Automatic diagnosis and biopsy classification with dynamic Full-Field OCT and machine learning

Olivier Thouvenin (olivier.thouvenin@espci.fr)
ESPCI Paris, CNRS, PSL University https://orcid.org/0000-0003-4853-7555

Jules Scholler
ESPCI Paris, CNRS, PSL University, Institut Langevin

Diana Mandache
AQUYRE Biosciences-LLTech SAS, 58 rue du dessous des berges, 75013 Paris, France// Bioimage Analysis Unit, Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France

Marie Christine Mathieu
Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus, 114 rue Edouard Vaillant, 94805, Ville-juif, France

Aïcha Ben Lakhdar
SCM Bichat, 59 rue Bichat, 75010 Paris.

Marie Darche
Institut de la Vision, Sorbonne Université, INSERM, CNRS, F-75012, Paris, France

Tual Monfort
ESPCI Paris, CNRS, PSL University, Institut Langevin

Claude Boccara
ESPCI Paris, CNRS, PSL University, Institut Langevin

Jean-Christophe Olivo-Marin
Institut Pasteur https://orcid.org/0000-0001-6796-0696

Kate Grieve
Institut de la Vision, Sorbonne Université, INSERM, CNRS, F-75012, Paris, France// Quinze-Vingts National Eye Hospital, 28 Rue de Charenton, Paris, 75012, France

Vannary Meas Yedid
Bioimage Analysis Unit, Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France

Emilie Benoit a la Guillaume
AQUYRE Biosciences-LLTech SAS, 58 rue du dessous des berges, 75013 Paris, France

Article

Keywords: Feature Engineering, Convolutional Neural Networks, Breast Biopsies, Automatic Diagnosis Algorithms, Tumor Margins

Posted Date: May 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-371207/v1

Automatic diagnosis and biopsy classification with dynamic Full-Field OCT and machine learning.

Jules Scholler[1,*], Diana Mandache[2,3,*], Marie Christine Mathieu[4], Aïcha Ben Lakhdar[5], Marie Darche[6], Tual Monfort[1], Claude Boccara[1], Jean-Christophe Olivo-Marin[3], Kate Grieve[6,7], Vannary Meas-Yedid[3], Emilie Benoit a la Guillaume[2], and Olivier Thouvenin[1]

[1]Institut Langevin, ESPCI Paris, CNRS, PSL University, 1 rue Jussieu, 75005 Paris, France
[2]AQUYRE Bioscences-LLTech SAS, 58 rue du dessous des berges, 75013 Paris, France
[3]Bioimage Analysis Unit, Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France
[4]Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus, 114 rue Edouard Vaillant, 94805, Villejuif, France
[5] SCM Bichat, 59 rue Bichat, 75010 Paris.
[6]Institut de la Vision, Sorbonne Université, INSERM, CNRS, F-75012, Paris, France
[7]Quinze-Vingts National Eye Hospital, 28 Rue de Charenton, Paris, 75012, France

[*]These authors contributed equally

Abstract

The adoption of emerging imaging technologies in the medical community is often hampered if they provide a new unfamiliar contrast that requires experience to be interpreted. Here, in order to facilitate such integration, we developed two complementary machine learning approaches, respectively based on feature engineering and on convolutional neural networks (CNN), to perform automatic diagnosis of breast biopsies using dynamic full field optical coherence tomography (D-FF-OCT) microscopy. This new technique provides fast, high resolution images of biopsies with a contrast similar to H&E histology, but without any tissue preparation and alteration. We conducted a pilot study on 51 breast biopsies, and more than 1,000 individual images, and performed standard histology to obtain each biopsy diagnosis. Using our automatic diagnosis algorithms, we obtained an accuracy above 88 % at the image level, and above 96% at the biopsy level. Finally, we proposed different strategies to narrow down the spatial scale of the automatic segmentation in order to be able to draw the tumor margins by drawing attention maps with the CNN approach, or by performing high resolution precise annotation of the datasets. Altogether, these results demonstrate the high potential of D-FF-OCT coupled to machine learning to provide a rapid, automatic, and accurate histopathology diagnosis.

1 Introduction

In 2018, cancer was the most frequent cause of premature death in 48 countries in the world [1]. Breast cancer was the most frequently diagnosed cancer among females in 154 countries [1]. When early diagnosis is made, conservative surgery is preferred in order to partially preserve organ function, and maintain the patient's body image and quality of life [2, 3]. Nonetheless, these surgeries could be associated with higher risk of recurrence, risk of re-excision and potential delay to the initiation of other therapies [3,4]. The difficulty partly lies in the efficient evaluation of tumor margins directly during surgery. Techniques such as intraoperative touch-preparation cytology, or frozen-section analysis can be performed during the surgery and have shown promising potential in reducing the re-excision rate after primary breast conserving surgery down to 10 % instead of 35 % using permanent section histopathology [5]. Nonetheless, these techniques still require about 30 minutes, are resource intensive, are associated with various artifacts (sampling, freezing artifacts), and offer a less precise diagnosis compared to standard histopathological assessment [6–8]. A recent review compared frozen sections and touch-preparation cytology to regular histology on respectively 4300 and 1900 breast cancer cases and reported a sensitivity/specificity of 0.86/0.96 for frozen sections and 0.91/0.95 for cytology [8]. This review also pointed out that these intraoperative margin assessment techniques are not commonly routinely used in hospitals. Thus, improvement of intraoperative management towards a real-time process combining, large-scale assessment and cell resolution measurement to accurately evaluate cancer margins is an important step towards reducing the re-excision rate, the risk of recurrence, as well as patient stress and use of medical resources.

It has been the reason for the development of many optical techniques over the past few years. Among others, quantitative phase imaging [9], optical coherence tomography (OCT) [10–12], Raman spectroscopy [13, 14], fluorescence lifetime imaging [15, 16], optical elastography [17], and multiphoton microscopies [18, 19] showed high potential to offer an accurate histopathological diagnosis. Nonetheless, none of these techniques have emerged as a gold standard, and they all

provide slightly different contrasts that often do not directly compare to standard hematoxylin-eosin (H&E) histology. As a consequence, these promising techniques are hardly adopted by the medical community since adapting to a new contrast is challenging and time consuming, and such efforts are not necessarily transferable from one technique to the other. For this reason, several groups have developed machine learning tools, including deep learning, associated with either standard histopathology [20–22], or new optical technologies [12, 16, 19] in order to provide a direct diagnosis or a set of comprehensive features the pathologists can use more directly.

Among these new optical techniques, we investigate here the combination of static and dynamic full field optical coherence tomography (FF-OCT), and develop machine learning algorithms to automate diagnosis. One advantage of this combination is that it captures non-destructively *en face* sections of tissues with a contrast resembling standard H&E histology, hence supposedly easily interpretable by histopathologists. FF-OCT is a variant of OCT [23] with superior lateral resolution and is able to acquire *en face* views of the sample in a single camera frame [24, 25]. FF-OCT captures light backscattered by a sample at a given depth [23, 26], which carries information on the 3D tissue architecture (mostly extracellular matrix), which is significantly disorganized in cancerous biopsies [27–29]. Recently, we developed dynamic FF-OCT (D-FF-OCT) that takes advantage of the intracellular dynamics of cells to add a new contrast depending on cell motility, metabolism [30], and can also reveal cell mitotic state [31]. Combined with static FF-OCT, both the 3D structure and cell distribution and shape are recovered, offering a view of the sample resembling standard H&E histology [32]; Static FF-OCT provides specific signatures of extracellular matrix, hence being similar to Eosin staining, while D-FF-OCT allows visualization of nuclei (similarly to Hematoxylin staining), cytoplasm, and red blood cells. In contrast to histology, FF-OCT does not require tissue fixation or staining, and allows measurement of a 1.3×1.3 mm^2 region in a few seconds. It has the potential to offer a fast, 3D, label-free, wide field, non-destructive characterization of tumors and detection of tumor margins directly in the operating room. A recent first study on 173 breast samples demonstrated that combining static and D-FF-OCT allowed rapid diagnosis of breast tumors during surgery with high accuracy around 90 % [33].

In this paper, we first performed a clinical pilot study in 35 patients (and N=51 biopsies) with breast cancer to evaluate the clinical potential of combining static and dynamic FF-OCT in comparison to H&E histology. For each patient, we imaged the entire breast biopsy (several cm^2) with static FF-OCT, and chose for each biopsy between 5 and 20 smaller regions of interest (ROIs) of area $2mm^2$, where we recorded static and dynamic FF-OCT. Altogether, less than 10 minutes were required to record a dataset from each biopsy, which is compatible with intraoperative diagnosis. We demonstrated that FF-OCT offers enough similarities with standard histology to be straightforwardly translated to histopathology diagnosis. After a short training, two pathologists obtained a total accuracy of 90 % (91.5 % sensitivity, 86.5 % specificity) by looking at the static and dynamic FF-OCT small ROIs.

More importantly, we developed and compared two machine learning strategies to propose an automatic diagnosis of the breast tumors at different scales. A first approach, based on feature engineering (FE), aims to measure precise metrics like collagen fiber disorder or cell density, and classify samples based on the combination of these metrics. This approach strongly depends the choice of metrics and the level of accuracy in their computation. Nonetheless, it is more interpretable and the same classifier can potentially be used for different cancer types. Our second approach, performs an automatic feature extraction using a custom convolutional neural network (CNN). It is potentially more efficient than the FE approach if the dataset has adequate size, homogeneity, and annotation. We applied both approaches on all the 2 mm^2 ROIs, and obtained similar classification accuracy of respectively 88 % (89 % sensitivity, 86 % specificity) and 90 % (92 % sensitivity, 85 % specificity). By considering all ROIs from each original sample, we were able to make a prediction at the biopsy level. We obtained an improved accuracy of respectively 100 % and 96 % (94 % sensitivity, 100 % specificity), hence demonstrating that both machine learning approaches can perform efficient automatic classification.

Next, we investigated how to reduce the spatial scale at which the classification is performed, with the objective of sharpening the tissue assessment and allowing precise identification of tumor margins. First we modified the algorithms to use smaller ROIs as inputs, but obtained a reduced accuracy below 80 %, probably because of annotation issues. As a proof-of-concept experience allowing circumventing the annotation issue, we built a different dataset on 5 mice retinas imaged with D-FF-OCT, which could be annotated more easily thanks to the highly organized structure of the retina. Using the CNN approach on this dataset, we demonstrated that the algorithm was able to predict the boundaries between the retinal layers in a folded retina, with a spatial resolution of 25 μm, a problem similar to tumor margin detection. Finally, we built attention maps

Table 1: Summary of techniques and results obtained in the study. Only the accuracy is reported here, while the specificity and sensitivity are shown latter. N.A. means not applicable. The ground truth mention means that all results are compared to the histological measurement, assumed to be without error here.

Imaging technique	Diagnosis Method	Supervision Level	Training Time	Computation time per image	Accuracy at sample scale ($2cm^2$)	Accuracy at meso. scale ($2mm^2$)	Accuracy at micro. scale ($10^{-2}mm^2$)
H&E	Histopathology diagnosis(Gold standard)	Expert	Years	N.A.	100 % (Ground truth)	N.A.	N.A.
Static and dynamic FF-OCT	FF-OCT diagnosis by medical expert (Also shown in [33])	Expert	Years + Hours of new training for FF-OCT	N.A.	85 % (only static) / 90 % (with dynamic)	N.A.	N.A.
Static and dynamic FF-OCT	Feature engineering by OCT expert + SVM classifier on medical expert annotation	Semi-automatic	Minutes	1-2 min	100 %	88 %	80-71 %
Dynamic FF-OCT	CNN on medical expert annotation	Automatic	Hours	Seconds	96 %	90 %	76 %

using the GradCAM algorithm [34] in the CNN. These attention maps highlighted key regions that influenced the network diagnosis prediction, hence showing the boundaries between healthy and tumorous parts within each image.
Altogether, these results demonstrate that combining static and dynamic FF-OCT allows manual and automatic accurate cancer diagnosis, comparable with the H&E histology gold standard without tissue preparation and modification. We demonstrated proof of principle experiments to increase the spatial resolution of the automatic classifications, aiming to produce crucial biomedical diagnosis from the tissue scale down to the single cell level and showing promising results for tumor margin detection. This also paves the way for faster and easier integration of optical imaging techniques based on the endogenous dynamic scattering of light in the biomedical community.

2 Results

D-FF-OCT allows cancer diagnosis by medical experts with 90 % accuracy

In this study, we imaged 51 samples from 35 fresh lumpectomy or mastectomy surgical specimens, excised by the pathologist during intra-operative examination (Figure 1 A1, A2). For all specimens, one biopsy was taken in the tumorous area and, when possible, a second biopsy was taken far from the tumorous area to obtain healthy tissue. Immediately following the excision, the biopsy was directly inserted in the FF-OCT sample holder (Figure 1 B1), and imaged with the LLTech microscope (Figure 1 B2). The static FF-OCT image of the entire biopsy (typically $2 \times 1cm^2$) is acquired (Figure 1 B3) in less than 5 minutes. Then, between 5 and 20 ROIs corresponding to a single field of view of $1.3 \times 1.3mm^2$ were manually selected to be representative of the different areas of the biopsy. One static and one dynamic FF-OCT image were acquired at each ROI. Insertion and imaging of the sample typically took 10 minutes. Once the imaging had been performed, the samples were placed in histology cassettes (Figure 1 C1), were fixed in formalin, included in paraffin, sliced and H&E stained (Figure 1 C2 - See Methods). A diagnosis was performed by the pathologist on the entire biopsy.
After 3 months of data collection, the two pathologists who performed the sample selection went through a blind analysis of the images, deciding on the presence of cancer or not in the sample based on the FF-OCT and D-FF-OCT images only. Pathologist P1 was already familiar with FF-OCT images since she participated in a study to evaluate FF-OCT performance on head and neck cancer diagnosis, but used the D-FF-OCT for the first time. The second pathologist P2 had no prior experience in any optical biopsy technique. During the data collection phase, training was organized on 4 samples, where the two pathologists could share their interpretation of the FF-

Table 2: Sensitivity, specificity and accuracy of P1 and P2 in the differential diagnosis between non tumorous and tumorous FF-OCT images

	Whole static FF-OCT		Partial Static and Dynamic FF-OCT	
	P1	P2	P1	P2
Sensitivity	78 %	89 %	86 %	92 %
Specificity	100 %	79 %	93 %	86 %
Accuracy	84 %	86 %	88 %	90 %

OCT and D-FF-OCT images and compare them with the H&E stained image.Correlating manually the large field static FF-OCT image with the histology slide allowed side by side comparison of the static and dynamic FF-OCT ROIs (Figure 1 D1, D2, E1, E2 in healthy and tumoral regions respectively) to their corresponding histology image region (Figure 1 D3, E3). It became clear that the static FF-OCT image reveals the extracellular matrix (Figure 1 D1, E1), hence being similar to Eosin labeling, in pink, and Dynamic FF-OCT allows visualization of cells cytoplasm and nuclei (Figure 1 D1, E1), hence giving a similar contrast to Hematoxylin. As one month separated the training from the start of the blind analysis, providing the pathologists enough time to forget, the 4 samples used in the training phase were included in the blind analysis. First, a high resolution static FF-OCT image of the entire specimen (Fig. 1 B3) was used to propose the diagnosis. Immediately afterwards, the high resolution static and dynamic images of all the ROIs taken from the sample were presented together with a low resolution version of the static FF-OCT entire image to locate the ROIs within the full sample.

The histology-based diagnosis was used as the gold standard to measure the accuracy of the pathologist's diagnosis performance using the FF-OCT images which is reported in table 2. The addition of D-FF-OCT improved the performance of the two pathologists both on the measured sensitivity and specificity, showing that D-FF-OCT highlights tissue features that are closer to the histology criteria the pathologists rely on to base their diagnosis. However, D-FF-OCT also caused misinterpretation from both P1 and P2 on two cases that were previously diagnosed correctly using FFOCT only, which demonstrates the interest of using both modalities, and the drawback of partial imaging. In total, five wrong diagnosis were counted for each pathologist among which only three were common to P1 and P2. Therefore, improvement of the pathologists' performance is expected with more training or with the help of automatic annotation tools. Nonetheless, isolated invasive cancer cells are more difficult to visualize with D-FF-OCT compared to histology, which was likely the cause of some errors of diagnosis, and one particular target of automated algorithms.

