HHS Public Access

Author manuscript

IEEE Trans Biomed Eng. Author manuscript; available in PMC 2024 October 01.

Published in final edited form as:

IEEE Trans Biomed Eng. 2023 October ; 70(10): 2863–2873. doi:10.1109/TBME.2023.3266678.

Anatomy-Specific Classification Model Using Label-free FLIm to Aid Intraoperative Surgical Guidance of Head and Neck Cancer

Mohamed Abul Hassan,
Department of Biomedical Engineering, University of California, Davis

Brent W Weyers,
Department of Biomedical Engineering, University of California, Davis

Julien Bec,
Department of Biomedical Engineering, University of California, Davis

Farzad Fereidouni,
Department of Biomedical Engineering, University of California, Davis

Jinyi Qi [Fellow, IEEE],
Department of Biomedical Engineering, University of California, Davis

Dorina Gui,
Department of Pathology and Laboratory Medicine, University of California, Davis

Arnaud F. Bewley,
Department of Otolaryngology – Head & Neck Surgery, University of California, Davis

Marianne Abouyared,
Department of Otolaryngology – Head & Neck Surgery, University of California, Davis

D. Gregory Farwell,
University of Pennsylvania – Department of Otorhinolaryngology – Head and Neck Surgery, *University of California, Davis*

Andrew C. Birkeland,
Department of Otolaryngology – Head & Neck Surgery, University of California, Davis

Laura Marcu [Senior Member, IEEE]
Department of Biomedical Engineering, University of California, Davis

Abstract

Intraoperative identification of head and neck cancer tissue is essential to achieve complete tumor resection and mitigate tumor recurrence. Mesoscopic fluorescence lifetime imaging (FLIm) of intrinsic tissue fluorophores emission has demonstrated the potential to demarcate the extent of the tumor in patients undergoing surgical procedures of the oral cavity and the oropharynx. Here, we report FLIm-based classification methods using standard machine learning models that account for the diverse anatomical and biochemical composition across the head and neck anatomy to

Corresponding Author Laura Marcu (lmarcu@ucdavis.edu).

improve tumor region identification. Three anatomy-specific binary classification models were developed (i.e., “base of tongue,” “palatine tonsil,” and “oral tongue”). FLIm data from patients (N=85) undergoing upper aerodigestive oncologic surgery were used to train and validate the classification models using a leave-one-patient-out cross-validation method. These models were evaluated for two classification tasks: (1) to discriminate between healthy and cancer tissue, and (2) to apply the binary classification model trained on healthy and cancer to discriminate dysplasia through transfer learning. This approach achieved superior classification performance compared to models that are anatomy-agnostic; specifically, a ROC-AUC of 0.94 was for the first task and 0.92 for the second. Furthermore, the model demonstrated detection of dysplasia, highlighting the generalization of the FLIm-based classifier. Current findings demonstrate that a classifier that accounts for tumor location can improve the ability to accurately identify surgical margins and underscore FLIm’s potential as a tool for surgical guidance in head and neck cancer patients, including those subjects of robotic surgery.

Keywords

Head and Neck Cancer; Positive Surgical Margin; Classification; Machine Learning; Fluorescence Lifetime Imaging; Trans-Oral Robotic Surgery

I. INTRODUCTION

IN 2021, approximately 54000 new cases of head and neck cancer (HNC) were diagnosed in the United States; representing 2.8% of all new cancer cases [1], [2]. HNC comprises the 3rd highest rate of positive surgical margins (PSMs) across all oncology fields [2]. Currently, 20–30% of patients present with PSMs at final histopathology (reported within a week of surgery) [3]. The sustained prevalence of PSMs has prompted the development of new tools for cancer margin discrimination during surgical procedures. Intraoperative identification of surgical margins during HNC procedures is key to ensure that both gross and microscopic complete tumor resection is achieved [4]. The HNC surgical margin is the edge or border of the tissue removed in cancer surgery [5]. When successfully resected, this border encloses the cancerous tissue with the rim of normal surrounding tissue.

Intraoperative margin assessment is currently based on preoperative imaging and visual/tactile identification of tumor tissue by the surgeon. Several factors limit the effectiveness of this approach. These include the inherent complexity of head and neck anatomy, the diversity of cancer types and locations, and limitations of standard histopathology (e.g., waiting time and sampling error). To quantify the extent of the tumor and minimize PSM, new techniques capable of providing real-time tumor-specific feedback to quantify the extent of the tumor are needed. These should mitigate cancer recurrence and enable more conservative tissue resection practices, thereby preserving uninvolved healthy tissue.

Label-free mesoscopic fluorescence lifetime imaging (FLIm) has demonstrated promise for identifying primary tumors of the oral cavity and the oropharynx [6] in addition to occult primary tumors of the oropharynx [7]. FLIm relies on tissue endogenous fluorophores emission upon ultraviolet light excitation; where collagen, nicotinamide adenine dinucleotide (NADH), flavin adenine dinucleotide (FAD), and porphyrins are

recognized as the main contributors to head and neck autofluorescence emission due to their high quantum yield and their prevalence in the oral cavity and oropharynx.[8] The radiative decay (lifetime) of endogenous fluorophores is influenced by numerous biochemical and physical factors such as pH, oxygenation, temperature, solvent polarity, and molecular binding. By taking advantage of alterations in biochemical and metabolic characteristics related to neoplastic processes [9], [10], FLIm measurements can provide contrast between healthy tissue, dysplasia, and neoplasms.

Previously, we reported [6] a FLIm-based classifier (N=53 patients) for margins assessment of tumor in the oral cavity and oropharynx. The model was agnostic to the anatomic location of the tumor, combined FLIm data obtained from various anatomies (i.e., pooled anatomies), and did not account for the inherent structural and biochemical differences within the diverse anatomies in the head and neck. Intrinsic differences such as the presence of lymphoid tissue in the base of tongue and tonsils of the oropharynx, stratified skeletal muscle and keratinization of the oral tongue pose a challenge for tissue types classification [11]. We hypothesized that anatomy-specific classification models would better account for the tumor environment within each anatomic location, hence improving the discriminatory power of the proposed classification model.

Intraoperative margin assessment requires identifying the correct tissue types in the tumor area and its surroundings. A model that aims to classify healthy from cancer tissue presumes a first approximation to the problem. However, incorporating dysplastic tissue (low- and high-grade dysplasia) with healthy and cancer could improve the overall reliability of the classification model, since from a clinical standpoint, high-grade dysplasia is typically resected while low-grade dysplasia is not. Dysplasia is the presence of abnormal cells within the tissue that may or may not become cancer. Due to its transitional state, dysplastic tissue can be found between the border of the tumor and the normal tissue, that is the resection margin. Hence, a classification model that incorporates healthy tissue, cancer, and dysplasia may more accurately represent the tissue type involved in tumor margins providing clinically relevant information to the surgeons.

An additional endpoint of interest is differentiating HPV-associated cancer (p16+ HNC) from neoplasms that do not exhibit HPV (p16- HNC). HPV-associated carcinogenesis is clinically assessed by p16 immunohistochemistry, and is important as the involvement of HPV is known to modulate tumor molecular signatures, immune responses, and result in different clinical prognosis [7]. The goals of this study are as follows: (1) to evaluate the effectiveness of the anatomy-specific training on classification performance (2) to optimize the anatomy-specific classifier to differentiate between dysplasia states (i.e., low-grade dysplasia (LGD), and high-grade dysplasia (HGD), (3) to investigate the effect of HPV involvement on classification performance.

II. Methods

The classification model reported here leverages the inherent structural and metabolic tissue characteristics assessed from tissue autofluorescence. Three anatomy-specific classifiers were modeled as follows: (1) oral tongue [oral cavity], (2) base of tongue [oropharynx], and

(3) palatine tonsil [oropharynx] (see Fig. 1). The model classifies healthy and cancer tissue, along with LGD and HGD, as a probability of the tissue being cancer; where 1% indicates a low probability of cancer (i.e., healthy), and 99% indicates a high probability of cancer.

A. Description of FLIm Hardware and Data Acquisition

A custom-built, fiber-based, point-scanning mesoscopic FLIm instrument was used to acquire data for this study. The hardware[12] and the processing [13] of data are described extensively elsewhere. Briefly, a 355 nm pulsed excitation laser (120 Hz repetition rate) was used to induce tissue autofluorescence. All measurements are averaged 4 times (to account for potential erroneous readings), resulting in 30 averaged measurements per second. The instrument operates with four spectral channels: CH1: 390±20 nm, CH2: 470±14 nm, CH3: 542±25 nm, and CH4: 629±26.5 nm, which are designed to match the autofluorescence emission maxima of collagen, NAD(P)H, FAD, and porphyrins, respectively. The signal from each channel was time-multiplexed onto a single microchannel plate photomultiplier tube, amplified, and time-resolved by a high sampling frequency digitizer at 80 ps intervals [8]. A separate visible 445 nm continuous-wave diode laser was integrated into the wavelength selection module to highlight the location where FLIm point measurements were acquired [14].

