Improved Resection Margins in Surgical Oncology Using Intraoperative Mass Spectrometry

Amoon Jamzad[1(✉)], Alireza Sedghi[1], Alice M. L. Santilli[1], Natasja N. Y. Janssen[1], Martin Kaufmann[2], Kevin Y. M. Ren[3], Kaitlin Vanderbeck[3], Ami Wang[3], Doug McKay[4], John F. Rudan[4], Gabor Fichtinger[1], and Parvin Mousavi[1]

[1] School of Computing, Queen's University, Kingston, ON, Canada
a.jamzad@queensu.ca
[2] Department of Medicine, Queen's University, Kingston, ON, Canada
[3] Department of Pathology, Queen's University, Kingston, ON, Canada
[4] Department of Surgery, Queen's University, Kingston, ON, Canada

Abstract. PURPOSE: Incomplete tumor resections leads to the presence of cancer cells on the resection margins demanding subsequent revision surgery and poor outcomes for patients. Intraoperative evaluations of the tissue pathology, including the surgical margins, can help decrease the burden of repeat surgeries on the patients and healthcare systems. In this study, we propose adapting multi instance learning (MIL) for prospective and intraoperative basal cell carcinoma (BCC) detection in surgical margins using mass spectrometry. METHODS: Resected specimens were collected and inspected by a pathologist and burnt with iKnife. Retrospective training data was collected with a standard cautery tip and included 63 BCC and 127 normal burns. Prospective data was collected for testing with both the standard and a fine tip cautery. This included 130 (66 BCC and 64 normal) and 99 (32 BCC and 67 normal) burns, respectively. An attention-based MIL model was adapted and applied to this dataset. RESULTS: Our models were able to predict BCC at surgical margins with AUC as high as 91%. The models were robust to changes in cautery tip but their performance decreased slightly. The models were also tested intraoperatively and achieved an accuracy of 94%. CONCLUSION: This is the first study that applies the concept of MIL for tissue characterization in perioperative and intraoperative REIMS data.

Keywords: Surgical margin detection · Multiple instance learning · Rapid evaporative ionization mass spectrometry · Intraoperative tissue characterization · Non-linear analysis · Basal Cell Carcinoma

A. Jamzad, A. Sedghi and A.M.L. Santilli—Joint first authors.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12263, pp. 44–53, 2020.
https://doi.org/10.1007/978-3-030-59716-0_5

1 Introduction

A main step in the clinical management of major cancers includes surgical resection of the tumor. Incomplete resection of tumors and the presence of cancer cells at the resection margins, otherwise known as "positive margins", often demands repeat surgery [10]. In some cancers, such as breast cancer, the positive surgical margin rates can be as high as 20% [4]. The subsequent revision surgeries burden the health care system with extra costs and wait times. They can also affect the cosmetic outcome of the patient, causing distress and potentially delaying life saving treatments such as chemotherapy or radiation therapy. Intraoperative evaluation of the tissue, including surgical margins, is currently a challenging task. Recent efforts have resulted in innovative perioperative and intraoperative technologies that can assess tissue in a high throughput manner, and provide surgeons with critical information on tissue pathology. The Intelligent Knife, iKnife (Waters Corp., MA), a mass spectrometry-based technology, is one such modality [2]. This technology is able to provide enriched feedback about the chemical properties of the tissue at the surgical tooltip, in real time [7,9,11,12]. The smoke created by the surgical electrocautery device is used and its molecular profiles such as lipids, fatty acids and small molecules, are analyzed through rapid evaporative ionization mass spectrometry (REIMS) [5]. iKnife can be seamlessly integrated into surgical workflows, as REIMS does not require sample preparation [13], and only a connection to the exhaust of the electrocautery device is needed for molecular profiling of the tissue.