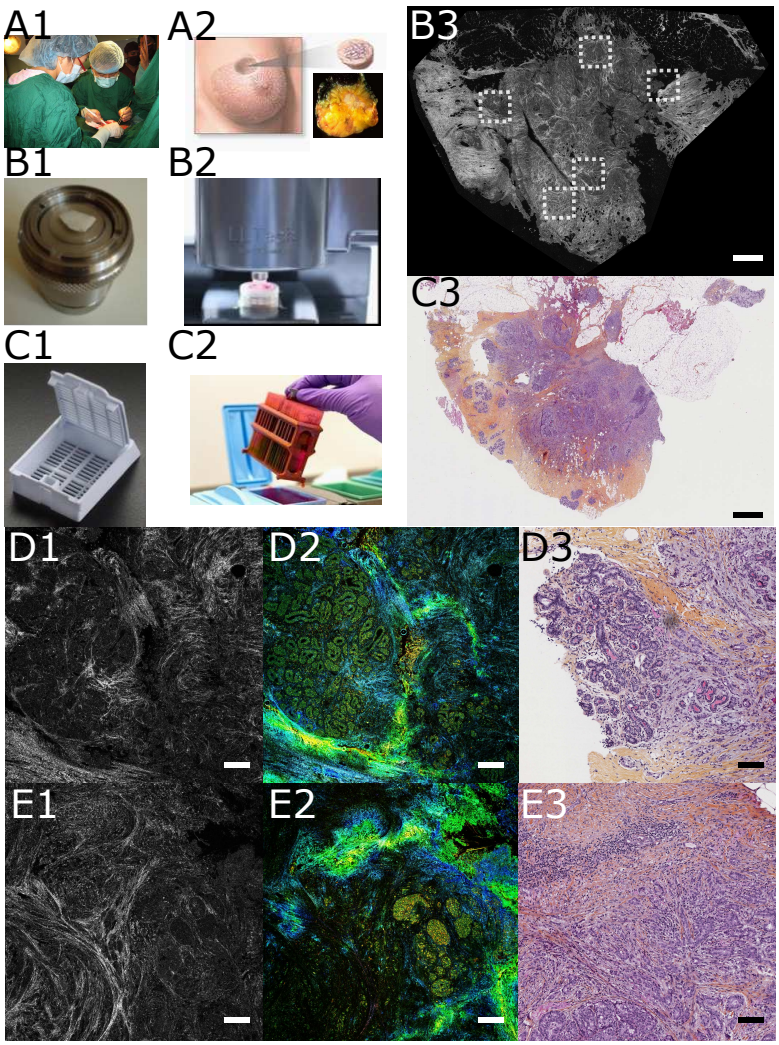

Figure 1: **Experimental protocol and correlation between FF-OCT, D-FF-OCT, and histology images**. During surgery (Panel A1), breast biopsies (Panel A2) are collected, inserted into the FF-OCT sample holder (Panel B1), and imaged under the FF-OCT microscope (B2). A large field static FF-OCT image of the entire biopsy is acquired (Panel B3). 5 to 20 ROIs (dashed white box) are manually selected and imaged with both static and dynamic FF-OCT.Then, the biopsy is removed from the sample holder and is inserted into a regular histology cassette (Panel C1), where it is fixed, sliced, and stained (Panel C2), according to H&E histology standards (Panel C3). Examples of 2 ROIs (Panels D and E) imaged from sample diagnosed with invasive ductal carcinoma (IDC) in a healthy and tumoral regions respectively. Static FF-OCT (Panel D1), D-FF-OCT (Panel D2), and H&E (Panel D3) images of a normal lobule were acquired. Similar images (Panels E1, E2, and E3) in an IDC area are displayed. Scale bars are 1 mm (B3, C3) and 0.1 mm (D1-E3) respectively.

Dynamic FF-OCT and machine learning allow an automatic cancer diagnosis at mesoscopic scale (2 mm^2) and macroscopic scale (2 cm^2) with up to 100 % accuracy.

Although the previous results are promising, they required training and time from expert histopathologists and could be improved by computer-based assistance. Adoption of combined static and dynamic FFOCT in the clinical workflow thus requires efficient automatic diagnosis to refine, facilitate and accelerate intraoperative interpretation. Hence, we developed and tested two machine learning algorithms to perform automatic diagnosis of the individual ROIs (mesoscopic scale - $2mm^2$), and then aggregated the results from all ROIs of each biopsy to propose a global diagnosis of the biopsy (macroscopic scale - $2cm^2$). Although both algorithms display similar results described below, they use complementary approaches, and different characteristics that will be further compared in the discussion.
The first approach was to define features of interest that will be measured automatically in the

dataset via image analysis. These features are then used in a multidimensional machine learning model, linear support vector machine (SVM), to separate healthy from tumoral ROIs. We first selected (see Methods) only the images that have a corresponding static FF- OCT and D-FF-OCT image from the same ROI (Figure 2A1 and B1), with at least 5 ROIs for each sample. From each image, we aimed to measure features describing cells and the extracellular matrix organization. We trained two random forest classifiers (one for static, and one for dynamic FF-OCT) to segment individual cells and fibers (Figure 2A2 and B2) using the *ILastik* software [35]. These classifiers were evaluated by manual inspection and are rather imprecise, but, because many cells and fibers can be found in each image, we believe the errors are averaged out. The segmented images were then analyzed using a second filtering step (see Methods) to exclude some misclassified pixels. We also calculated mesoscale features, by segmenting regions of high-fiber, or high-cell, densities (Figure 2A3, B3). In total, combining static and dynamic FF-OCT images, we measured 44 features (listed in the Methods with associated references), including cell parameters (size, intensity), fiber parameters (density, organization), mesoscale organization, and fat content (Figure 2C) that were previously characterized as potential cancer biomarkers (see Methods). The histological diagnosis of the sample was attributed to all the ROIs acquired in this sample. Establishing this *ground truth* is rather imprecise because healthy regions can be found in cancerous samples but it spares the tedious frame by frame annotation work. Also, if healthy regions can be found in cancerous samples, the opposite should not be found, so that our *ground truth* has inhomogeneous accuracy as well. We trained a SVM to classify each ROI (Figure 2C). We used a 5-fold cross validation and penalized by a factor 3 the possibility of having false positives (healthy samples found as cancerous) to balance the dataset inhomogeneity. With a linear SVM, we achieved a mean accuracy of $88 \pm 3\%$ (sensitivity $89 \pm 4\%$ and specificity $86 \pm 3\%$) and an area under the ROC curve (AUC) of 0.90 ± 0.03 at the individual image level over the 5 models. To classify at the sample scale, the ratio of ROIs classified as normal versus all ROIs was calculated for each sample (Figure 2D). Segmenting samples with a ratio above 0.5 of healthy ROIs resulted in a 100 % accuracy segmentation between healthy and cancerous samples at the macroscopic scale.

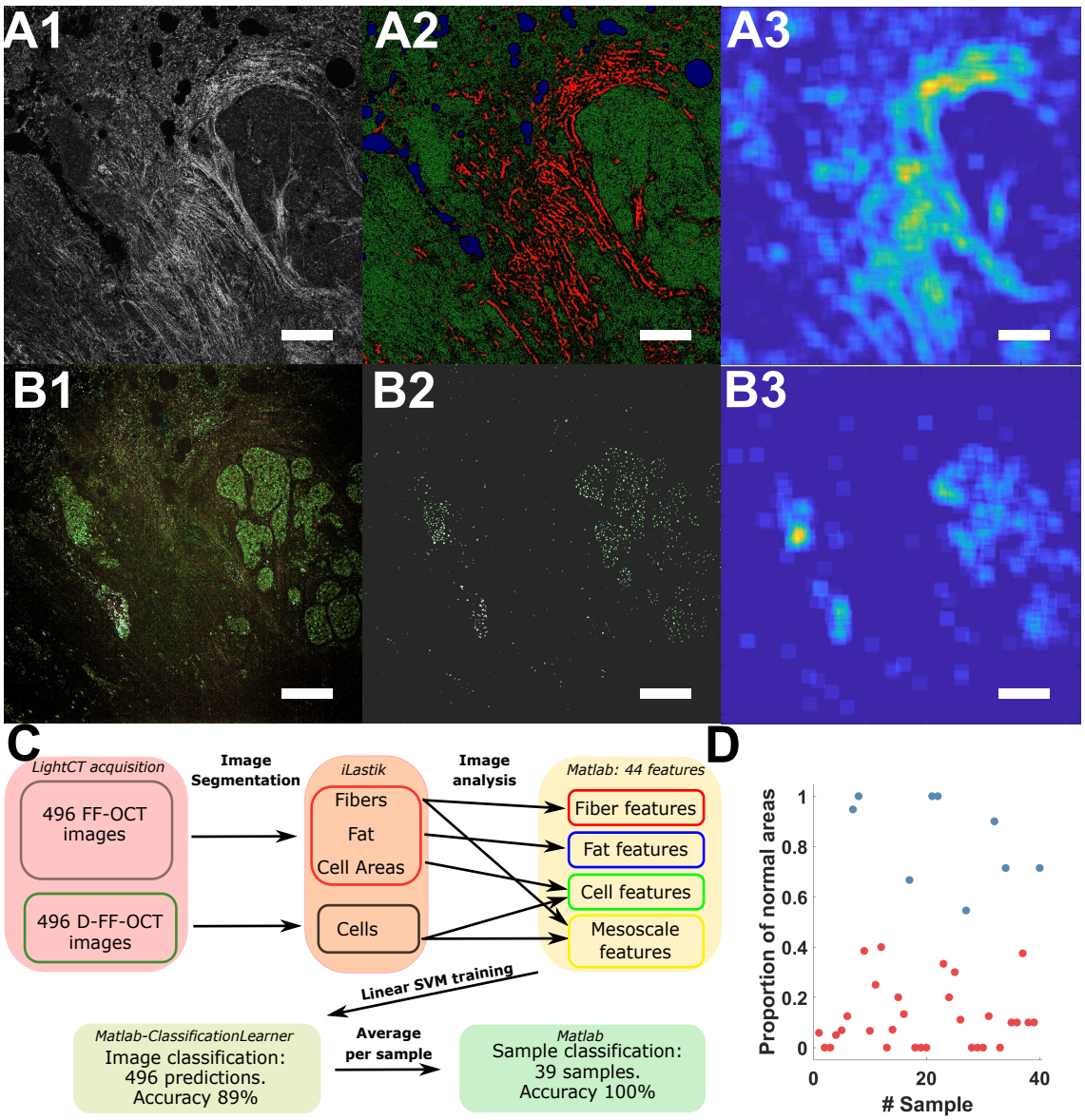

Figure 2: **Feature engineering and SVM classification of breast cancer D-FF-OCT images.** Static (A1) and dynamic (B1) FF-OCT images of a cancerous breast sample are analyzed using two random forest classifiers (A2,B2). FF-OCT image is segmented (A2) into fibers (red), cells (green), and fat/holes (blue). D-FF-OCT image is segmented (B2) into cells. Mesoscale fiber regions (A3) describe the region of high fiber density in the FF-OCT image. Mesoscale cell regions (B3) describe regions of high cell density in the D-FF-OCT image. The segmented images (A2, B2) and the mesoscale images (A3,B3) are used to calculate engineered features, such as cell and collagen fiber characteristics and size and shape of regions of high cell density. (C) Chart summarizing the processing of FF-OCT and D-FF-OCT images in order to classify each image and each sample using SVM. (D) Proportion of normal areas found for each healthy (blue) and cancerous sample (red) showing 100 % separability between the two classes. Scale bars: 200 μm.

Our second approach was to explore a purely data-driven approach facilitated by the Deep Learning paradigm. We chose to fine-tune a pre-existing Convolutional Neural Network (CNN), namely the VGG16 [36] architecture with weights pre-trained on the ImageNet dataset [37], in order to directly classify each ROI from the D-FF-OCT images. We modified the VGG16 in order to adapt its architecture to our problem (see Methods and figure 4C). The network was trained on full resolution $1440 \times 1440 \times 3$ RGB D-FF-OCT ROIs. In this section, we used in total 373 individual ROIs from 47 samples from which frame by frame annotation was performed by the histopathologist P2 (34 samples with a tumor, 13 without- See Methods for extended dataset description). We used the ROIs from 80 % of the samples (37 samples; 286 ROIs, 185 positive and 101 negative) for training the network, and used data augmentation (see Methods) to obtain a 6-fold increase of the training set size. The 20 % remaining (10 samples) were used to test the

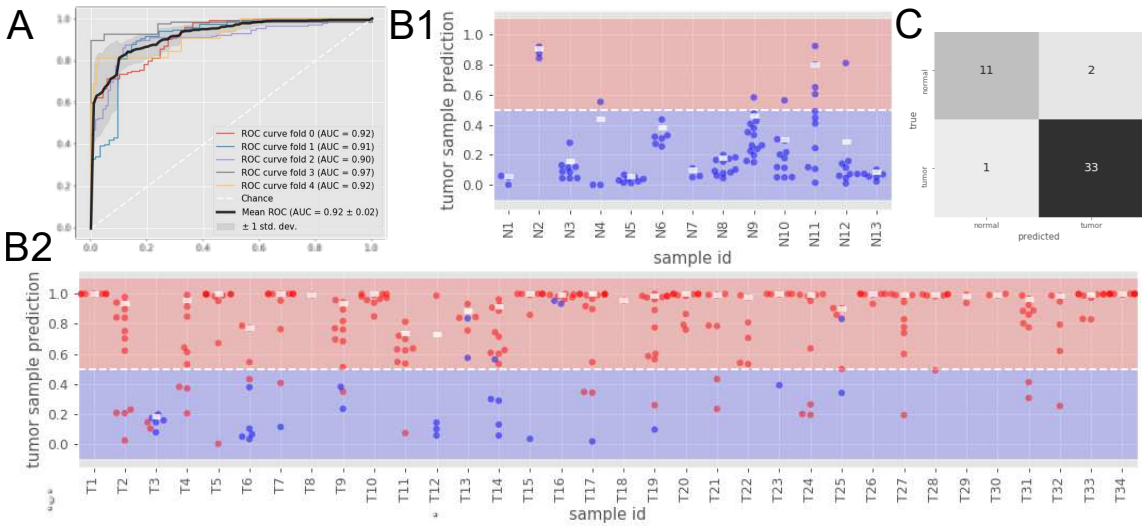

Figure 3: **CNN classification of breast cancer** (A) ROC curves and AUCs of the 5-fold cross validated models. (B) Tumor probability prediction by the CNN for each ROI in each sample, either healthy (B1) or tumoral (B2). Each point represents a ROI which ground truth diagnosis is either healthy (blue) or tumoral (red). The white rectangle represents the aggregated tumor probability per sample, computed as the 90th percentile of the probabilities of the ROIs in a sample : (B1) normal samples prediction (B2) tumoral samples prediction.

CNN (see Methods for details) with a 5-fold cross-validation in order to test the entire dataset ultimately. The CNN outputs a probability for the ROI to be in the tumoral part, and we defined a threshold at 50% above which the ROI is considered as positive. Using the 5-fold cross-validation, we obtained a mean accuracy of $89 \pm 4\%$ (sensitivity $88 \pm 4\%$ and specificity $86 \pm 6\%$) and the area under the ROC curve (AUC) of 0.92 ± 0.02 at the individual image level (Figure 3 A-B) over the 5 models. To classify at the sample scale, we concatenated the tumor prediction probability from all models for all ROIs. For each sample, we computed the 90^{th} percentile of the probabilities of the presence of a tumor. If the probability is above 50% the sample is considered as cancerous. This results in a sample-wise accuracy of 94%, sensitivity of 97% and specificity of 85%, translating to 2 false positives and one false negative (Figure 3 C) . Interestingly, the misclassified cases were also wrongly diagnosed by the pathologists at the blind D-FF-OCT assessment.