FLIm data was acquired during two distinct surgical situations: (1) Transoral robotic surgical procedures (TORS) using the da Vinci SP (for oropharyngeal cancer), and (2) non-TORS procedures performed by hand (oral cavity cancer). The average FLIm scan duration was approximately 45 seconds, therefore resulting in approximately 1,350 averaged point measurements per surgical field scanned. All scans were performed prior to commencing surgical excision of tissue. Typically, in such situation no- or limited-amount of blood is present in the operative field. For a small subset of patients that had bleeding due to tongue retractors or bleeding from nasal intubation the region of interest for FLIm scanning was rinsed with isotonic saline followed by aspiration prior to optical scan.

The method for acquiring and visualizing the data was adapted to distinct surgical scenarios. For example, the non-TORS approach combines a hand-held fiber probe (Omniguide Laser Handpiece) and endoscopic camera (Stryker), whereas TORS involved the actuation of the fiber optic probe using the robotic instruments, where the surgical field was visualized using the integrated da Vinci camera.

A fiber probe was used for TORS procedures; in the latter setup, a 3D printed stainless steel grasper tab was added to the distal end of the fiber probe to enable grasping and maneuvering by the da Vinci instruments [15]. The FLIm parameters were overlaid on the endoscopic/robotic video of the surgical field at the position of each point measurement to generate a parametric map of the surgical area. Real-time implementation of this visual augmentation aided the surgeon in comprehensively sampling the surgical field during the data collection process.

B. FLIm Pre-processing and Feature Extraction

Spectral channels 1–3 were used for analysis in this study. Channel 4, attributed to the emission maximum of porphyrins, was not used in the present investigation due to low

signal-to-noise ratio (SNR). A threshold of ≥30 dB was applied as a filtering criterion to select FLIm points with good SNR. The raw FLIm waveform was pre-processed for background subtraction. Then, two non-parametric methods (Laguerre - based deconvolution [13] and lifetime and spectral phasors [16]) were used to retrieve the fluorescence decay characteristics for each spectral channel. Each method generated a set of features that were used in further analysis,

For simplicity, in this paper, the average lifetime, spectral intensity ratio and 12 Laguerre coefficients derived from the former method are referred to as Laguerre features (LG). While the 3 harmonics of the lifetime phasors and spectral phasors derived from the latter method are referred as phasors-based features (PH).

C. Description of the Patient Cohort & Tumor Type

Eligible patients were recruited in the study over a 5-year period from 2016 – 2021. The research was performed under the approval of the UC Davis Institutional Review Board (IRB) and with the patient's informed consent. FLIm was performed on a total of 100 patients. Only patients with a newly diagnosed preoperative diagnosis of HNC of either the oral cavity or oropharynx were enrolled in this study. FLIm data from 15 patients were excluded from the analysis for the following reasons. Data from 8 patients were unable to be analyzed due to either accurate registration with tissue histopathology or instrumentation issues (e.g., malfunctioning software or damaged fiber optic probe) during procedure. Data from 7 additional patients with tumor presenting limited-to no-mucosal involvement (i.e., the extent of tumor comprised only the lamina propria or deeper into the submucosa, see limitation section).

Accordingly, data from only N=85 patients (49 TORS and 36 non-TORS) with distinct anatomies (Table I) were used for classifier training and testing. The 'other' anatomy label included glossotonsillar sulcus, pharynx, floor of the mouth, lip, retromolar trigone, gingivae, palate, and vallecula. Among these patients, N=40 patients presented with HPV mediated oropharyngeal cancer confirmed by immunohistochemistry. Where, p16+ is used as nomenclature for HPV-mediated tumor and p16- for tumor uninvolved with HPV [17].

Tumor type confirmed by pathology included squamous cell carcinoma (N=73), basaloid carcinoma (N=4), polymorphous adenocarcinoma (N=2), and verrucous carcinoma (N=1).

D. FLIm data labeling from Histology

Each FLIm datapoint used in classification was linked to ground-truth pathology labels (i.e., healthy tissue, lymphoid tissue, LGD, HGD, and cancer) using a process described in our work [18]. In brief, FLIm scan was performed prior to the surgical removal of a patient's cancer. Following tumor removal, the resulting tissue specimen was transferred to a surgical pathology grossing room where the tissue was serially grossed to generate tissue slices, which were then formalin-fixed, paraffin embedded, sectioned, and stained with Hematoxylin and Eosin (H&E). A pathologist (DG), blind to FLIm data, digitally annotated the H&E slides using Aperio Imagescope (Leica Biosystems, United States) and assigned the pathology labels according to histologic features observed within 250 μm of the tissue's epithelial surface. This thickness represents an estimate of the excitation light

penetration depth [19]. The annotation labels were used to group tissue as follows: healthy tissue, LGD, HGD, lymphoid tissue, and cancer. For p16+ tumors of the oropharynx, no dysplasia was marked by the pathologist in concordance with the American Joint Committee on Cancer (AJCC), due to the lack of a basement membrane in Waldeyer's ring [20]. Dysplasia however was designated by the pathologist for p16- tumors of the oropharynx, and tumors of the oral cavity.

After annotation, tissue labels were overlaid onto grossed specimen slices (Fig. 1 (c)). Changes in morphology (i.e., compression, expansion, bending) between the H&E section and gross slice were considered to ensure accuracy when registering tissue labels from H&E onto the gross slice. The same process was then performed to overlay tissue labels from the gross slice onto the original intact excised specimen. Then, labels from the excised specimen were overlaid *in vivo* video images. To perform this overlay, first, an *in vivo* recording of the surgical excision process was used to determine the boundaries of the excised specimen, and landmarks (identifying clear visible features) were used to facilitate registration of the *ex vivo* specimen *in vivo*. Next, labels from the *ex vivo* specimen were transferred to the *in vivo* image. To account for motion between the TORS surgical/endoscopic camera and the evaluated tissue, a motion compensation algorithm was applied [21]. Training and evaluation of the classifier were based only on regions with directly registered histopathology, and regions outside of the surgical margins that displayed the absence of cancer based on patient radiology scans (PET, CT, MRI).

The distribution of FLIm data points associated to pathology labels for the 85 patients included in this study is tabulated in Table I. Among the set of five tissue histopathology labels generated by the pathologist, lymphoid tissue was not investigated and was therefore omitted in classifier training and computations due to its low occurrence. In total, lymphoid tissue represented 0.7% of the acquired FLIm data. Lymphoid tissue presents with unique FLIm optical signatures [11] which requires further investigation due to its disparate histological and physiological composition relative to the mucosa of oral cavity and oropharynx [11].

E. Training Data

The classifier was trained based on FLIm data labeled using the binary classes of "healthy" and "cancer" (i.e., "0", "1"). A binary probabilistic classification model was developed to predict the probability 'f' for each FLIm point (see Eq. 1) labeled as healthy or cancer. The 'healthy' class comprised epithelium of varying thicknesses, in addition to keratinized mucosa, inflammation, and reactive tissue. The cancer class consisted of squamous cell carcinoma (SCC), basaloid carcinoma, and verrucous carcinoma. Dysplasia was categorized as LGD and HGD. The LGD was grouped with healthy tissue, whereas HGD was grouped with cancer (in concordance with malignant transformation potential).

$$D = \left\{(x_i, f_i) \mid x_i \in \mathbb{R}^d, f_i \in \left\{0, 1\right\}\right\}_{i=1}^{n} \tag{1}$$

The training dataset 'D' contains the labeled class variable 0,1 for the i^{it} sample corresponding to the features vector 'x_i'. 'd' is the dimension of the features vector and 'n' is the number of optical point measurements in the training dataset. The training dataset 'D' was separated into three main training datasets: (1) pooled anatomy (i.e., data from all patients, irrespective of anatomy), (2) anatomy specific (i.e., data of only a specific anatomy such as base of tongue), and (3) anatomic region-specific (data from a combination of anatomies with similar tissue composition and structure such as oropharynx comprising lymphoid components).

Distinct training sets (Table II) were configured for the base of the tongue, oral tongue, and palatine tonsil. The anatomy-specific training dataset yielded superior classification performance for the oral tongue. Whereas the pooled anatomy training dataset yielded superior performance for the base of the tongue, and the anatomic region training dataset consisting of palatine tonsil and lingual tonsil yielded superior performance for the palatine tonsil anatomy.

