Identification of colorectal malignancies enabled by phasor-based autofluorescence lifetime macroimaging and ensemble learning

João L. Lagarto,[a,*] Alberto I. Herrando,[a,b] Rafaela Rego,[c] Laura Fernández,[b] José Azevedo,[b] Hugo Domingos,[b] Pedro Vieira,[b] Amjad Parvaiz,[b] Vladislav I. Shcheslavskiy,[d,e] Pedro G. Silva,[a] Mireia Castillo-Martin,[c]

[a]Champalimaud Foundation, Biophotonics Platform, Lisbon, Portugal
[b]Champalimaud Foundation, Digestive Unit, Lisbon, Portugal
[c]Champalimaud Foundation, Pathology Service, Biophotonics Platform, Lisbon, Portugal
[d]Becker and Hickl GmbH, Berlin, Germany
[e]Privolzhsky Research Medical University, Nizhny Novgorod, Russia

Abstract

Significance: Colorectal cancer (CRC) remains one of the most frequent cancers and a leading contributor to cancer-associated mortality globally. CRCs are often diagnosed at an advanced stage, which leads to high mortality and morbidity. This outcome is exacerbated by high rates of recurrence and postoperative complications that contribute substantially to poor prognosis. Advancements in endoscopic assessment have improved CRC prevention, early detection, and surveillance over the years. Yet, CRC remains one of the most significant health challenges of the 21st century. Label-free optical spectroscopy methods have long been explored as potential partners to endoscopy, not only to enhance diagnostic accuracy but also to confer predictive capabilities to endoscopic evaluations.

Aim: We investigated the potential of time-resolved autofluorescence measurements excited at 375 nm and 445 nm to correctly classify benign and malignant tissues in CRC surgical specimens from 117 patients.

Approach: Multiparametric autofluorescence lifetime data were collected in two distinct datasets, which were used for training (n = 73) and testing (n = 44) a supervised ensemble learning classification model, with standard histopathology assessment serving as ground truth.

Results: Using 5-fold cross-validation, we achieved 82.6 ± 0.02% sensitivity, 90.4 ± 0.01% specificity, 87.4 ± 0.01% accuracy, and 0.941 ± 0.004 area under the curve (AUC) for training data. Evaluation on unseen test data yielded similar results, with 85.2% sensitivity, 84.5% specificity, 84.8% accuracy, and 0.915 AUC.

Conclusions: While preliminary, our findings underscore the potential impact of AI-assisted autofluorescence lifetime measurements in advancing CRC prevention, early detection, and surveillance efforts.

Keywords: autofluorescence lifetime, colorectal cancer, machine learning, phasors, label-free, diagnosis, disease monitoring.

***João L. Lagarto, E-mail: joao.lagarto@research.fchampalimaud.org

1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and ranks second in cancer-related mortality. This sobering statistic is largely attributed to nearly 70% of patients being

diagnosed at an advanced stage of the disease, when treatment options are more restricted and less effective. Despite some encouraging trends indicating an overall decrease in incidence and mortality in recent years, there has been a steady rise in the incidence of CRC among younger individuals under 50 years of age [1]. Currently, this demographic accounts for 13% of colon cancers and 16% of rectal cancers, and these figures are expected to nearly double by 2030, as screening remains particularly low among younger individuals [2]. As CRC stands as a significant health challenge for the foreseeable future, it is imperative to improve prevention and early detection of precursor lesions that can become malignant over time and, in this way, decrease CRC-associated mortality.

Endoscopic assessment is the cornerstone of CRC prevention, early detection, and surveillance, enabling detection of pre-malignant lesions and early-stage cancer across the anus, rectum, and the entire length of the colon [3]. Despite its effectiveness, the information obtained from the standard endoscopic evaluation is limited and may not fully characterize the structural, molecular, and metabolic features of detected lesions. This limitation can hinder the ability to accurately predict which polyps will progress to cancer, leading to challenges in determining the optimal surveillance strategy for patients. Chromoendoscopy and other advanced imaging techniques such as Narrow Band Imaging (NBI) or Flexible Spectral Imaging Color Enhancement (FICE) can enhance the discriminatory potential of standard endoscopy [4]. Yet, characterization of lesions is still challenging and often correlated to the endoscopist's training and experience [5].

To overcome these limitations, advanced label-free optical spectroscopy methods have emerged as promising partners to standard white light endoscopy, owing to their increased sensitivity to molecular, structural, and metabolic alterations in tissues, without requiring the introduction of potentially toxic exogenous labels. Among them, multispectral autofluorescence

lifetime imaging and spectroscopy have been widely exploited with demonstrated success in a broad range of clinical applications, with particular emphasis on cancer detection and margin assessment [6–10]. Moreover, multiple studies have shown that autofluorescence lifetime measurements can provide substantial diagnostic information that goes beyond classification of benign and malignant tissues [11]. In recent years, translation of this technique into clinical practice has been accelerated by the flourishing of Artificial Intelligence (AI) systems and the development of increasingly sophisticated machine learning and deep learning models. AI models have been employed in data processing to outperform traditional methods [12], tissue classification [13,14], delineation of margins [15,16], or determination of metabolic phenotypes [17].

Clearly, AI-assisted autofluorescence lifetime measurements can have tremendous impact in clinical decision-making, by enabling rapid tissue assessment, precise delineation of margins, or categorization of lesions, beyond what conventional systems can offer. In a previous study, we characterized the tissue-level autofluorescence signatures of normal, adenoma, and tumor tissues considering the underlying clinicopathological features [18]. While we reported significant differences between tissues in various characteristics, there was considerable overlap in the data owing to a large intra- and interpatient variability. Moreover, in a selected number of cases, we observed opposite autofluorescence patterns that we could not explain, making the interpretation of the data and its categorization quite challenging with the naked eye. Here, we continued our work towards practical clinical implementation, focusing on quantifying the differences previously observed and demonstrating that fiber-based autofluorescence lifetime measurements can provide rapid diagnostic information with high accuracy. To that end, a supervised ensemble learning model was applied to multiparametric autofluorescence lifetime data obtained from CRC surgical specimens for classification of colorectal tissues and identification of malignant lesions, using

histopathologic assessment as ground truth. Data were collected from two distinct datasets (total of 117 patients), which were used independently for training (n = 73 patients) and testing (n = 44 patients) of the supervised model. We tested and evaluated various predictive classification models, each based on different sets of spectroscopic features derived from phasor analysis of autofluorescence decay data. The results presented here, while preliminary, are indicative of the potential of AI-assisted autofluorescence lifetime measurements for classification of colorectal lesions and identification of malignancies.

