Deployment of deep models for intra-operative margin assessment using mass spectrometry

Amoon Jamzad
School of Computing
Queen's University
Kingston, ON, Canada
a.jamzad@queensu.ca

Laura Connolly
School of Computing
Queen's University
Kingston, ON, Canada

Fahimeh Fooladgar
Department of Electrical
and Computer Engineering
University of British Columbia
Vancouver, BC, Canada

Martin Kaufmann
Department of Surgery
Queen's University
Kingston, ON, Canada

Kevin Yi Mi Ren
Department of Pathology
and Molecular Medicine
Queen's University
Kingston, ON, Canada

Shaila Merchant
Department of Surgery
Queen's University
Kingston, ON, Canada

Jay Engel
Department of Surgery
Queen's University
Kingston, ON, Canada

Sonal Varma
Department of Pathology
and Molecular Medicine
Queen's University
Kingston, ON, Canada

Purang Abolmaesumi
Department of Electrical
and Computer Engineering
University of British Columbia
Vancouver, BC, Canada

Gabor Fichtinger
School of Computing
Queen's University
Kingston, ON, Canada

John Rudan
Department of Surgery
Queen's University
Kingston, ON, Canada

Parvin Mousavi
School of Computing
Queen's University
Kingston, ON, Canada

Abstract

Real-time margin assessment in breast cancer surgeries is critical to reduce positive margin rates. The iKnife is an intra-operative modality that captures the molecular signature of tissues and can be paired with AI to facilitate real-time tissue characterization. As training these AI models is typically done with homogeneous ex-vivo iKnife data, intra-operative deployment is challenging because of tissue heterogeneity and unseen classes. In this study, we explore different mechanisms to address the intra-operative deployment challenge. Using cross validation and comparison to baseline methods, we show that the intermediate attention of graph transformer model as well as the uncertainty estimation of Bayesian neural network can be used to to reduce false positive rate of breast cancer surgery. We conclude that the class prediction output is not enough for successful deployment and additional interpretability features are needed to improve the performance.

1 Introduction

It is estimated that 1 in 8 women will develop breast cancer in their lifetime [1]. Breast conserving surgery (BCS) is a preferred treatment option for breast cancer as it minimizes healthy tissue loss. Complete cancer resection in BCS is critical to prevent cancer recurrence and is quantified by histopathology assessment of the resection margins of the surgical specimen, post-operatively. It is

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

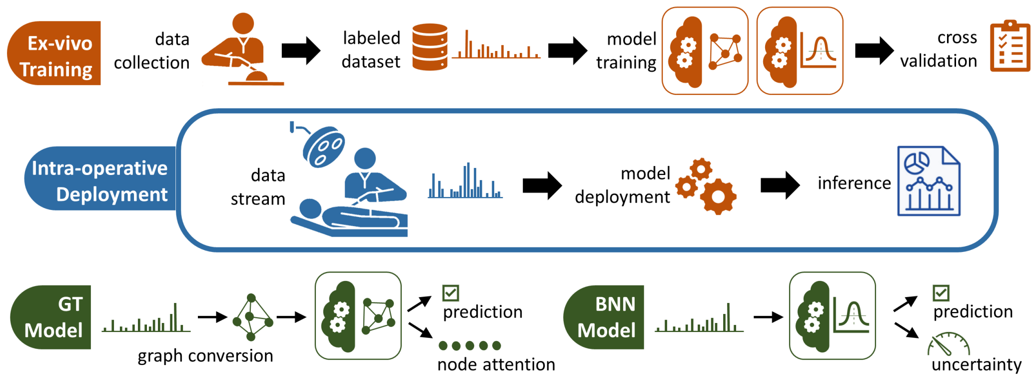

Figure 1: Overview of the methodology for training models on ex-vivo dataset, and deploying the trained model intra-operatively.

estimated that as many as 35% of patients who undergo BCS end up with positive margins, which indicates that cancer was left behind [9]. Various imaging and metabolomic-sensing modalities can be used intraoperatively to augment the surgeon's ability to identify tumor boundaries and reduce positive margins. The intelligent knife (iKnife) is one such modality, composed of a surgical electrocautery device, and a smoke evacuation tube that is attached to a mass spectrometer. The surgical byproduct (smoke) is used to capture the metabolomic signature of tissue in real-time [3]. It has been demonstrated that deep learning models can be used on iKnife data to characterize tissue, as either healthy or cancerous [8][2][10][7][5].