F. Anatomy-Specific Classification Model

Handcrafted and non-handcrafted feature models are investigated. The models '$\widehat{H}$' were evaluated on multiple combinations of features 'x' from the FLIm data (Eq. 2) based on the ROC-AUC performance metric.

$$f = \widehat{H}(x) \tag{2}$$

Forty-two LG features (1 average lifetime, 1 intensity ratio, and 12 Laguerre coefficients for each of the 3 channels) were used, in addition to 4 PH features per channel (12 total). Combining LG and PH, referred to as LGPH, created 54 features. One non-handcrafted feature 'x'-based classification model was also evaluated on both the fluorescence decay waveform and the reconstructed deconvolved waveform. The anatomy-specific classification model was developed in two steps: (1) Modeling classifiers for each anatomy based on the dataset consisting of healthy and cancer (*Healthy vs. Cancer*) labels while excluding LGD and HGD labels. (2) The best-performing classification model for each of the anatomies are tested on LGD and HGD labels. The model was tested for dysplasia labels under two configurations (a) transfer learning model classifying unseen dysplasia labels (*Dysplasia Tested*) and (b) classifying on retrained model with dysplasia labels (*Dysplasia Trained/Tested*).

1) Handcrafted Feature-Based Classification Model—Decision tree (DT), support vector machine (SVM), multi-layer perception model (MLP), and applying bootstrap aggregation of the best performing models to generate an ensemble learning model. The DT model used an adaptive boosting for binary classification with 100 estimators, each with a maximum depth of 10 and a learning rate of 0.01. The SVM model used a linear kernel with a sequential minimal optimization, the regularization parameter (i.e., C parameter) was set to 1.0. The MLP model consisted of a feedforward and fully connected neural network

with 10 fully connected layers. ReLU activation functions were used at the output of each layer while using SoftMax as the final classification layer.

2) Non-Handcrafted Feature-Based Classification Model—Two non-handcrafted-based feature classification were studied; specifically, convolutional neural network (CNN) and optimal transport (OT). The CNN model followed the model and training strategy reported in previous work [6]. The raw FLIm signals and FLIm decay curves were utilized for model development. The OT model was chosen due to its ability to perform accurate classification while training with limited labeled data [22]. This model involved transforming the FLIm signal to an orthogonal space using cumulative distribution transform and applying a near subspace classifier to perform classification based on the nearest distance. The normalized difference of the nearest distance between the two classes was taken as the probability score of the prediction (Eq. 3–5).

The assigned class (i.e., cancer and healthy) is denoted by 'k' Term 'd' denoted the distance between the subspaces, '$\hat{s}$' is the transform space, and A^k is the orthogonal projection matrix. The OT-based classification was formulated in two strategies: First, an ensemble learning model consisting of OT models for the of the three FLIm spectral channels. Second, OT model which concatenates the three FLIm channels across time dimension.

$$d^{(k)} = \| \hat{s} - A^{(k)}\hat{s} \|^2 \tag{3}$$

$$d_\delta = d^{(1)} - d^{(2)} \tag{4}$$

$$f = \frac{d_\delta - min(d_\delta)}{max(d_\delta) - min(d_\delta)} \tag{5}$$

G. Tissue Region-Based Prediction Refinement

A region-based prediction was defined to increase the classification accuracy by refining the point-level classifier prediction score 'f' across the tissue region. For this, we leveraged the high number of point measurements obtained from an area of interest by spatial averaging each FLIm point measurement with both inverse distance and SNR weighting. The spatial averaging of point measurements allowed rectification of isolated point-level misclassifications (see Fig. 2). To preserve the lateral resolution of the prediction map, the averaging was limited to a 15-pixel radius (approximately 0.55 mm) of at least five point-measurement centers.

H. Evaluation Metrics

Evaluation of the classification model was performed for both point-measurement level and tissue region level. The latter enables assessment of the performance of the method over an entire tissue surface. A leave-one-patient-out cross-validation approach was adopted. The classification model was first validated by the area-under-receiver operator curve (ROC-AUC). The ROC-AUC was computed at the point-measurement level for cases in which at least 100 data points from healthy regions and 100 from cancer regions were acquired to prevent misleading patient-level scores caused by highly imbalanced data. A threshold of 0.5 was applied as the decision boundary to compute the sensitivity and specificity using all patients in which cancer or healthy point measurements were acquired, respectively.

III. Results

A. Classification Performance for Healthy vs. Cancer

Figure 4 presents the classification performance of the anatomy-specific classifier for healthy vs. cancer for all tissue types: base of the tongue (red), oral tongue (blue), and palatine tonsil (green). The sensitivity, specificity, and AUC obtained from the best-performing classification model among the various classifiers tested are presented in Fig. 3 (a). The DT model trained with LG features reported the highest performance for the base of the tongue. While the ensemble learning model (DT-LG+SVM-LGPH) reported the highest classification performance for the oral tongue and palatine tonsil (Table. III).

For the base of the tongue, the SVM and DT models based on Laguerre features reported similar classification performances; with the DT model reporting a slightly higher ROC-AUC (0.94±0.09) when compared to the SVM. The MLP and ensemble learning models reported lower performance. For the oral tongue, the DT, SVM, and MLP showed similar performance. The highest performance was reported for the ensemble learning model. For palatine tonsils, the DT, SVM, and MLP reported similar performance. The ensemble learning model based on DT and SVM achieving the highest performance (Fig. 3 (a)).

The ensemble learning model based on DT and SVM provided the strongest discrimination compared to the DT, SVM, MLP, and OT for the oral tongue and palatine tonsil anatomy sites. The CNN-based classification approach reported the lowest classification performance: ROC-AUC of 0.64±0.6, 0.63±0.11, and 0.62±0.8 for the base of the tongue, oral tongue, and palatine tonsil, respectively. The OT model based on concatenating FLIm decay curves enabled the model to discriminate healthy tissue from cancer (see Fig. 3 (a)). In contrast, the ensemble learning OT models failed to generalize. Indicating that the concatenating FLIm decay curves maximize the signal correlation across all three channels of FLIm compared to using separate channels to generate meaningful information.

The point-level and region-level classification performance for anatomy-specific classifiers are reported in Table. III. The region-level refinement improves the point-level classification performance of the anatomy-specific classification model by ~7–12%. The tissue region-based prediction takes advantage of multiple closely located point measurements to improve the accuracy of the predictor. This refinement increases tissue prediction for all anatomies

with the greatest impact observed for the oropharynx (i.e., the base of the tongue and palatine tonsil).

Figure 3 (b) illustrates the characteristic of the anatomy-specific classifiers through the overall prediction probability for each site. Oral tongue, and palatine tonsil models were trained on specific anatomy data and combined-specific anatomy data (i.e., oral tongue model was trained on oral tongue data, and the palatine tonsil model on palatine tonsil and lingual tonsil). These two models reported a higher density of prediction probability closer to '0' and '1', compared to the base of tongue model that was trained on pooled anatomy data (i.e., data from all eighty-five patients irrespective of anatomy). The pooled training strategy contributed to the lower density of prediction probability close to '0' and '1'.

B. Classification Performance for Healthy vs. Cancer vs. Dysplasia

From a surgical standpoint, dysplasia is subjective and somewhat difficult diagnosis, and is considered a precancerous lesion. In otolaryngology, LGD is generally observed closely due to lower rates of progression to cancer, whereas HGD is resected due to the risk of cancer progression. Thus, the highest performing anatomy-specific classification model for healthy vs. cancer is applied to classify healthy vs. cancer vs. dysplasia (see Fig. 5). The transfer learning model for dysplasia (*Dysplasia Tested*) reported a higher ROC-AUC of 0.92±0.08 compared to retrained model (*Dysplasia Trained/Tested*). The "*Dysplasia Tested*" model specificity improved by ~3% while diminishing the sensitivity by ~5% compared to the "*Healthy vs. Cancer*" model (Table. IV). The "*Dysplasia Tested*" model maintained the classification model performance compared to the "*Healthy vs. Cancer*" model with a ~ 1% drop in sensitivity and specificity. The better tradeoff between sensitivity and specificity achieved by "*Dysplasia Tested*" were best in classifying healthy, cancer, dysplasia tissue types (see Fig. 5(b)).