Dynamic FF-OCT with CNN, activation maps, and adequate data annotation allow reducing the spatial scale of the segmentation

So far, the spatial resolution of the predictions were limited by the size of the chosen ROIs of $1.3 \times 1.3 mm^2$. This spatial sampling might be a limiting factor in applications where the tumor margins need to be drawn to make sure the entire tumor was removed. In this section, we propose and compare two different strategies to increase the spatial resolution of the predictions.
First, with both feature engineering and CNN classification, we separated each ROI into several smaller thumbnail images with a possible overlap. We evaluated different trimmings starting from dividing each ROI into 4 subregions to dividing into 256 subregions. The difficulty of this approach was that the ground truth was not defined at this scale, and we had to apply the histopathologist expert diagnosis (at the sample level in the case of feature engineering, and at the ROI level in the case of the CNN classification) to all the subregions. It likely increases the ratio of images with false labeling as the size of the thumbnail image decreases. We indeed observed the prediction accuracy decrease at both the thumbnail image level and the macroscopic level when all the predictions were aggregated (see supplementary document). The decrease is more important as the subregion size becomes smaller, ultimately leading to a random classification. The CNN approach also showed a reduced accuracy of 76 % when the subregion size was reduced by a factor 5 (see supplementary file).

To test the hypothesis that the limiting factor to increase the spatial resolution and draw boundaries between regions of different physiological states is the correct image labeling at smaller scale, we performed a proof-of-principle experiment in retinal explants imaged with a custom high resolution FF-OCT microscope (see Methods). Here, we aim to draw the boundaries between retinal layers using a modified CNN. Retinal samples were used because D-FF-OCT can efficiently

detect all retinal cells [38], and they are easy to annotate since all layers have significantly different organizations and cell content. 5 different mouse retinas were imaged with D-FF-OCT among which, one was used to create a training set (Figure 4A) by annotating 395 random thumbnail images (27x27 μm^2) using their corresponding retinal layer. The annotations defined 5 classes: ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer nuclear layer (ONL) and photoreceptor inner and outer segments (IS/OS). The test set was created by annotating manually 292 random thumbnail images on a second retina in order to avoid over-fitting problems. In order to validate the trained CNN the 3 last retinas were used, especially on folded areas were several layers appear on each image (figure 4B). To do so, the input images were cropped in $27 \times 27 \mu m^2$ thumbnail images with an overlap of 20 μm between images in order to obtain a $7 \times 7 \mu m^2$ resolution for the final classification map. Each thumbnail image was then classified using the trained CNN resulting in a coarse classification of the input images (Figure 4B1, B2, B3 in respectively the GCL, and two depths at the frontier between the INL and the ONL). The accuracy reached 92 % on the 3 held out retinas (that were not used during training and testing), and the CNN only failed to make a good prediction in low signal areas (corresponding to deep regions and weakly reflective cells). This suggests that the high resolution determination of tissue boundaries, including tumor margins could be performed efficiently as long as an accurate and time-consuming labeling at the microscopic scale is performed.

Finally, because an accurate labeling at the microscopic scale is not always possible and meaningful, as well as to increase CNN interpretability, we developed another strategy based on the computation of activation maps of the CNN. Typically, in CNNs, convolutional layers naturally retain spatial information, whereas fully-connected layers usually present at the end of the neural network performing the classification task does not, causing the loss of spatial information. However, this spatial information can be restored by locating the image area that triggered the CNN to decide which prediction to make (Figure 4C). In practice, we used the Grad-CAM technique [34] (see Methods). This results in a coarse localization of the class presence (here tumoral areas) in the initial ROI (figure 4D1), without the need for annotation at smaller scale. Positive and negative attention maps can be computed, showing either cancerous cells (figure 4D2) that invaded the stroma, or healthy lobules (figure 4D3). Besides, building attention maps can also serve other purposes, including verifying that the model is not biased, and drawing attention to specific parts of the image to assist the surgeon to classify the sample. The activation map has a spatial resolution of the same size as the last convolutional feature map of the network (here $45 \times 45 \mu m^2$).

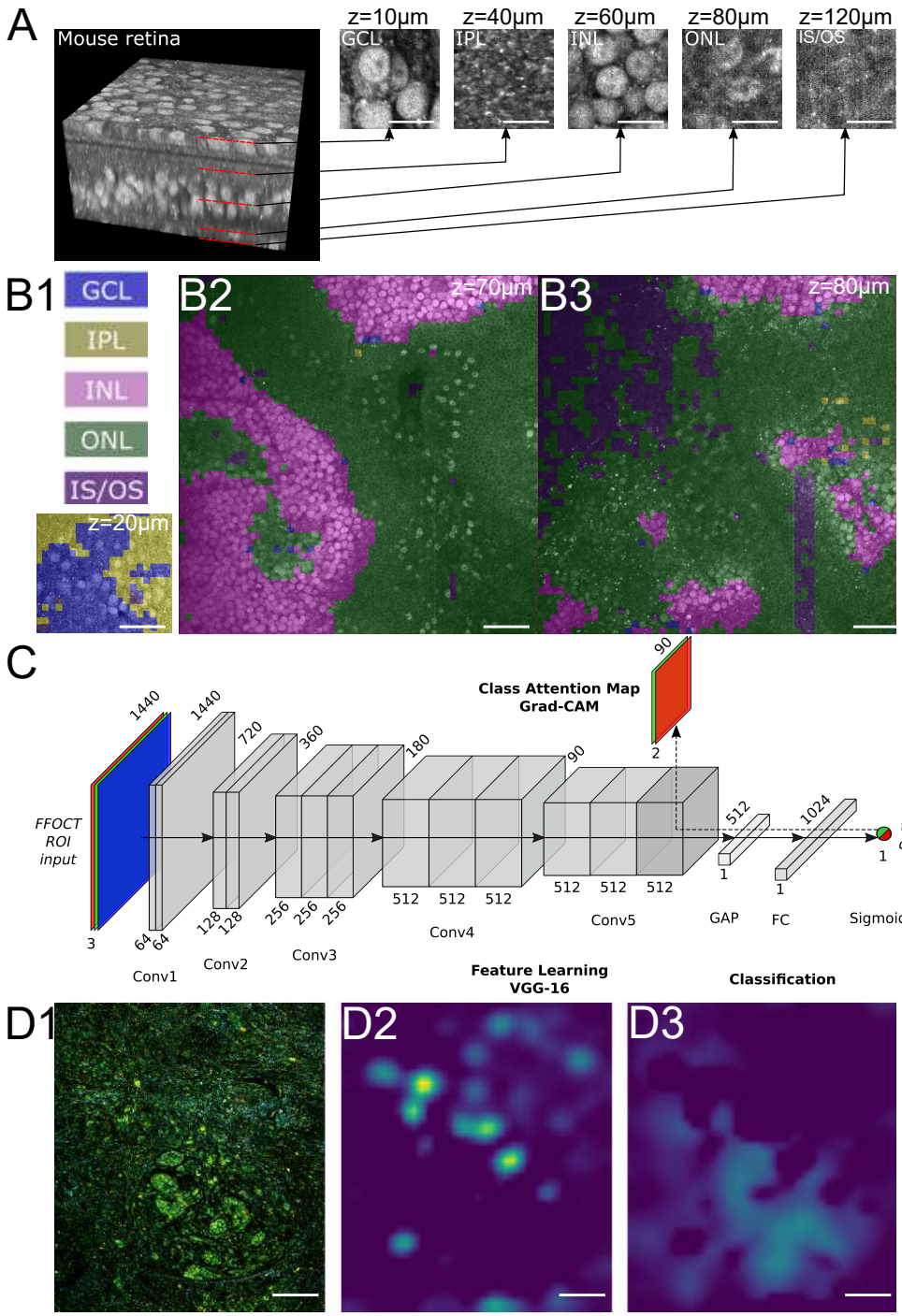

Figure 4: **Proposed strategies to reduce the spatial scale of the predictions**. A first strategy simply proposes to reduce the size of the starting images, but requires precise labeling at the microscopic scale. We could only achieved this labeling in mice retinas (Panel A), where retinal layers are well organized and show distinct and clearly separable morphological features. Using a modified CNN, we could draw the boundaries between retinal layers in folded retinas where several layers overlap at the same depth (Panels B1, B2, B3 at different depths from retinal surface - 20, 70, and 80 μm respectively). A second strategy using CNN is to propagate the gradients from the last convolutional layer using the GradCAM algorithm (Panel C) in order to retrieve the spatial information from the CNN. From a ROI showing healthy breast lobule surrounded by isolated infiltrating cancerous cells, correctly predicted as cancerous with 97 % confidence by the CNN (Panel D1), the tumor positive attention map (panel D2) focuses on the regions with the infiltrating cells, while the the tumor negative attention map (panel D3) shows healthy regions, including the healthy breast lobule. Scale bars are 25 μm in A, 50 μm in B and 200 μm in D.

3 Discussion

To summarize, we demonstrate the potential of combining static and dynamic FF-OCT to perform an accurate label-free and nondestructive histopathological diagnosis in breast samples, either by manual or automatic image interpretation. Interestingly, static FF-OCT offers a view of the extracellular matrix and tissue organization of the sample, while D-FF-OCT reveals cells within this matrix. The combination of both techniques hence produces images similar to H&E staining. Together with another recent clinical study in breast samples [33] as well, the pilot study presented here really demonstrates the strength and potential of using FF-OCT for histopathology and intraoperative diagnosis. Therefore, we additionally developed two methods of image processing and statistical learning and successfully brought more specificity and interpretability to these images. We believe this work is interesting because this is the first time, to our knowledge, that machine learning algorithms are developed for D-FF-OCT. In the histopathology field, many machine learning algorithms have been developed on traditional stainings, but applying them to FFOCT aims to facilitate the introduction of a high-potential imaging technique with a similar, but label-free and richer contrast mechanism in the field. In the end, although these results are preliminary and should be confirmed on bigger cohorts and different cancer types, we demonstrated that automatic diagnosis with FF-OCT compares well with manual diagnosis given by histopathological experts, and should increase its performances as more and more precise datasets will be generated.

In this study, we developed and tested two machine learning algorithms which strategies and outputs are complementary. If both techniques are hard to compare quantitatively because of deviations on the dataset content and corresponding labeling, they are likely to target different applications and datasets. The feature engineering (FE) approach is based on the computation of known meaningful metrics such as collagen fiber characteristics, or cell morphological parameters, that could be used directly on top of the classification. Hence, it is likely to be more appropriate for exploratory diagnosis, where the quantification of these metrics can discriminate between different diagnostics (e.g. different tumor types). Because the FE approach relies on averaging cell and fiber parameters, this approach would converge more easily if the tumor is homogeneous with such parameters centered around well defined values. In this study, we realized that the computation of these metrics is still rather imprecise (*e.g.* it is extremely complicated to make sure that 100 % of the cells are segmented and that their diameter is correctly sampled at the optical resolution). Besides, some parameters (e.g. extracellular matrix disorganization) are not necessarily well defined at the microscopic scale. So that the FE approach did not perform well when the spatial scale is reduced. This is expected given that the FE approach is quite similar to the reading method of the histopathologist, who most often requires reviewing large portions of the sample to provide an accurate diagnosis. In general, improving the metric computation, defining new useful metrics, and increasing the spatial resolution might help to increase the sensitivity of this approach. Among others, the FE approach could be greatly improved if the aspect of cell nuclei and the number of mitotic cells, key metrics in standard histopathology [39, 40], could be automatically quantified. In previous work, we demonstrated that D-FF-OCT performed at high resolution is able to measure the aspect and dynamics of nuclei [32, 38] and measure the mitotic state [31], suggesting that this measurement could be automated. Nonetheless, increasing the resolution decreases the field of view and thus increases the imaging time if a full sample is acquired. The best compromise between resolution (and more precise computation of cell metrics) and imaging time should be defined depending on the application.

Our second strategy was based on supervised learning and convolutional neural networks (CNN) and we demonstrated this approach could be used to increase the spatial resolution of the predictions if adequate labeling is obtained. The CNN approach works without *a priori* knowledge of the tumor biology, and the same network structure can be applied to any other D-FF-OCT dataset as long as sufficient annotation is available. The CNN would reach better accuracy with a greater dataset size and homogeneity. It would therefore be more recommended for large and homogeneous datasets, or when drawing the tumor margins with high resolution is required. Nonetheless, the CNN approach is less interpretable, and because it does not rely on any general feature of tumors, it is likely that a new network should be retrained from scratch for any new tumor type, while the FE approach could potentially identify general biomarkers. We also believe that the Grad-cam algorithm that defines attention maps, *i.e.* subregions that were the most meaningful for the network to choose a diagnosis, could be an important help to rapidly draw the attention of surgeons or histopathologists on the regions of interest. Because the CNN approach is unbiased as compared to human interpretation, the use of attention maps could also improve our knowledge of tumorous processes by pointing out small details as potential tumor markers.

Currently, one of the key limitation is that the D-FF-OCT image of large samples would require about half an hour of acquisition, while we aim to keep acquisition time below 10 minutes. Within this time constraint, we had to image only a few ROIs per sample. This partial imaging procedure suffers from sampling bias if the ROIs are selected randomly. However, as static FFOCT imaging is faster, it could be used as a first row of interpretation and drive the selection of the ROIs where to apply D-FF-OCT for further investigation. For future studies, we aim to develop a first CNN trained on static FF-OCT image of the whole sample [41] to automatically define a few ROIs where the D-FF-OCT will be acquired to increase the accuracy of the diagnosis. We can also expect that progress in camera technologies and in real time processing of the D-FF-OCT image via GPU computing will help reducing the imaging time and will soon enable to record the D-FF-OCT of the entire sample in a few minutes.

Overall, we demonstrated that the computer-aided combination of static and D-FF-OCT reaches the standards of breast histopathology, while being nondestructive, label-free, and free from freezing and slicing artefacts. We forecast two main contributions in histopathology. First, because the two automated methods produce a probabilistic prediction (which we here mostly thresholded at 50 % to perform a prediction), it is easy to obtain a confidence measurement on the prediction. Therefore, computer-aided FF-OCT could reduce the workload of pathologists by automatically sorting out cases with high confidence predictions and concentrate their attention to samples, or ROI, with less confidence for manual inspection. Second, it could improve intraoperative diagnosis during organ preserving surgeries and replace expensive, time consuming, and less precise frozen sections as it provides a fast and accurate enough diagnosis, while preserving the samples for later analysis. More importantly, we believe the nondestructive aspect of FF-OCT is fundamental to go even beyond standard histology, because it allows combination with other techniques, including molecularly-specific optical techniques (such as fluorescence or Raman spectroscopy), or destructive tumor genotyping. Multimodal imaging could thus identify both the spatial organization of a tumor and its molecular or genetic characteristics. As the cells are still alive during FF-OCT imaging, and as D-FF-OCT can monitor cell metabolism, an interesting individualized follow-up approach would be to perform chemotherapy treatment on the extracted biopsy and evaluate the response from the metabolic changes of the cells within. Depending on the response of the biopsy, it could orient the use of one molecule instead of another before starting to treat the patient. As histopathology is being moved by digital transformation, computer-aided FF-OCT must quickly develop and confirm its potential on larger datasets and various tumor types to be part of the transformation.

Funding

"HELMHOLTZ" (European Research Council (ERC) (#610110), PI Mathias Fink and José-Alain Sahel) "OREO" [ANR-19-CE19-0023], (PI Kate Grieve) ANRT fellowship (grant CIFRE 2018/0139).

Acknowledgments

The authors wish to thank the Centre de Ressources Biologiques and the technicians of the Biopathology Department of Gustave Roussy for the processing of histological samples.

Disclosures

Claude Boccara is one of the founders of LLTech and is a part-time LLTech employee. Emilie Benoit is an employee of LLTech. All the other authors declare that they have no conflict of interest.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] F. Steinbach, M. Stöckle, S. Müller, J. Thüroff, S. Melchior, R. Stein, and R. Hohenfellner, "Conservative surgery of renal cell tumors in 140 patients: 21 years of experience," *The Journal of urology*, vol. 148, no. 1, pp. 24–29, 1992.

[3] U. Veronesi, N. Cascinelli, L. Mariani, M. Greco, R. Saccozzi, A. Luini, M. Aguilar, and E. Marubini, "Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer," *New England Journal of Medicine*, vol. 347, no. 16, pp. 1227–1232, 2002.