Due to the destructive nature of electrocautery that generates the smoke for iKnife, pathology validated labels are difficult to attain. In practice, the histopathology of the surrounding tissue to a "burn" is analyzed and the burn is labelled based on an educated estimate of a pathologist. Since the data labels are not conclusively determined, they are referred to as weak labels. The problem of weakly annotated data is common to pathology where images and reports are either grossly outlined, or annotations of collected data are vague. Introduced two decades ago, multiple instance learning (MIL) [3], is a strategy for dealing with weakly labeled data. Here, a single label is assigned to a "bag" of multiple instances. The bag label is positive if it contains at least one positive instance, and negative otherwise. Using bags with different proportions of positive instances, MIL methods learn signatures of positive instances. As weak annotations result in noisy instance labels, considering a *bag* of instances, rather than each individually, helps compensate for the effect of the weak labels. It is important to identify instances that play a prominent role in predicting the overall label of a bag. This is referred to as the *attention* of an instance-[8]. Recently, attention-based MIL has been used with deep learning for whole-slide annotation of histopathology images of breast and colon cancer [6].

In this paper, for the first time, we propose to extend the concept of attention-based MIL to learn from weakly labelled REIMS data for detection of perioperative surgical margins. To create surgical smoke, the iKnife burns the tissue in contact with the tool tip. The data created in a constant stream lends itself well to the concept of *bags*. Each mass spectrum from a burn is considered as an

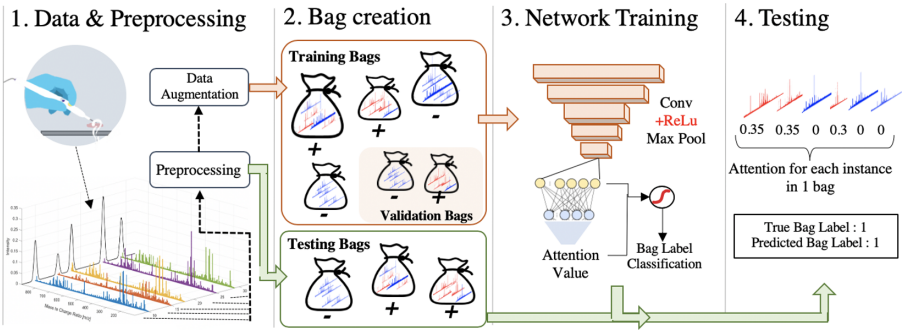

Fig. 1. Overview: Molecular profiles of the cautery smoke aspirated at the tip of the iKnife is preprocessed and augmented. The data is randomly divided into *bags* which are labelled based on instances they contain. MIL models are trained for margin classification using the training and validation bag sets. Using the trained model, the test bags are then predicted and evaluated. (Color figure online)

instance, and the stream of data from multiple burns are packed in bags. The prediction of a positive bag is, hence, an indication of the presence of positive margins. We demonstrate the accuracy of our models in prospective peroperative data, and their robustness to changes in surgical cautery tips. Finally, we investigate the feasibility of our developed approach to real time tissue typing, intraoperatively. Our methods are presented in the context of surgical margin detection for Basal Cell Carcinoma (BCC). BCC is the most commonly diagnosed cancer with a worldwide incidence rate of 2.75 million cases, and low metastasis rate [14]. Therefore, it is an ideal application for evaluation of our proposed surgical margin detection approach. Methods built on BCC data, can be translated to other cancers where surgical margins are crucial to patient outcomes.

2 Materials and Methods

Figure 1 depicts an overview of the proposed workflow. The cautery smoke aspirated at the tip of iKnife is collected for each specimen. Its spectra are selected, labelled, preprocessed, and augmented. The data is then represented as bags of instances and used for training of a deep model that is capable of predicting the bag label as well as the attention values of each instance of input spectra.