Figure 6(a) illustrates the prediction probability of the anatomy-specific classifier "*Dysplasia Tested*" compared to the histological co-registration (i.e., ground truth) while predicting healthy, LGD, HGD, and cancer tissues. The prediction probability of the classifier indicates a high confidence while predicting most cancer and healthy tissue. While most dysplasia tissues were predicted at the boundary between healthy and cancerous tissue (Fig. 6). LGD tissues were predicted with a low probability of cancer and HGD tissue was predicted with a relatively high probability of cancer.

C. Classification Performance for p16+ Patients

Tumors with p16+ HNC exhibit unique molecular signatures, immune responses, and disparate clinical prognosis from p16-HNC tumors. Hence, we investigated the effect of p16+ clinical characteristics data on optimizing classifier performance. The patient cohort used for this analysis included 31 p16+ patients distributed across the three anatomies as follows: 10 base of the tongue, 1 oral tongue, and 20 palatine tonsil. We investigated palatine tonsil anatomy since the patient distribution presented the least imbalance of p16+ patients and p16- patients. Table V shows how the classifier performance diminishes for p16+ patients compared to p16-patients. The same classifier (with the same boundary) reported an improved sensitivity of ~24% for p16-patients compared to the sensitivity reported for

p16+ patients. We used patient HPV clinical characteristics data to improve the sensitivity of the classifier for the p16+ patients by further optimizing the classification boundary with a threshold of 0.4. The optimized classifier reported a sensitivity of 0.88±0.18 and specificity of 0.81±21 for p16+ patients, therefore improving the probability of classifying cancer tissue types for p16+ patients by ~14% sensitivity.

IV. Discussion

In this work, we demonstrated the performance of probabilistic classifiers for head and neck squamous cell carcinomas delineation at the resection margins which account for tumor anatomic location (i.e., oral and oropharynx). Also, we investigated the effectiveness of the classification models in discriminating dysplastic tissue. Moreover, the study examined the relative contributions of various FLIm parameters and machine learning models to the classification performance. Current results demonstrate improved performance when compared to previously reported optical cancer detection approaches for HNC [6], [23]–[25].

A. Anatomy-specific classifier

We demonstrated that developing an anatomy-specific classifier to increase the homogeneity of the training set can improve the classification performance. Both oral tongue and palatine tonsil models for which training data is abundant benefited from the increased homogeneity. In contrast, the best classification performance for the base of the tongue, where limited data were available, was obtained from the pooled model. This underscores the need to strike a balance between the training set size and its homogeneity.

B. Dysplasia

The low representation of dysplasia tissue in the head and neck poses a challenge for modeling a dysplasia classifier. Grouping dysplasia subtypes to the main classes resolved the imbalance in the dataset. Due to the different likelihood of LGD and HGD evolving into cancer [26], LGD was grouped with healthy tissue and HGD was grouped with cancer tissue. The generalization of the FLIm-based classification model to low- and high-grade dysplasia is a promising outcome for general tissue region classification and margin assessment. The probability of cancer assigned by the model (see Fig. 5 (a) and Fig. 6) implies the ability of FLIm to capture the evolving tissue changes associated with dysplasia. These findings are very relevant from a clinical standpoint as HGD is typically resected while LGD is not.

C. Impact of Machine Learning

The hand-crafted FLIm features contributed to achieving the highest discrimination. The LG features exhibit a similar trend to our previously reported work. The combination of the average lifetime, intensity ratio, and Laguerre coefficients yields the highest and most consistent discriminatory power. The phasors, as an individual set of features, reported low classification performance. However, the combination of phasors and LG features positively affected the classification performance for the current patient cohort. This grants the need for further evaluation of the added value of phasors-based features.

The classification performance of the hand-crafted feature models was sensitive to feature dimension. The increased dimensionality (i.e., combining Laguerre and phasors) positively affects the performance of SVM and MLP models, but did not improve the DT model performance. This suggests that the high-dimensional hyper-rectangles of the DT model may not be able to transform the input feature space efficiently.

The comparison of the discriminatory power between handcrafted models and non-handcrafted models (i.e., OT and CNN) emphasizes that the features generated from domain knowledge are superior in extracting meaningful information from FLIm. Nevertheless, the classification performance of the optimal transport model is promising. Further investigation to improve non-handcrafted feature generation while leveraging domain knowledge may result in much more desirable classification performance.

D. Comparison with the State-of-the-Art

Most recent instruments (optical spectroscopy and imaging-based) used for HNC margin assessment [27]–[29] were validated on ex-vivo post-resection tissue specimens. Although the evaluation of the cancer resection margins on excised specimens holds its own merits; the ability to perform margin assessment in-vivo (in situ) enables direct identification of remaining cancer tissue which greatly facilitates surgical workflow. A few studies reported methods enabling in vivo assessment of surgical margins based on both exogenous fluorescence, endogenous fluorescence, or other source of endogenous optical contrast.

Tumor assessments by near-infrared imaging of exogenous fluorescence [30]–[33] require injecting a molecular contrast agent. While promising, the clinical translation of these fluorescent molecular probes is hindered by the difficulties of regulatory approval, and none have received regulatory approval for clinical use [34]. In contrast, autofluorescence imaging approaches [6], [23]–[25], narrow-band imaging systems [35], [36], and hyperspectral imaging [37] do not require the addition of any exogenous markers, and are thus much easier to use in clinics.

Narrow-band imaging uses the absorption spectrum of hemoglobin to determine the neoangiogenic patterns inside and surrounding the tumor. However, determining neoangiogenic patterns is challenging due to the varying tissue characteristics and only applies to the epithelial surface, leading to low-diagnostic accuracy [38]. Hyperspectral imaging [37] uses surface reflectance to identify the tumor. These approaches are time-consuming requires controlled lighting environment, which has impacted their clinical adaptation [38].

Autofluorescence imaging (AFI) leverages the endogenous fluorescence properties of biomolecules in tissue and cells. VELscope, a commercially available AFI technique, is widely used as an early detection screening tool for the oral cavity. Facilitated by the commercial availability, multiple studies have investigated this device’s performance with mixed results; a systematic review [25] reported HNC detection varying from the sensitivity of ~0.2 – 1 and specificity of ~0.08 – 1. However, no difference in local recurrence rate for the patient undergoing fluorescence visualization-guided surgery, and non-fluorescence visualization-guided surgery was observed [39]. The lack of improvement

in local recurrence reduction may also be due to the fact that the device is only able to map the lateral extent of the tumor and not the residual tumor that may be present in the surgical cavity. This limitation is also valid for the narrow-band imaging systems [38].

As shown in this study, another approach to improve the performance of fluorescence-based imaging is based on time-resolved measurements that provide an additional source of contrast [8]. Earlier studies, performed on a small number of patients undergoing HNC resection surgery, employing both point scanning [40] and endoscopic imaging [41], have reported the added value of lifetime information. More recent studies using advanced FLIm research instrumentation, performed on a larger patient population, have further underscored the earlier finding; e.g. HNC classification performance of AUC 0.88 [6] and AUC 0.81 [23] were reported. While in the current study, we adopted a method similar machine-learning model as in previous reports, the implementation of domain-specific features (e.g., LG) and anatomy-specific have contributed to superior HNC detection (AUC 0.94) compared to earlier anatomy-agnostic models. In addition, the AUC reported in [23] is computed by averaging the image-level predictions while in present study the AUC is computed from each point. The prediction from each individual point allows for a more accurate delineation of the surgical margins. Moreover, the classification model reported in [23] considers precancerous and cancerous oral lesions such as LGD, moderate dysplasia, HGD, and SCC as HNC. The model developed in the current study considers LGD as healthy and HGD as cancerous since, from a clinical standpoint, HGD is typically resected while LGD is not [42].

A significant subset of HNC is located in the oropharynx, where tumor assessment is challenging due to the limitations of visual and tactile identification of the tumor [43]. Therefore, surgical guidance tools suitable for imaging of the oropharynx are needed. This cannot be achieved with a rigid forward-viewing endoscope design AFI [23], [41] or handheld-camera-based AFI techniques [36] that are limited to the interrogation of the oral cavity. The freehand scanning FLIm approach employed in current study utilizes a flexible fiber optic probe and continuous tracking of the probe’s location enabling image reconstruction [12] from all head and neck anatomies regardless of their location. As demonstrated here, when combined with TORS, this point-scanning approach allows for interrogation of HNC in the oropharynx. In addition, the cancer probability map generated by the system could provide the surgeon with visual identification of the tumor and surrounding healthy regions overlaid on the surgical field including display on the robot’s console. Such implementation will enable real-time guidance of surgical resection.