2 Methods

2.1 Autofluorescence lifetime setup

The optical instrument used for collection of autofluorescence data consisted of a time and spectrally resolved autofluorescence macro-imaging system with multiplexed excitation at 375 nm and 445 nm. A complete description of this system is provided elsewhere [18,19]. Briefly, two ps-pulsed laser sources (BDS-SM-375 and BDS-SM-445, Becker and Hickl GmbH, Germany) operating at 20 MHz repetition rate were multiplexed at 50 Hz, so that each wavelength excited the sample alternately in 20 ms periods. Excitation light was delivered to the sample via a custom optical fiber bundle (FiberTech Optica, Canada) with a single 300 µm fiber for excitation (NA=0.22). The resulting autofluorescence signal was collected by a set of 200 µm optical fibers (NA=0.22) circularly arranged around the excitation fiber and subsequently delivered to the detection system consisting of a wavelength selection module and three photon-counting hybrid detectors (HPM-100-40-CMOUNT, Becker and Hickl GmbH). The wavelength selection module included of a set of mirrors and filters that divided the autofluorescence signal in three spectral bands of interest, selected to isolate the autofluorescence from key endogenous fluorophores: 380

± 20 nm, 472 ± 14 nm, and 525 ± 25 nm. This excitation-collection geometry provides a total of five detection channels, see Supplementary Table 1. The hybrid detectors were connected to a single router (HRT-41, Becker and Hickl GmbH) that serialized photon events to a time-correlated single photon counting (TCSPC) acquisition card (SPC-130 EM, Becker and Hickl GmbH) that recorded the fluorescence decay curve for each detection channel.

A key feature of this system is the ability to record TCSPC autofluorescence lifetime data in bright conditions, owing to the synchronization of the fluorescence acquisition with an external light source that provides periodic white illumination to the sample at 50 Hz [19,20]. A USB color camera (FFY-U3-16S2C-C, FLIR, USA) was also used to record and provide spatial feedback of the measurements. The spatial resolution of the system is primarily determined by the geometry of the fiber arrangement, NA of the fibers, and probe-to-target distance. Variations in probe-to-target distance can cause fluctuations in the excitation spot size and collection efficiency, while inconsistent scanning speed may lead to spatial undersampling and reduced signal quality [20]. While it is challenging to accurately determine the spatial resolution of our system, we estimate it to be ~1 mm, under ideal conditions.

It is noteworthy that throughout this study we used two slightly different fiber bundles, consisting of either six or seven collection fibers. We used this difference as basis for splitting the data into training and test sets, as described in detail below.

2.2 Sample collection and data acquisition

Colorectal surgical specimens were collected under the Champalimaud Foundation Biobank Informed Consent, clinical protocol 2021020203 approved by the Champalimaud Foundation Ethics Committee. Following surgical resection, samples were immediately transported to the Pathology Service laboratory, where they were opened and cleaned for autofluorescence

133 measurements, which were typically carried out within 1 hour of the final vascular ligation and 15
134 minutes of complete resection [18]. The autofluorescence acquisition consisted of hand-scanning a
135 large region of the specimen using the fiber optic probe (areas as large as 25 cm^2 were measured),
136 including the suspicious malignant lesion and the surrounding normal tissue. The position of the
137 fiber was determined for each TCSPC measurement using the images captured with the color
138 camera, thus permitting autofluorescence maps to be generated, as illustrated in Fig. 1. Following
139 measurements, specimens were routinely processed for diagnosis. A total of 118 colorectal
140 samples were collected from 117 patients.

141 *2.3 Data preprocessing*

142 Pre-processing of data encompassed three crucial steps. First, raw single-point
143 autofluorescence lifetime data were analyzed and the corresponding autofluorescence maps
144 generated. Second, regions of interest (ROI) were defined in each specimen in both lesional and
145 non-lesional areas, guided by histopathology diagnosis. Finally, pixel-wise autofluorescence data
146 were averaged within each ROI (feature extraction).

147 *2.3.1 Analysis of autofluorescence data*

148 Autofluorescence intensity decays were processed according to the phasor transformation [21]
149 using the instrument response function (IRF) as reference (τ_{ref} = 0 ns). In detail, each
150 autofluorescence decay *I(t)* was transformed from the time-domain to the Fourier space and the
151 corresponding phasor coordinates *g* and *s* were calculated according to Eq. 1 and 2, where *T* is the
152 period of the laser repetition (*T* = 50 ns).

$$g = \int_0^T I(t)\cos\left({}^{2\pi t}/_{T}\right)dt \Big/ \int_0^T I(t)dt \qquad (1)$$

$$s = \int_0^T I(t)\sin\left({}^{2\pi t}/_{T}\right)dt \Big/ \int_0^T I(t)dt \quad (2)$$

153

154 From the coordinates g and s, characteristic phase and modulation lifetimes (τ_p and τ_m,
155 respectively) were obtained from the following relations:

$$\tau_p = \frac{T}{2\pi}\frac{s}{g} \quad (3)$$

$$\tau_m = \frac{T}{2\pi}\sqrt{\frac{1}{g^2+s^2}-1} \quad (4)$$

156

157 All fluorescence decays were processed for background correction prior to the phasor
158 transformation. Measurements containing fewer than 50 photons and 10 peak counts were
159 excluded from the analysis to guarantee the accuracy and precision of lifetime estimation [21]. The
160 IRF was measured for all channels (by removing band-pass filters) using excitation light scattered
161 off a reflective surface. Fluorescence lifetime measurements of the system were validated using
162 POPOP (τ = 1.36 ns in ethanol [22]) and Coumarin 6 (τ = 2.72 ns in ethanol [23]), for 375 nm and 445
163 nm excitation, respectively.

164 In addition to the fluorescence lifetime analysis, we calculated the fractional autofluorescence
165 intensity F measured in each detection channel with respect to the total autofluorescence signal for
166 each excitation wavelength. Hence, we extracted five parameters representative of the
167 autofluorescence decay curve in each channel (g, s, τ_p, τ_m, F), i.e. a total of 25 parameters for five
168 detection channels. The feature pool was completed by the normalized optical redox ratio (RR),
169 calculated as follows,

$$RR = \frac{I_2}{I_2 + I_5} \equiv \frac{NAD(P)H}{NAD(P)H + FAD} \quad (5)$$

170

where I_2 and I_5 denote the absolute autofluorescence intensity in detection channels 2 and 5, respectively.

2.3.2 Diagnosis and data labelling

Following optical measurements, surgical specimens were sent for routine histological processing and diagnosis, which established the ground truth for the supervised learning model. The process of obtaining ground truth data (labelling) from fresh surgical specimens and subsequent feature extraction is depicted in Fig. 1. Because specimens were sectioned vertically and perpendicular to the surface of measurements (i.e. to the optical map) according to the bread loaf technique, it was impossible to obtain a one-to-one correlation of histological and optical features. Accordingly, the exact origin of each slice on the measured specimen, i.e. prior to fixation in formalin, could not be determined either. These limitations cannot be easily circumvented without causing major disruptions to the histopathology workflow; thus, our analysis relied on the macroscopic assessment of the lesions by experienced pathologists and subsequent confirmation by microscopic examination of the most representative histology slide. Based on this assessment, we traced and identified the lesions as accurately as possible in the white light image captured during optical measurements from which we selected ROIs, as illustrated in Figure 1. To mitigate the limitations, ROIs were drawn conservatively within the lesion, following simple guidelines: 1) ROIs were kept small (~20×20 pixels); 2) in case the borders of the lesion were ill-defined, ROIs were drawn closer to the center of the lesion; 3) ROIs in normal tissues were drawn as remotely as possible from any visible lesion. Where possible, a maximum of three ROIs were drawn per sample and tissue type, as sparsely distributed as possible. This was to capture the spatial heterogeneity of the tissue. Optical data were averaged within in each ROI. ROIs were labelled as *benign* or *malignant* according to histopathological diagnosis, where *benign* included benign

lesions (i.e. adenomas) and normal tissues. We opted for a binary rather than a multiclass classification model given the low number of adenoma lesions in our dataset. The number of patients and ROIs included in our analysis is indicated in Table 1.