2.1 Data

Data was collected from 65 patients in 8 surgical clinic days, over a period of 10 months. Patients were recruited from the skin clinic at our institution, according to a protocol approved by the institutional HREB. BCC lesions presented on patients' head, neck or back. The suspected BCC region was first outlined by

the surgeon on the skin and then resected. The resected specimen was inspected by a pathologist for perioperative point based data collection. Point burns were acquired from a cross section of the specimen containing BCC, by contact of the cautery tip with the tissue. Each burn was labelled by a derma-pathologist based on a visual inspection of the tissue at the location of the burn. A standard cautery tip was used for most of the data acquisition. To increase the specificity of the burn location, a non-insulated fine tip was used for some of the resected specimens. In Sect. 2.5, we describe how the data is divided for training and prospective testing. Experiments were performed in a controlled environment with the same operator, pathologist and surgeon at every clinic. An external lock mass of leucine enkephalin (1 ng/µl) was used for iKnife calibration (mass in negative-ion mode m/z 554.2615). The electrocautery was used on cut mode with the generator at 35 W. A sample iKnife recording consisting of five burns is shown in step 1 of Fig. 1. The chromatogram, in black, represents the total ion current (z axis), recorded over the acquisition time (x axis). Each peak in the chromatogram represents a burn. From each burn, the scan with the highest signal to noise ratio is chosen as the representative of that burn. Each scan is a mass spectral profile where the ion count, a measure of intensity, is plotted along ion mass charge ratios m/z. Five mass spectral scans are also shown in color.

2.2 Preprocessing

Using the Offline Model Builder (OMB), a Waters Corp. software platform, each scan was individually processed by normalizing, lock mass correcting and binning the intensity values. REIMS typically ionizes small molecules with mass to charge ratio of less than m/z 1200. Previous literature has reported that the majority of the total ion current to be present below m/z 900 [2]; we determined that 85% of the total ion current was between m/z 100–900 in our data. Therefore, we focused on this range for further analysis. Max binning was performed on this region with a bin size of 0.1, meaning that for a spectrum, the maximum intensity value for a m/z bin of size 0.1 is chosen to represent that range. For the range of m/z 100–900, each spectra is represented by 8000 peaks. To reduce the number of trainable parameters in the final model, we further applied max pooling with window and stride size of 10 to reduce the number of peaks (features of the spectra), to 800.

2.3 Intensity-Aware Augmentation

To avoid overfitting models to training data, it is essential to have a large number of data samples. This is not always clinically feasible. We propose a new data augmentation method for REIMS spectra that uses the inherent calibration error and background noise to create new data. First, a random shift sampled from a uniform distribution is added to the location of each peak in a spectrum. To increase the variability between multiple augmentations from the same spectrum, the shift range is also selected randomly from 0 to 3 bin widths. Next, random high-frequency Gaussian noise, multiplied by a random spline-smoothed

low-frequency envelope is generated to mimic background noise. The standard deviation of the Gaussian is randomly selected proportional to the standard deviation of original spectrum. The generated noise is only added to the low-intensity peaks in the data while for the high-intensity a different Gaussian noise with half the standard deviation of the initial one is added. This intensity-aware approach ensures that the inherent molecular signatures and peak-ratios of the spectrum are preserved during augmentation.

2.4 MIL Model and Attention Mechanism

In the formulation of the MIL, a bag of instances is defined as $X = \{x_1, ..., x_n\}$, where the instances are not ordered or related to one another, and the size of the bag n may vary. Every instance in a bag has an individual binary label y, where $X_{labels} = \{y_1, ..., y_n\}$. Each bag is also assigned a single binary label $Y = max(X_{labels})$. Positive bags, $Y = 1$, have one or more instances of the target class while negative bags, $Y = 0$, have none. Considering the goal of margin classification for BCC, all BCC spectra are labeled 1 as the target class. In our architecture, similar to [6], we use a weighted average of instances where the weights are determined by a neural network and indicate each instance's attention value. The weights sum up to 1 and are invariant to the size of the bag. The proposed structure allows the network to discover similarities and differences among the instances. Ideally, instances within a positive bag that are assigned high attention values are those most likely to have label $y_i = 1$. The attention mechanism allows for easy interpretation of the bag label predictions by the models.

Every bag is fed to the network with all of its instances. The overall structure of the attention based MIL network consists of convolutional layers followed by fully connected dense layers. As visualized in step 3 of Fig. 1, every bag is passed through 5 convolutional layers (kernel size of 10) of 10, 20, 30, 40 and 50 filters. ReLu activation and max pooling is performed between every convolution. The final array is then flattened and passed through two dense layers of size 350 and 128 to get the final attention for each instance. The first dense layer and the generated attention weights are then combined in final layer and using Sigmoid activation the prediction of the bag label is outputted.

