Development of a Research Testbed for Intraoperative Optical Spectroscopy Tumor Margin Assessment

David Morton, Laura Connolly, Leah Groves, Kyle Sunderland, Tamas Ungi, Amoon Jamzad, Martin Kaufmann, Kevin Ren, John F. Rudan, Gabor Fichtinger, and Parvin Mousavi

Queen's University, 99 University Ave, K7L 3N6, Kingston, Canada

16djm10@queensu.ca, 15lpc1@queensu.ca, lg94@queensu.ca, 1krs1@queensu.ca, ungi@queensu.ca, aj90@queensu.ca, kaufmann@queensu.ca, kevin.ren@kingstonhsc.ca, john.rudan@kingstonhsc.ca, fichting@queensu.ca, mousavi@queensu.ca

Abstract: Surgical intervention is a primary treatment option for early-stage cancers. However, the difficulty of intraoperative tumor margin assessment contributes to a high rate of incomplete tumor resection, necessitating revision surgery. This work aims to develop and evaluate a prototype of a tracked tissue sensing research testbed for navigated tumor margin assessment. Our testbed employs diffuse reflection broadband optical spectroscopy for tissue characterization and electromagnetic tracking for navigation. Spectroscopy data and a trained classifier are used to predict tissue types. Navigation allows these predictions to be superimposed on the scanned tissue, creating a spatial classification map. We evaluate the real-time operation of our testbed using an ex vivo tissue phantom. Furthermore, we use the testbed to interrogate ex vivo human kidney tissue and establish a modeling pipeline to classify cancerous and non-neoplastic tissue. The testbed recorded latencies of 125 ± 11 ms and 167 ± 26 ms for navigation and classification respectively. The testbed achieved a Dice similarity coefficient of 93%, and an accuracy of 94% for the spatial classification. These results demonstrated the capabilities of our testbed for the real-time interrogation of an arbitrary tissue volume. Our modeling pipeline attained a balanced accuracy of 91% ± 4% on the classification of cancerous and non-neoplastic human kidney tissue. Our tracked tissue sensing research testbed prototype shows potential for facilitating the development and evaluation of intraoperative tumor margin assessment technologies across tissue types. The capacity to assess tumor margin status intraoperatively has the potential to increase surgeon confidence in complete tumor resection, thereby reducing the rates of revision surgeries.

Keywords: computer-assisted intervention; tumor margin assessment; intraoperative optical spectroscopy; tissue conserving surgery; machine learning

1 Introduction

Surgical intervention is a common approach for treating early-stage cancers [1]. However, the visual distinction between healthy and cancerous tissue can be challenging, necessitating reliance on preoperative planning and imaging for interventional guidance. While various tissue detection technologies show promise, there is currently no effective, accurate, and real-time standard of care for intraoperative tumor margin assessment.

There are several emerging technologies that differentiate tissues by measuring properties such as chemical composition, acoustic interactions, and electromagnetic interactions. An example of a chemical tissue characterization technique is rapid evaporative ionization mass spectrometry which uses an electrocautery device to vaporize tissue and a mass spectrometer to measure the chemical composition of the vapor [2]. An example of an acoustic tissue characterization technique is temporally enhanced ultrasound which uses a multi-second series of raw radiofrequency ultrasound data to measure the acoustic properties of a tissue [3].

Spectroscopy-based technologies are among these emerging technologies. Spectroscopy is a non-invasive technology that electromagnetically characterizes tissue by observing its interactions with various wavelengths of light. Tissues with different chemical compositions and physical structures will reflect, absorb, and transmit light differently, resulting in a unique electromagnetic profile for each tissue type. Various spectroscopy technologies have shown promise for use in tumor margin assessment. P. Gao et al. (2017) reviewed the clinical applications of Raman spectroscopy for the detection of breast cancer [4]. D. Mojahed et al. (2020) used optical coherence tomography to assess breast margins post lumpectomy [5]. J. Wang et al. (2023) evaluated the impact of radiofrequency spectroscopy in breast-conserving surgery and found a reduction in the rate of revision surgeries [6] [7]. L. De Boer et al. (2015) showed that diffuse reflection spectroscopy held promise in the detection of breast tumor boundaries [8].

Rapid acquisition times combined with the non-invasive, and repeatable nature of diffuse reflection broadband optical spectroscopy (DRS) makes it a promising candidate for both in vitro and ex vivo margin assessment. DRS relies on 3 main components: a broadband light source, an optical fiber reflection probe, and a spectrometer. A broadband light source is used to illuminate a tissue sample in a desired wavelength band. In optical spectroscopy, this is a subset of the optical spectrum, composed of the visible, near-ultraviolet, and near-infrared bands. The optical fiber reflection probe is used to transmit the radiated light to the tissue surface. When the light interacts with the tissue, the reflected intensity at each wavelength is dependent on the electromagnetic properties of the tissue. In the case of cancerous and benign tissues, each will reflect and absorb light in varying proportions, depending on the wavelength. The light is collected by the optical fiber probe and transmitted to the spectrometer. Frequency-specific diffraction gratings are used to separate the light into component wavelengths [9]. The result is a

broadband spectrum that electromagnetically characterizes the tissue. The non-invasive, repeatable, and rapid nature of spectroscopy makes it a promising technology for use in intraoperative tumor margin assessment.