E. Limitations

The development and validation of classifiers rely on the availability of accurate tissue labels. As seen in Fig. 1, labeling of in vivo data is a multi-step process prone to errors due to, among other sources, misregistration between the ex vivo specimen and the surgical field. This was partially addressed by implementing an exclusion radius to reject points that were close to the tumor boundary. But this limits the data obtained from the tumor margin and may lead to an underestimation of the FLIm’s performance. The retrospective classifier development and validation methods reported here are well-suited to the investigation of

various classifier types and training sets. In future work, a more accurate estimation of the classifier will be performed in a prospective study where the classifier's output, reported on tissue at the time of surgery, can be directly compared with histological findings.

This study presents an HNC margin assessment approach within approximately a ~150–250 μm of the tissue's epithelial surface, which corresponds to the penetration depth of the laser's excitation wavelength used in this study (335 nm). While this could be perceived as a limitation, in the context of the current study the shallow penetration depth of FLIm maximizes the signal that can be obtained from a small amount of residual tumor tissue. The epithelial thickness of the H&N anatomy is on average between 100 to 200 μm and more than 90% of all HNC is present in the mucosa [7]. Thus, FLIm is well-suited for the interrogation of these surface-presenting tumors for surgical guidance.

The current dataset demonstrates an imbalanced ratio of p16- vs. p16+ patients, where p16- cancers are predominantly represented. It is understood that the proportion of p16+ tumors are increasing in the United States population, and that p16+ tumors present with distinct clinical, histological, molecular, and prognostic characteristics from p16- tumors. On average, patients with p16+ tumors tend to be younger than patients with classical p16- HNC, and p16+ patients often do not exhibit the classical risk factors for HNC, such as alcohol and tobacco abuse. It is conceivable that the reduced classifier performance for p16+ patients may be related to the reduced training data available. This may make an impact on the results as a greater proportion of the training data for p16- HNC comes from an older patient population on average. Correspondingly, as the database is expanded, it will be important to revisit this aspect given the increasing incidence of this p16+ patient population. It will also be impactful to evaluate if FLIm can predict p16 status in our future work, as this information may better guide surgeons in their approach to patient treatment from a radiosensitivity and prognostic standpoint.

V. Conclusion

Intraoperative tumor assessment is essential for the surgeon to quantify the extent of resection and mitigate cancer recurrence. This work demonstrated that label-free FLIm has the ability to provide meaningful information for machine learning used to identify the extent of the tumor. The results reported herein establishes the importance of accounting for the anatomic site when developing the classifier. The free-hand scanning approach used by our FLIm system allows the surgeon to scan the desired region of interest within the surgical field, including areas with a complex tissue geometry only accessible with TORS. The superior discrimination observed for the anatomy-specific classification model based on FLIm-derived parameters provides the surgeon with a cancer probability map that identifies the tumor region and underscores FLIm's potential as a label-free margin assessment tool in the operating room.

Acknowledgment

The authors thank our clinical research coordinators Angela Beliveau, M.P.H., CCRP and Randev Sandhu, CCRP at the University of California, Davis Medical Center Department of Otolaryngology for their many contributions to enroll and consent patients in our study, maintain patient files and medical history documents, and disseminate

research study information to our team. We would like to also thank Dr. Xiangnan Zhou for improving the visualization software, data processing, and stability of the FLIm system software. We are grateful to Dr. Alba Alfonso Garcia for providing technical feedback on manuscript and contributing to the figures. We would like to acknowledge Dr. Jonathan Sorger (Intuitive Surgical, Sunnyvale CA) for his support for our ongoing industry collaboration; key areas of his industry support include FLIm visualization aspects and integration of FLIm fiber optic probes into the da Vinci SP TORS platform. Finally, we are grateful for Roberto P. Frusciante's involvement with data collection and performing histopathology registration for part of the dataset featured in this manuscript. The authors declare no conflict of interest for the research.

This work was supported by the National Institutes of Health under Grant 2R01CA187427 in collaboration with Intuitive Surgical, Inc; and P41 - EB032840-01.

References

[1]. "Oral Cavity and Pharynx Cancer — Cancer Stat Facts," NIH National Cancer Institute, Surveillance, Epidemiology, and End Results Program (SEER), 2021. [Online]. Available: https://seer.cancer.gov/statfacts/html/oralcav.html. [Accessed: 18-Mar-2021].

[2]. Orosco RK et al. , "Positive Surgical Margins in the 10 Most Common Solid Cancers," Sci. Reports 2018 81, vol. 8, no. 1, pp. 1–9, Apr. 2018.

[3]. van Keulen S. et al. , "Intraoperative Tumor Assessment Using Real-Time Molecular Imaging in Head and Neck Cancer Patients," J. Am. Coll. Surg, vol. 229, no. 6, pp. 560–567.e1, Dec. 2019. [PubMed: 31568855]

[4]. Tringale KR, Pang J, and Nguyen QT, "Image-guided surgery in cancer: A strategy to reduce incidence of positive surgical margins," Wiley Interdisciplinary Reviews: Systems Biology and Medicine, vol. 10, no. 3. Wiley-Blackwell, 01-May-2018.

[5]. Meier JD, Oliver DA, and Varvares MA, "Surgical margin determination in head and neck oncology: Current clinical practice. The results of an International American Head and Neck Society Member Survey," Head Neck, vol. 27, no. 11, pp. 952–958, Nov. 2005. [PubMed: 16127669]

[6]. Marsden M. et al. , "Intraoperative Margin Assessment in Oral and Oropharyngeal Cancer Using Label-Free Fluorescence Lifetime Imaging and Machine Learning," IEEE Trans. Biomed. Eng, vol. 68, no. 3, pp. 857–868, 2021. [PubMed: 32746066]

[7]. Weyers BW et al. , "Intraoperative delineation of p16+ oropharyngeal carcinoma of unknown primary origin with fluorescence lifetime imaging: Preliminary report," Head Neck, May 2022.

[8]. Marcu L, French PMW, and Elson DS, Fluorescence lifetime spectroscopy and imaging : principles and applications in biomedical diagnostics. CRC Press, 2014.

[9]. Suhling K. et al., "Fluorescence Lifetime Imaging," in Handbook of Photonics for Biomedical Engineering, Dordrecht: Springer Netherlands, 2014, pp. 1–50.

[10]. Berezin MY and Achilefu S, "Fluorescence lifetime measurements and biological imaging," Chem. Rev, vol. 110, no. 5, pp. 2641–2684, May 2010. [PubMed: 20356094]

[11]. Weyers BW et al., "Investigating sources of FLIm data variability in head & neck cancer," 10.1117/12.2609864, vol. PC11949, p. PC1194902, Mar. 2022.

[12]. Gorpas D. et al. , "Autofluorescence lifetime augmented reality as a means for real-time robotic surgery guidance in human patients," Sci. Rep, vol. 9, no. 1, Dec. 2019.

[13]. Liu J, Sun Y, Qi J, and Marcu L, "A novel method for fast and robust estimation of fluorescence decay dynamics using constrained least-squares deconvolution with Laguerre expansion," Phys. Med. Biol, vol. 57, no. 4, pp. 843–865, Feb. 2012. [PubMed: 22290334]

[14]. Yankelevich DR et al. , "Design and evaluation of a device for fast multispectral time-resolved fluorescence spectroscopy and imaging," Rev. Sci. Instrum, vol. 85, no. 3, 2014.

[15]. Gorpas D, Ma D, Bec J, Yankelevich DR, and Marcu L, "Real-Time Visualization of Tissue Surface Biochemical Features Derived from Fluorescence Lifetime Measurements," IEEE Trans. Med. Imaging, vol. 35, no. 8, pp. 1802–1811, Aug. 2016. [PubMed: 26890641]

[16]. Fereidouni F, Gorpas D, Ma D, Fatakdawala H, and Marcu L, "Rapid fluorescence lifetime estimation with modified phasor approach and Laguerre deconvolution: a comparative study," Methods Appl. Fluoresc, vol. 5, no. 3, p. 035003, Sep. 2017. [PubMed: 28644150]

[17]. Lewis JS, Shelton J, Kuhs KL, and Smith DK, “p16 Immunohistochemistry in Oropharyngeal Squamous Cell Carcinoma Using the E6H4 Antibody Clone: A Technical Method Study for Optimal Dilution,” Head Neck Pathol., vol. 12, no. 4, pp. 440–447, Dec. 2018. [PubMed: 29190003]

[18]. Weyers BW et al. , “Procedure for Histopathology Labeling of Intraoperative Optical Imaging Data,” (under review), Jan. 2023.

[19]. Avci P. et al., “Low-level laser (light) therapy (LLLT) in skin: stimulating, healing, restoring,” 2013.