To define the sensitivity of the bag labels, we explored the minimum number of cancerous instances that would be required in a bag during training to learn the distinction between BCC and normal burns. Sweeping 2 parameters, we adjusted mean length of the bag, between 3 and 10 (standard deviation of 1), as well as the maximum involvement of cancerous instances in the positive bags between 0.1 and 0.8. This created 64 trained models and their performances are discussed in Sect. 3.1.

2.5 Experiments

To evaluate the models, we used the data from the first four clinic days as retrospective set and stratified them into 5 training/validation folds, all collected

Table 1. Table displaying the division of data into training/validation before augmentation and the separate testing folds.

	Fold	#BCC Burns	#Normal Burns	**Total Burns**	#Patients
Train	fold-1	14	24	**38**	9
	fold-2	6	30	**36**	7
	fold-3	14	20	**34**	5
	fold-4	12	25	**37**	6
	fold-5	17	28	**45**	7
Test	standard cautery	66	64	**130**	17
	fine tip cautery	32	67	**99**	11

using the standard cautery tip. The division of this data and the separation of the test sets is displayed in Table 1. The complete training set consisted of 63 BCC and 127 normal scans from 34 patients. All scans from a particular patient were kept within the same fold. Before augmentation, each fold contained approximately 1:2 ratio of cancer to normal spectra. Employing the proposed data augmentation technique, the size of each cross-validation fold increased to around 500 BCC and 500 normal scans.

For all of the experiments, the training folds were converted to a collection of 600 bags (300 negative and 300 positive bags). Bags were randomly formed in a way that an instance within a fold may be placed in more than one bag, but no two bags may have the same combination of instances. An ensemble of the 5 models was used to predict the labels of test data. Each model in the ensemble used 4 folds of the data for training and one for validation. The final label of a test bag was predicted by averaging the bag probabilities over these five models.

The data collected from clinic days 5 through 8 was used to generate two prospective test sets. The first test set contained burns collected with standard cautery tip. The second set contained burns collected with the fine cautery tip.

Intraoperative: Intraoperative data was collected from patients recruited similarly to the others in the study. During intraoperative resection, the surgeon only uses the standard cautery blade. The iKnife was connected to the cautery and smoke was collected throughout the procedure. To assess the feasibility of real time deployment and the performance of our model, an intraoperative case of neck lesion removal with continuous cut duration of 1.5 min was selected. The data was processed similar to perioperative burns and was used to the test our trained models.

3 Results and Discussion

3.1 Model Performance

The 64 trained models were tested on the prospective datasets of 500 bags generated with same parameters as their training equivalents. The AUC for each

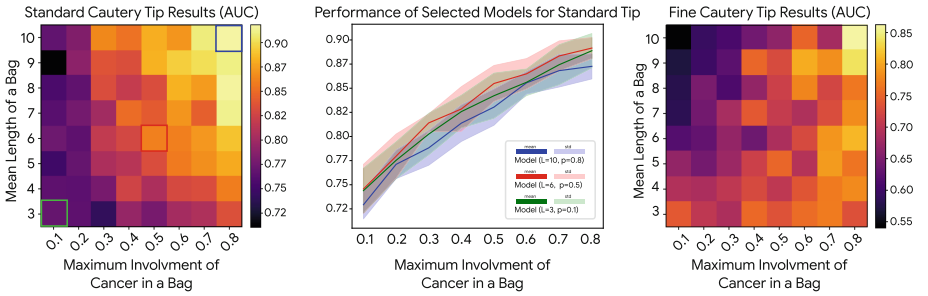

Fig. 2. *Left:* Exploring the two parameters of mean bag length and maximum involvement of cancerous instances in a positive bag, 64 models were trained on the augmented standard tip data. These models were tested on 500 standard tip bags and the mean AUC of the ensemble models are displayed. Metrics from the 3 outlined models can be seen in Table 1. *Middle:* Three outlined models from the left colormap, (L = length, p = max involvement), tested on data with mean length of 7 and standard deviation of 3 with varying max cancer involvement. *Right:* Same trained models from the left graph but tested on 500 bags of data from fine tip cautery. Mean AUC of the ensemble models are displayed.