While there are many point-based optical technologies with adequate acquisition times and promising tumor margin assessment capabilities, they are currently limited in their intraoperative usability. Practical deployment of such technologies necessitates the development of software to process their signals and provide informative and actionable insights to a surgeon in real time. Additionally, many of these point-based methods do not leverage spatial location information, which is important for visualizing and navigating to detected cancerous tissue for resection. We see promise in the development of a tracked tissue sensing testbed for the rapid testing and deployment of intraoperative technologies. We address these gaps in this paper through the development and evaluation of a research testbed for real-time navigated tumor margin assessment. The paper is organized accordingly. Section 2 outlines the design and implementation of our research testbed. Section 3 outlines the evaluation of real-time testbed performance using a biological ex vivo tissue phantom. Section 4 outlines the application of the research testbed for the interrogation of human kidney tissue. Section 5 provides a summary of the key conclusions of this work.

2 Design and Implementation of Research Testbed

This section presents our prototype for a tracked tissue-sensing research testbed. It provides an overview of the system's physical architecture and the software module we developed. Section 2.1 reviews the physical design detailing the system architecture and hardware components. Section 2.2 details the specifics of our developed software module that facilitates hardware interaction, data collection, and real-time operation.

2.1 Testbed Architecture

Our tracked tissue sensing testbed is designed for intraoperative tumor margin assessment. Enabling a surgeon to perform a freehand or assisted scan of the tissue surface post-resection. This includes the ex vivo inspection of tumor margins, as well as the in vitro inspection of the tumor bed. The latter is visualized in Figure 1.

The testbed design consists of four main components: a tissue sensor for tissue characterization, a position tracker for recording sensor pose, a classifier for predicting tissue type, and a navigation computer for processing and presenting information in an intuitive graphical representation. Figure 1 provides an overview of the testbed design.

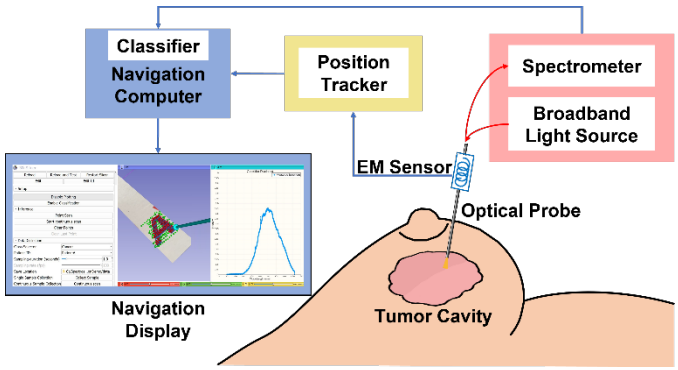

Figure 1
A diagram of the tracked tissue sensing testbed

The tissue sensor, represented in red in Figure 1, is a diffuse reflection optical spectroscopy probe. The probe consists of an SLS201L stabilized tungsten broadband light source (Thorlabs, USA), a CCS200 compact charge-couple-device spectrometer (Thorlabs, USA), and an FCR-7UVIR200-2 optical fiber reflection probe (Aventes, NL). Light is reflected off the tissue and measured using a spectrometer. The light is separated into individual wavelengths, characterizing the tissue. The tissue sensor operates in the range of 420 nm to 1000 nm. The position tracker, depicted in yellow in Figure 1, is an electromagnetic tracking system. The position tracker used is the Northern Digital Inc. (NDI) trakSTAR 3D Guidance system (NDI, USA), with the mid-range transmitter (NDI, USA), and Model 800 6-degree of freedom (DoF) 8mm sensors (NDI, USA). An electromagnetic sensor is rigidly attached to the spectroscope, allowing for precise determination of the probe position and orientation in 3D space. A trained machine learning classifier is employed within the software to predict the tissue type based on the acquired spectrum.

The navigation software, depicted in blue in Figure 1, is responsible for communicating sensor information to the user in an intuitive manner. It enables the collection, storage, processing, classification, and display of hardware information.

2.2 Software

We develop an open-source software module in 3D Slicer to facilitate real-time hardware communication, streamline data collection of novel tissue types, and enable real-time spatial classification with an intuitive data visualization system (Figure 2).

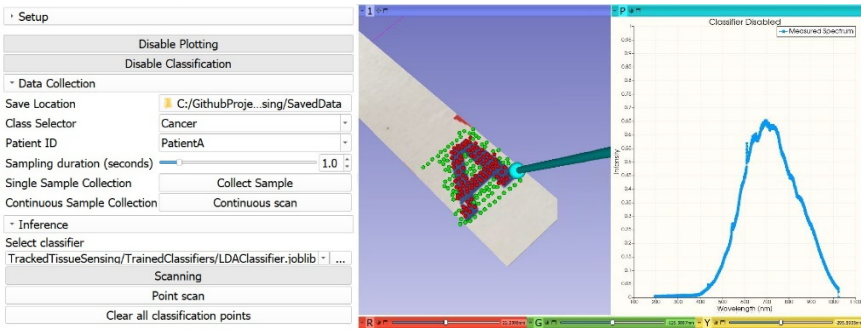

Figure 2

An overview of the developed software module. (Left) The graphical user interface. (Centre) The 3D navigation display. Red and green points represent predicted classes, and a blue model displays the position and orientation of the optical probe. (Right) The spectrum viewer. Displays a real-time visualization of the optical spectra.

2.2.1 Real-Time Hardware Communication

To facilitate hardware communication, we leverage the open-source software Plus and OpenIGTLink protocol [10] [11]. The communication pipeline is visualized in Figure 3. Hardware information is collected in real time using the Plus device interface. The spectrometer data is formatted as a 2D matrix, where the first row represents the wavelengths, and the second row denotes the corresponding observed intensity at each wavelength. The navigation data is formatted as a homogeneous transformation matrix, providing information about the location and orientation relative to a reference sensor. Finally, OpenIGTLink is employed to allow our 3D Slicer module to access the data in real time.