[20]. Amin MB et al. , “The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more ‘personalized’ approach to cancer staging,” CA. Cancer J. Clin, vol. 67, no. 2, pp. 93–99, Mar. 2017. [PubMed: 28094848]

[21]. Marsden M. et al. , “FLImBrush: dynamic visualization of intraoperative free-hand fiber-based fluorescence lifetime imaging,” Biomed. Opt. Express, vol. 11, no. 9, p. 5166, Sep. 2020. [PubMed: 33014606]

[22]. Shifat-E-Rabbi M. et al. , “Radon Cumulative Distribution Transform Subspace Modeling for Image Classification,” J. Math. Imaging Vis, vol. 63, pp. 1185–1203, 2021. [PubMed: 35464640]

[23]. Duran-Sierra E. et al. , “Machine-learning assisted discrimination of precancerous and cancerous from healthy oral tissue based on multispectral autofluorescence lifetime imaging endoscopy,” Cancers (Basel)., vol. 13, no. 19, Oct. 2021.

[24]. Awais M. et al. , “Healthcare professional in the loop (HPIL): Classification of standard and oral cancer-causing anomalous regions of oral cavity using textural analysis technique in autofluorescence imaging,” Sensors (Switzerland), vol. 20, no. 20, pp. 1–25, Oct. 2020.

[25]. Cicciù M. et al. , “Early Diagnosis on Oral and Potentially Oral Malignant Lesions: A Systematic Review on the VELscope® Fluorescence Method,” Dent. J, vol. 7, no. 3, p. 93, Sep. 2019.

[26]. Yang Y, xiu Li Y, Yang X, Jiang L, jun Zhou Z, and qin Zhu Y, “Progress risk assessment of oral premalignant lesions with saliva miRNA analysis,” BMC Cancer, vol. 13, Mar. 2013.

[27]. Matthies L. et al. , “Optical diagnosis of oral cavity lesions by label-free Raman spectroscopy,” Biomed. Opt. Express, vol. 12, no. 2, p. 836, Feb. 2021. [PubMed: 33680545]

[28]. Hurskainen MO, Sarin JK, Myllymaa S, González-Arriagada WA, Kullaa A, and Lappalainen R, “Feasibility of near-infrared spectroscopy for identification of l-fucose and l-proline-towards detecting cancer biomarkers from saliva,” Appl. Sci, vol. 11, no. 20, Oct. 2021.

[29]. Ma L. et al. , “Pixel-level tumor margin assessment of surgical specimen in hyperspectral imaging and deep learning classification,” in Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, 2021, p. 34.

[30]. van Keulen S. et al. , “Rapid, non-invasive fluorescence margin assessment: Optical specimen mapping in oral squamous cell carcinoma,” Oral Oncol., vol. 88, pp. 58–65, Jan. 2019. [PubMed: 30616798]

[31]. Krishnan G. et al. , “Fluorescent Molecular Imaging Can Improve Intraoperative Sentinel Margin Detection in Oral Squamous Cell Carcinoma,” J. Nucl. Med, vol. 63, no. 8, pp. 1162–1168, Aug. 2022. [PubMed: 35027369]

[32]. Wang J. et al. , “A c-MET-Targeted Topical Fluorescent Probe cMBP-ICG Improves Oral Squamous Cell Carcinoma Detection in Humans,” Ann. Surg. Oncol, 2022.

[33]. Zhang RR et al. , “Beyond the margins: Real-time detection of cancer using targeted fluorophores,” Nature Reviews Clinical Oncology, vol. 14, no. 6. Nature Publishing Group, pp. 347–364, 01-Jun-2017.

[34]. Lee YJ et al. , “Intraoperative Fluorescence-Guided Surgery in Head and Neck Squamous Cell Carcinoma,” Laryngoscope, vol. 131, no. 3. John Wiley and Sons Inc, pp. 529–534, 01-Mar-2021. [PubMed: 33593036]

[35]. Guida A. et al. , “Oral lichen planus and other confounding factors in narrow band imaging (NBI) during routine inspection of oral cavity for early detection of oral squamous cell carcinoma: A retrospective pilot study,” BMC Oral Health, vol. 19, no. 1, Apr. 2019.

[36]. Ota A, Miyamoto I, Ohashi Y, Chiba T, Takeda Y, and Yamada H, “Diagnostic Accuracy of High-Grade Intraepithelial Papillary Capillary Loops by Narrow Band Imaging for Early

Detection of Oral Malignancy: A Cross-Sectional Clinicopathological Imaging Study," Cancers (Basel)., vol. 14, no. 10, May 2022.

[37]. Eggert D. et al. , "In vivo detection of head and neck tumors by hyperspectral imaging combined with deep learning methods," J. Biophotonics, vol. 15, no. 3, Mar. 2022.

[38]. Wu C, Gleysteen J, Teraphongphom NT, Li Y, and Rosenthal E, "In-vivo optical imaging in head and neck oncology: Basic principles, clinical applications and future directions review-Article," International Journal of Oral Science, vol. 10, no. 2. Sichuan University Press, 2018.

[39]. Durham JS et al. , "Effect of Fluorescence Visualization-Guided Surgery on Local Recurrence of Oral Squamous Cell Carcinoma: A Randomized Clinical Trial," JAMA Otolaryngol. - Head Neck Surg, vol. 146, no. 12, pp. 1149–1155, Dec. 2020. [PubMed: 33034628]

[40]. Meier JD et al. , "Time-resolved laser-induced fluorescence spectroscopy as a diagnostic instrument in head and neck carcinoma," Otolaryngol. - Head Neck Surg, vol. 142, no. 6, pp. 838–844, Jun. 2010. [PubMed: 20493355]

[41]. Sun Y. et al. , "Endoscopic fluorescence lifetime imaging for in vivo intraoperative diagnosis of oral carcinoma," in Microscopy and Microanalysis, 2013, vol. 19, no. 4, pp. 791–798. [PubMed: 23702007]

[42]. Lorini L. et al. , "Clinical and Histological Prognostic Factors of Recurrence and Malignant Transformation in a Large Series of Oral Potentially Malignant Disorders," Front. Oncol, vol. 12, Apr. 2022.

[43]. Turner L, Mupparapu M, and Akintoye SO, "Review of the complications associated with treatment of oropharyngeal cancer: a guide for the dental practitioner.," Quintessence Int., vol. 44, no. 3, pp. 267–79, Mar. 2013. [PubMed: 23444208]

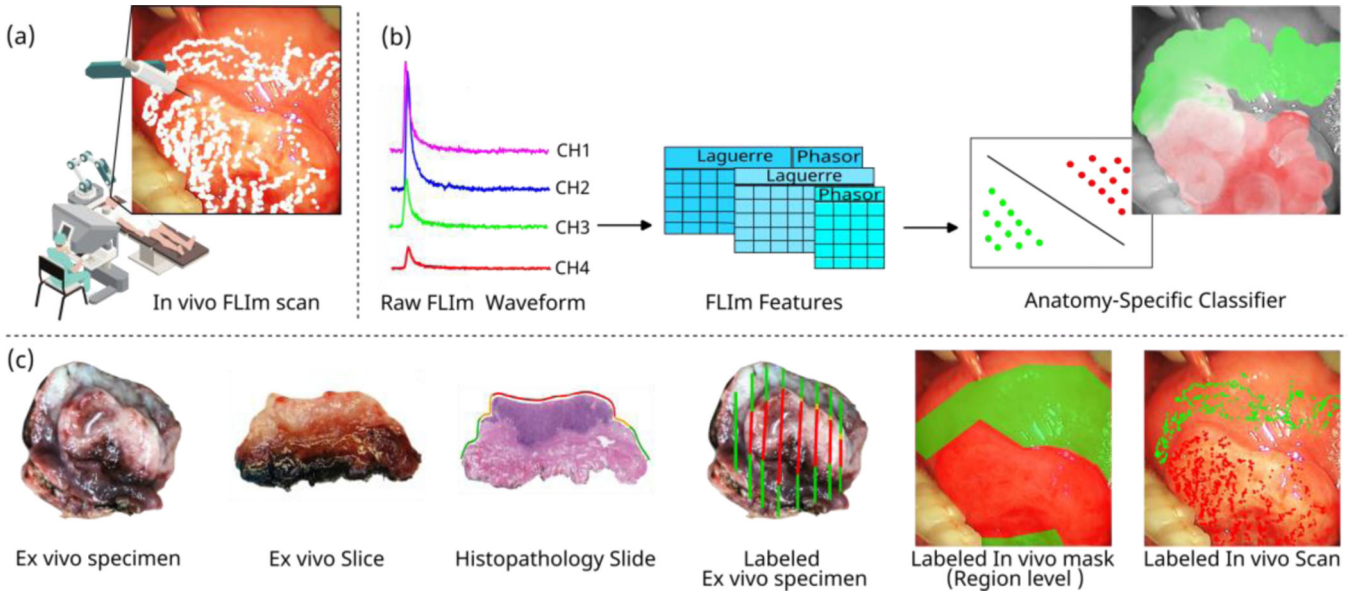

Fig. 1.