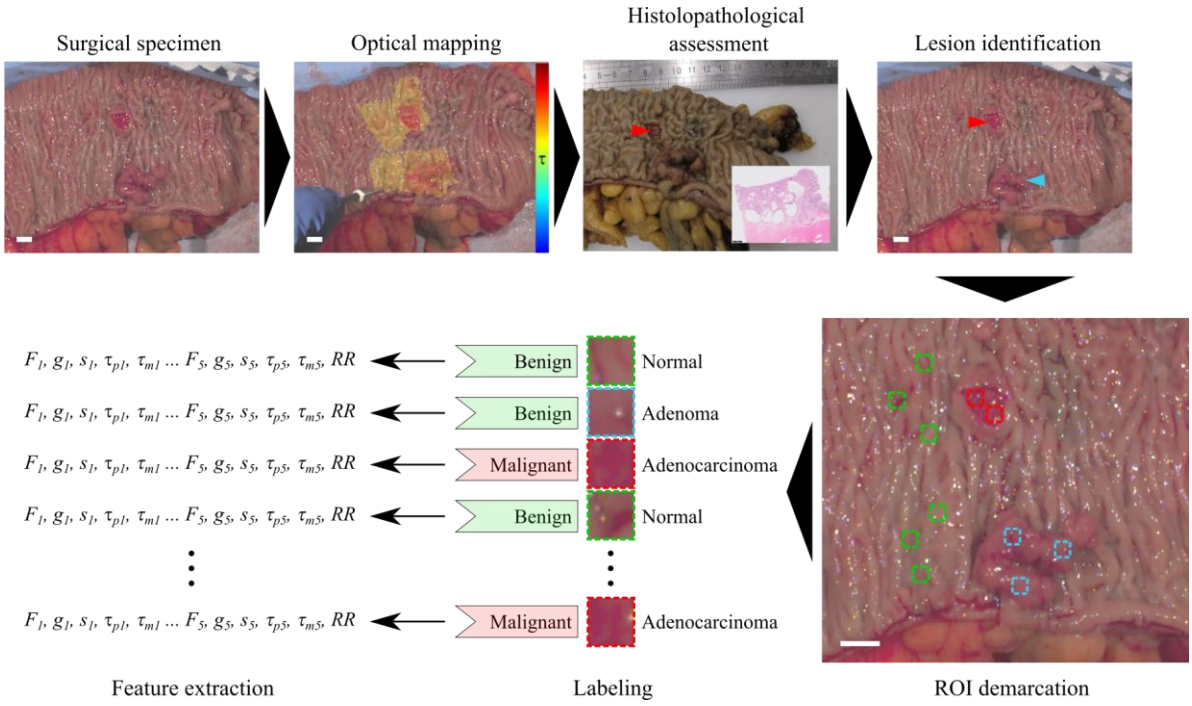


Figure 1. Study workflow for every collected sample. The surgical specimen is transported to the optical laboratory immediately after resection and typically within 1 hour of the last vascular ligation. Optical measurements are carried out in fresh tissue before histopathological processing and assessment. Following diagnosis, benign and malignant lesions are identified in the white light image, and regions of interest (ROI) are demarcated. Autofluorescence parameters are averaged within each ROI and mean values are taken as characteristic features. Each ROI is labelled as benign or malignant according to histopathological diagnosis. Scale bars 10 mm.

2.4 Tissue classification

2.4.1 Training and test datasets

To ensure validity and reliability of the results, data were divided into two independent groups, as indicated in Table 1. Data were assigned to either training or test group according to the fiber

bundle used during the acquisition: training data were acquired with a fiber bundle consisting of seven collection fibers and test data were acquired with an identical fiber bundle, but with six collection fibers only. As expected, difference between fibers had no impact in the measured fluorescence lifetimes (see Fig. 2A) given the ratiometric nature of the fluorescence lifetime measurement. A total of 672 observations (ROIs) were included in this study (see Table 1 for details). The training dataset included 428 observations (63.7%) of which 167 (39.0%) were labelled as malignant. The test dataset consisted of 244 observations (36.3%), including 108 (44.2%) labeled as malignant. The number of observations in both datasets was relatively well-balanced between the two classes (benign and malignant), thus requiring no additional balancing corrections. The raw data from the training dataset served as basis for a recent study on optical characterization of CRC [18].

Table 1. Patient numbers and observations in training and test datasets.

	Training data		**Test data**	
	N	**%**	**N**	**%**
Number of patients	73	-	44	-
Number of samples measured	74	-	44	-
Number of observations (ROIs)	428	(63.7%)	244	(36.3%)
Normal	237	(55.4%)	128	(52.5%)
Adenoma	24	(5.6%)	8	(3.3%)
Adenocarcinoma	167	(39.0%)	108	(44.2%)
Benign	261	(61.0%)	136	(55.8%)
Malignant	167	(39.0%)	108	(44.2%)

2.4.2 Hyperparameter optimization and training

An Adaptive Boosting algorithm (AdaBoostM1) was evaluated for binary classification of autofluorescence data in CRC samples [24]. Boosting is an ensemble learning method that

sequentially combines the prediction of multiple weak learners (decision trees) to obtain a strong classifier. The model was implemented in Matlab 2022a (Mathworks, USA) using the *fitcensemble* function. The machine learning pipeline is illustrated in Fig. 2B. The model was first explored with training data for various combinations of hyperparameters using 5-fold cross validation. Model hyperparameters (learning rate, number of weak learners, and maximum nodes per weak learner) were tuned iteratively using Bayesian optimization to minimize the cross-validation loss function over 60 evaluation cycles. A range of values was specified for each hyperparameter with the aim of overcoming underfitting and overfitting (see Table S2). The optimal hyperparameter vector was found at the minimum cross-validation loss and used to estimate performance of the model. The model was then trained using the complete training dataset, i.e. without partitioning data for cross-validation. Supplementary Fig. S1 shows the variation of the objective function over 60 optimization cycles.