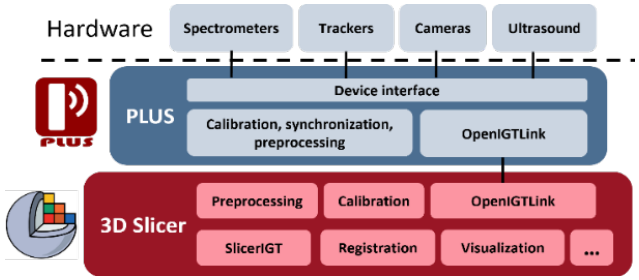

Figure 3

An overview of the real-time hardware communication pipeline

2.2.2 Data Collection

A software interface was created to facilitate novel tissue interrogation and the collection of labeled datasets for rapid training of machine learning classification models (Figure 2). The spectrum viewer, located on the right-hand side of Figure 2, has been developed to provide real-time visualization of the incoming spectra. This feature enables the assessment of signal quality and ensures the acquisition of a strong and clear signal, minimizing noise and artifacts. Additionally, the spectrum viewer offers a preliminary exploration of new tissue features, allowing for early insight for later analysis. If a machine learning model is uploaded and active, the title of the chart is dynamically set based on the predicted classification, offering immediate context and interpretation of the investigated tissue type. The user interface, located on the left-hand side of Figure 2 offers various labeling and data collection parameters. The class selector enables the user to specify the histopathology or tissue type being scanned. Patient ID allows the assignment of the scan to a specific patient, with anonymization according to a reference sheet. Sample ID differentiates between samples when multiple tissue segments are obtained from a single patient. Continuous sample collection mode allows continuous collection of spectra from a single class, useful for scanning large, homogenous tissue samples. Single sample collection mode in conjunction with sampling duration, enables scanning of a single location for a consistent duration, useful for heterogeneous samples.

2.2.3 Navigation and Spatial Classification

The final software objective is the development of a navigation display to facilitate an intuitive, real-time visual representation of classification predictions in 3D space. An example navigation display is shown in the center of Figure 2. A blue cylindrical model represents the probe location and orientation. A static image of the scanned area is registered to the scene. The location and classification data are combined to generate a point with color denoting class. For example, locations predicted as cancer are colored red, and locations predicted as benign are colored green. The user can toggle between scanning a single point or performing a continuous scan. The user can input the system path to import a custom-trained machine-learning classifier. When classification is enabled, the module observes incoming spectra, applies relevant data processing, and inputs each spectrum into the model for prediction. The model prediction is passed to the navigation display for visualization. The navigation display requires calibration before operation to ensure accurate visualization. The electromagnetic sensor is fixed to the handle of the probe, rather than the tip of the probe directly. The location of the probe tip is calibrated using the Pivot Calibration module in 3D Slicer. An overhead image of the region of interest is then registered to the 3D Slicer scene. Four reference points are marked on the region of interest and visible in the overhead image. The calibrated probe is placed on each reference point and the position is recorded.

The Fiducial Registration Wizard module is used to align the points on the imported image with their physical position in the tracker space.

3 Evaluation of Testbed Functionality

The purpose of this section is to evaluate the real-time operation of our tracked tissue sensing testbed. We measure the latency of the navigation and classification portions of the testbed's navigation display. We then perform a freehand scan of an ex vivo biological tissue phantom to observe the testbed's spatial classification capabilities. The tissue phantom is composed of porcine and bovine tissue.

3.1 Experimental Setup

To operate the testbed, we required a trained tissue classifier. As the classification performance was not the primary focus of this experiment a simplified classifier was trained and deployed. The classifier used in this experiment was a k-Nearest-Neighbor (kNN) model with a k value of 3. To train the kNN, bovine and porcine tissue were acquired and prepared for sampling. The surface of each tissue was scanned and labeled, which resulted in a dataset of 200 spectra, 100 bovine, and 100 porcine. Each spectrum was cropped to 360-1000 nm and a min-max normalization was performed. A binary label of 0 and 1 was assigned to bovine and porcine tissue spectra respectively. The trained kNN classifier was imported into the testbed.

3.2 Testbed Latencies

Before evaluating the spatial classification performance of the testbed, the latency of the navigation and classification was experimentally measured. We determined its suitability for real-time use and ensured temporal consistency between the classification and location data. The classification and navigation latencies were determined using a 240-frame-per-second slow-motion camera to record the experimental workspace and navigation display. To determine the classification latency, the probe was then repeatedly transitioned between two solid colors for 20 cycles. The slow-motion camera was analyzed frame by frame to approximate the delay between the probe reaching the color, and the display displaying the class. To determine the electromagnetic tracking latency the probe was repeatedly lifted from a rigid surface and returned for 20 cycles. The slow-motion video was analyzed, and the frames were counted between when the physical probe and virtual probe model stopped moving.

3.3 Spatial Classification Performance

To evaluate the spatial classification performance of the testbed, an ex vivo tissue phantom was created using bovine and porcine tissue to simulate a tumor cavity. The phantom layout is visible in Figure 4. Porcine tissue is used as the main phantom body with bovine tissue inserted to represent a tumor. The phantom was prepared fresh and frozen immediately. The tissue phantom measured approximately 10 cm in diameter and 3 cm in depth. Before experimentation, the phantom was removed from the freezer and allowed to thaw completely.