Overview methodology of the label-free FLIm based intraoperative surgical guidance, data collection, histopathology registration, and data processing. (a) Schematic of the In-vivo FLIm scan acquisition during surgical procedure. (b) Depiction of the FLIm fluorescence decay waveforms, extracted parameters, incorporation into the classifier, and associated visualization of probability of cancer. The application of anatomy-specific classification model to classify tissue regions based on the FLIm features (i.e., LG: Laguerre features and PH: Phasor based features). (c) The labels for classifier training and testing were derived directly from histopathology evaluated and annotated by a clinical pathologist (DG). Each annotated H&E section was registered with the image of the *ex vivo* and *in vivo* FLIm scan. The red annotations correspond to cancer labels, green annotations correspond to healthy, and orange corresponds to dysplasia.

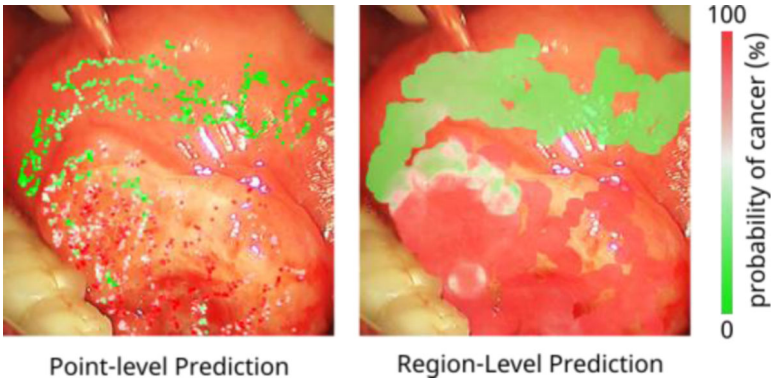

Fig. 2.

Depiction of the tissue region-based prediction refinement from the point-level prediction using interpolation approach based on inverse distance weighting and SNR weighting, The label "*probability of cancer (%)*" is defined by the prediction probability of the classification model

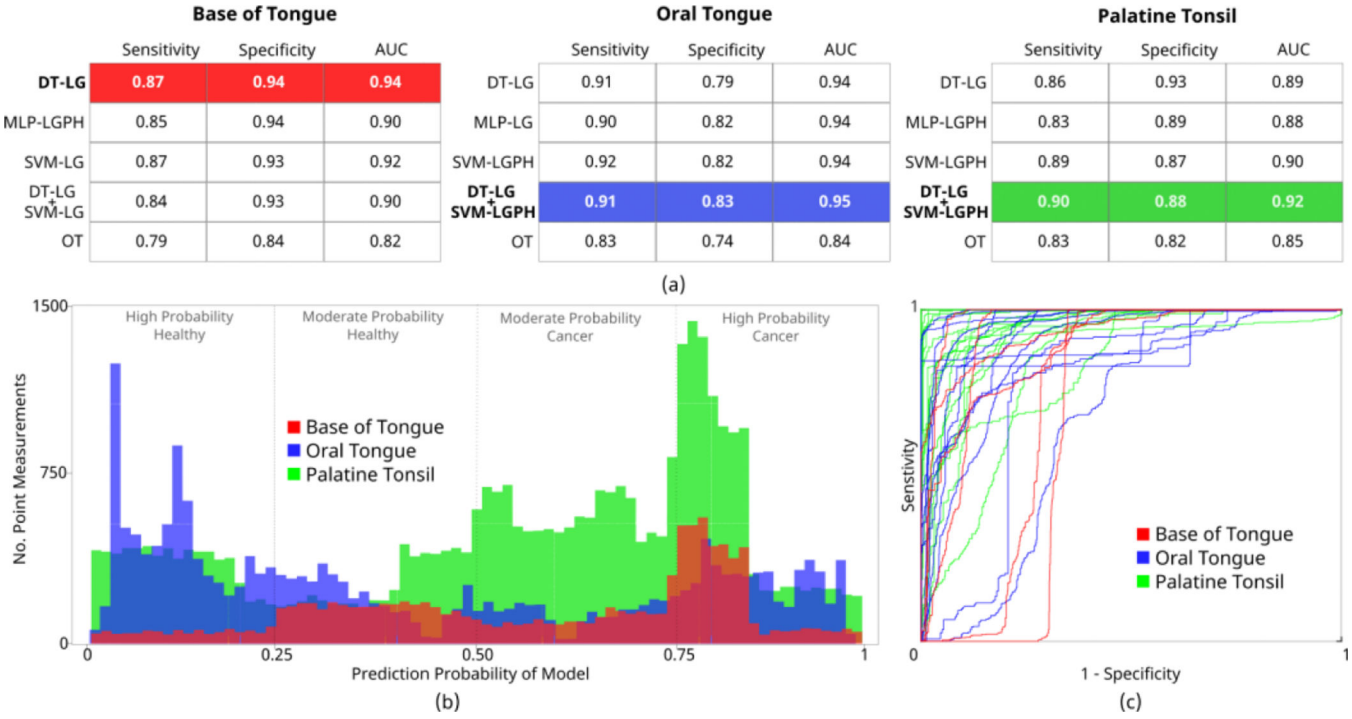

Fig. 3.

Region-level classification performance of the anatomy-specific classifier in discriminating healthy vs. cancer for the base of the tongue (red), oral tongue (blue), and palatine tonsil (green). (a) Sensitivity, specificity, and AUC were obtained from the best-performing classifiers for handcrafted and non-handcrafted-based models. The highlighted row indicates the best-performing classification model for each anatomy. (b) Prediction probability histogram of the three anatomy-specific classifiers, the base of the tongue (DT-LG), the oral tongue (DT-LG + SVM-LGPH), and the palatine tonsil (DT-LG + SVM-LGPH) illustrating the confidence of the model while making predictions. The prediction probability closer to '0' and '1' infers very confident predictions of either cancer or healthy tissue, whereas the prediction probability between '0.25' and '0.75' represents a prediction with less confidence. (c) Patient-level ROC curves for healthy vs. cancer classification. Each curve is color-coded by anatomy. Patients with a minimum acquisition of 100 healthy points and 100 cancer point measurements were included in the ROC curves. Abbreviations: *DT-LG* = Decision Tree & Laguerre Features, *MLP-LGPH* = Multi-Layer Perception Model, Laguerre, & Phasor-Based Features, *SVM-LG* = Support Vector Machine & Laguerre Features, *OT*: Optimal Transport.

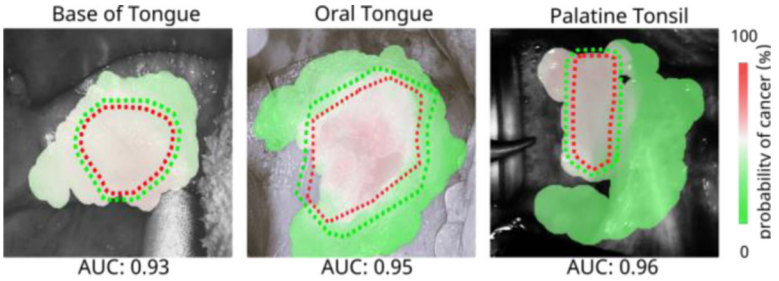

Fig. 4.