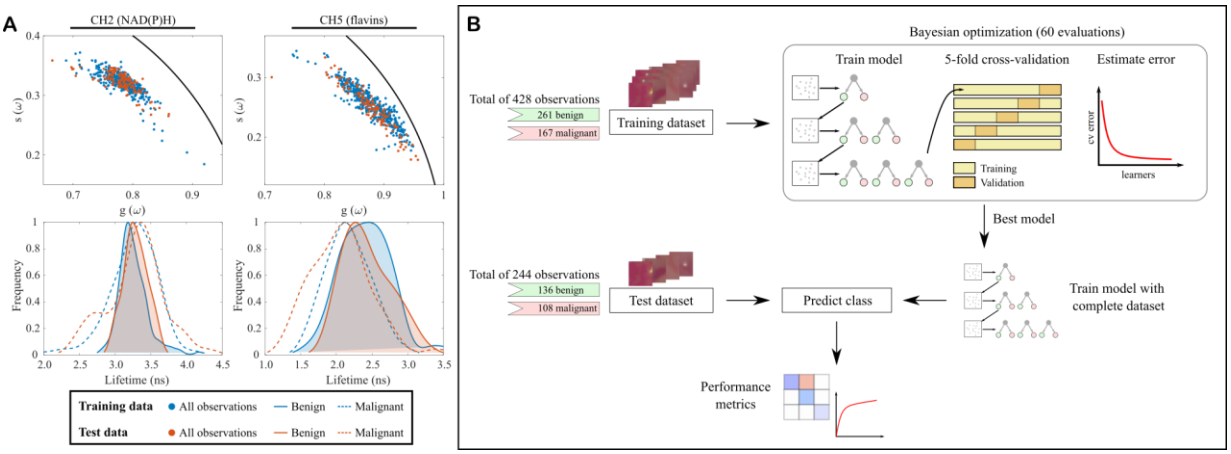


Figure 2. A) Phasor plots (top row) and phase lifetime (bottom) distribution in benign and malignant tissues for training and test dataset, measured in CH2 (left) and CH5 (right), respectively. B) Machine learning pipeline. A dataset consisting of 428 observations is used to train a boosting algorithm. The best hyperparameter vector was found through Bayesian optimization with 5-fold cross-validation. Final model performance was evaluated against an independent test set consisting of 244 observations.

2.4.3 Model evaluation and performance metrics

Performance of the model was first estimated on training data using the optimal set of hyperparameters with 5-fold cross validation. The model was trained 20 times and performance metrics were averaged out. Independent evaluation of the model was conducted on the test dataset. Common metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic curve (AUC-ROC) were computed. In addition, we measured the Mathews correlation coefficient (MCC), which has been shown to be more informative than other metrics for evaluation of binary classification models [25]. Since our aim was to identify malignant lesions from surrounding normal tissues and benign lesions, such as adenomatous polyps, the following definitions were used: true positive (TP) rate was defined as the percentage of correctly classified malignant lesions; true negative (TN) rate was defined as the percentage of correctly classified normal tissues or benign lesions; false positive (FP) rate was defined as the percentage of normal tissue classified as malignant; and false negative (FN) rates were defined as the percentage of malignant tissues classified as normal.

3 Results

3.1 Class separation

We first investigated the autofluorescence lifetime signatures of benign and malignant tissues obtained from test data. Fig. 3A shows the distribution of data in phasor coordinates for all detection channels, where each observation corresponds to a single point in the phasor cloud. The corresponding phase and modulation lifetimes are plotted in Fig. 3B. Redox ratio and normalized fluorescence intensity data are provided in Supplementary Fig. S2. These data are consistent and

support our previous results on the training data [18]. Specifically, malignant tumors exhibit shorter lifetimes in channels 1 and 5, whereas lifetime differences in the remaining channels are more subtle. Likewise, the autofluorescence signals emanating from NAD(P)H do not appear to offer a relevant source of contrast between tissues. Rather, in detection channels 2 and 3, results depict a wider distribution of data originating from malignant tissues. Such findings suggest increased heterogeneity, potentially associated to diverse metabolic phenotypes harbored in the tumors. This heterogeneity is best illustrated in the phasor plots but also evident in Fig. 3B.

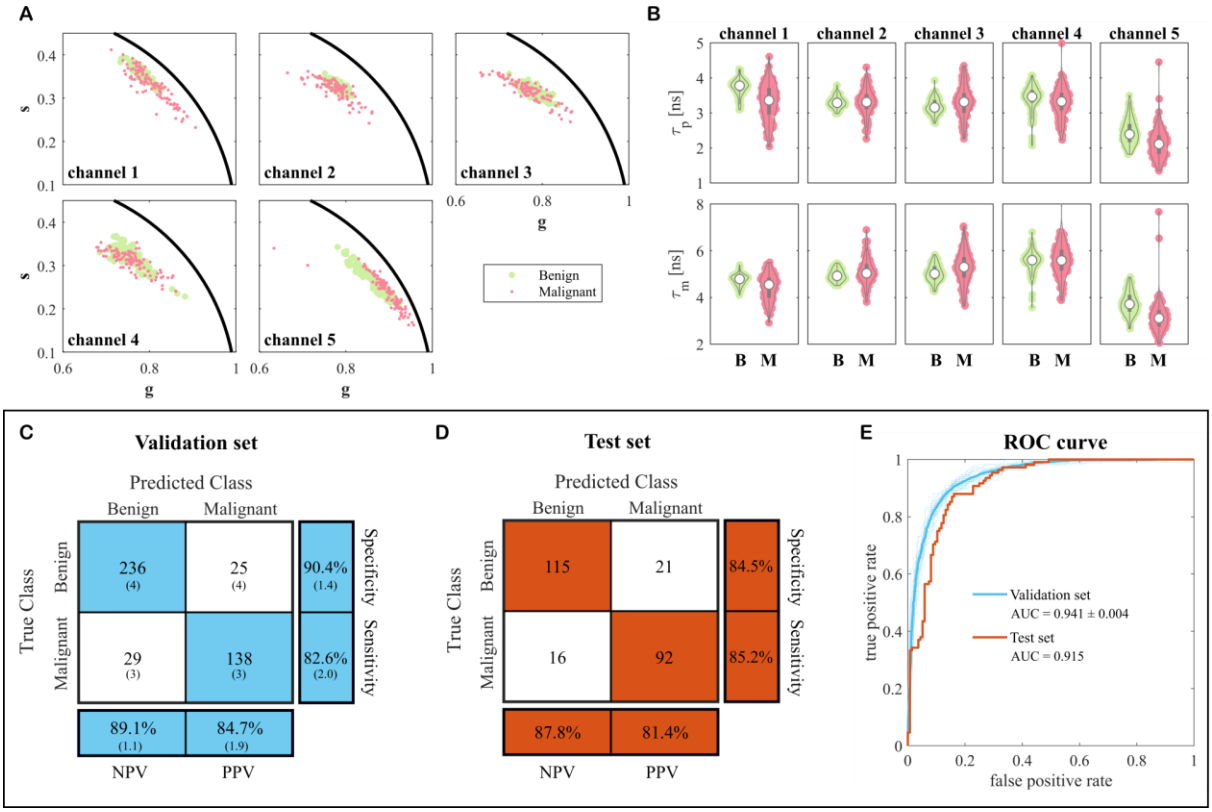


Figure 3. A) Phasor distribution of test data in all detection channels. B) Average phase and modulation lifetimes measured in benign (B) and malignant (M) ROIs. Panels C and D show confusion matrices for training (cross validated) and test sets, respectively. TP, TN, FP, and FN values are rounded to the closest integer following an averaging of 20 iterations. Percentage values are calculated using the original "unrounded" values. Values in parenthesis represent SD. E) ROC curves for cross validation (in cyan) and test (in orange) sets. Solid lines indicate

the average ROC curves for 20 iterations of the model. Shaded lines in blue represent the ROC curve for each iteration.