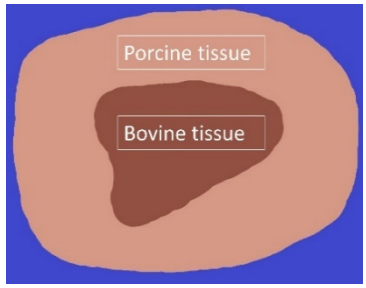

Figure 4
An illustration of the biological tissue phantom design used in this experiment

The thawed tissue phantom was placed in the workspace as shown in Figure 5. An overhead optical image of the tissue phantom was taken prior to data collection for use as the ground truth image. The optical probe was calibrated, and the ground truth image was registered to the coordinate system. The trained classifier was inputted into the testbed, and a continuous data collection mode was selected. The tissue phantom was scanned freehand in a rough grid pattern with 1 mm spacing between displayed points. The spatial classification was overlaid on the ground truth image of the tissue phantom. The number of correctly and incorrectly classified points was recorded for each class. The testbed visualization was evaluated using the Dice similarity coefficient and accuracy.

Figure 5
Experimental workspace for the testbed evaluation using an ex vivo biological tissue phantom

3.4 Results

This section details the results of our testbed evaluation process in terms of testbed latency and spatial classification performance.

3.4.1 Testbed Latency Results

The latency of the navigation and classification components of the navigation display were determined to be 125 ± 11 ms and 167 ± 26 ms respectively (Table 1).

Table 1
The testbed latencies for the navigation and classifier portions of the display visualization. The mean and standard deviation is reported in ms and is calculated over 20 trials.

Navigation latency [ms]	125 ± 11
Classification latency [ms]	167 ± 26

A study by Xu et al. (2013) assessed the effects of latency on real-time tool manipulation in a dV-Trainer simulator environment [12]. System latencies were progressively increased in 100 ms increments from ~ 0 to 1000 ms. They concluded that latencies of <200 ms facilitated easy tool manipulation, with latencies <300 ms deemed safe by all participants [12]. Our testbed achieved this threshold for both the classification network and the navigation latencies. This result also allows for improved temporal alignment of the position and classification data streams through an estimated offset, increasing the accuracy of the navigation display. The effect of this increased alignment will be particularly prevalent when either the probe location or tissue class is changing rapidly. Overall, the results indicate that our testbed architecture is capable of real-time classification, localization, and visualization. Facilitating quick and accurate assessment of tissue margins.

3.4.2 Spatial Classification Results

The output of our spatial classification experiment is displayed in Figure 7. The red points are classified as bovine tissue, and the green points are classified as porcine tissue. The spatial classification is overlaid on the ground truth image of the tissue phantom. Our testbed achieved a Dice similarity coefficient of 93%, and an accuracy of 94%. The confusion matrix can be seen in Figure 6.

Our spatial classification results demonstrate potential for the real-time classification and visualization of a 3D tissue surface. The visualization in Figure 7 met basic qualitative metrics, and the system's modularity enables task-specific enhancements. The testbed effectively generated an intuitive visualization of the predicted classes, with the navigation display aligning well qualitatively. Red and green points were primarily located in their respective ground truth regions, although minor discrepancies were observed along tissue boundaries and throughout the navigation display. The presence of false negatives and positives within bulk

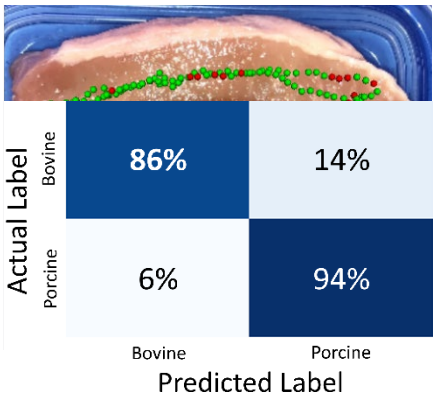

Figure 6
Confusion matrix for the spatial classification of bovine and porcine tissue [16]

tissue regions were most likely due to the simplified classifier and training regimen used. Discrepancies at the tissue boundaries were most likely a result of various tracking and registration errors within the testbed. The modularity of the testbed allows it to adapt to specific task requirements. The tissue sensor, tracking system, and classifier can be modified according to the procedure requirements. For instance, if a procedure requires higher navigation accuracy, the trakSTAR (NDI, USA) electromagnetic tracker could be replaced with a Plus-compatible optical tracker.

The tissue classifier is interchangeable and can be tailored for a specific application, allowing for rapid experimentation. The system would greatly benefit from the exploration of more complex classification models to enable the robust analysis of diverse tissue representations with increased classification accuracy. Potential avenues of exploration include deep learning methods such as convolutional neural networks (CNNs), and transformer-based networks. Additionally, the system could benefit from further methods to minimize classification errors. A robust data preprocessing pipeline where extraneous information is filtered out has the potential to improve performance. The exploitation of domain-specific knowledge would also be beneficial. For example, it is unlikely that a single cancerous detection will be completely isolated and surrounded by healthy tissue as seen in the bulk scan in Figure 7. Isolated cancerous detections a distance from the main tumor body may be filtered out as false positives, culminating in a much cleaner display.

Overall, the results show that our tracked tissue sensing testbed is able to accurately interrogate tissue samples with real-time navigation and classification. Further work focuses on the adoption of more comprehensive classification pipelines for the application of the testbed to clinically relevant tissue types.

4 Application of Testbed for Interrogation of Human Kidney Tissue

This section focuses on the application of our testbed to interrogate cancerous and non-neoplastic human kidney tissue. Section 4.1 details the procedures used for the collection and analysis of ex vivo human kidney data. Section 4.2 presents a modeling pipeline designed for the classification of cancerous and non-neoplastic tissue. It details the model used, the preprocessing methods experimented with, and the evaluation methods used. Section 4.3 details the results of our experimentation and analysis.