Classifier visualizations of "*Healthy vs. Cancer*" over the healthy and cancer boundary from the histological slice annotations for in vivo patient scan for the base of the tongue, oral tongue, and palatine tonsil anatomies for healthy (green) and cancer (red) tissue. The semi-transparent classification probability map is overlaid on the grayscale image of the in vivo tissue. The green and red dotted lines highlight the healthy and cancer boundary derived from the histopathology via annotated H&E staining images. The classification probability map is visualized as a spectrum of cancer probability to aid the surgeon with intraoperative cancer margin. The label "*probability of cancer (%)*" is defined by the prediction probability of the classification model

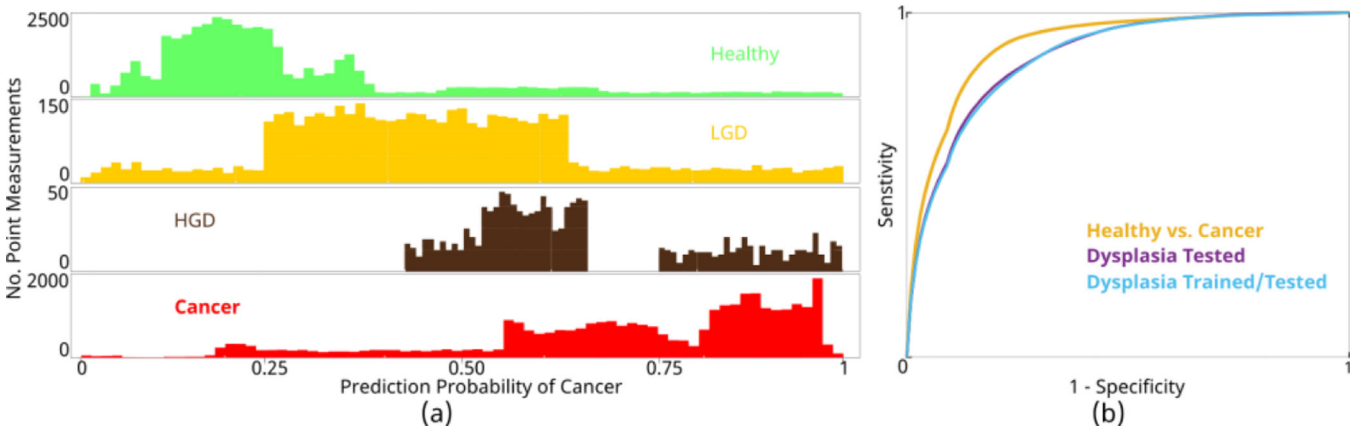

Fig. 5.

Classification performance of the anatomy-specific classifier in discriminating healthy vs. cancer vs. dysplasia for head and neck patients. (a) Prediction probability histogram of the anatomy-specific classifier "*Dysplasia Tested*" while predicting healthy, low-grade dysplasia, high-grade dysplasia, and cancer tissues. (b) Average ROC-curve of the anatomy-specific classifiers. "*Healthy vs. Cancer*" is the average ROC-curve of the anatomy-specific classifier trained on healthy vs. cancer and tested on healthy vs. cancer "*Dysplasia Tested*" is the average ROC-curve of the anatomy-specific classifier trained on healthy vs. cancer and tested on healthy vs. cancer vs. dysplasia. "*Dysplasia Trained/Tested*" is the average ROC-curve of the anatomy-specific classifier trained on vs. cancer vs. dysplasia and tested on healthy vs. cancer vs. dysplasia.

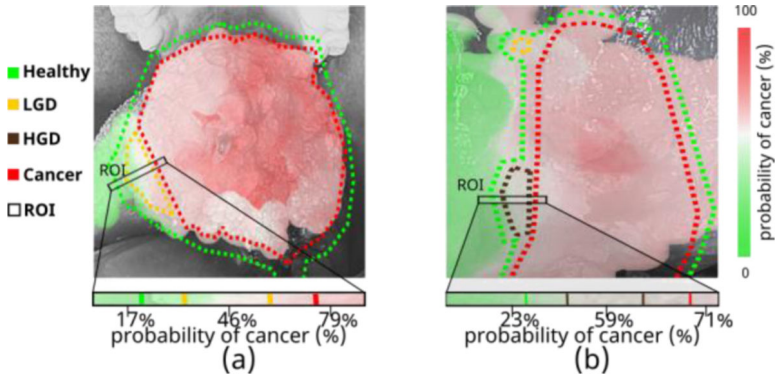

Fig. 6.

Classifier visualizations of "*Dysplasia Tested*" over the healthy, LGD, HGD, and cancer boundary from the histological slice annotations for in vivo patient scans with (a) low-grade dysplasia and (b) high-grade dysplasia. The semi-transparent classification probability map is overlaid on the grayscale image of the in vivo tissue. The green and red dotted lines highlight the healthy and cancer boundary, while the yellow and brown dotted lines highlight the LGD and HGD boundary derived from the histopathology via annotated H&E staining images. The classified tissue is visualized as a spectrum of cancer probability to aid the surgeon with intraoperative cancer margin assessment. The ROI on the tissue sample illustrates the probability of cancer assigned for dysplasia tissue as the "gray area" between healthy and cancerous tissue types. The label "*probability of cancer (%)*" is defined by the prediction probability of the classification model

TABLE I

ANATOMY AND TISSUE LABEL BREAKDOWN FOR THE 85 PATIENTS

Anatomy	No. Patient (N)	Tissue Label	FLIm Point (n)
Palatine tonsil	27	*Healthy*	54,762
Oral tongue	26	*LGD*	4844
Base of tongue	17	*HGD*	969
Other anatomy	15	*Lymphoid*	1510
		Cancer	36,316

Abbreviations: *LGD* = Low-Grade Dysplasia, *HGD* = High-Grade Dysplasia

TABLE II

TRAINING SET SELECTION FOR ANATOMY-SPECIFIC CLASSIFICATION MODEL DEVELOPMENT. DIFFERENT TRAINING SET GROUPINGS ARE TESTED USING THE DECISION TREE (DT) MODEL BASED ON LAGUERRE (LG) FEATURES. THE CONFIGURATION WITH BOLD TEXT INDICATES THE BEST-PERFORMING TRAINING SET USED FOR DEVELOPING THE CLASSIFICATION MODEL

Training Set	Configuration	Testing
Pooled anatomy	**BOT+OT+PT+OA**	*Base of Tongue*
Anatomy specific	BOT	
Anatomic region	BOT+PT	
Pooled anatomy *Anatomy specific*	BOT+OT+PT+OA **OT**	*Oral Tongue*
Pooled anatomy	BOT+OT+PT+OA	*Palatine Tonsil*
Anatomy specific	PT	
Anatomic region	**PT+LT** * BOT+PT	

* Two patients from the ‘base of the tongue’ subset represent ‘lingual tonsil.’ Abbreviations: *BOT* = Base of Tongue, *OT* = Oral Tongue, *PT* = Palatine Tonsil, *LT* = Lingual Tonsil, *OA* = Other Anatomy.

TABLE III

POINT-LEVEL AND REGION-LEVEL HEALTHY VS. CANCER CLASSIFICATION FOR ANATOMY-SPECIFIC CLASSIFIER ON FLIM SCANS, MEAN(SD).

Anatomy	*Model*	*Point*			*Region*		
		Sensitivity n=29260	**Specificity n=44064**	**AUC**	**Sensitivity N=59**	**Specificity N=70**	**AUC**
Base of Tongue	DT-LG	0.77(0.21)	0.85(0.13)	0.82(0.08)	0.87(0.13)	0.94(0.11)	0.94(0.09)
Oral tongue	DT-LG+SVM-LGPH	0.81(0.18)	0.74(0.21)	0.88(0.08)	0.91(0.16)	0.83(0.20)	0.95(0.04)
Palatine Tonsil	DT-LG+SVM-LGPH	0.74(0.20)	0.84(0.14)	0.82(0.11)	0.90(0.11)	0.88(0.18)	0.92(0.06)

Abbreviations: *DT-LG* = Decision Tree & Laguerre Features, *SVM-LGPH* = Support Vector Machine, Laguerre Features, & Phasor-Based Features

TABLE IV

REGION-LEVEL HEALTHY VS. CANCER VS. DYSPLASIA CLASSIFICATION PERFORMANCE FOR ANATOMY-SPECIFIC CLASSIFIER ON FLIM SCANS, MEAN(SD).

Classifier	Sensitivity	Specificity	AUC
Healthy vs. Cancer	0.90(0.18)	0.88(0.19)	0.94(0.10)
Dysplasia Tested	**0.89(0.17)**	**0.87(0.17)**	**0.92(0.08)**
Dysplasia Trained/Tested	0.85(0.20)	0.91(0.15)	0.91(0.09)

Author Manuscript

Author Manuscript

Author Manuscript

TABLE V

REGION-LEVEL HEALTHY VS. CANCER CLASSIFICATION PERFORMANCE FOR HPV+ PATIENTS OF PALATINE TONSIL, MEAN (SD).

Condition	Sensitivity	Specificity
p16- (Boundary = 0.5)	0.98(0.03)	0.85(0.29)
p16+ (Boundary = 0.5)	0.74(0.29)	0.90(0.16)
p16+ (Boundary = 0.4)	**0.88(0.18)**	**0.81(0.21)**

Abbreviations: *p16+ SCC* = HPV-Positive Squamous Cell Carcinoma, *p16- SCC* = HPV-Negative Squamous Cell Carcinoma