[4] I. Gage, S. J. Schnitt, A. J. Nixon, B. Silver, A. Recht, S. L. Troyan, T. Eberlein, S. M. Love, R. Gelman, J. R. Harris, *et al.*, "Pathologic margin involvement and the risk of recurrence in patients treated with breast-conserving therapy," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 78, no. 9, pp. 1921–1928, 1996.

[5] K. Esbona, Z. Li, and L. G. Wilke, "Intraoperative imprint cytology and frozen section pathology for margin assessment in breast conservation surgery: a systematic review," *Annals of surgical oncology*, vol. 19, no. 10, pp. 3236–3245, 2012.

[6] T. S. Menes, P. I. Tartter, H. Mizrachi, S. R. Smith, and A. Estabrook, "Touch preparation or frozen section for intraoperative detection of sentinel lymph node metastases from breast cancer," *Annals of surgical oncology*, vol. 10, no. 10, pp. 1166–1170, 2003.

[7] T. Khamechian, J. Alizargar, and T. Mazoochi, "The value of touch preparation for rapid diagnosis of brain tumors as an intraoperative consultation," *Iranian journal of medical sciences*, vol. 37, no. 2, p. 105, 2012.

[8] E. R. St John, R. Al-Khudairi, H. Ashrafian, T. Athanasiou, Z. Takats, D. J. Hadjiminas, A. Darzi, and D. R. Leff, "Diagnostic accuracy of intraoperative techniques for margin assessment in breast cancer surgery," *Annals of surgery*, vol. 265, no. 2, pp. 300–310, 2017.

[9] M. Takabayashi, H. Majeed, A. Kajdacsy-Balla, and G. Popescu, "Disorder strength measured by quantitative phase imaging as intrinsic cancer marker in fixed tissue biopsies," *PloS one*, vol. 13, no. 3, 2018.

[10] F. T. Nguyen, A. M. Zysk, E. J. Chaney, J. G. Kotynek, U. J. Oliphant, F. J. Bellafiore, K. M. Rowland, P. A. Johnson, and S. A. Boppart, "Intraoperative evaluation of breast tumor margins with optical coherence tomography," *Cancer research*, vol. 69, no. 22, pp. 8790–8796, 2009.

[11] O. M. Carrasco-Zevallos, C. Viehland, B. Keller, M. Draelos, A. N. Kuo, C. A. Toth, and J. A. Izatt, "Review of intraoperative optical coherence tomography: technology and applications," *Biomedical optics express*, vol. 8, no. 3, pp. 1607–1637, 2017.

[12] A. C. Sullivan, J. P. Hunt, and A. L. Oldenburg, "Fractal analysis for classification of breast carcinoma in optical coherence tomography," *Journal of biomedical optics*, vol. 16, no. 6, p. 066010, 2011.

[13] A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, J. Lyons, D. Hicks, M. Fitzmaurice, R. R. Dasari, J. P. Crowe, and M. S. Feld, "In vivo margin assessment during partial mastectomy breast surgery using raman spectroscopy," *Cancer research*, vol. 66, no. 6, pp. 3317–3322, 2006.

[14] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, L. Bernstein, M.-C. Guiot, K. Petrecca, and F. Leblond, "Intraoperative brain cancer detection with raman spectroscopy in humans," *Science translational medicine*, vol. 7, no. 274, pp. 274ra19–274ra19, 2015.

[15] Y. Sun, J. Phipps, D. S. Elson, H. Stoy, S. Tinling, J. Meier, B. Poirier, F. S. Chuang, D. G. Farwell, and L. Marcu, "Fluorescence lifetime imaging microscopy: in vivo application to diagnosis of oral carcinoma," *Optics letters*, vol. 34, no. 13, pp. 2081–2083, 2009.

[16] J. Unger, C. Hebisch, J. E. Phipps, J. L. Lagarto, H. Kim, M. A. Darrow, R. J. Bold, and L. Marcu, "Real-time diagnosis and visualization of tumor margins in excised breast specimens using fluorescence lifetime imaging and machine learning," *Biomedical Optics Express*, vol. 11, no. 3, p. 1216, 2020.

[17] W. M. Allen, K. Y. Foo, R. Zilkens, K. M. Kennedy, Q. Fang, L. Chin, B. F. Dessauvagie, B. Latham, C. M. Saunders, and B. F. Kennedy, "Clinical feasibility of optical coherence micro-elastography for imaging tumor margins in breast-conserving surgery," *Biomedical optics express*, vol. 9, no. 12, pp. 6331–6349, 2018.

[18] N. D. Kirkpatrick, M. A. Brewer, and U. Utzinger, "Endogenous optical biomarkers of ovarian cancer evaluated with multiphoton microscopy," *Cancer Epidemiology and Prevention Biomarkers*, vol. 16, no. 10, pp. 2048–2057, 2007.

[19] S. You, Y. Sun, L. Yang, J. Park, H. Tu, M. Marjanovic, S. Sinha, and S. A. Boppart, "Real-time intraoperative diagnosis by deep neural network driven multiphoton virtual histology," *npj Precision Oncology*, vol. 3, no. 1, pp. 1–8, 2019.

[20] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[21] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[22] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, *et al.*, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell reports*, vol. 23, no. 1, pp. 181–193, 2018.

[23] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and A. Et, "Optical coherence tomography," *Science*, vol. 254, pp. 1178–1181, Nov. 1991.

[24] E. Beaurepaire, A. C. Boccara, M. Lebec, L. Blanchot, and H. Saint-Jalmes, "Full-field optical coherence microscopy," *Optics letters*, vol. 23, no. 4, pp. 244–246, 1998.

[25] A. Dubois, K. Grieve, G. Moneron, R. Lecaque, L. Vabre, and C. Boccara, "Ultrahigh-resolution full-field optical coherence tomography," *Appl. Opt.*, vol. 43, pp. 2874–2883, May 2004.

[26] W. Drexler and J. G. Fujimoto, *Optical coherence tomography: technology and applications*. Springer Science & Business Media, 2008.

[27] M. Jain, N. Shukla, M. Manzoor, S. Nadolny, and S. Mukherjee, "Modified full-field optical coherence tomography: A novel tool for rapid histology of tissues," *Journal of pathology informatics*, vol. 2, 2011.

[28] K. Grieve, K. Mouslim, O. Assayag, E. Dalimier, F. Harms, A. Bruhat, C. Boccara, and M. Antoine, "Assessment of sentinel node biopsies with full-field optical coherence tomography," *Technology in cancer research & treatment*, vol. 15, no. 2, pp. 266–274, 2016.

[29] O. Assayag, M. Antoine, B. Sigal-Zafrani, M. Riben, F. Harms, A. Burcheri, K. Grieve, E. Dalimier, B. Le Conte de Poly, and C. Boccara, "Large field, high resolution full-field optical coherence tomography: a pre-clinical study of human breast tissue and cancer assessment," *Technology in cancer research & treatment*, vol. 13, no. 5, pp. 455–468, 2014.

[30] C. Apelian, F. Harms, O. Thouvenin, and C. Boccara, "Dynamic full field optical coherence tomography: subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis," *Biomedical Optics Express*, vol. 7, p. 1511, Mar. 2016.

[31] J. Scholler, K. Groux, O. Goureau, J.-A. Sahel, M. Fink, S. Reichman, C. Boccara, and K. Grieve, "Dynamic full-field optical coherence tomography: 3d live-imaging of retinal organoids," *Light: Science & Applications*, vol. 9, no. 1, pp. 1–9, 2020.

[32] O. Thouvenin, C. Apelian, A. Nahas, M. Fink, and C. Boccara, "Full-field optical coherence tomography as a diagnosis tool: Recent progress with multimodal imaging," *Applied Sciences*, vol. 7, p. 236, 03 2017.

[33] H. Yang, S. Zhang, P. Liu, L. Cheng, F. Tong, H. Liu, S. Wang, M. Liu, C. Wang, Y. Peng, *et al.*, "Use of high-resolution full-field optical coherence tomography and dynamic cell imaging for rapid intraoperative diagnosis during breast cancer surgery," *Cancer*, vol. 126, pp. 3847–3856, 2020.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, Oct 2019.

[35] C. Sommer, C. Straehle, U. Koethe, and F. A. Hamprecht, "Ilastik: Interactive learning and segmentation toolkit," in *2011 IEEE international symposium on biomedical imaging: From nano to macro*, pp. 230–233, IEEE, 2011.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[38] O. Thouvenin, C. Boccara, M. Fink, J.-A. Sahel, M. Paques, and K. Grieve, "Cell motility as contrast agent in retinal explant imaging with full-field optical coherence tomography," *Investigative Ophtalmology & Visual Science*, vol. 58, p. 4605, Sept. 2017.

[39] F. Clayton, "Pathologic correlates of survival in 378 lymph node-negative infiltrating ductal breast carcinomas. mitotic count is the best single predictor," *Cancer*, vol. 68, no. 6, pp. 1309–1317, 1991.

[40] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.

[41] D. Mandache, E. Dalimier, J. Durkin, C. Boceara, J.-C. Olivo-Marin, and V. Meas-Yedid, "Basal cell carcinoma detection in full field oct images using convolutional neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 784–787, IEEE, 2018.

[42] J. Scholler, V. Mazlin, O. Thouvenin, K. Groux, P. Xiao, J.-A. Sahel, M. Fink, C. Boccara, and K. Grieve, "Probing dynamic processes in the eye at multiple spatial and temporal scales with multimodal full field OCT," *Biomedical Optics Express*, vol. 10, pp. 731–746, Feb. 2019.

[43] J. Scholler, "FFOCT control and acquisition software," 2019. https://doi.org/10.5281/zenodo.3137245.

[44] C. Apelian, F. Harms, O. Thouvenin, and A. C. Boccara, "Dynamic full field optical coherence tomography: subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis," *Biomed. Opt. Express*, vol. 7, pp. 1511–1524, Apr 2016.

[45] J. Scholler, "Motion artifact removal and signal enhancement to achieve in vivo dynamic full field oct," *Opt. Express*, vol. 27, pp. 19562–19572, Jul 2019.

[46] G. M. Sharif and A. Wellstein, "Cell density regulates cancer metastasis via the hippo pathway," *Future Oncology*, vol. 11, no. 24, pp. 3253–3260, 2015.

[47] H. Dolznig, F. Grebien, T. Sauer, H. Beug, and E. W. Müllner, "Evidence for a size-sensing mechanism in animal cells," *Nature cell biology*, vol. 6, no. 9, pp. 899–905, 2004.

[48] D. T. Rosenthal, H. Iyer, S. Escudero, L. Bao, Z. Wu, A. C. Ventura, C. G. Kleer, E. M. Arruda, K. Garikipati, and S. D. Merajver, "p38γ promotes breast cancer cell motility and metastasis through regulation of rhoc gtpase, cytoskeletal architecture, and a novel leading edge behavior," *Cancer research*, vol. 71, no. 20, pp. 6338–6349, 2011.

[49] X. Lin, N. Wan, L. Weng, and Y. Zhou, "Light scattering from normal and cervical cancer cells," *Applied optics*, vol. 56, no. 12, pp. 3608–3614, 2017.

[50] H. Abu-Tayeh, K. Weidenfeld, A. Zhilin-Roth, S. Schif-Zuck, S. Thaler, C. Cotarelo, T. Z. Tan, J. P. Thiery, J. E. Green, G. Klorin, *et al.*, "'normalizing'the malignant phenotype of luminal breast cancer cells via alpha (v) beta (3)-integrin," *Cell death & disease*, vol. 7, no. 12, p. e2491, 2016.

[51] P. P. Provenzano, D. R. Inman, K. W. Eliceiri, J. G. Knittel, L. Yan, C. T. Rueden, J. G. White, and P. J. Keely, "Collagen density promotes mammary tumor initiation and progression," *BMC medicine*, vol. 6, no. 1, p. 11, 2008.

[52] D. Arifler, I. Pavlova, A. Gillenwater, and R. Richards-Kortum, "Light scattering from collagen fiber networks: micro-optical properties of normal and neoplastic stroma," *Biophysical journal*, vol. 92, no. 9, pp. 3260–3274, 2007.

[53] G. Hinton, "Overview of mini-batch gradient descent." `http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`, 2016. Accessed: 2019-11-26.

Methods

3.1 Description of the breast datasets and *ground truth* definition.

In total, 35 patients admitted at Gustave Roussy breast surgery department between June and August 2017 for a mastectomy or lumpectomy were enrolled in this study, from which 51 samples were collected. Once the surgical specimen arrived at the pathology department, when confirmed that tissue extraction would not impact standard processing of the specimen, none, one or several biopsies were taken in the tumorous area and none, one or several biopsies were taken far from the tumorous area to sample healthy tissue. The heterogeneous tissue sampling resulted in 37 tumoral samples, and 14 healthy biopsies according to the histology diagnosis. The most represented cancer type was Invasive Ductal Carcinoma, but three other types were reported (See Table 3).

Table 3: Study set composition according to histology-based diagnosis.

Main Diagnosis	Number of samples
Healthy	14
Invasive Ductal Carcinoma	28
Invasive Lobular Carcinoma	5
Ductal Carcinoma In Situ	3
Lobular Carcinoma In Situ	1

For all samples, the static FF-OCT image of the entire biopsy (typically $2 \times 1cm^2$) is acquired first. Then, on average 11 regions of interest (ROIs) corresponding to a single field of view of $1.3 \times 1.3mm^2$ were manually selected by the person in charge of imaging to be representative of the different areas of the biopsy. Unless an error or oversight was made by the imaging person, one static FFOCT image and one dynamic FF-OCT (D-FF-OCT) image were acquired at each ROI and the raw data of the D-FF-OCT images was saved to allow custom processing, as required in the feature engineering method. In total, 540 pairs of static and dynamic FF-OCT were acquired from the 51 samples, among which 500 were used for the blind analysis at the sample scale and later annotated by pathologist P2 at the ROI scale. The performance of the pathologists' interpretation based on the FF-OCT images was measured as referred to the standard H&E histology-based diagnosis of the same sample. As explained hereafter, only subsets of the original dataset could be used for each automated analysis.

For the feature engineering analysis, we selected the images under 3 criteria. A sample was excluded if less than 5 ROIs had been acquired with complete data. Furthermore, for each ROI, one static and its corresponding dynamic FF-OCT image should be available (without translation), as well as the raw D-FF-OCT data. It resulted in 496 static and 496 dynamic FF-OCT images from 40 samples (31 tumoral, and 9 healthy). In the feature engineering analysis, the sample histological diagnosis is passed on to all the ROIs to define the ground truth for each ROI, resulting in a ground truth with some errors.

In contrast, the CNN approach only relies on the dynamic FF-OCT images. Nonetheless, in order to obtain convincing results with this approach, histopathologist expert P2 was asked to label all the individual dynamic FF-OCT images based on his interpretation and correlation with the H&E histological slide. We converted the expert P2 annotation into a binary label indicating the presence of a tumoral part in each ROI. However, not all images could be labeled with high confidence, so that the dataset was reduced to 373 dynamic FF-OCT images from 47 samples (34 with a tumor, 13 without).