model is visualized in the color maps of Fig. 2. The left color map displays the 64 models tested against the standard cautery tip test set. This figure demonstrates that as the bag length and cancer involvement increases, the models better learn the underlying patterns of the instances. Three models were selected (outlined) from the left color map and their performance metrics including accuracy, specificity, sensitivity, and AUC are listed in Table 2.

To mimic the uncertainty of intraoperative predictions, we tested the three selected models on bags of greater variability. All of the test bags in this case had a mean length of size 7 with a standard deviation of 3, therefore ranging from bags of size 1 to 15. The AUC, seen in the middle figure of Fig. 2, has the same upwards trend with proportion of cancerous instances as in the left colormap, even with the increased complexity of the test set. Finally, to determine the robustness of the model against potential changes in the input spectra, we examined the performance against the test dataset collected with the fine tip cautery. The fine tip was used during data collection to ideally be more precise with our burns and therefore attain better pathology guided labels. However, we do not conclusively know how the tip change may affect the signal recorded and therefore the ability of a model to perform. Trained on the same augmented standard cautery tip as both previous experiments, the right colormap in Fig. 2 visualizes the results of this experiment. The AUC trend is similar to that of the prospective standard cautery tip. However, the overall AUC is lower for most models suggesting that there may be a difference in the recorded mass spectral signal between the two tips.

As a comparison to baseline, we also implemented a standard MIL model known as multi instance support vector machine (mi-SVM). This model was

Table 2. Performance metrics of the 3 highlighted models from the left graph of Fig. 2, with comparisons to the mi-SVM baseline model [1]. Each model was tested with the prospective test set taken with the standard cautery tip. To reduce the sensitivity of the results to bag selection, each evaluation was performed 10 times with 500 randomly generated test bags.

	Accuracy	Sensitivity	Specificity	AUC
MIL (L = 3, p = 0.1)	0.72 ± 0.01	0.53 ± 0.02	0.92 ± 0.01	0.76 ± 0.02
MIL (L = 6, p = 0.5)	0.81 ± 0.01	0.71 ± 0.03	0.90 ± 0.01	0.86 ± 0.01
MIL (L = 10, p = 0.8)	0.81 ± 0.02	0.90 ± 0.01	0.72 ± 0.03	0.91 ± 0.01
mi-SVM (L = 3, p = 0.1)	0.69 ± 0.02	0.52 ± 0.04	0.85 ± 0.02	0.77 ± 0.02
mi-SVM (L = 6, p = 0.5)	0.71 ± 0.02	0.53 ± 0.06	0.89 ± 0.03	0.83 ± 0.01
mi-SVM (L = 10, p = 0.8)	0.75 ± 0.02	0.51 ± 0.03	0.99 ± 0.01	0.91 ± 0.01

presented by Andrews *et al.* in 2003 [1], based on an alternative generalization of the maximum margin idea used in SVM classification. The goal of mi-SVM, is to find both the optimal instance labelling as well as the optimal hyperplane. The performance of this baseline on the same 3 models and set of test bags is also listed in Table 2.

To demonstrate the true weakness of our labels, we performed a supervised method of principal component and linear discriminant analysis, PCA/LDA. Using the same training set, a PCA/LDA model was trained and tested on each individual scan in the standard cautery testing set. This linear approach performed at an accuracy of 75.7%. Although this method cannot be compared directly to our results as it does not utilize the bagging approach, it demonstrated the drop in performance when trying to use the instance level labels for direct classification.