Similar to other applications in computer-assisted surgery, mainly ex-vivo datasets are used to train deep learning models from iKnife data. The mass spectra are collected using precise point burns from homogeneous regions of surgically removed specimens, under the supervision of a pathologist. The class label of spectra are further validated according to the order and location of the point burns by histopathology assessment. This collection protocol results in datasets with clean samples/labels, which facilitate training and exploration of the underlying molecular mechanisms. Deploying these methods intra-operatively is more challenging due to tissue heterogeneity and the presence of unseen pathology during surgery (e.g. skin and vessles in BCS). These data points can be considered out of distribution (OOD) and can disrupt the performance of deep models during intra-operative deployment. Therefore, it is critical to use models that benefit from learning from ex-vivo data but are also equipped with mechanisms to handle unseen intra-operative labels.

In this paper, we explore the ex-vivo (quantitative) and intra-operative (qualitative) performances of two approaches for iKnife margin assessment. In addition to class predictions, these models generate secondary outputs (intermediate attention and estimated uncertainty, respectively) to increase the interpretability of the results and hence, improve intra-operative deployment and utility.

2 Methodology

Dataset Figure 1 shows an overview of the proposed methodology. As shown in the figure, two set of data are used in this study. The ex-vivo dataset is collected from resected breast tissues extracted in BSC and consists of 41 cancer and 103 normal spectra with pathology validated labels. The intra-operative data is a consecutive iKnife recording of a full BSC case for around 27 minutes at a sampling rate of 1Hz (1616 spectra). Each spectrum is associated with the time sequence of the operation and labelled qualitatively based on call-outs from the surgeon and matched with pathology reporting. Each spectrum undergoes standard preprocessing steps including lockmass correction, normalization, and binning, and is truncated to a range of mass-to-charge ratio (m/z) of 100-1000.

Model Training As mentioned, we use and evaluate two deep models for margin evaluation using mass spectrometry data that provide the uncertainty of the predictions, as well as identify the attention mechanisms of the models. We demonstrate that these complementary approaches have the potential to enable intraoperative deployment of the deep models: Graph transformers and Bayesian Neural

Table 1: Cross-validated performance of models on ex-vivo dataset

Model	Balanced Acc	Sensitivity	Specificity	AUC
GT	**91.1 ± 4.5**	85.7 ± 9.4	96.6 ± 4.6	0.95 ± 0.03
CNN (baseline)	84.1 ± 9.9	72.2 ± 18.9	96.0 ± 6.1	0.94 ± 0.04
BNN	**90.3 ± 7.5**	87.6 ± 8.3	93.0 ± 8.2	0.95 ± 0.04
MLP (baseline)	87.4 ± 6.8	77.1 ± 12.5	97.8 ± 4.2	0.93 ± 0.05

Networks. *Graph transformers* are graph-based deep models that utilize a multi-head (parallel) attention mechanism in each layer to improve model stability, performance, and interpretability [4]. Here, the intermediate attention paid to graph nodes are considered as a secondary output of the model to improve intra-operative deployment. In this study, a 3-layer GT with 10 attention heads and 10 hidden features is used. To be able to use this model on our data, each spectrum is converted to an individual graph. In this graph representation, each node contains ions from a specific spectral subband and is connected to other nodes with overlapping subbands. Different subband widths are used for nodes to create a multi-level hierarchical graph [7]. *Bayesian neural networks* use Bayesian inference to predict the distribution of the output [6]. By leveraging the Monte-Carlo dropout and placing a Gaussian distribution over the output layer of the network, BNN is capable of estimating prediction uncertainty in addition to the class [6]. These per-sample uncertainty values can be used during deployment to filter out low confidence predictions. For this study, a BNN with 3 fully connected layers (128, 64, and 2 units) is implemented.