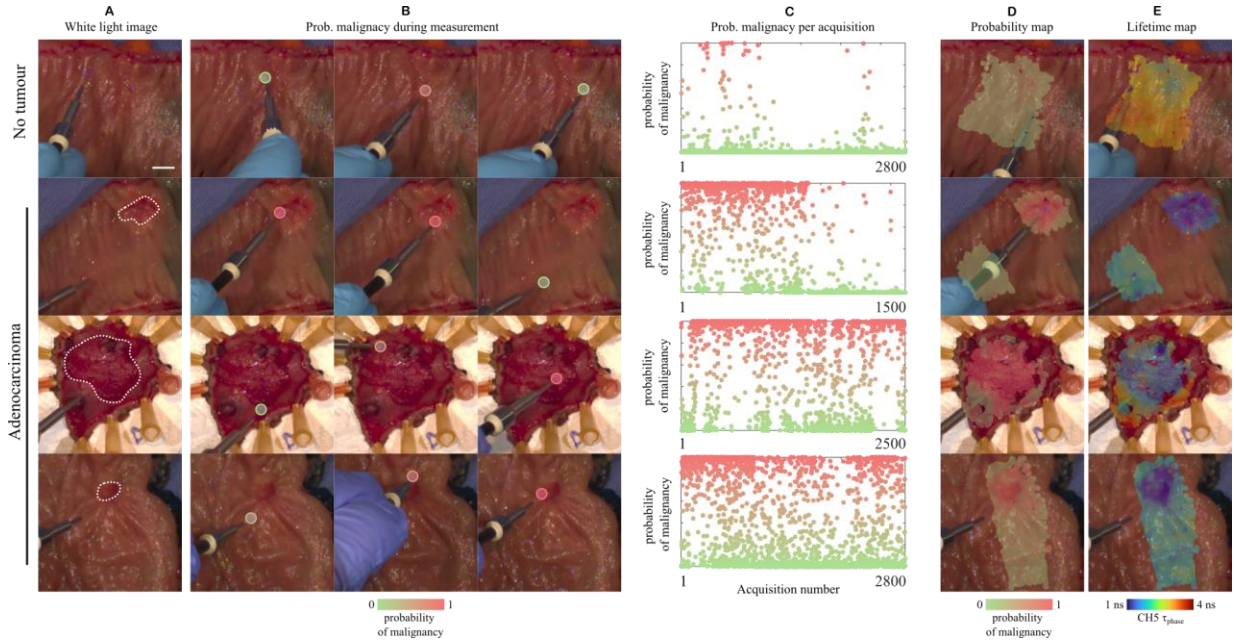
3.2 Tissue classification and model performance

Table 2. Performance metrics for training and test datasets. Metrics on training data were obtained by averaging over 20 iterations of cross validation.

	Training data	Test data
Number of observations	428	244
Sensitivity	0.826 ± 0.020	0.852
Specificity	0.904 ± 0.014	0.845
Accuracy	0.874 ± 0.008	0.848
PPV	0.847 ± 0.019	0.814
NPV	0.891 ± 0.011	0.878
MCC	0.733 ± 0.022	0.695
AUC-ROC	0.941 ± 0.004	0.915

Following initial characterization of the test dataset and confirmation of the trends previously obtained with the training dataset, we evaluated the ability of our autofluorescence lifetime system to predict malignancy from single point measurements using the supervised ensemble learning model. Performance of the model was first estimated on training data using the optimal set of hyperparameters with 5-fold cross-validation, as previously described in 2.4.2. The results presented in Fig. 3C and Table 2 indicate the model performs well on the training dataset, achieving an accuracy of 87.4% ± 0.01%, AUC of 0.941 ± 0.004, and MCC of 0.733 ± 0.022. It excels particularly in the classification of benign tissues, with NPV ranging from 88.0% to 91.2% and specificity ranging from 89.0% to 91.8%. The estimated performance is slightly lower in the classification of malignancies, with PPV ranging from 82.8% to 86.6% and sensitivity ranging from 80.6% to 84.6%. The model performed equally well on the test data (Fig. 3D and Table 2), indicating good generalization to new and unseen data. The model correctly predicted 84.8% of the new observations, with higher sensitivity (85.2%) compared to that estimated with training

297 data. The AUC and MCC on test data were 0.915 and 0.695, respectively, indicating a strong
298 positive correlation between the predicted and true classes.



299

300 Figure 4. Investigation of ensemble learning model-based probability of malignancy in tissue. A) White light image
301 at time zero of normal and tumor specimens. Dotted lines delineate malignant lesions. B) Probability of malignancy
302 estimated from single point measurements. C) Probability of malignancy for every fluorescence measurement during
303 an entire specimen acquisition. D) Probability and E) fluorescence lifetime maps. Scale bar 10 mm.

304 After evaluating the model's performance on the independent dataset, we next investigated its
305 applicability to individual single point measurements, as opposed to ROIs, while maintaining
306 comparable performance. This investigation aimed to explore the applicability of the model in a
307 scenario that more closely mimics the real-world application (e.g. single-point endoscopic
308 navigation for identification of adenomas) and the generation of probability maps of malignancy
309 from which resection margins can be objectively determined. Figure 4 illustrates representative
310 results, showcasing one sample with no malignancy in the top row, contrasted with three samples
311 exhibiting malignancy. Measurements report higher probability of malignancy as the fiber probe
312 is moved over the tumoral regions, as depicted in Fig. 4B. As the fiber moves away from the tumor,

the probability of malignancy decreases accordingly. In regions devoid of malignant tumor (as in the top row), the model fails the prediction in several measurements, as depicted in Fig. 4C. This is because the model is not 100% accurate. Irrespective of that, the averaging of sequential single-point measurements mitigates these inaccuracies, resulting in relatively smooth probability maps of malignancy, from which regions of tumor can be clearly identified.

3.3 Towards simplification of the optical setup

We next investigated the impact of specific sets of features on model performance. Specifically, we aimed to understand how each channel impacted the results and whether we could achieve comparable performance with a less complex optical system. We first estimated the importance of each feature on the optimized model, over 20 iterations of cross-validation, as described in section 2.4. This was achieved by summing estimate predictor importances over all weak learners in the ensemble. Model predictors are ranked by their importance score in Fig. 5A and averaged out by channel in Fig. 5B. As hinted by our data presented in Fig. 3, channels 1 and 5 are the most relevant contributors to the binary classification, except for RR. NAD(P)H autofluorescence signals (channels 2 and 3) were found to be the least important contributors to the classification.