3.2 Attention

Ideally a positive bag alert would root from a positive instance being given a high attention value. In a practical setting, this would alert the surgeon of a positive margin. To quantify the performance of the attention network, we evaluated the accuracy of correctly placing the highest attention value in a positive bag on a positive instance. Using the model with the highest AUC on the standard cautery tip test data (mean bag length of 10 and maximum cancerous involvement of 0.8), we were able to reach attention accuracy of 0.88 ± 0.04 on 500 test bags.

3.3 Intraoperative Trial

We demonstrated the applicability of our model on intraoperative data. Before deploying the model in the operating room, we wanted to evaluate its performance and sensitivity. The intraoperative case selected was comprised of 83 scans, including burns and no-burns spectra acquired continuously. Prospective

pathology validation of the excised specimen labelled all of the margins negative, implying the absence of BCC in the scans. We created bags using a sliding window of size 10 to mimic bag creation from a continuous stream of data. For a model trained on bags with a mean size of 10 and positive bag's cancer portion of 0.8, the test on intraoperative data resulted an accuracy of 94% with a standard deviation of 6%.

4 Conclusion

In this study we adapted the concept of attention-based MIL to REIMS data analysis for perioperative margin evaluation for the first time. The framework consisted of preprocessing, intensity-aware augmentation, instance/bag representation of mass spectrmetry data, and model training. Training on retrospective BCC data, the performance of models with different bagging parameters was investigated on prospective data collected with standard and fine tip cautery blades. The feasibility of using the trained model on itraoperative data for margin assessment was also demonstrated. For future work, we plan to acquire more fine tip data to investigate the effect of transfer learning on improving the model predictive power. In practice, we are also looking at adaptive bag length selection during intraoperative data stream using the chromatogram signal. Another challenge to address is the presence of non-burn spectra, recorded during time intervals where the surgeon is not burning any tissue, along with burn signals intermixed in the interaoperative data stream. Implementation of a real-time burn screening algorithm to disregard the non-burn periods will increase the model accuracy in margin detection.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems 15, pp. 577–584. MIT Press (2003)
2. Balog, J., et al.: Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. Sci. Transl. Med. **5**(194), 194 (2013)
3. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1), 31–71 (1997)
4. Fisher, S.L., Yasui, Y., Dabbs, K., Winget, M.D.: Re-excision and survival following breast conserving surgery in early stage breast cancer patients: a population-based study. BMC Health Serv. Res. **18**, 94 (2018)
5. Genangeli, M., Heeren, R., Porta Siegel, T.: Tissue classification by rapid evaporative ionization mass spectrometry (REIMS): comparison between a diathermic knife and CO2 laser sampling on classification performance. Anal. Bioanal. Chem. **411**, 7943–7955 (2019)
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 2127–2136 (2018)

7. Kinross, J.M., et al.: iKnife: rapid evaporative ionization mass spectrometry (REIMS) enables real-time chemical analysis of the mucosal lipidome for diagnostic and prognostic use in colorectal cancer. Cancer Res. **76**(14 Suppl.), 3977 (2016)
8. Liu, G., Wu, J., Zhou, Z.H.: Key instance detection in multi-instance learning. In: Asian Conference on Machine Learning, vol. 25, pp. 253–268 (2012)
9. Marcus, D., et al.: Endometrial cancer: can the iknife diagnose endometrial cancer? Int. J. Gynecol. Cancer **29**, A100–A101 (2019)
10. Moran, M.S., et al.: Society of surgical oncology-American society for radiation oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer. J. Clin. Oncol. **32**(14), 1507–1515 (2014)
11. Phelps, D.L., et al.: The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). Br. J. Cancer **118**(10), 1349–1358 (2018)
12. St John, E.R., et al.: Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. Breast Cancer Res. **19**(59) (2017)
13. Strittmatter, N., Jones, E.A., Veselkov, K.A., Rebec, M., Bundy, J.G., Takats, Z.: Analysis of intact bacteria using rapid evaporative ionisation mass spectrometry. Chem. Commun. **49**, 6188–6190 (2013)
14. Verkouteren, J., Ramdas, K., Wakkee, M., Nijsten, T.: Epidemiology of basal cell carcinoma: scholarly review. Br. J. Dermatol. **177**(2), 359–372 (2017)