Experiments For *ex-vivo training*, the dataset is first divided into 4 folds, preserving cancer to normal ratio per fold and restricting all data from each patient to one fold only. The GT and BNN models are then trained and validated in a 4-fold cross-validated scheme. In addition, two baselines including a convolutional network (CNN) and a dense network (MLP) are also cross-validated and the average and standard deviation of the performance metrics are monitored for comparison. After optimizing the ex-vivo prediction performance, the models are used for *intra-operative deployment*. First, only the model predictions are visualized and compared with the reference labels created from surgeon callouts. Then, the secondary outputs (node attentions for GT and uncertainty for BNN) and their effect on deployment performance are explored. While for BNN the predictions with low confidence are dropped, we propose a node dropout for GT that drops the nodes that are attracting attentions from the uncertain spectral subbands.

3 Results and discussions

The ex-vivo performance of models and baselines are summarized in Table 1. As can be seen, both proposed models outperform the baselines. The improvement in balanced accuracy is statistically significant (all p-values<0.001). The model with highest test performance (among cross-validation configuration) is selected for the intra-operative deployment. The reference surgery callout for the intraoperative case is visualized in Fig 2 *A*. In addition to healthy and cancer breast, the skin was also cut in some instances of the surgery, which is OOD for the ex-vivo models. The direct prediction results of BNN and GT models are shown in Fig 2 *B* and *D*, respectively. As can be seen, there are different false positive instances where normal breast or skin burns are detected as positive margin. By discarding the prediction instances whose uncertainty is high, the intra-operative performance of BNN is improved as shown in Fig 2 *C*. For the GT model, the attention patterns show that the nodes covering m/z 650-750 are activated mainly when the skin is cut. By dropping out these nodes from graphs during deployment, the false positive range improves as shown in Fig 2 *E*.

4 Conclusion and Future Work

Training medical deep models with ex-vivo data is practical as the data is more readily available, homogeneous, and clean. However, deployment of these models in intra-operative setups can be challenging due to the presence of unseen label, which cannot be addressed solely by class prediction. We showed that additional mechanism like attention and uncertainty estimation can improve the performance of intra-operative deployment. Particularly, the GT and BNN models improve false

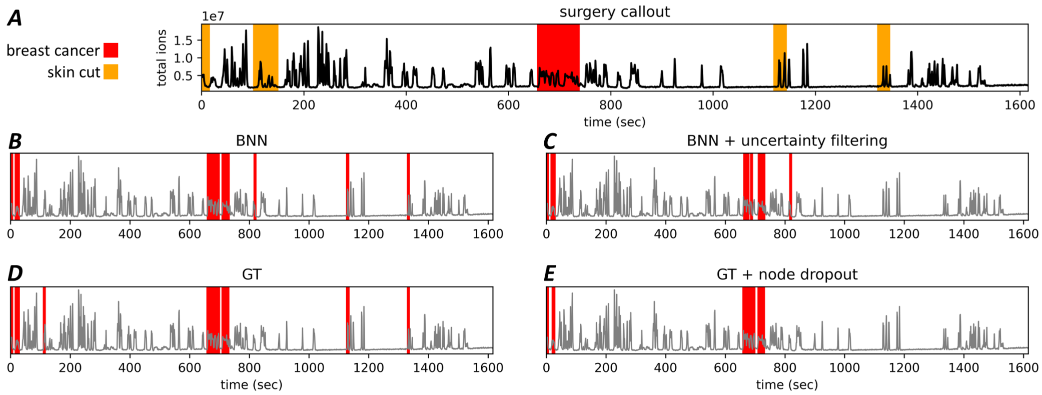

Figure 2: Intra-operative model deployment. Qualitative labeling of breast surgery based on surgeon's callouts (A) Predictions of BNN model before (B) and after uncertainty-based filtering (C). Predictions of GT model before (D) and after attention-based node dropout (E).

positive rates for BCS. our future aim is to explore more OOD and uncertainty compatible models for surgical margin assessment.