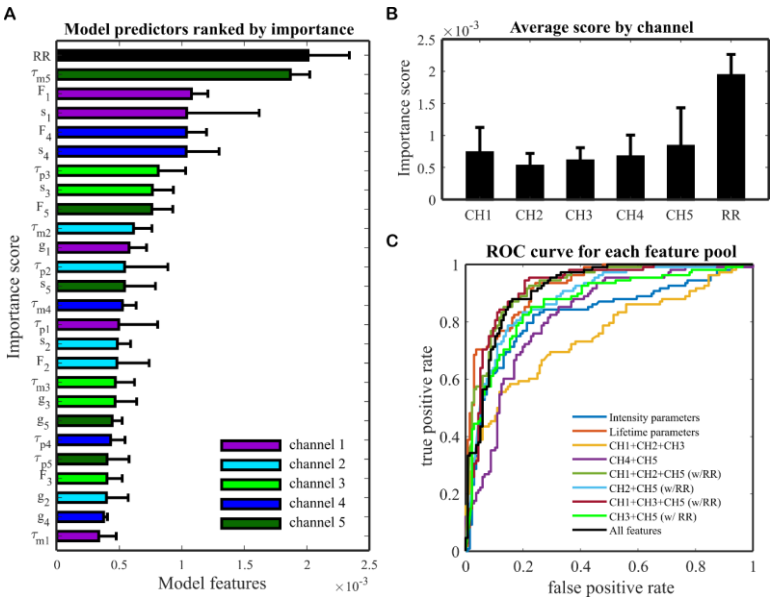


Figure 5. A) Predictor importance on the optimized model, estimated over 20 iterations of cross validation. B) Average prediction importance in each detection channel of the optical setup and redox ratio. C) ROC curves illustrating performance of classification models for different sets of features. Performance of the classifiers was measured on test data.

To evaluate the impact of each subset of features in the classification performance, the model was retrained with different feature pools. Since our aim is to simplify the optical system, we selected groups of features that would reduce the number of optical components, detectors, or excitation sources. For example, we investigated whether an optical system that only probes NAD(P)H and flavins could achieve similar performance compared to a 5-channel system. The different groups of features are listed in Table 3, together with the performance metrics on training and test datasets. The best performance is achieved when the autofluorescence signals from key endogenous molecules (collagens, NAD(P)H, and flavins) are all taken into consideration. Indeed, our data indicates that channels 1, 2 or 3, and 5 can achieve similar classification performance compared to a system with five detection channels. This is also clearly evidenced in the ROC curves presented in Fig. 5C. If data from one of these channels is not included, the performance drops considerably. In this context, it is also worth pointing out that two excitation wavelengths

perform considerably better than excitation with a single wavelength. This is because dual excitation at 375 nm and 445 nm enables optimal excitation of key endogenous fluorophores, which cannot be achieved with single excitation wavelength, whether that is at 375 or 445 nm.

Table 3. Classifier performance for different feature pools averaged over 20 training iterations. The ensemble was optimized for each feature pool, within the range of hyperparameters investigated.

		Training data			Test data		
Feature pool	**$N_{features}$**	**Accuracy**	**AUC-ROC**	**MCC**	**Accuracy**	**AUC-ROC**	**MCC**
Intensity parameters only (all channels and RR)	6	0.797 ± 0.008	0.836 ± 0.009	0.567 ± 0.018	0.725	0.825	0.467
Lifetime parameters only (all channels)	20	0.826 ± 0.009	0.904 ± 0.006	0.630 ± 0.018	0.824	0.923	0.646
375 nm excitation only (CH1, CH2, CH3)	15	0.809 ± 0.012	0.875 ± 0.009	0.594 ± 0.026	0.660	0.753	0.334
445 nm excitation only (CH4, CH5)	10	0.764 ± 0.014	0.832 ± 0.009	0.500 ± 0.031	0.762	0.822	0.520
Collagen, NAD(P)H and flavins (CH1, CH2, CH5, and RR)	16	0.866 ± 0.009	0.935 ± 0.006	0.716 ± 0.020	0.853	0.924	0.700
NAD(P)H and flavins (CH2, CH5, and RR)	11	0.833 ± 0.010	0.903 ± 0.006	0.646 ± 0.022	0.816	0.883	0.626
Collagen, NAD(P)H and flavins (CH1, CH3, CH5, and RR)	16	0.864 ± 0.012	0.933 ± 0.005	0.713 ± 0.013	0.848	0.920	0.694
NAD(P)H and flavins (CH3, CH5, and RR)	11	0.864 ± 0.008	0.915 ± 0.006	0.713± 0.017	0.800	0.865	0.593
All features	26	0.874 ± 0.008	0.941 ± 0.004	0.733 ± 0.022	0.848	0.915	0.695

4 Discussion and Conclusions

The potential of multiparametric time-resolved autofluorescence imaging and spectroscopy has been extensively showcased across various clinical applications, offering label-free tissue characterization and the ability to discriminate between benign and malignant lesions.

Undoubtedly, this technique - assisted by AI - is well positioned to address various gaps in current clinical and surgical practice, providing a quantitative readout that can make the clinical decision more informed and objective. The work presented here provides yet another demonstration of the clinical utility and versatility of this technology. We can clearly envisage the application of this work across various scenarios and levels: integrating the technology into the surgery-to-pathology workflow to complement the efforts of pathologists, providing rapid identification of positive margins ex vivo; delineating surgical margins in vivo; and swiftly identifying and characterizing lesions during endoscopic evaluations. Our model was developed having the latter in mind, i.e. to use AI-enabled autofluorescence lifetime measurements to rapidly determine the probability of malignancy of colorectal lesions during colonoscopy, not only at the diagnostic stage, but also during treatment follow up, integrated in surveillance protocols.

The diagnosis of CRC traditionally relies on endoscopic evaluation followed by histopathological analysis of biopsied tissue samples. However, obtaining adequate samples, especially from large or extensive lesions, can pose challenges. Multiple biopsies are often needed, which can be time and resource consuming, and may strain pathology laboratory resources even further. Moreover, these biopsies may not fully represent the entire lesion. To address these critical limitations and enhance the speed and accuracy of diagnostic assessments during endoscopy, the application of machine learning algorithms has been explored, either on endoscopic images alone [26] or in combination with optical spectroscopy techniques. Among the latter, hyperspectral imaging and diffuse reflectance spectroscopy have been the preferred methods, with reported accuracy of over 90% in the classification of cancerous tissues obtained from CRC specimens [27,28] and in vivo [29]. It is however overly simplistic to view optical spectroscopic techniques solely as tools for identifying cancerous tissues. These systems offer a significant advantage by harnessing

spectroscopic data, offering insights that extend beyond mere tissue classification and delve into specific clinicopathological features such as staging, microsatellite instability, or depth of invasion [18,30,31], which can potentially offer clues regarding oncological outcomes and response to therapy. Access to this information during endoscopic evaluation would not only streamline downstream processes, saving time and resources, but would also enable earlier and more personalized treatment decisions. This would in turn enhance the likelihood of a favorable clinical outcome.

In this context, we consider our study to be preliminary, with the primary objective of evaluating the potential of label-free autofluorescence lifetime measurements in distinguishing between benign and malignant colorectal lesions.