FF-OCT setups

In this study we used two different FF-OCT systems. The first one, used for breast biopsy imaging, is a commercial Light-CT scanner, manufactured by LLTech SAS (Paris, France). It is a Linnik interferometer equipped with microsope ojectives (Olympus, UMPLFLN 10XW) used immersed in silicone oil and a broadband LED source (104 nm full width at half maximum, 565 nm central wavelength), thus providing micrometer 3D resolution. X, Y and Z motorization allows axial and transverse scanning. FF-OCT acquisition is performed using 4-phase modulation induced by a piezoelectric actuator controlling small translations of the reference mirror. As a consequence, a FF-OCT single field of view requires the acquisition of at least 4 images and typically 20 images with 5-times accumulation. Including scanning time, the FF-OCT imaging speed of the Light-CT scanner is $1min.cm^{-2}$. For D-FF-OCT acquisition, the mirror position is not modulated, and the

system records the fluctuations arising from the motion of scatterers inside the coherence volume over a few seconds. Typically, 1000 images are acquired. Then, the time series of each pixel is processed by FFT and results in a RGB value assignment per pixel, where the R-value is the FFT summed contribution over the high frequency range (5.4 - 25 Hz), the G-value is the FFT summed contribution over the medium frequency range (0.6 – 5.4 Hz), the B-value is the (0 – 0.6 Hz). Including scanning time, the D-FF-OCT imaging speed of the Light-CT scanner is 14 $min.cm^{-2}$. Saving raw data further lengthen the D-FF-OCT imaging duration. During the study, as several biopsies could be resected at the same time, it was important limiting the imaging duration to reduce the elapsed time between resection and imaging so that the tissue freshness was preserved and good quality D-FF-OCT signal was guaranteed. This is why we performed ROI sampling instead of acquiring the whole D-FF-OCT image.
The second system, used for retinal sample imaging, is a laboratory setup, which consists of a Linnik interferometer where both arms contain identical microscope objectives (Olympus UPlanSApo 30x), see Fig. 5(a). Both sample and reference arms are mounted on translation stages (X-NA08A25, Zaber Technologies) for optical path length matching and axial scanning, respectively. In addition, the reference mirror is mounted on a piezoelectric translation stage (STr-25, Piezomechanik) for phase-shifting. Indeed, in order to construct an FFOCT image, at least two frames are acquired with different phase modulations and appropriate phase-shifting algorithms are used to build the final image [25, 42]. This setup is purposely built in an inverted configuration where the sample is directly placed on top of a coverslip and imaged from beneath with high-numerical-aperture oil-immersed objectives [42]. In this configuration the sample is held motionless by gravity and is naturally flattened against the coverslip. The characteristics of both setups are summarized in Table 4.

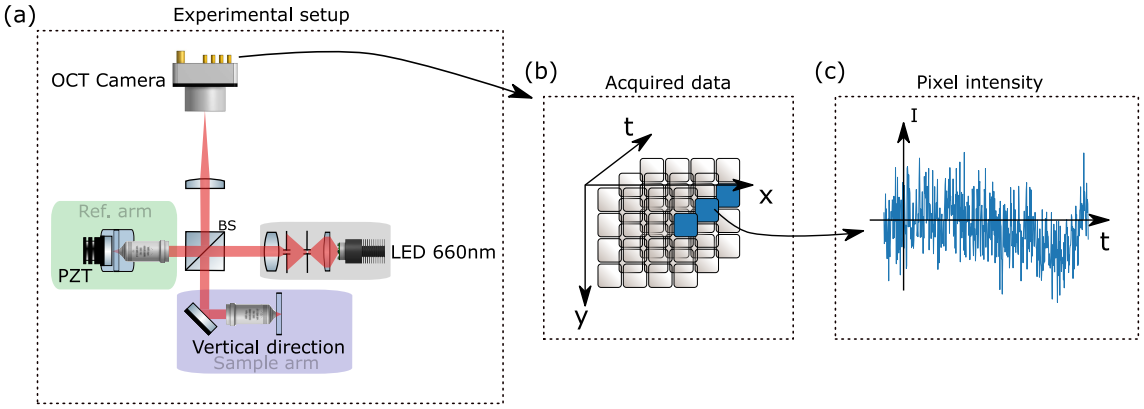

Figure 5: PZT: piezoelectric translation - BS: Beam splitter. Experimental D-FFOCT setup in an inverted configuration optimized for tissue imaging. 512 images are acquired by the CMOS camera (a). The resulting $(1440, 1440, 512)$ 3D tensor (b) is then processed independently for each pixel (c).

Table 4: **Setups characteristics.**

Setup	Lab. setup Fig. 5(a)	LightCT (LLTech SAS)
Datasets	Retina	Breast
Transverse resolution $[\mu m]$	0.4	1.2
Axial resolution $[\mu m]$	1.4	1
Field of view $[\mu m \times \mu m]$	390×390	1260×1260
Camera Framerate $[Hz]$	150	150

Acquisition protocol and signal processing to construct D-FF-OCT images for feature engineering and retinal imaging.

In order to construct a D-FF-OCT image, a stack of 512 direct images (1440×1440 pixels) is acquired (Fig. 5(b)) using custom software [43] where each pixel is processed independently (Fig. 5(c)). In the first report on D-FF-OCT, grayscale dynamic images were computed using a running standard deviation averaged over the whole acquisition [44]. Computing the intensity standard deviation in time removes the signals from highly scattering stationary structures such

as collagen and myelin fibers and enhance signals arising from cells.
Since recently, we calculate color images [31] in the HSV (hue-saturation-value) color space in which, contrary to the red-green-blue (RGB) colour space, it is possible to assign a physical property to each of the three channels for quantitative visual interpretation. The idea is then to attribute a colour for each pixel depending on the characteristic time period or frequency of the dynamic signal. Each individual pixel can be thought of as a sum of subcellular random walks with a typical covariance function depending on the motion type (e.g., diffusive, hyperdiffusive).
We started by computing the power spectrum density (PSD) using Welch's method for each pixel and then used an L1 normalization on each PSD. Then, the hue channel, was computed as the mean frequency (which is simply the dot product between the normalized PSD and the frequency array). The values were then inverted and rescaled between 0 and 0.66 to go from blue (low frequencies) to red (high frequencies).
Saturation was computed as the inverse of the normalized PSD bandwidth. As a consequence, the saturation channel carries the frequency bandwidth information. The saturation map is then inverted and rescaled between 0 and 0.8. The broader the spectrum, the lower the saturation. White noise has a broader bandwidth and will therefore appear greyish instead of coloured.
Finally, the value is calculated using the cumulative sum on small time windows to improve the signal to noise ratio [45]:

$$\bar{I}_{dyn}(\boldsymbol{r}) = \frac{1}{N}\sum_{i} max\left(\left|CumSum\left(I(\boldsymbol{r}, t_{[i,i+\tau]}) - \bar{I}(\boldsymbol{r}, t_{[i,i+\tau]})\right)\right|\right) \quad (1)$$

where $CumSum$ is the cumulative sum operator, N is the total number of sub-windows, τ is the sub-windows length so that $t_{[i,i+\tau]}$ is the time corresponding to one sub-window and $\bar{I}(\boldsymbol{r}, t_{[i,i+\tau]})$ is the signal mean on the sub-window. Images were computed with $\tau = 16$.
The resulting HSV image is finally converted into a RGB image for display purposes.

H&E histology preparation and imaging

. All imaged samples were fixed in 10 % buffered formalin, embedded in paraffin, sectioned at 3 μm, and stained with hematoxylin, eosin, and saffron (HES).
Histological sections were scanned at 20× magnification with a NanoZoomer C9600 scanner (Hamamatsu Photonics, Massy, France) for interpretation and archiving. For the correlation, histopathologist P2 together with a FFOCT technology expert from LLTech, decided on the angle to apply to the histology image to obtain the same orientation as the FFOCT large-field image. Then, well-oriented histology image was displayed on one screen while the ROIs were consecutively displayed on a second screen. Both histology and FFOCT images were zoomed in and out to establish the correlation.

Segmentation of cells and fibers using iLastik

The first step of the engineered features analysis on breast samples was to perform cell and fiber segmentation in the images, using *iLastik*, a free segmentation software [35]. *iLastik* is a relatively intuitive machine-learning tool based on random forest classifiers. The labels are manually drawn on the training images in a user interface. Each pixel neighborhood is characterized by a set of generic nonlinear spatial transformations calculated by the software and applied to each channel (R, G, or B) of the D-FF-OCT image. The same transformations were applied to the greyscale FF-OCT images. The following image transformations empirically gave the best contrast in our case:

1. **Gaussian smoothing** with a increasing standard deviation of 0.3, 0.7, 1.0, 1.6, 3.5, 5.0.
2. **Laplacian calculation after Gaussian smoothing** with a standard deviation of either 0.7, 1.0, 1.6, 3.5, or 5.0. This computes the edge of the objects.
3. **Difference of Gaussians** with a standard deviation of either 0.7, 1.0, 1.6, 3.5, or 5.0. It also computes the edge of the objects, by subtracting two images after Gaussian smoothing with almost similar standard deviations and is supposed to approximate a derivative.
4. **Hessian of Gaussian Eigenvalues** with a standard deviation of either 0.7, 1.0, 1.6, 3.5, 5.0. It computes the local texture of the image by calculating the determinant of the Hessian matrix.

In total, for all the D-FF-OCT images in the dataset, 21 transformed images are calculated per color, hence 63 grayscale images, plus 21 images for FF-OCT images. For both FF-OCT and D-FF-OCT datasets, the training step was performed on 8 images (5 from cancerous samples and 3 from healthy samples). For D-FF-OCT, we defined three classes: *Cells, Between Cells, Not Cells (Fibers, Noise, Fat, etc...)*, out of which only the *Cell* class was used. A few pixels corresponding to each class are drawn manually, as shown in figure 6A. The best practice is to draw a small number of pixels to minimize constraints, and to draw pixels at the interface between classes (meaningful pixels). The classifier (figure 6B) should be created as soon as a few pixels of each class are annotated (typically a few cells, and a few spaces between cells in the first image), and then finely tuned by manually changing the class of misclassified pixels in all 8 images. The whole training procedure can last for around 30 minutes, and is not highly accurate since a few cells and cell contours are missed. High accuracy is not necessary though given than the expected outcome is not a real metric of cell sizes or shapes, but how these parameters compare from one sample to another.
Segmentation of FF-OCT was performed similarly, although we used 4 classes: *Fibers, Around Fibers, Cells (and Noise), Fat (and holes/ slicing artifacts).* The *Around Fibers* calss is used to be able to segment individual fibers, and not only the regions of high fiber density. The *Cells* class contains most of what is not fibers, including extracellular matrix and cells. We later split this data into bright pixels (50% brightest pixels), most likely forming the extracellular matrix, darker pixels (between 25% and 50% brightest pixels)being part of cells, and we excluded the darkest pixels (¡25%). The *Fat* class segments dark pixels in circular regions, as observed in fat regions of the breast (Figure 2A1 and A2 in blue). Unfortunately, it also finds a few holes in the sample, caused by slicing, folding artifacts, or bubbles.
In total after these learning steps, each FF-OCT and D-FF-OCT image can be segmented in about 10 seconds.

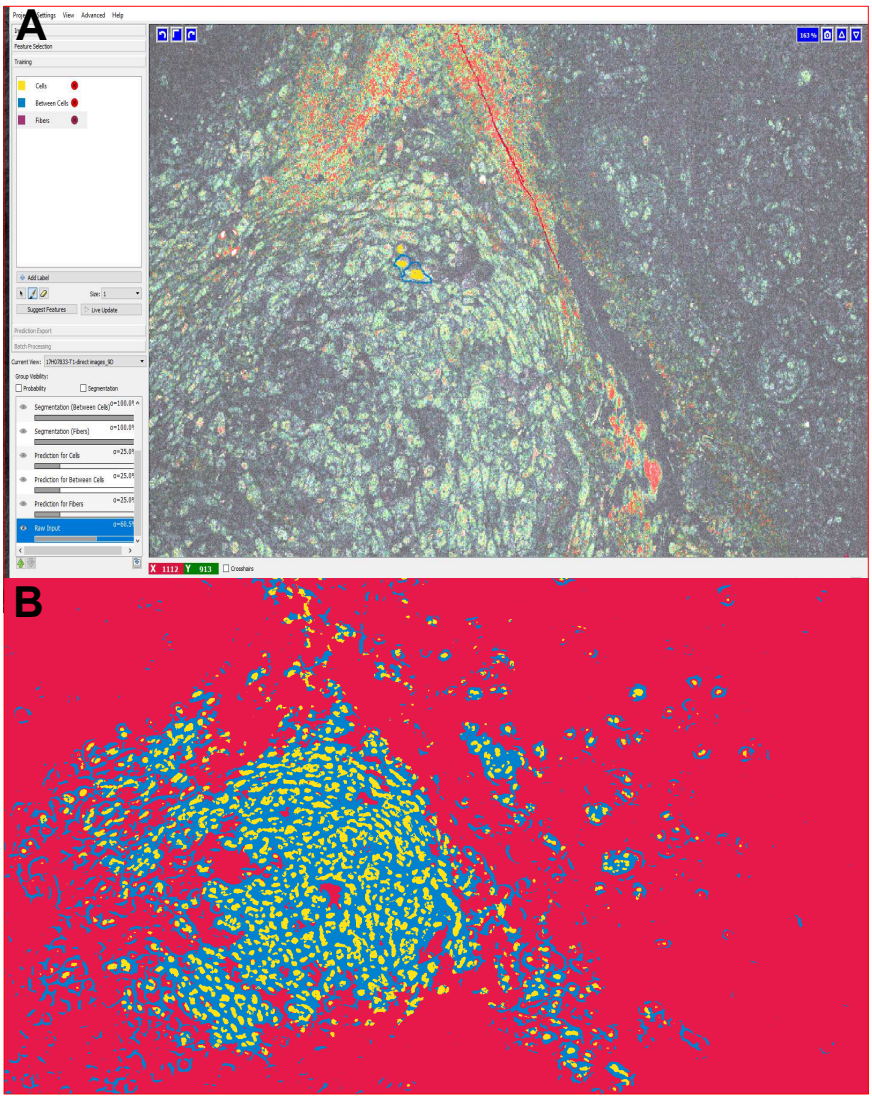

Figure 6: Machine learning based segmentation with iLastik on D-FF-OCT images. Panel A shows the user interface of iLastik, and illustrates the learning procedure. Pixels corresponding to each class of interest are manually drawn (opaque pixels). To start with initial prediction, a small number of pixels in the first image were drawn, and then the prediction is refined step by step by correcting misclassified pixels. Panel B shows the segmentation result after the learning process. For DCI images, only the cell segmentation is used (class 1 - in yellow here), but the others two classes were used to segment the cells with higher precision.

Further image analysis and feature calculation

The two sets of segmented images were then opened and analyzed using image processing (*Matlab*, MathWorks). From the D-FF-OCT images, we target individual cells, and some of their morphological parameters that we expect to be modified in a tumor, such as cell density [46], cell size [47] or shape [48]. We also measured their light scattering properties [49], and their motility [48]. From *iLastik* segmented images, we detected and characterized all 8-connected objects using the *regionprops* function (Matlab, Image Processing Toolbox), which we estimate to correspond to single cells (See figure 7. We additionally filtered all regions with an area below 20 pixels (20 μm^2, corresponding to a cell radius below 2.5 μm) to filter out some classification errors. We then extracted each object diameter, eccentricity, mean intensity on each channel, the spatial heterogeneity of intensity, as well as the total cell density. For each image, we could obtain hundreds of

measurements for each of these parameters, but in order to have a reasonable number of features for the SVM analysis, we only used the average and standard deviation of the obtained histograms.