4.1 Data Collection

Three fresh human kidney tissue samples were collected from a single patient, denoted Patient A. The tissue samples remained unfixed and were stored at -80°C. The samples were heterogeneous and were composed of cancerous and non-neoplastic regions of an excised kidney. The front and back of the tissue samples are shown in Figure 8. Each tissue sample was approximately 25 mm by 25 mm by 5 mm in size. The experiment was performed in a biosafety cabinet to minimize contamination.

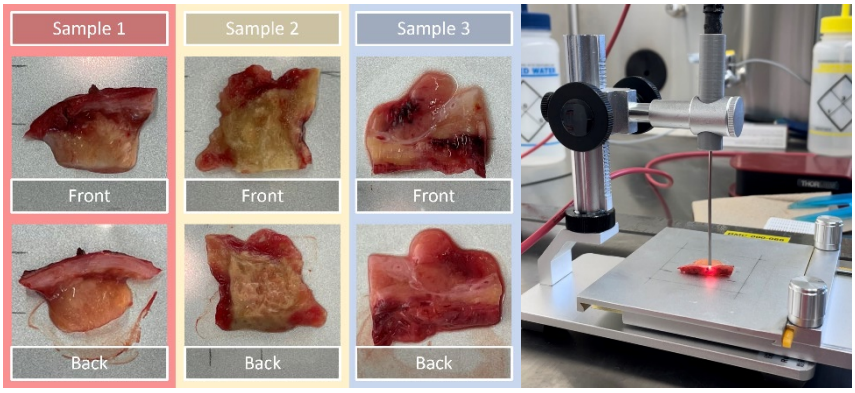

Figure 8
(Left) Images of the resected ex vivo human kidney tissue specimens (Right) The experimental setup for the interrogation of tissue specimens

The experimental setup can be seen in Figure 8. The setup consisted of the tissue sensor, a 2-DoF stage, and a 1-DoF stand. The data collection process is as follows. A sample was removed from the freezer and thawed completely for 20 minutes. The sample was placed in the center of the stage. An overhead image was taken of the specimen. The image was manually segmented by a clinician. The data collection module was used in single-point collection mode with an acquisition time of 1 second. Due to time constraints associated with drying tissue, the light source was not manually toggled before each reading. Reflected ambient light was instead recorded at multiple locations on each pathological region before the scan to estimate the ambient light at a later time. Each pathological region of the tissue sample was scanned in a grid pattern. The sample was allowed to rehydrate in saline before the second side was scanned.

Figure 9 displays the resulting ground truth segmentations of the cancerous and non-neoplastic tissue, as visually determined by a trained pathologist. The green and red overlays represent non-neoplastic and cancerous tissue respectively. The white overlay denotes regions that were not scanned due to ambiguity. The resulting dataset contained 315 cancerous and 84 non-neoplastic unique spectra. The associated patient ID, sample ID, scan side, and histopathological metadata were recorded.

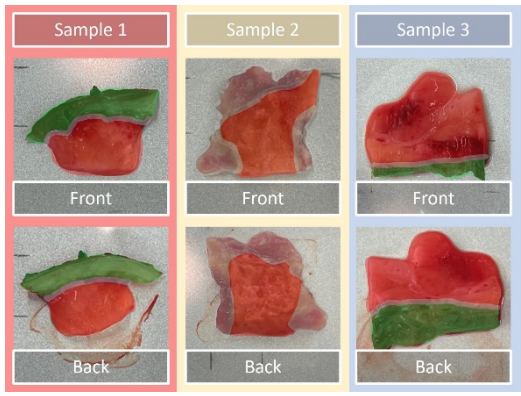

Figure 9
Approximate regions of cancerous (red) non-neoplastic (green) and ambiguous (white) tissue determined visually by a trained pathologist

4.2 Tissue Classification Pipeline

We present a modeling pipeline designed to train a classifier to accurately differentiate between cancerous and non-neoplastic spectra. We detail the model used, the preprocessing methods implemented, and the methods used to evaluate the pipeline performance.

4.2.1 Model

We conducted an ablation study involving three distinct machine learning techniques: a Linear Discriminant Analysis (LDA) classifier, a k-Nearest-Neighbors (kNN) classifier, and a Support Vector Machine (SVM). LDA aims to find a linear combination of features that maximizes the separation between different classes while minimizing the variance within each class. kNN classifies a data point based on the majority class among its k nearest neighbors in the feature space. SVM calculates a hyperplane in the feature space that best separates the classes. The classification performance of each model is evaluated with baseline data processing applied to determine the optimal model for use in the remainder of the experiment. Machine learning models were used exclusively in this study. The limited size of our dataset was a confounding factor in this selection. Machine learning can be used with less data compared to deep learning methods, which contain a large number of trainable parameters.

4.2.2 Preprocessing

We explored the effects of various preprocessing methods on classification performance. These methods include min-max normalization, spectroscope signal normalization, ambient light compensation, and dimensionality reduction.

Min-max normalization

When conducting a scan of a tissue surface, the distance between the probe and the tissue is variable. Min-max normalization was introduced to normalize the intensity changes that result from this variability.

Spectroscope signal normalization

The light source and spectrometer chosen during data collection will have unique output intensity curves and receiver sensitivity curves across the interrogated wavelength range. This information is imposed on the dataset during data collection. Spectroscope signal normalization was introduced to ensure the dataset is invariable to the choice of spectroscope hardware. To normalize the dataset, each spectrum undergoes division by the receiving sensitivity curve and the output intensity curve. The spectrometer manual specifies an amplitude-corrected sensitivity curve for our operating band[1]. Thus, the theoretical receiving curve is assumed to be flat. The broadband light source manual specifies a theoretical output intensity curve[2].

