and non-graph convolution network. Four-fold cross validation is used for comparison of the different approaches, to increase the generalizability (3 folds for train/validation, test on remaining unseen fold, report average test performance). Separate ablation studies are performed for the baseline models to fine tune their structural parameters. All experiments are implemented using PyTorch with Adam optimizer, learning rate of 10^{-4}, batch size of 32, and early stopping based on validation loss. To demonstrate the robustness of the model and ensure it is not overfitting, we also report the performance of the ensemble model from the 4-fold cross validation study on the 5th unseen prospective test fold.

Clinical Relevance: Hormone receptor status plays an important role in determining breast cancer prognosis and tailoring treatment plans for patients [6]. Here, we explore the correlation of the attention maps generated by EGT with the status of HER2 and PR hormones associated with each spectrum. These hormones are involved in different types of signaling that the cell depends on [5].

Intra-operative Deployment: To explore the intra-operative capability of the models, we deploy the ensemble models of the proposed method as well as the baselines from the cross-validation study to the BCS iKnife stream.

3 Results and Discussion

Ablation Study and Ex-vivo Evaluation: According to our ablation study, hyper parameters of 11 attention heads, 11 hidden features per attention head, the cross entropy loss function, and annealing coefficient of 30, result in higher performances when compared to other configurations (370k learnable parameters). The performance of EGT in comparison with the mentioned baselines are summarized in Table 1. As can be seen, the proposed EGT model with average accuracy of 94.1% outperformed all the baselines statistically significantly (maximum p-values of 0.02 in one-tail paired Wilcoxon Signed-Rank test). The lower standard deviation of parameters shows the robustness of EGT compared to other baselines. The regularization term in EGT loss prevents overconfident estimation of incorrect predictions [18] that could lead to superior results, compared to GTN, without overfitting. Lastly, when compared to other state-of-the-art baselines with uncertainty estimation mechanisms, the proposed Evidential Graph Transformer network (average balanced accuracy of $91.6 \pm 4.3\%$ in Table 1) outperforms MC Dropout [9], Deep Ensembles [15], and Masksembles [7] ($86.1 \pm 5.7\%$, $88.5 \pm 6.8\%$, and $89.2 \pm 5.4\%$ respectively [19]).

The estimated probabilities in evidence based models are directly correlated with model confidence and therefore more interpretable. To demonstrate this,

Table 1. Average(standard deviation) of accuracy (ACC), balanced accuracy (BAC) Sensitivity (SEN), Specificity (SPC), and the area under the curve (AUC) for the proposed Evidential Graph Transformer in comparison with graph transformer (GTN), graph convolution (GCN), and non-graph convolution (CNN) baselines.

Model		ACC%	BAC%	SEN%	SPC%	AUC
Evidential Graph Transformer		**94.1(3.2)**	**91.6(4.3)**	**97.2(3.2)**	**85.9(7.9)**	**0.96(0.03)**
Graph	GT	90.4(6.3)	88.7(8.3)	92.6(6.4)	84.8(15.1)	0.96(0.05)
	GCN	91.6(4.0)	88.1(5.9)	96.3(4.0)	80.0(11.7)	0.92(0.07)
Non-graph	CNN	91.5(4.3)	87.6(7.5)	96.2(4.4)	79.0(16.1)	0.89(0.08)

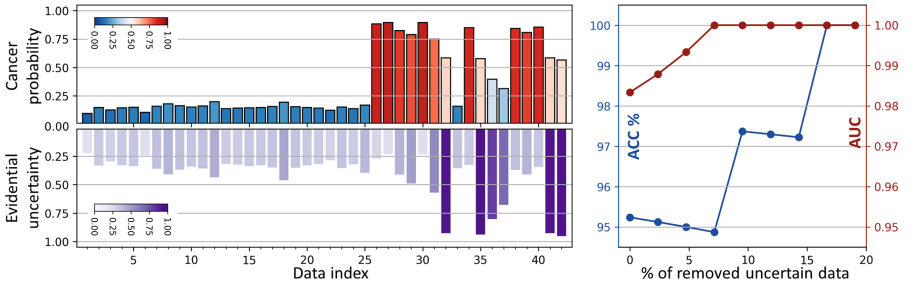

Fig. 3. *Left* Estimated probabilities and uncertainty scores for data samples in test set. *Right* Effect of uncertain data exclusion on accuracy and AUC during model deployment.