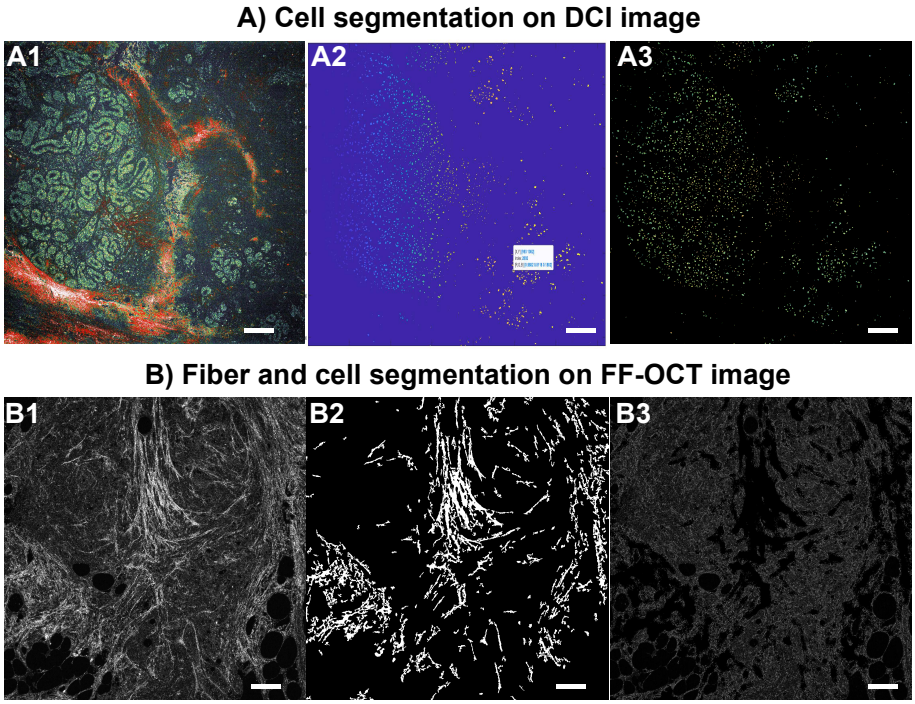

Figure 7: Cell and fiber segmentation output. Panel A shows the cell segmentation using iLastik and Matlab. From original DCI image (A1), after iLastik segmentation, the first class (cells) is selected, and separated in independent regions with Matlab (A2). Here, the increasing color from blue to red gives the number of the identified region (from 1 to 3951 here).Each region of interest (ROI) can be analyzed independently, and only the regions of area above 20 pixels are kept and considered as cells. The obtained mask image is multiplied by the D-FF-OCT image to obtain panel A3. Panel B shows the fiber and cell segmentation results on FF-OCT image. The original image (B1) is processed by iLastik and classified between fibers, pixels between fibers, and cells. The first and third classes are extracted and filtered using Matlab to obtain fibers (B2) and cells (B3)

Additionally, we measured what we called mesoscale cell features to have a measurement of the spatial organization of cells at an intermediate level. For example, in healthy breast samples, relatively high cell density can be found in acinii, but cells are highly organized in a rather circular region with a central lumen. In *Ductal carcinoma In Situ* (DCIS), the acinii show a modified shape with a progressive invasion of the central lumen [50]. It is therefore important to quantify whether the high cell density regions contain small, well organized cells (as in healthy samples) or extended cells with an abnormal shape. To do this, we used the D-FF-OCT segmented image, and convolved it with a 60 pixel wide square, which thus gives an image of local cell density (Figure 2). From this map, we segmented all pixels with a local density above 0.05 (arbitrary), and detected all 8-connected objects. We then extracted the number of high density areas, their surface, and eccentricity. We finally counted the number of all cells whose position are outside these regions of high cell density, aiming to track for tumor cells migrating outside the ducts.

From the FF-OCT images, we mostly target the stromal collagen fibers since their organization [51] and light scattering properties [52] are expected to change in cancerous samples. We also measured some properties linked with cell regions and fat regions to add information to segment the images. We first used the segmented images corresponding to fibers. We detected and characterized all 8-connected objects, corresponding to single fibers, and filtered out all regions of area below 100 pixels. We measured each fiber area, eccentricity, orientation, and intensity, as well as the fiber density, and extracted the mean and standard deviation of each parameter in an image. We then calculate mesoscale fiber features by convolving the segmented fiber image with a 60 pixel wide square, and by segmenting all pixels with local fiber density above 0.1. We kept the area, eccentricity, intensity, and number of these regions of high fiber density. We then used the image segmented by cell region to calculate the intensity histogram in these regions. We finally used the

segmented image of fat to calculate the proportion of fat in the image.

All these features are summarized in the table 5:

SVM training

From the previous section, we obtained a matrix with 44 features for all 496 images in our dataset. The idea behind the SVM classification is to find a (N-1) dimensional hyperplane that best separates the dataset between healthy and cancerous images. If none of these features alone can clearly separate the dataset, the SVM will find a linear combination of all these parameters that best separates the dataset.
When a feature was missing (e.g. if no region of high cell density was found in a ROI), the value 0 was attributed.
For training the SVM, we attributed the histological diagnosis to all images of each sample. We could easily test several models to perform the image classification (*Matlab Classification Learner* toolbox, MathWorks), but decided to keep a simple linear SVM. As mentioned in the main text, the dataset labeling is highly asymmetric; first, we have about three times more cancerous images than healthy images. Second, images from healthy samples should all be healthy, but images from cancerous samples can be either healthy or cancerous. In order to reduce overfitting (e.g. always predicting cancer should give 75% accuracy), we made the SVM learning asymmetric by penalizing the false positives by a factor 3.

Tumor detection using fine-tuned CNN

Description of the CNN architecture.

The CNN architecture is similar to the VGG16 [36] architecture with weights pre-trained on the ImageNet dataset [37], but with small modifications. We removed the classifier part and added a global average pooling (GAP) layer followed by a fully-connected layer of 1024 neurons and an output neuron with sigmoid activation. The GAP layer allows network inputs of different sizes since it reduces each activation map to a single value, it results in a 512-dimensional vector bottleneck between the feature extracting convolutional layers and the dense classifier layers. Another advantage of GAP is the reduction of network parameters which improves generalization and is particularly useful in the case of small scale data sets. Following the same reasoning we chose only one fully-connected layer. With the presented configuration we obtain a network with approximately 15 million parameters of which 500 thousand correspond to the classifier and the rest to the pre-trained weights.

Training

The network was fine-tuned by minimizing the binary cross-entropy loss using the stochastic gradient descent (SGD) optimizer with a learning rate of 1e$-$4 and momentum of 0.8 on mini-batches of size 3 (due to memory constraints).

Since we opted for fine-tuning, training time is significantly reduced compared to a training from scratch approach: consequently the phenomenon of overfitting occurs much faster. To avoid training beyond the optimal model, we have set two stopping conditions: (i) validation loss has not improved in the last 100 epochs or (ii) training accuracy has already reached 100%. With this, training lasts around 200 epochs and the optimal model is found somewhere between epoch 80 and 150 (depending on the data split and initialization, so depending on the random state). Training time is around 6.5 minutes per epoch, and 20 hours per experiment. Thus conducting a 5-fold cross-validation experiment took around two and a half days (64 hours); note that in these delays we also included the lag introduced by logging performance metrics, as well as the overhead introduced by reading the image batches from the disk, and not only training (i.e. forward and backward propagation). Experiment tracking was made possible with the software *Neptune* (Neptune.ai), which helped organize and compare the performance of over 200 experiments conducted for this project, therefore allowing us to choose the optimal hyperparameters in an exploratory fashion.

Data

The network was trained on full resolution $1440 \times 1440 \times 3$ RGB fields of view with binary labels indicating the presence or absence of tumorous tissue, obtained from the pathologist's refined annotation per ROI. An important detail is that in tumorous ROIs, there might be portions of the

Table 5: Calculated features from image analysis of static and dynamic FF-OCT images. For some features (e.g. cell diameter), a distribution is first calculated, but the extracted features are the mean and standard deviation (STD) of the distribution.

Number	Type of feature	Description of the feature	Additional comment
1 & 2	Cell features from D-FF-OCT	Diameter of segmented cells (Mean and STD)	Measure Cell size
3 & 4	Cell features from D-FF-OCT	Eccentricity of segmented cells (Mean and STD)	Measure Cell shape
5	Cell feature from D-FF-OCT	Total cell density	$Density = \frac{N_{seg.cells}}{N_{pixels}}$
6 & 7	Cell features from D-FF-OCT	Average intensity within each segmented cell (Mean and STD)	Scattering is expected to increase in cancerous cells
8	Cell feature from D-FF-OCT	Mean intensity of the red channel of all pixels classified of the *cell* class	Fast dynamics : Cancerous cells are expected to have increased metabolic activity, hence faster and stronger fluctuations.
9	Cell feature from D-FF-OCT	Mean intensity of the green channel of all pixels classified of the *cell* class	Intermediate dynamics
10	Cell feature from D-FF-OCT	Mean intensity of the blue channel of all pixels classified of the *cell* class	Slow dynamics
11 & 12	Mesoscale features from D-FF-OCT	Local cell density in regions of high cell density (Mean and STD)	High cell density : At least 5% of the surrounding pixels belong to the *cell* class.
13& 14	Mesoscale features from D-FF-OCT	Area of regions of high cell density (Mean and STD)	In healthy samples, the lobules are high cell density regions but they are well organized and small.
15& 16	Mesoscale features from D-FF-OCT	Eccentricity of regions of high cell density (Mean and STD)	
17	Cell feature from D-FF-OCT	Number of cells measured outside regions of high cell density	Migrating cells should not be found in healthy samples.
18-21	Cell features from D-FF-OCT	Spatial STD of intensity and STD normalized by intensity) within each segmented cell (2 means and 2 STDs)	Measure a the intensity heterogeneity (STD(I) and STD(I)/I) inside each cell.
22 & 23	Fiber features from FF-OCT	Diameter of segmented fibers (Mean and STD)	Measure collagen fiber size
24 & 25	Fiber features from FF-OCT	Eccentricity of segmented fibers (Mean and STD)	Measure collagen fiber shape
26 & 27	Fiber features from FF-OCT	Angular distribution of segmented fibers (Mean and STD)	Measure collagen fiber organization, possibly altered in cancerous samples.
28	Fiber features from FF-OCT	Density of segmented fibers	Possibly increasing in cancerous samples
29 & 30	Fiber features from FF-OCT	Mean intensity of each segmented fiber (Mean and STD)	Scattering of the extracellular matrix is possibly increasing in cancerous samples.
31 & 32	Cell features from FF-OCT	Mean and STD of intensity of all pixels that belongs to the *cell* class.	Scattering of cancerous cells are possibly increasing.
33	Cell features from FF-OCT	Mean intensity of the 25% brightest pixels that belongs to the *cell* class.	
34	Mesoscale features from FF-OCT	Number of pixels belonging to the regions of high fiber density	
35 & 36	Mesoscale features from FF-OCT	Mean intensity within each region of high fiber density (Mean and STD)	
37 & 38	Mesoscale features from FF-OCT	Area of each region of high fiber density (Mean and STD)	
39 & 40	Mesoscale features from FF-OCT	Eccentricity of each region of high fiber density (Mean and STD)	
41	Mesoscale features from FF-OCT	Number of regions of high fiber density	
42	Fat features from FF-OCT	Average intensity of all pixels belonging to the *Fat* class	
43	Fat features from FF-OCT	Number of pixels belonging to the *Fat* class	
44	Cell features from FF-OCT	Mean intensity of the 25%-50% least bright pixels that belongs to the *cell* class.	

image resembling healthy tissue.

The dataset partitioning into training and test sets was performed in a stratified manner with respect to the global per-sample diagnosis ; in terms of size, 80% of the samples ($n_{train} = 37, n_{train-tumor} = 27, n_{train-normal} = 10$) used for training and the remaining 20% ($n_{test} = 10, n_{test-tumor} = 7, n_{test-normal} = 3$) for evaluating the performance. However, this stratification strategy does not ensure the exact same distribution of classes at ROI level, as each sample has a different number of acquired ROIs; this results in 286 fields of view for training (185 positive and 101 negative) and 87 for testing (60 positive and 27 negative). To compensate for the class imbalance, an importance penalization of the loss function was applied for each ROI, which is also known in the literature as class weight. In our case the healthy ROIs are less numerous so they will have a higher weight (i.e. 1.5 for healthy ROIs and 0.75 for cancerous ROIs).

In terms of data augmentation i.e. artificially increasing the number of data points by applying relevant transformations to the existing data, we applied contrast stretching, with 3 look up tables per FOV, together with vertical and horizontal flips, which expanded the training set by up to 6 times.

Quantitative Performance

With a probability threshold set at 50% for ROI diagnosis, we obtained a per-FOV accuracy of 90% which corresponds to 92% sensitivity and 85% specificity. Another metric that is worth mentioning, due to its lack of dependence on the probability threshold, is the area under the ROC curve (AUC), which is equal to 0.94. For per-sample metrics, out of 10 samples in the test set 7 of which were cancerous and 3 healthy, there was only one misclassification, which is a normal sample predicted as tumorous; which translates to 90% accuracy, 100% sensitivity and 67% specificity

To aggregate the per-ROI predictions to a global per-sample diagnosis, assigning the maximum tumor probability prediction to the sample would be the most straightforward approach. This would translate to "if a sample contains at least one ROI with a tumor, then the sample is cancerous". This approach is however overly sensitive to outliers. On the other hand, the average or median are not suitable either because a bimodal distribution is expected (i.e. a sample most likely contains both healthy and tumorous FOVs), in other words small tumorous areas would be missed, or cancel out good prediction. Therefore, we chose the 90^{th} percentile as a good trade-off between the mean and the maximum aggregations. This would translate into the probability value that 90% of the ROIs fall into.

In order to validate the method and ensure model correctness, we ran a 5-fold cross-validation experiment with the same hyper-parameters and trained 5 models on partitions of 4/5 of the samples and tested their performance on 1/5 of the samples, respectively, hence keeping the same 80/20 train/test ratio at each run.

GradCAM algorithm and attention maps.

The receptive field represents the zone (or patch) of the input image that a CNN feature is computed upon. In classical convolutional arhictectures, like VGG16 used in the present work, that are composed of convolutional and pooling layers, the size of the receptive field is directly proportional to its corresponding layer depth, allowing to learn progressively more complex features. Receptive fields of pooling layers: 6×6, 16×16, 44×44, 100×100 and 212×212. The GradCAM algorithm uses the gradient information flowing into the last convolutional layer of the CNN (Figure 4C) to assign importance values to each neuron for a particular class. The averaged gradients flowing back from a chosen class output neuron to a previous layer (usually last convolutional layer) act as weighting factors for each activation map, the final result being a linear combination between the weights and filter activation maps.

Attention maps : sum of last convolutional layers of the last 2 convolutional blocks in vgg i.e. the third layer of the fifth (last) and fourth blocks. We chose to add the attention map of the second-to-last block to offer a finer detail of the segmentation due to its increased resolution i.e. 180×180 as compared to 90×90 respectively.

Thumbnail image classification on mouse retina

In order to perform both feature extraction and classification the VGG16 network was used and adapted to this problem. The input size was changed so that it accepts 100x100x3 thumbnail

images. The convolutional part, depicted in orange in Fig. 8(b) was not retrained so that weights correspond to the VGG16network trained on the ImageNet data-set [37]. To perform the classification, 2 fully-connected layers of size 1024 were added with a Rectified Linear Unit (ReLU) activation followed by a fully connected layer of size 5 with a SoftMax activation. These last layers were trained using dropout with a rate of 0.5. The final model consists of 20.5 million parameters of which 5.8 million are trainable. Data augmentation was used to generate more training data by using translation, rotations, mirror flips and a small amount of shear (maximal shear angle was 10°), hence improving the generalization. The network was trained with 30 epochs with a batch size of 32 images and a total of 500 images per epoch using a categorical cross-entropy loss and a RMSProp optimizer [53] which adapts the learning rate using a running average of the past gradient magnitudes.

5 different mouse retinas were imaged with D-FF-OCT and a training set was created on a single mouse retina Fig. 8(a) by manually annotating 395 random thumbnail images in 5 classes: GCL, IPL, INL, ONL and IS/OS. The test set was created by manually annotating 292 random thumbnail images on a second retina in order to avoid over-fitting problems.