Ambient light compensation

The ambient light contains information specific to the data collection environment, as well as probe-to-surface distance information. The ambient light within a signal should be removed to improve the robustness of the dataset to these factors. We interrogated two methods of ambient light compensation: ambient light estimation (ALE) and ambient peak zeroing (APZ). The ALE technique uses spectra collected prior to a tissue scan, with the light source on and off, to estimate the ambient light reflection profile. The ambient light profile is scaled for each incoming spectrum using the ratio of the ambient light peak to the broadband signal peak in the spectrum. The scaled ambient light profile is then subtracted from the spectrum. To limit overprocessing, we calculated a general ambient light profile for the dataset rather than for each scan. This ensured no scan-specific or pathological information was embedded into the spectra. The APZ technique looked to estimate the location of the peak ambient light and to eliminate it from each spectrum. The prerecorded spectra were analyzed, and the ambient light spectra were extracted. The location of the peak ambient light was determined, and the associated wavelengths were zeroed with a buffer of ∼±3 nm around the peak.

Dimensionality reduction

Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Binning, were introduced to address the disparity between the size of the dataset, 399 spectra, and the dimensionality of the dataset, 2578 features. The dimensionality reduction was chosen using 0.9999 explained variance within the PCA as guidance. The number of components varied for each fold, with 80 to 140 features explaining approximately 0.9999 explained variance. To ensure direct

[1]Documentation available: www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=3482
[2]Documentation available: www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=7269

comparability between dimensionality reduction methods, the number of bins in the Binning method, and the number of components in the PCA method were fixed at 100 for all folds. We used a PCA implementation with the number of components set to 100. For Binning, we calculated 100 wavelength bins by dividing the length of the signal by 100. For each bin, we found the mean signal intensity for the wavelength range. We repeated this process for all remaining bins to produce a lower-resolution signal.

4.2.3 Pipeline Evaluation

We evaluated the performance of the modeling pipeline using an intrinsically determined 4-fold cross-validation based on sample side data groupings[3]. We separated the front and back of each scan into 6 separate groups. Sample 1 and Sample 2 were strategically combined to ensure each fold contained cancerous and non-neoplastic tissue.

This resulted in 4 folds consisting of Sample 1 Front/Sample 2 Front, Sample 1 Back/Sample 2 Back, Sample 3 Front, and Sample 3 Back. To perform cross-validation, a single fold was chosen as the test set, and the remaining folds were combined for the train set. This process was repeated until each fold was chosen as the test set. An overview of the train and test sets are shown in Table 2. To address the significant class imbalance in the dataset, random oversampling was performed during model training. The metrics reported to evaluate model performance were balanced accuracy, sensitivity, specificity, F1 score, area under the receiving operating curve (AUC), and accuracy. The mean and standard deviation were reported over the test data. The experiment was averaged over 28 trials to account for any statistical outlier performances.

Table 2

An overview of the 4-fold-cross-validation regimen. S denotes Sample, and [ ] denotes grouped data

Train set	Test set
[S1Front, S2Front], [S1Back, S2Back], S3Front	S3Back
[S1Front, S2Front], [S1Back, S2Back], S3Back	S3Front
[S1Front, S2Front], S3Front, S3Back	[S1Back, S2Back]
[S1Back, S2Back], S3Front, S3Back	[S1Front, S2Front]

[3] Splitting the data strictly by the tissue sample number theoretically provides the most robust results. This method maximally preserves statistical variation between the training and testing sets while minimizing signal contamination. However, the size of our dataset limits its application. The inter-sample variation in pathological presentation greatly hinders the classification performance, and the size of our dataset limits the generalization of the model between samples. Thus, leave-one-sample-out was not feasible for this experiment.

4.3 Results

This section details the results of our tissue classification pipeline for the differentiation of cancerous and non-neoplastic kidney tissue. We performed an ablation study of various state-of-the-art machine learning models including a k-Nearest-Neighbor classifier, a Linear Discriminant Analysis classifier, and a Support Vector Machine (Table 3). We performed an ablation study of the preprocessing methods including spectroscope signal normalization, ambient light compensation, min-max normalization, and dimensionality reduction (Table 4). The evaluation was performed using 4-fold cross-validation. The balanced accuracy, sensitivity, specificity, F1 score, AUC, and accuracy are reported. The ambient light compensation methods are labeled ALE (Ambient light estimation) and APZ (Ambient peak zeroing). The min-max normalization is denoted MM.

Table 3

The results of our classifier ablation study using baseline processing methods. The mean ($\pm$ standard deviation) balanced accuracy, sensitivity, specificity, F1, AUC, and accuracy are reported (%).

Experiment	Bal Acc	Sens	Spec	F1	AUC	Acc
SVM	63 ± 10	73 ± 31	89 ± 11	75 ± 15	63 ± 10	67 ± 14
kNN	73 ± 7	77 ± 21	91 ± 9	82 ± 9	73 ± 7	74 ± 11
LDA	$\mathbf{80 \pm 13}$	$\mathbf{89 \pm 17}$	$\mathbf{92 \pm 9}$	$\mathbf{89 \pm 8}$	$\mathbf{80 \pm 13}$	$\mathbf{84 \pm 12}$

In our evaluation of various machine learning classification models, the Linear Discriminant Analysis (LDA) method demonstrated superior performance compared to other approaches. This statistical significance was established through a one-tailed Wilcoxon rank-sum test, with maximum p-values of 0.007 for the balanced accuracy measure. The LDA method projects the data onto a lower dimensional latent space while maximizing the separation between classes. This approach optimizes the ratio of between-class variance to within-class variance, enhancing data clustering within the latent space. Moreover, projecting to a lower-dimensional feature space may improve the handling of noisy features compared to alternative methods. We use the LDA classifier for the remainder of the experiments.