Real-time differentiation of benign and malignant adenomas during endoscopy is critical to improve diagnosis and treatment, as malignant adenomas signal cancer progression with risks of invasion or metastasis. Real-time detection would allow for immediate and precise intervention, reducing the need for additional procedures and preventing cancer progression. Label-free autofluorescence lifetime spectroscopy could enhance diagnostic accuracy, guide targeted polyp removal and optimize follow-up strategies, ultimately improving patient outcomes. The results of this study demonstrate that cancer tissues can be identified with high accuracy (90.4% ± 1.4% specificity and 82.6% ± 2.0% sensitivity on training data; 84.5% specificity and 85.2% sensitivity on test data; see Table 2) using an optimized ensemble learning model on 26 spectroscopic parameters obtained from multidimensional autofluorescence lifetime measurements. This performance is comparable to that reported in similar studies on CRC surgical specimens [27,28,32,33], even though our study was conducted on a larger cohort. Moreover, our findings were tested on an independent and previously unseen data set, further validating the reliability of the results. As we continue to gather more data and expand the database, we expect to achieve higher performance

and better generalization of the model, as suggested by the results obtained when training and test data are combined into a single dataset (see Fig. S3). This will also permit further stratification of malignant data considering cancer subtype, staging, microsatellite instability, and other characteristics. Increasing the number of adenoma lesions is also clinically significant, since it was notably lower compared to the count of normal and malignant tissues (see Table 1). Consequently, we opted to group normal tissues and adenoma lesions in a single "benign" class to avoid significant imbalance of classes. We note however that adenomas have slightly different autofluorescence signatures compared to normal and malignant tissues [18] and this can be a confounding factor that limits the performance of the model. Notwithstanding, our model performed remarkably well in the identification of benign tissues ($NPV_{training}$ = 88.0 – 89.2%; NPV_{test} = 87.8%). Therefore, we will conduct *ex vivo* measurements in biopsy specimens collected during index and surveillance endoscopies in order to increase the count of adenomas included in the model, as there is an increased likelihood of obtaining them from these procedures.

The findings of this study further validate earlier observations regarding the limitations of single wavelength excitation in capturing the full autofluorescence signature of colorectal tissues. Previously, we demonstrated that dual excitation at 375 nm and 445 nm enhances the specificity of the autofluorescence signal related to clinicopathological features [18]. This observation is consistent with the reduced performance of the model when relying solely on features obtained from excitation at either 375 nm or 445 nm (see Table 3). As expected, the classification accuracy increases when spectroscopic features linked to collagens, NAD(P)H, and flavins are all integrated into the model. In our system, this corresponds to detection channels 1, 2 or 3, and 5. However, when utilizing solely NAD(P)H and flavins spectroscopic features, including RR, we observe a slight decrease in performance, while the accuracy remains above 80% in both training and test

datasets. This result suggests that, if the goal is merely the identification of malignant tissues, the system could be substantially simplified without compromising the performance, thereby improving its commercial viability.

One limitation of this work is the inability to obtain histology images that precisely correspond to the autofluorescence maps. This is due to the sample processing technique which aligns with the bread loaf method, where tumors are sliced vertically to the horizontal plane of the specimen, resulting histological slides are perpendicular to the autofluorescence map plane. Consequently, it is nearly impossible to trace features in the autofluorescence maps back to the histology section. As a result, our ROIs cannot be entirely validated as absolute ground truth. They are, nonetheless, the most approximate representation of the ground truth to the best of our expertise and experience. This limitation was mitigated by drawing conservative ROIs away from the borders of the lesions. Yet, finding a solution remains challenging, as it necessitates deviating from the standard workflow and adding further strain to the already stressed pathology service.

To conclude, in this study we demonstrate the potential of multiparametric time-resolved autofluorescence measurements in combination with ensemble learning to classify benign and malignant colorectal lesions obtained from a total of 117 patients. The classification model, leveraging spectroscopic features derived from phasor analysis across five detection channels, achieved high tumor classification accuracy (84.8%), sensitivity (84.5%), and specificity (85.2%). Additionally, the AUC-ROC of 0.915 and MCC of 0.695 further underscore its excellent performance. We envisage application of this method not only in vivo during endoscopic evaluations, but also in the rapid identification of positive margins ex vivo, thereby complementing the work of pathologists. Future work will focus on collecting more data and expanding our

447 database to enable further stratification, extending the range of our investigation beyond binary
448 classification benign vs. malignant.

449 *Disclosures*

450 The authors declare that there are no financial interests, commercial affiliations, or other potential
451 conflicts of interest that could have influenced the objectivity of this research or the writing of this
452 paper

453 *Acknowledgements*

454 The authors thank the colorectal surgical team for their assistance in this project, including
455 collection of samples. The authors also thank all personnel of the Champalimaud Surgical Center
456 involved in sample collection, and all technicians from the Pathology Service and the
457 Champalimaud Foundation Biobank for their assistance with sample preparation.
458 Alberto I. Herrando was supported by the European Union's Horizon 2020 research and innovation
459 programme under Marie Skłodowska-Curie grant 857894 – CAST. Vladislav Shcheslavskiy was
460 supported by RSF 23-15-00294.

461 *Code, Data and Materials availability*

462 The code and dataset underlying this work are not publicly available at the time but may be
463 obtained from the corresponding author upon reasonable request.

464 *References*

465 1.Saraiva, M. R., Rosa, I. & Claro, I. Early-onset colorectal cancer: A review of current
466 knowledge. *World Journal of Gastroenterology* vol. 29 1289–1303 Preprint at
467 https://doi.org/10.3748/wjg.v29.i8.1289 (2023).

2.Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA Cancer J Clin* **73**, 233–254 (2023).

3.Saito, Y. *et al.* Colonoscopy screening and surveillance guidelines. *Digestive Endoscopy* vol. 33 486–519 Preprint at https://doi.org/10.1111/den.13972 (2021).

4.Pamudurthy, V., Lodhia, N. & Konda, V. J. A. Advances in endoscopy for colorectal polyp detection and classification. *Baylor University Medical Center Proceedings* vol. 33 28–35 Preprint at https://doi.org/10.1080/08998280.2019.1686327 (2020).

5.Robertson, D. J., Kaminski, M. F. & Bretthauer, M. Effectiveness, training and quality assurance of colonoscopy screening for colorectal cancer. *Gut* **64**, 982–990 (2015).

6.Alfonso-Garcia, A. *et al.* Real-time augmented reality for delineation of surgical margins during neurosurgery using autofluorescence lifetime contrast. *J Biophotonics* 1–12 (2019) doi:10.1002/jbio.201900108.

7.Alfonso-García, A. *et al.* First in patient assessment of brain tumor infiltrative margins using simultaneous time-resolved measurements of 5-ALA-induced PpIX fluorescence and tissue autofluorescence. *J Biomed Opt* **27**, (2022).

8.Adams, A. C. *et al.* Fibre-optic based exploration of lung cancer autofluorescence using spectral fluorescence lifetime. *Biomed Opt Express* **15**, 1132 (2024).

9.Fernandes, S. *et al.* Fibre-based fluorescence-lifetime imaging microscopy: a real-time biopsy guidance tool for suspected lung cancer. *Transl Lung Cancer Res* **13**, 355–361 (2024).

10.Phipps, J. E. *et al.* Automated detection of breast cancer in resected specimens with fluorescence lifetime imaging. *Phys Med Biol* **63**, 015003 (2017).

11.Komarova, A. D. *et al.* Metabolic heterogeneity of colorectal cancer as a prognostic factor: insights gained from fluorescence lifetime imaging. *bioRxiv* 2024.01.02.573874 (2024) doi:10.1101/2024.01.02.573874.