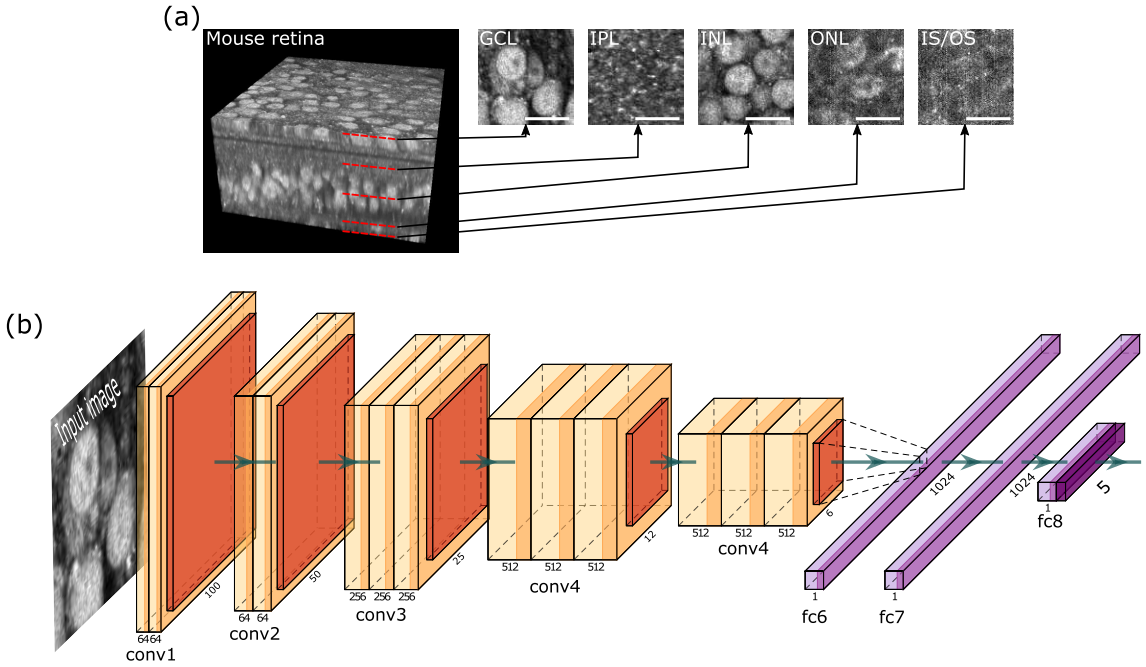

Figure 8: (a) **3D stack of mouse retina** on 80 μm depth. 697 thumbnail images of 100×100 pixel (corresponding to 27×27 μm) were randomly selected and manually classified on a single retina in 5 different classes: GCL, IPL, INL, ONL and IS/OS. (b) Neural network architecture based on VGG16 used for classifying thumbnail images in 5 different classes. Only the lasts fully connected layers (depicted in purple) were trained. Scale bar: (a) 25 μm.

3.2 Mice retina preparation

Wild-type C57BL/6JRj female mice, from eight to ten weeks old, were purchased from Janvier SA. Mice were maintained at the Institut de la Vision animal facility under pathogen-free conditions. All animals were housed under a 12-h light-dark cycle, with food and water available ad libitum. Mice were sacrificed by cervical dislocation. The cornea was cut through and the lens removed by using a thin scalpel blade, introduced at the level of the ora serrata. Curved forceps were then used to delicately extract the entire retina. Retinas were then rinsed in sterile PBS (phosphate-buffered saline 1X, Gibco), cut radially four times (in a "flower" shape) to be whole mounted. Time between time of death of the mouse and retinal imaging never exceeded five minutes. All manipulations were performed in accordance with the association for research in vision and ophthalmology (ARVO) Statement for the Use of Animals in Ophthalmic and Vision Research. No experimental procedure was realized on the mice before sacrifice.

Supplementary Material

Additional details regarding the SVM.

Put figures from methods here (MatlabSeg.pdf + Ilastik.pdf) because I don't think they can go in Methods. (Similar for the big table with all the features

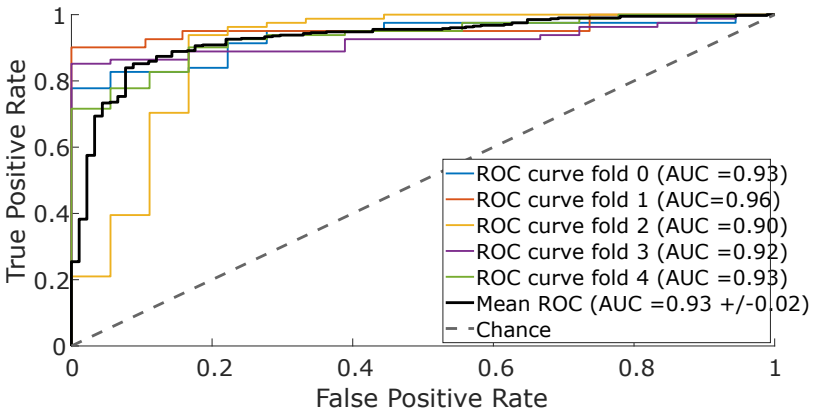

Figure 9: AUC Curves linear SVM

Reducing the spatial scale with the feature engineering approach.

In order to reduce the spatial scale of the predictions in the case of the feature engineering approach, we reduced the size of the input images. In practice, each $1.3 \times 1.3mm^2$ ROI was subdivided into respectively 1x1 (full ROI), 2x2, 4x4, 8x8, or 16x16 smaller images. The features were calculated on the original full ROI but were then attributed to the corresponding sub-image associated with the center of the segmented object (e.g. cell, fiber, region of high cell density, etc...). We obtained 5 matrices of 44 features by $(496xN^2)$ images, with N the number of ROI subdivision ranging from 1 to 16, that were used to perform 5 independent SVM trainings. The phenotype of each ROI was attributed to all subimages derived from the ROI. A 5-fold cross validation was performed for all 5 SVMs, and the accuracy, specificity, sensitivity, and AUC were measured and compared (Figure 10A). For all 4 metrics, the prediction scores decrease with the ROI subdivision ratio N, reaching a low accuracy of $71.3\% \pm 0.4\%(Sensitivity : 77.0\% \pm 0.6\%, Specificity : 49.2\% \pm 0.8\%)$ and low AUC of 0.704 ± 0.04 for N=16, which corresponds to a sub-image size of $82x82\mu m^2$ (Figure 10 B1 and B2 for N=4 and N=16 respectively). When aggregated for each sample, these lower prediction scores at the sub-image level also translates into lower prediction score at the tissue scale with a measured accuracy of 92.5% for N=4, and 80% for N=16 (Figure 10 C1 and C2 for N=4 and N=16 respectively).

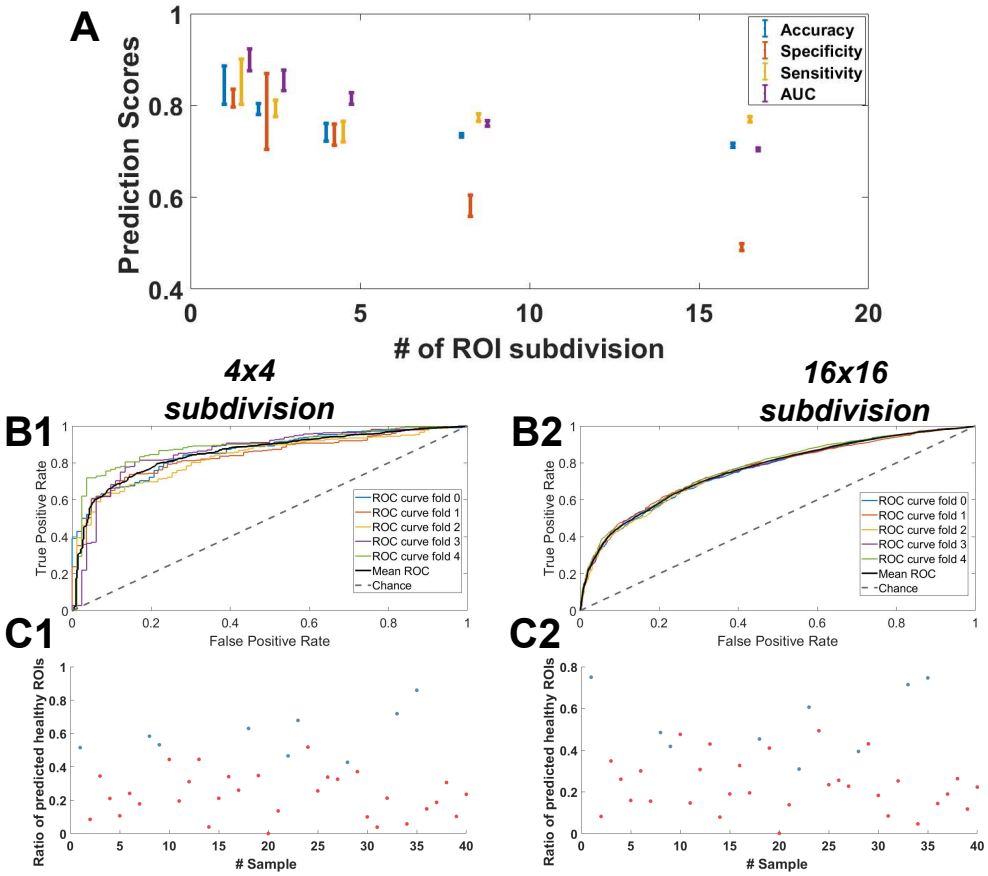

Figure 10: **Classification results with feature engineering approach with ROI subdivisions. A**. Predictions scores (Accuracy, Sensitivity, Specificity, and AUC) obtained for SVMs trained with subimages with increasing subdivision ratios N.**B.** ROC curves obtained for N=4 (Panel B1) and N=16 (Panel B2). **C.** Ratio of healthy subimages for each tissue measured for N= 4 (Panel C1) and N=16 (Panel C2).

Reducing the spatial scale with the CNN approach.

Following the same scheme, the classification between cancerous and healthy samples from breast biopsies was performed on D-FFOCT sub-images taken on a LightCT FFOCT commercial device. Each ROI (1440x1440 pixels) was partitioned into 12x12 = 144 thumbnail images of 224x224 pixels with a 50% overlap between the small images. Because of the small spatial scale, several thumbnail images were mostly showing noise, so that we had to create a third class corresponding to noisy thumbnail images. This is done to help the network focusing on interesting thumbnails rather than trying to classify noise when it is clearly not possible because no actual dynamic signals are present in the thumbnail. Therefore, we had to perform a first training step, during which we manually checked all the thumbnail images. If no signal was detectable on the thumbnail image, we labeled it as a noisy thumbnail. For other thumbnail images, we gave the class of the corresponding full ROI. Note that it may happen that healthy subregions could be found in cancerous ROIs, meaning that the dataset labeling is certainly biased in that regard.
In this experiment, we used the full D-FF-OCT dataset available composed of 650 ROIs taken on 51 samples from 35 patients. 517 ROIs were from cancerous samples and 133 ROIs were from healthy samples. During the training, adaptive sampling was used so that the neural network saw the same number of examples for each category. 10% of the thumbnail images were held out for final accuracy testing (these thumbnail images were selected among the same ROI, so that one ROI is not split across training, validation and testing sets). The network performed reasonably well with an overall accuracy of 76,41% (see Table 6) but had lower performances than the network trained on full ROI. In particular, the classifier tends to over predict cancer even-though adaptive sampling was used. Plotting the ratio between thumbnails classified as healthy and thumbnails classified as cancerous allows to separate most of the ROI, except for 3 false negatives and 1 false positive.

Table 6: **Accuracy results on breast cancer classification.**

	Accuracy on held out set
Healthy	45.41%
Cancerous	83.63%
Noise	66.98%
Total	76.41%

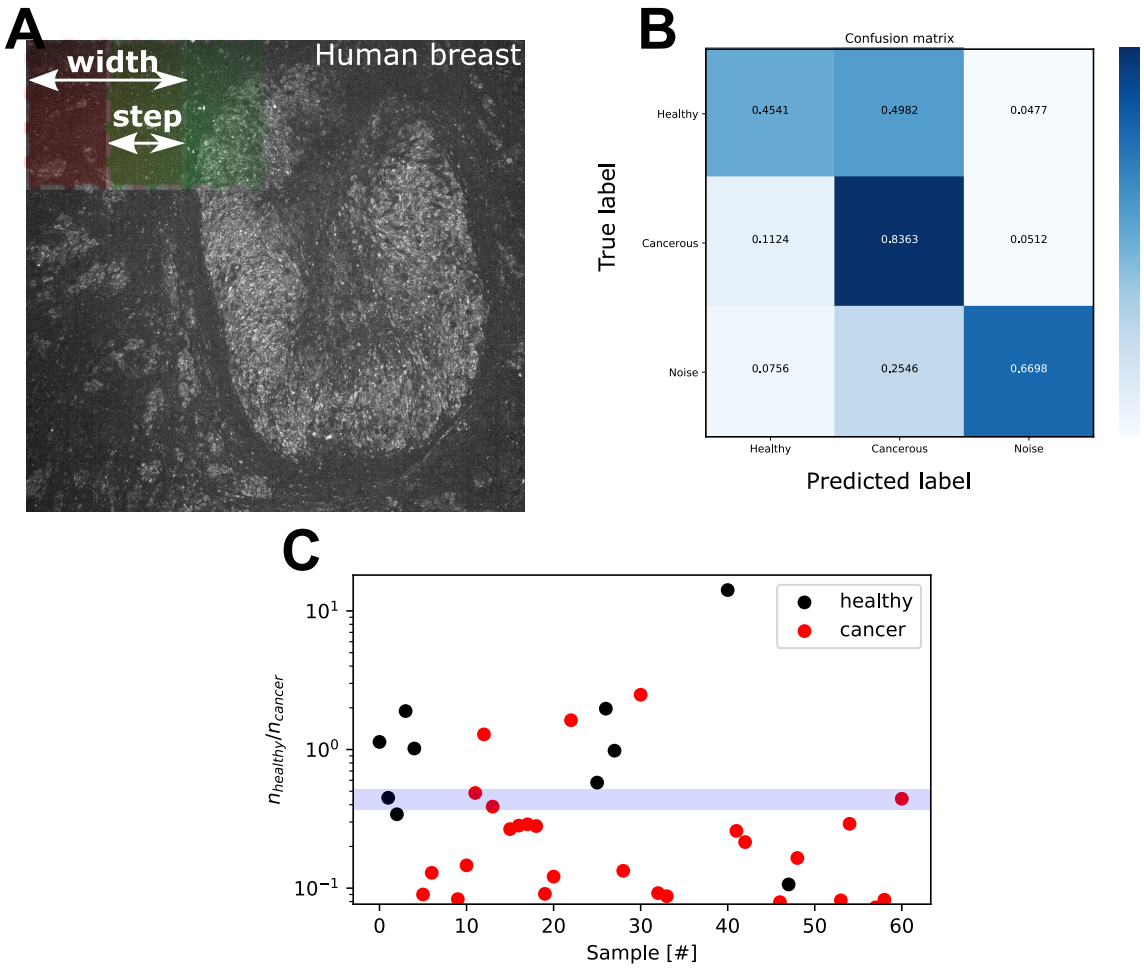

Figure 11: **CNN classification of breast cancer on small ROIs**. Each $1.3x1.3mm^2$ ROI was subdivided into thumbnails images, *i.e.* small square ROIs of width=224 pixels, with an overlap of 50% (step = 112 pixels), obtaining 12x12 = 144 thumbnail images per ROI (Panel A). A CNN was trained from the thumbnail images and by adding a third possible class corresponding to noisy thumbnail images (*e.g.* bottom right corner in Panel A). The confusion matrix of the CNN results is shown in panel B. When aggregated per sample by computing the ratio of predicted healthy thumnbail images over cancerous thumbnail images, the CNN results allow segmenting healthy from cancerous samples with reasonable accuracy (Panel C).