the probability of cancer predictions and uncertainty scores for all test samples are visualized in the left plot of Fig. 3. As seen, the higher the uncertainty score (bottom bar plot), the closer the estimated cancer probability is to 0.5 (top bar plot). This information can be provided during deployment to further augment surgical decision making for uncertain data instances. This is demonstrated in the right plot of Fig. 3, where the samples with high uncertainties are gradually disregarded. It can be seen that by not using the network prediction for up to 10% of most uncertain test data, the AUC increases to 1. Providing surgeons with not only the model decision but also a measure of model confidence will improve their intervention decisions. For example, if the model has low confidence in a prediction they can reinforce their decision by other means.

The result of our graph structure ablation shows the drop of average ACC to 85.6% by randomizing the edges in the graph (p-value 0.004). Dropping overlapping nodes further decreased the ACC to 82.3% (p-value 0.001). Although the model still trained due to node aggregation, random graph structure acts as noise and affects the performance. Multi-level graphs were shown to outperform other structures for masspect data [Akbarifar 2021] as they preserve the receptive field in the neighborhood of subbands (metabolites).

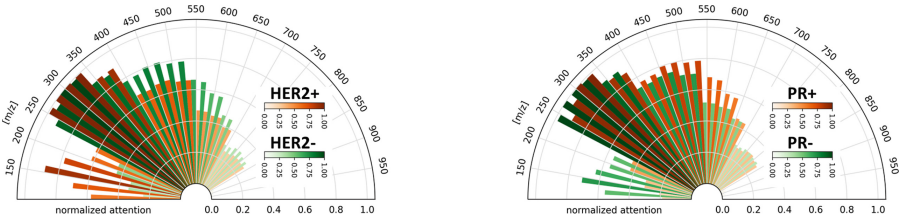

Fig. 4. Visualization of attention distribution for HER2 (*left*) and PR (*right*) hormone receptors in cancerous spectra.

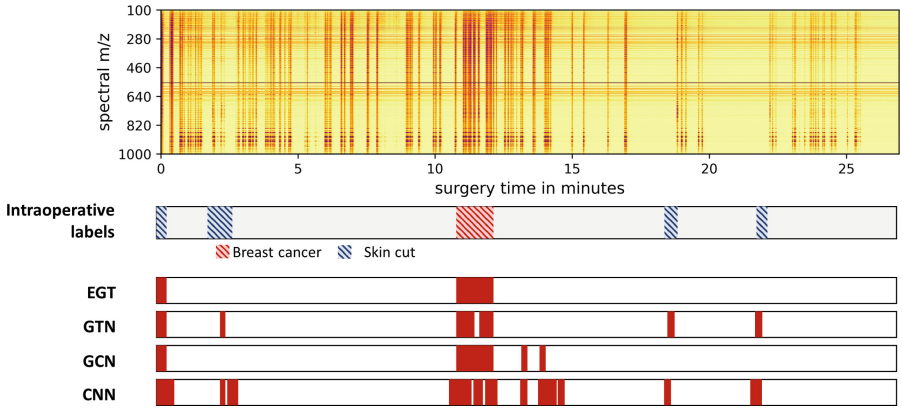

Fig. 5. Intra-operative data and label from a BCS case (top) and the temporal prediction of different ex-vivo models (bottom).

Clinical Relevance: An appropriate visualization of the attention map for samples can be used to help with this exploration. Accumulating the attentions maps from the cancerous burns based on their hormone receptor status results in the representative maps demonstrated in Fig. 4. The polar bars in this figure show the attention level paid to the nodes in the associated m/z subband. It can be seen that more attention is paid to the amino acids range (100–350 m/z) in HER2 positive breast cancer in comparison to HER2 negative breast cancer, which is in accordance with previous literature that has found evidence for higher glutamine metabolism activity in HER2+ [13]. we have also found that there's more attention in this range for PR negative breast cancer in comparison PR positive, which is in concordance with previous literature demonstrating that these subtypes have higher glutamine metabolic activity [4,5].