Statement on potentially negative Social Impact

Both attention mechanism and uncertainty estimation presented here can improve the interpretability of deep models. This becomes more crucial for clinical application where the additional transparency directly improves the clinical decision making. However, patient variability can negatively impact the interpretation of these mechanisms. The iKnife modality is a new technology with currently limited access worldwide. Extensive experiments and more patient recruitment is needed to eliminate patient-variability bias in decision making and to ensure the positive impact for all.

Acknowledgments and Disclosure of Funding

We would like to thank the following sources of funding: Natural Sciences and Engineering Council of Canada (NSERC), the Canadian Institute for Health Research (CIHR), Britton Smith Chair in Surgery to J. Rudan, and Canada Research Chair to G. Fichtinger.

References

[1] World cancer research fund. *American Institute for Cancer Research*, pages https://www.wcrf.org/dietandcancer/breast–cancer.

[2] Faranak Akbarifar, Amoon Jamzad, Alice Santilli, Martin Kauffman, Natasja Janssen, Laura Connolly, Kevin Y M Ren, Kaitlin Vanderbeck, Ami Wang, Doug Mckay, John Rudan, Gabor Fichtinger, and Parvin Mousavi. Graph-based analysis of mass spectrometry data for tissue characterization with application in basal cell carcinoma surgery. *SPIE Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, 11598, 2021.

[3] Júlia Balog, László Sasi-Szabó, James Kinross, Matthew R Lewis, Laura J Muirhead, Kirill Veselkov, Reza Mirnezami, Balázs Dezső, László Damjanovich, Ara Darzi, Jeremy K Nicholson, and Zoltán Takáts. Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Science Translational Medicine*, 5(2):194, 2013.

[4] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[5] Fahimeh Fooladgar, Amoon Jamzad, Laura Connolly, Alice Santilli, Martin Kaufmann, Kevin Ren, Purang Abolmaesumi, John F Rudan, Doug McKay, Gabor Fichtinger, and Parvin Mousavi.

Uncertainty estimation for margin detection in cancer surgery using mass spectrometry. *Int. J. Comput. Assist. Radiol. Surg.*, September 2022.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[7] Amoon Jamzad, Alice Santilli, Faranak Akbarifar, Martin Kaufmann, Kathryn Logan, Julie Wallis, Kevin Ren, Shaila Merchant, Jay Engel, Sonal Varma, Gabor Fichtinger, John Rudan, and Parvin Mousavi. Graph Transformers for Characterization and Interpretation of Surgical Margins. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12907 LNCS:88–97, sep 2021.

[8] Amoon Jamzad, Alireza Sedghi, Alice M.L. Santilli, Natasja N.Y. Janssen, Martin Kaufmann, Kevin Y.M. Ren, Kaitlin Vanderbeck, Ami Wang, Doug Mckay, John F. Rudan, Gabor Fichtinger, and Parvin Mousavi. Improved resection margins in surgical oncology using intraoperative mass spectrometry. *Medical Image Computing and Computer Assisted Intervention, MICCAI Lecture Notes in Computer Science vol 12263. Springer, Cham*, pages https://doi.org/10.1007/978–3–030–59716–0_5, 2020.

[9] Benjamin W. Maloney, David M. McClatchy, Brian W. Pogue, Keith D. Paulsen, Wendy A. Wells, and Richard J. Barth. Review of methods for intraoperative margin detection for breast conserving surgery. *Journal of Biomedical Optics*, 23:1, 2018.

[10] Alice M L Santilli, Amoon Jamzad, Alireza Sedghi, Martin Kaufmann, Kathryn Logan, Julie Wallis, Kevin Y M Ren, Natasja Janssen, Shaila Merchant, Jay Engel, Doug McKay, Sonal Varma, Ami Wang, Gabor Fichtinger, John F Rudan, and Parvin Mousavi. Domain adaptation and self-supervised learning for surgical margin detection. *International Journal of Computer Assisted Radiology and Surgery*, 10.1007/s11548-021-02381-6:1–9, 2021.