Table 4

The results of our preprocessing experimentation using a 4-fold sample-side-based cross-validation regimen. The mean ($\pm$ standard deviation) balanced accuracy, sensitivity, specificity, F1, AUC, and accuracy are reported (%).

Experiment	Bal Acc	Sens	Spec	F1	AUC	Acc
Baseline	80 ± 11	89 ± 15	92 ± 8	89 ± 7	80 ± 11	84 ± 10
ALE	81 ± 12	91 ± 12	93 ± 7	91 ± 5	81 ± 12	86 ± 7
APZ	83 ± 9	90 ± 11	93 ± 7	91 ± 4	83 ± 9	86 ± 6
Min-Max (MM)	80 ± 12	89 ± 15	92 ± 9	89 ± 7	80 ± 12	84 ± 10

PCA	84 ± 10	93 ± 12	93 ± 7	92 ± 6	84 ± 10	88 ± 9
Binning (Bin)	88 ± 8	92 ± 8	95 ± 6	93 ± 3	88 ± 8	89 ± 5
PCA+ALE+MM	85 ± 9	94 ± 10	94 ± 6	93 ± 5	85 ± 9	89 ± 7
PCA+APZ+MM	85 ± 9	94 ± 11	93 ± 6	93 ± 5	85 ± 9	89 ± 7
Bin+APZ+MM	90 ± 7	92 ± 8	96 ± 6	94 ± 4	90 ± 7	90 ± 5
Bin+ALE+MM	91 ± 4	93 ± 7	96 ± 4	95 ± 3	91 ± 4	91 ± 4

Our evaluations of the effect of dimensionality reduction techniques, Binning and PCA, showed they were beneficial to the performance of the classifier. Both techniques reduce the complexity of the input data which is significant for small datasets. Our dataset contains just 399 spectra whereas each spectrum contains 2578 features. This introduces significant variance and noise into the data which is difficult to generalize for many machine learning models. By reducing the dimensionality, we simplify the data, reducing high variance noise that can distract the model, and allowing it to focus on relevant features. Binning provided a greater performance increase compared to PCA. Binning preserves information related to the order of features along the electromagnetic spectrum. PCA projects the features to a latent space and orders the features according to their variance. This may be a contributing factor to explain the discrepancy in performance between the methods. However, we can confirm the importance of dimensionality reduction techniques. Min-max normalization and Spectroscope signal normalization did not find a significant impact on classifier performance. Min-max normalization compensated for the inconsistent probe-to-surface distance inherent to freehand scanning without diminishing classification performance. Spectroscope signal normalization increased the robustness of the trained model by compensating for differences in the light source and the spectrometer used for data collection. Ambient peak zeroing (APZ), and ambient light estimation (ALE) were able to successfully reduce the effects of ambient light while providing a small positive impact on classifier performance. However, each technique was limited. APZ does not compensate for the entire ambient light spectrum, focusing just on the peak range. In an effort to reduce over-processing, ALE was performed using average ambient light profiles for the entire dataset. Since different tissues interacted with the ambient light differently, this technique was less effective on a dataset scale. The ideal ambient light compensation would record the ambient light directly before each measurement to allow for accurate compensation. A light source that can be rapidly digitally toggled would be ideal for this application. We observed that Binning in conjunction with ALE and Min-Max produced the strongest and most consistent performance across folds. This configuration increased the balanced accuracy by 11%, from a baseline of 80% to 91% and the standard deviation decreased 8% from a baseline of 13% to 5%. Additionally, there was an average increase of 7% across all reported metrics.

Overall, our experimentation shows promising results for the ability to differentiate between cancerous and non-neoplastic human kidney tissue ex vivo using a simple

linear classification model. We show that our preprocessing pipeline provides significant increases in model performance while effectively compensating for the effects of external variables. However, these results are limited by the size and variability of our dataset. Further data collection is required to draw more robust conclusions. Further investigation of deep learning models and preprocessing steps would also be beneficial.

Conclusions

The objective of this work was the development and evaluation of a tracked tissue sensing research testbed for navigated tumor margin assessment. The results presented demonstrated the capabilities of our tracked tissue sensing testbed for accurate real-time navigated tissue inspection. The first contribution focused on the development and performance of the testbed for real-time tissue assessment. We showed that our tracked diffuse reflection spectroscope was capable of real-time spatial classification and navigation of an arbitrary tissue volume using handheld interrogation methods. The second contribution focused on the application of the testbed for a clinically relevant tissue type. The results of our interrogation of human kidney tissue demonstrated the viability of the testbed for the rapid application to novel tissue types. We showed our tracked tissue sensor was capable of real-time data collection and exploration while offering promising potential for the classification of cancerous and non-neoplastic kidney tissues.