Intra-operative Deployment: The raw intra-operative iKnife data (y-axis is m/z spectral range and x-axis is the surgery timeline) along with the temporal reference labels extracted from surgeon's call-outs and pathology report are shown in Fig. 5, top. As seen, the iKnife stream contains spectra from skin cuts, which is considered as an unseen label for the ex-vivo models. The results

of deploying the proposed models and baselines are presented in Fig. 5, bottom. When a spectrum is classified as cancer, a red line is overlaid on the timeline. Previous studies showed the similarity between skin and breast cancer mass spectrum that can confuse the binary models. Since our proposed EGT is equipped with uncertainty estimation, this information can be used to eliminate skin spectra from being wrongly detected as cancer. By integrating uncertainty, predictions for such burns are flagged as uncertain so clinicians can compensate for surgical decision making with other sources of information.

4 Conclusion

Intra-operative deployment of deep learning solutions requires a measure of interpretability as well as predictive confidence. These two factors are particularly importance to deal with heterogeneity of tissues which represented as mixed or unseen labels for the retrospective models. In this paper, we propose an Evidential Graph Transformer for margin detection in breast cancer surgery using mass spectrometry with these benefits in mind. This structure combines the attention mechanisms of graph transformer with predictive uncertainty. We demonstrate the significance of this model in different experiments. It has been shown that the proposed architecture can provide additional insight and consequently clearer interpretation of surgical margin characterization and clinical features like status of hormone receptors. In the future, we plan to work on other uncertainty estimation approaches and further investigate the graph conversion technique to be more targeted on the metabolic pathways, rather than regular conversion.

References

1. Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L.: Graph-based deep learning for medical diagnosis and analysis: past, present and future. Sensors **21**(14), 4758 (2021). https://doi.org/10.3390/S21144758
2. Akbarifar, F., et al.: Graph-based analysis of mass spectrometry data for tissue characterization with application in basal cell carcinoma surgery. In: SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 11598 (2021)
3. Balog, J., et al.: In vivo endoscopic tissue identification by rapid evaporative ionization mass spectrometry (REIMS). Angewandte Chemie Int. Ed. **54**(38), 11059–11062 (2015). https://doi.org/10.1002/anie.201502770
4. Budczies, J., et al.: Glutamate enrichment as new diagnostic opportunity in breast cancer. Int. J. Cancer **136**(7), 1619–1628 (2015). https://doi.org/10.1002/ijc.29152
5. Demas, D.M., et al.: Glutamine metabolism drives growth in advanced hormone receptor positive breast cancer. Front. Oncol. **9** (2019). https://doi.org/10.3389/fonc.2019.00686
6. Dunnwald, L.K., Rossing, M.A., Li, C.I.: Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. **9**(1), R6 (2007). https://doi.org/10.1186/bcr1639

7. Durasov, N., Bagautdinov, T., Baque, P., Fua, P.: Masksembles for uncertainty estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13539–13548 (2021)
8. Dwivedi, V.P., Bresson, X.: A generalization of transformer networks to graphs. In: Methods and Applications, AAAI Workshop on Deep Learning on Graphs (2021)
9. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA, 20–22 June 2016
10. Hargreaves, A.C., Mohamed, M., Audisio, R.A.: Intra-operative guidance: methods for achieving negative margins in breast conserving surgery. J. Surg. Oncol. **110**(1), 21–25 (2014). https://doi.org/10.1002/JSO.23645
11. Jamzad, A., et al.: Graph transformers for characterization and interpretation of surgical margins. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 88–97. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_9
12. Jsang, A.: Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer, Cham Verlag (2016). https://doi.org/10.1007/978-3-319-42337-1
13. Kim, S., Kim, D.H., Jung, W.H., Koo, J.S.: Expression of glutamine metabolism-related proteins according to molecular subtype of breast cancer. Endocrine-Related Cancer **20**(3), 339–348 (2013). https://doi.org/10.1530/ERC-12-0398
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations. ICLR (2017)
15. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Santilli, A., et al.: Domain adaptation and self-supervised learning for surgical margin detection. Int. J. Comput. Assist. Radiol. Surg. 1–9 (2021). https://doi.org/10.1007/s11548-021-02381-6
17. Santilli, A., et al.: Self-supervised learning for detection of breast cancer in surgical margins with limited data. In: Proceedings - International Symposium on Biomedical Imaging, April 2021, pp. 980–984, April 2021. https://doi.org/10.1109/ISBI48211.2021.9433829
18. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
19. Syeda, A.: Self-supervision and uncertainty estimation in surgical margin detection (2023)