12.Smith, J. T. *et al.* Fast fit-free analysis of fluorescence lifetime imaging via deep learning. *Proc Natl Acad Sci U S A* **116**, 24019–24030 (2019).

13.Vasanthakumari, P. *et al.* Discrimination of cancerous from benign pigmented skin lesions based on multispectral autofluorescence lifetime imaging dermoscopy and machine learning. *J Biomed Opt* **27**, (2022).

14.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)* **13**, (2021).

15.Unger, J. *et al.* Real-time diagnosis and visualization of tumor margins in excised breast specimens using fluorescence lifetime imaging and machine learning. *Biomed Opt Express* **11**, 1216 (2020).

16.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* **68**, 857–868 (2021).

17.Hu, L. *et al.* Label-free spatially maintained measurements of metabolic phenotypes in cells. *Front Bioeng Biotechnol* **11**, (2023).

18.Herrando, A. I. *et al.* Detection and characterization of colorectal cancer by autofluorescence lifetime imaging on surgical specimens. *Sci Rep* **14**, 24575 (2024).

19.Herrando, A. *et al.* Dual excitation spectral autofluorescence lifetime and reflectance imaging for fast macroscopic characterization of tissues. *Biomed Opt Express* (2023) doi:10.1364/BOE.505220.

20.Lagarto, J. L., Shcheslavskiy, V., Pavone, F. S. & Cicchi, R. Real-time fiber-based fluorescence lifetime imaging with synchronous external illumination: A new path for clinical translation. *J Biophotonics* **13**, (2020).

21.Ranjit, S., Malacrida, L. & Gratton, E. Differences between FLIM phasor analyses for data collected with the Becker and Hickl SPC830 card and with the FLIMbox card. *Microsc Res Tech* **81**, 980–989 (2018).

519 22.Hinsdale, T. *et al.* Optically sectioned wide-field fluorescence lifetime imaging microscopy
520 enabled by structured illumination. *Biomed Opt Express* **8**, 1455 (2017).
521 23.Lagarto, J., Hares, J. D., Dunsby, C. & French, P. M. W. Development of Low-Cost
522 Instrumentation for Single Point Autofluorescence Lifetime Measurements. *J Fluoresc* (2017)
523 doi:10.1007/s10895-017-2101-7.
524 24.Freund, Y. & Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an
525 Application to Boosting. *J Comput Syst Sci* **55**, 119–139 (1997).
526 25.Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over
527 F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, (2020).
528 26.Zhou, D. *et al.* Diagnostic evaluation of a deep learning model for optical diagnosis of
529 colorectal cancer. *Nat Commun* **11**, (2020).
530 27.Jansen-winkeln, B. *et al.* Feedforward artificial neural network-based colorectal cancer
531 detection using hyperspectral imaging: A step towards automatic optical biopsy. *Cancers (Basel)*
532 **13**, 1–14 (2021).
533 28.Saito Nogueira, M. *et al.* Diffuse reflectance spectroscopy for colorectal cancer surgical
534 guidance: towards real-time tissue characterization and new biomarkers. *Analyst* **149**, 88–99
535 (2023).
536 29.Baltussen, E. J. M. *et al.* Tissue diagnosis during colorectal cancer surgery using optical
537 sensing: An in vivo study. *J Transl Med* **17**, (2019).
538 30.Nogueira, M. S. *et al.* Evaluation of wavelength ranges and tissue depth probed by diffuse
539 reflectance spectroscopy for colorectal cancer detection. *Sci Rep* **11**, 1–17 (2021).
540 31.Blake, N. *et al.* Deep Learning Applied to Raman Spectroscopy for the Detection of
541 Microsatellite Instability/MMR Deficient Colorectal Cancer. *Cancers (Basel)* **15**, (2023).
542 32.Ollesch, J. *et al.* Clinical application of infrared fibre-optic probes for the discrimination of
543 colorectal cancer tissues and cancer grades. *Vib Spectrosc* **91**, 99–110 (2017).

33.Beaulieu, R. J. *et al.* Automated diagnosis of colon cancer using hyperspectral sensing. *The International Journal of Medical Robotics and Computer Assisted Surgery* **14**, (2018).

João L. Lagarto is the head of the Biophotonics Platform of Champalimaud Foundation. He received his MSc degree in Biomedical Engineering from NOVA University of Lisbon in 2008, and his PhD in photonics from Imperial College of London in 2014. He is the author of more than 25 journal papers. His current research interests include the development and application of optical instrumentation for oncological applications, with emphasis on time-resolved autofluorescence spectroscopy.

Biographies and photographs for the other authors are not available.

Caption List

Fig. 1 Study workflow for every collected sample. The surgical specimen is transported to the optical laboratory immediately after resection and typically within 1 hour of the last vascular ligation. Optical measurements are carried out in fresh tissue before histopathological processing and assessment. Following diagnosis, benign and malignant lesions are identified in the white light image, and regions of interest (ROI) are demarcated. Autofluorescence parameters are averaged within each ROI and mean values are taken as characteristic features. Each ROI is labelled as benign or malignant according to histopathological diagnosis. Scale bars 10 mm.

Fig. 2 A) Phasor plots (top row) and phase lifetime (bottom) distribution in benign and malignant tissues for training and test dataset, measured in CH2 (left) and CH5 (right), respectively. B) Machine learning pipeline. A dataset consisting of 428 observations is used to train a boosting

algorithm. The best hyperparameter vector was found through Bayesian optimization with 5-fold cross-validation. Final model performance was evaluated against an independent test set consisting of 244 observations.

Fig. 3 A) Phasor distribution of test data in all detection channels. B) Average phase and modulation lifetimes measured in benign (B) and malignant (M) ROIs. Panels C and D show confusion matrices for training (cross validated) and test sets, respectively. TP, TN, FP, and FN values are rounded to the closest integer following an averaging of 20 iterations. Percentage values are calculated using the original "unrounded" values. Values in parenthesis represent SD. E) ROC curves for cross validation (in cyan) and test (in orange) sets. Solid lines indicate the average ROC curves for 20 iterations of the model. Shaded lines in blue represent the ROC curve for each iteration.

Fig. 4 Investigation of machine learning model-based probability of malignancy in tissue. A) White light image at time zero of normal and tumor specimens. Dotted lines delineate malignant lesions. B) Probability of malignancy estimated from single point measurements. C) Probability of malignancy for every fluorescence measurement during an entire specimen acquisition. D) Probability and E) fluorescence lifetime maps. Scale bar 10 mm.

Fig. 5 A) Predictor importance on the optimized model, estimated over 20 iterations of cross validation. B) Average prediction importance in each detection channel of the optical setup and redox ratio. C) ROC curves illustrating performance of classification models for different sets of features. Performance of the classifiers was measured on test data.

Table 1 Patient numbers and observations in training and test datasets.

Table 2 Performance metrics for training and test datasets. Metrics on training data were obtained by averaging over 20 iterations of cross validation.

593 **Table 3** Classifier performance for different feature pools averaged over 20 training iterations.
594 The ensemble was optimized for each feature pool, within the range of hyperparameters
595 investigated.
596