The current system, with diffuse reflection spectroscopy, is limited to a superficial scan of the tissue surface. However, cancerous tissue may lie under a layer of healthy tissue and may not be detected. An important area for development is the integration of multimodal imaging. The testbed's modularity and flexibility enable the implementation and combination of various margin detection techniques. Of particular interest would be integrating the superficial broadband spectroscope with depth information from ultrasound. Various works have shown the potential for combining multiple modalities. For example, L. Connolly et al. (2022) [13] experimented with using throughput broadband spectroscopy and temporally enhanced ultrasound for multilayer tissue differentiation. S. Wilson et al. (2022) [14] created a device to focus point-based electromagnetic techniques within a 2D ultrasound slice. Both may provide insight into potential future directions for our testbed. Another clear direction for future research involves exploring and implementing state-of-the-art deep-learning classification models. For example, enhancing the testbed by integrating attention CNNs or transformer-based models with self-supervised learning and uncertainty estimation has the potential to greatly improve system usability and interoperability.

A substantial gap exists between preoperative and intraoperative information for the surgical treatment of early-stage cancer. The capacity to intraoperatively assess tumor margin status has the potential to increase surgeon confidence in complete tumor resection, mitigating the probability of revision surgery. Mitigation of the rate of revision surgeries also minimizes their challenges and risks for patients

including reducing the likelihood of postoperative complication, improving cosmesis, reducing psychological distress, and decreasing healthcare costs [15].

The development of accurate intraoperative tumor margin assessment tools has a crucial role to play in improving outcomes for early-stage cancer patients around the world.

References

[1] M. Arruebo, N. Vilaboa, J. Sáez-Gutierrez Berta and Lambea, A. Tres, and A. Valladares Mónica and González-Fernández, "Assessment of the evolution of cancer treatment therapies," *Cancers (Basel)*, Vol. 3, No. 3, pp. 3279-3330, Aug. 2011, doi: 10.3390/cancers3033279

[2] J. Balog *et al.*, "Intraoperative tissue identification using rapid evaporative ionization mass spectrometry," *Sci Transl Med*, Vol. 5, No. 194, Jul. 2013, doi: 10.1126/SCITRANSLMED.3005623

[3] S. Azizi *et al.*, "Deep recurrent neural networks for prostate cancer detection: Analysis of temporal enhanced ultrasound," *IEEE Trans Med Imaging*, Vol. 37, No. 12, pp. 2695-2703, 2018, doi: 10.1109/TMI.2018.2849959

[4] P. Gao *et al.*, "The Clinical Application of Raman Spectroscopy for Breast Cancer Detection," *Journal of Spectroscopy*, Vol. 2017, p. 5383948, 2017, doi: 10.1155/2017/5383948

[5] D. Mojahed *et al.*, "Fully Automated Postlumpectomy Breast Margin Assessment Utilizing Convolutional Neural Network Based Optical Coherence Tomography Image Classification Method," *Acad Radiol*, Vol. 27, No. 5, pp. e81-e86, 2020, doi: 10.1016/j.acra.2019.06.018

[6] M. Thill, "MarginProbe®: Intraoperative margin assessment during breast conserving surgery by using radiofrequency spectroscopy," *Expert Rev Med Devices*, Vol. 10, No. 3, pp. 301-315, 2013, doi: 10.1586/erd.13.5

[7] J. Wang, L. Zhang, and Z. Pan, "Evaluating the impact of radiofrequency spectroscopy on reducing reoperations after breast conserving surgery: A meta-analysis," *Thorac Cancer*, Vol. 14, No. 16, pp. 1413-1419, Jun. 2023, doi: 10.1111/1759-7714.14890

[8] L. L. De Boer *et al.*, "Fat/water ratios measured with diffuse reflectance spectroscopy to detect breast tumor boundaries," *Breast Cancer Res Treat*, Vol. 152, 2015, doi: 10.1007/s10549-015-3487-z

[9] C. Palmer, *Diffraction Grating Handbook (8th edition)*. Richardson Gratings, Newport Corporation, 2014

[10] A. Lasso, T. Heffter, A. Rankin, C. Pinter, T. Ungi, and G. Fichtinger, "PLUS: Open-source toolkit for ultrasound-guided intervention systems," *IEEE Trans Biomed Eng*, Vol. 61, No. 10, pp. 2527-2537, Oct. 2014, doi: 10.1109/TBME.2014.2322864

[11] J. Tokuda *et al.*, "OpenIGTLink: an open network protocol for image-guided therapy environment," *The International Journal of Medical Robotics and Computer Assisted Surgery*, Vol. 5, No. 4, pp. 423-434, Dec. 2009, doi: 10.1002/RCS.274

[12] S. Xu, M. Perez, K. Yang, C. Perrenot, J. Felblinger, and J. Hubert, "Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator," *Surg Endosc*, Vol. 28, No. 9, pp. 2569-2576, 2014, doi: 10.1007/s00464-014-3504-z

[13] L. Connolly *et al.*, "Feasibility of combined optical and acoustic imaging for surgical cavity scanning," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, C. A. Linte and J. H. Siewerdsen, Eds., SPIE, 2022, p. 120341H. doi: 10.1117/12.2611964

[14] S. L. M. Wilson *et al.*, "Development of a novel, dual-modality image guidance system by combining a focused gamma probe with ultrasound imaging," in *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, C. A. Linte and J. H. Siewerdsen, Eds., SPIE, 2022, p. 120341F. doi: 10.1117/12.2612730

[15] J. L. Oh, "Multifocal or Multicentric Breast Cancer: Understanding Its Impact on Management and Treatment Outcomes," *Methods of Cancer Diagnosis, Therapy and Prognosis*, pp. 583-587, Nov. 2008, doi: 10.1007/978-1-4020-8369-3_40

[16] D. Morton *et al.*, "Tracked tissue sensing for tumor bed inspection," in *Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*, C. A. Linte and J. H. Siewerdsen, Eds., SPIE, 2023, p. 124661K. doi: 10.1117/12.2654217