Figures

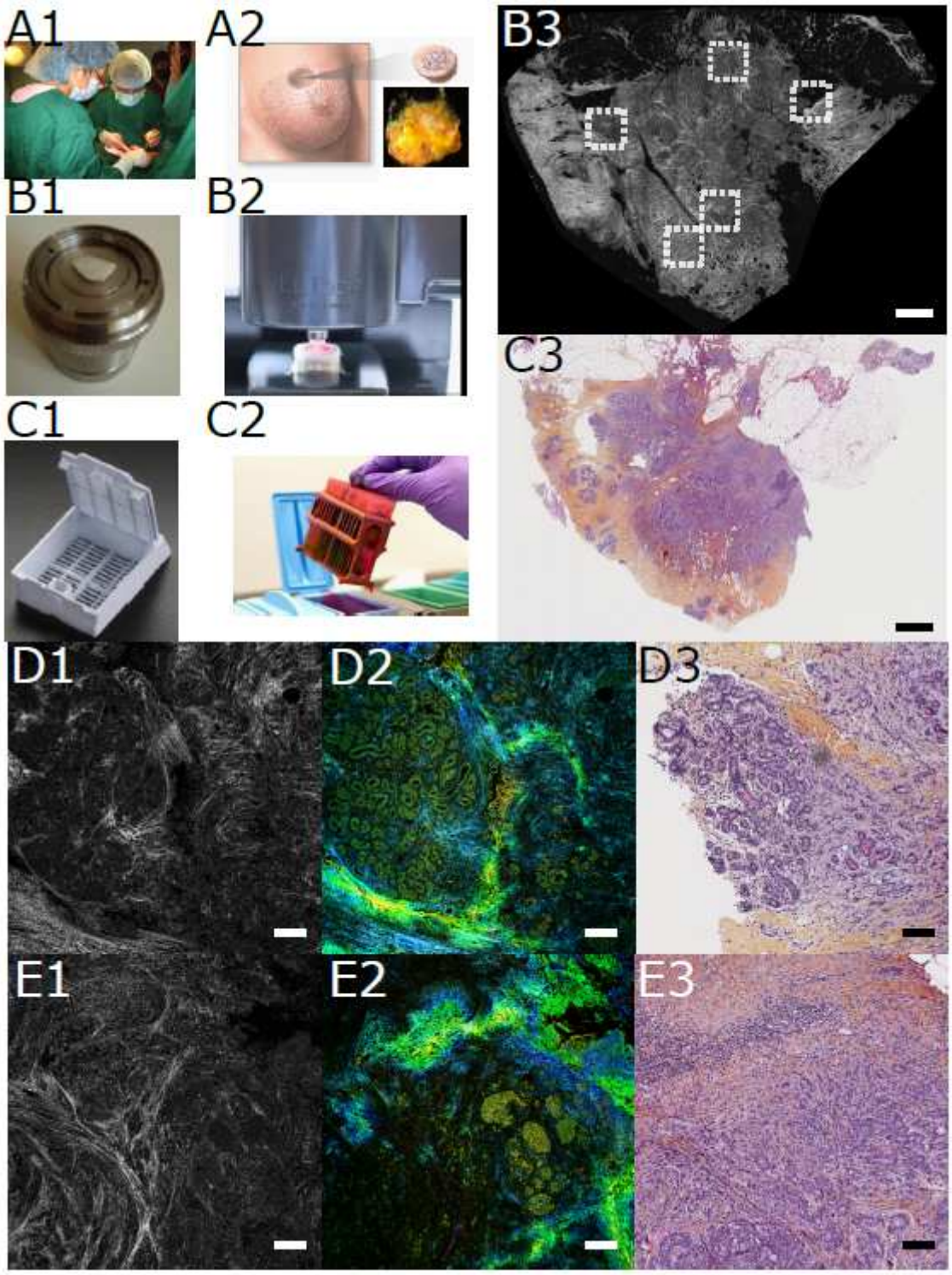

Figure 1

Experimental protocol and correlation between FF-OCT, D-FF-OCT, and histology images. During surgery (Panel A1), breast biopsies (Panel A2) are collected, inserted into the FF-OCT sample holder (Panel B1), and imaged under the FF-OCT microscope (B2). A large field static FF-OCT image of the entire biopsy is acquired (Panel B3). 5 to 20 ROIs (dashed white box) are manually selected and imaged with both static

and dynamic FF-OCT. Then, the biopsy is removed from the sample holder and is inserted into a regular histology cassette (Panel C1), where it is fixed, sliced, and stained (Panel C2), according to H&E histology standards (Panel C3). Examples of 2 ROIs (Panels D and E) imaged from sample diagnosed with invasive ductal carcinoma (IDC) in a healthy and tumoral regions respectively. Static FF-OCT (Panel D1), D-FF-OCT (Panel D2), and H&E (Panel D3) images of a normal lobule were acquired. Similar images (Panels E1, E2, and E3) in an IDC area are displayed. Scale bars are 1 mm (B3, C3) and 0.1 mm (D1-E3) respectively.

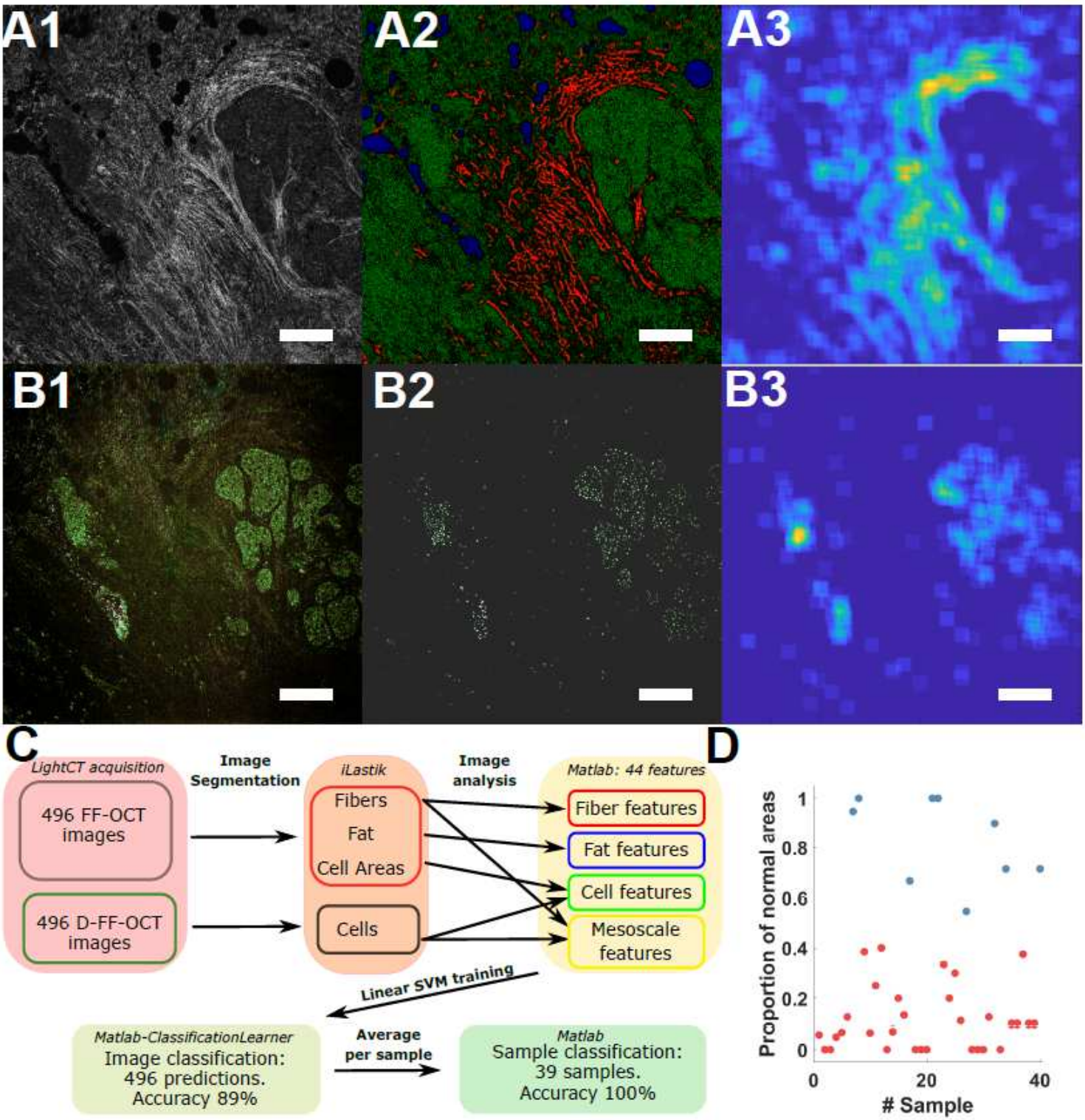

Figure 2

Feature engineering and SVM classification of breast cancer D-FF-OCT images. Static (A1) and dynamic (B1) FF-OCT images of a cancerous breast sample are analyzed using two random forest classifiers (A2,B2). FF-OCT image is segmented (A2) into fibers (red), cells (green), and fat/holes (blue). D-FF-OCT image is segmented (B2) into cells. Mesoscale fiber regions (A3) describe the region of high fiber density in the FF-OCT image. Mesoscale cell regions (B3) describe regions of high cell density in the D-FF-OCT image. The segmented images (A2, B2) and the mesoscale images (A3,B3) are used to calculate engineered features, such as cell and collagen fiber characteristics and size and shape of regions of high cell density. (C) Chart summarizing the processing of FF-OCT and D-FF-OCT images in order to classify each image and each sample using SVM. (D) Proportion of normal areas found for each healthy (blue) and cancerous sample (red) showing 100 % separability between the two classes. Scale bars: 200 µm.

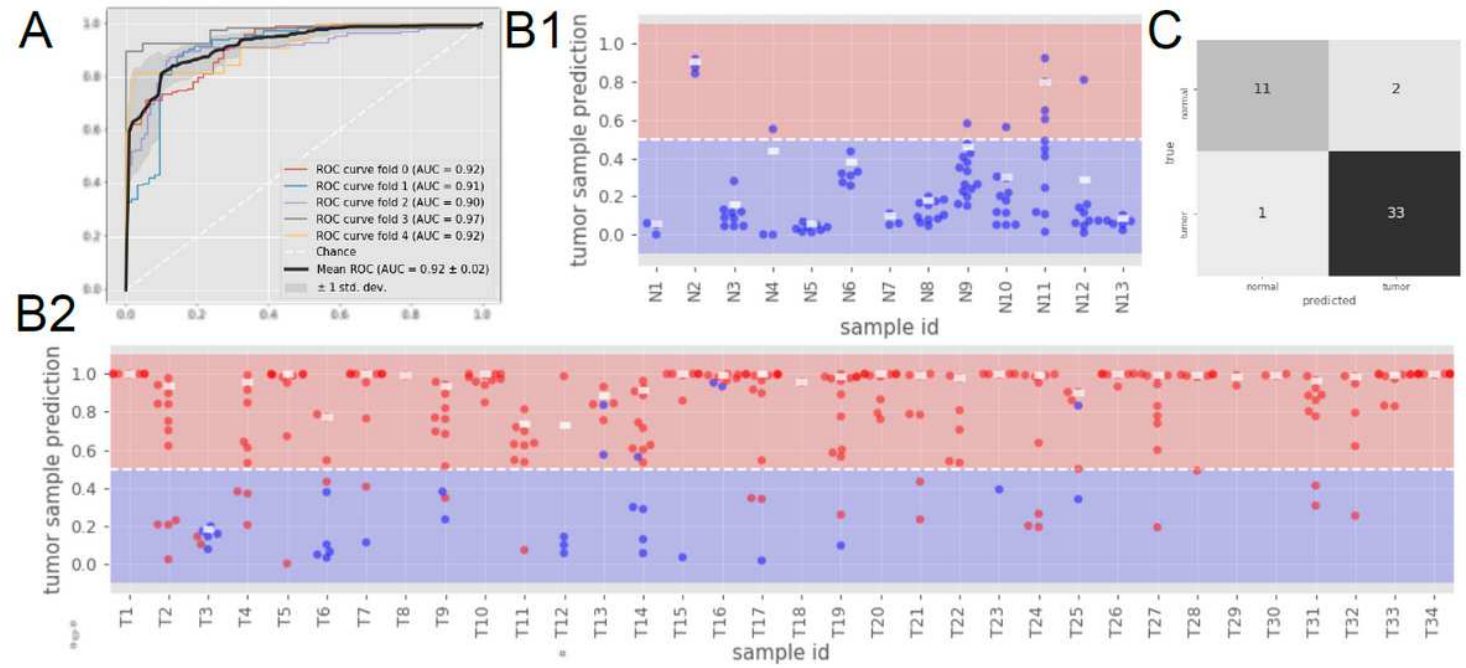

Figure 3

CNN classification of breast cancer (A) ROC curves and AUCs of the 5-fold cross validated models. (B) Tumor probability prediction by the CNN for each ROI in each sample, either healthy (B1) or tumoral (B2). Each point represents a ROI which ground truth diagnosis is either healthy (blue) or tumoral (red). The white rectangle represents the aggregated tumor probability per sample, computed as the 90th percentile of the probabilities of the ROIs in a sample : (B1) normal samples prediction (B2) tumoral samples prediction.

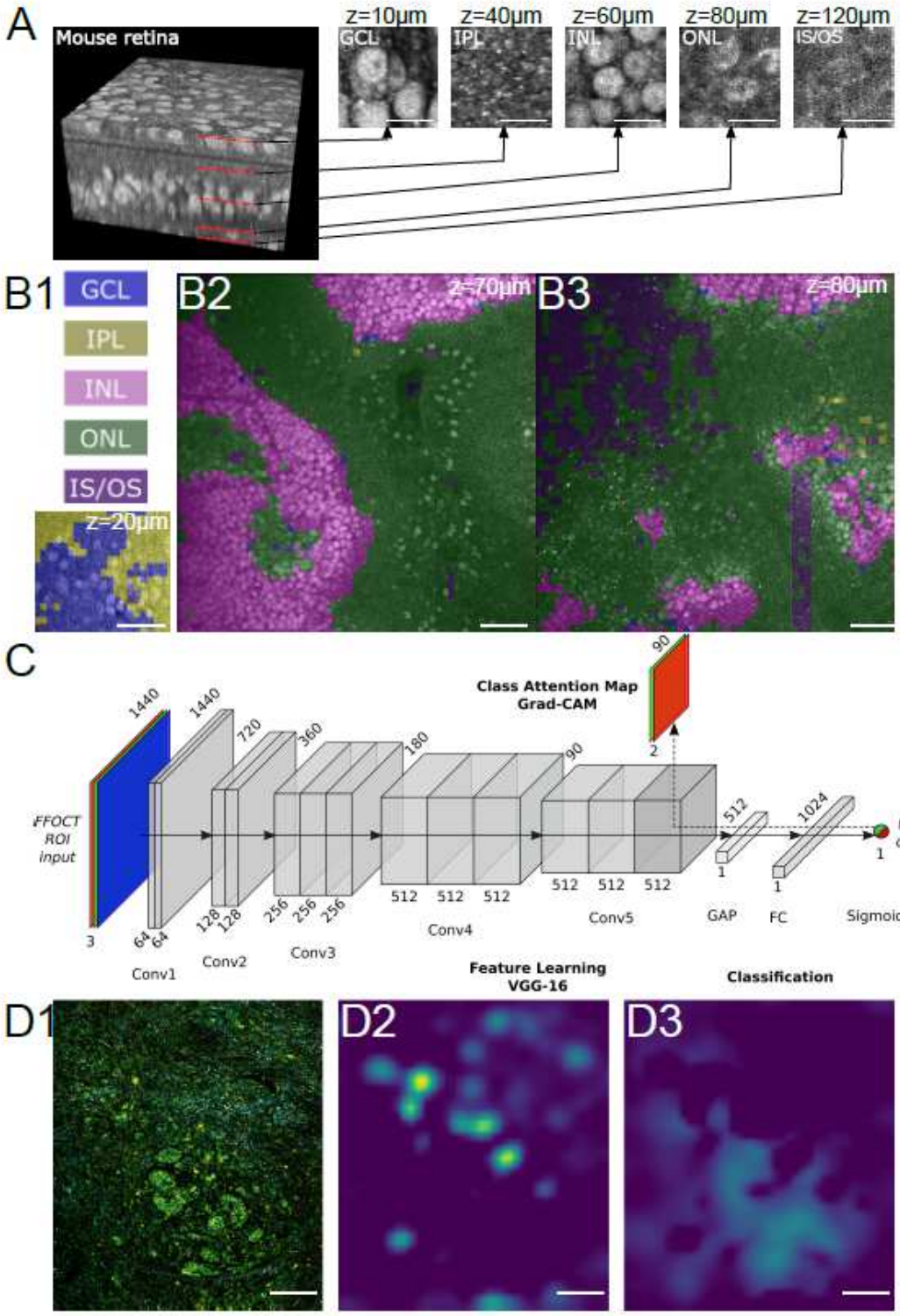

Figure 4

Proposed strategies to reduce the spatial scale of the predictions. A first strategy simply proposes to reduce the size of the starting images, but requires precise labeling at the microscopic scale. We could only achieved this labeling in mice retinas (Panel A), where retinal layers are well organized and show distinct and clearly separable morphological features. Using a modified CNN, we could draw the boundaries between retinal layers in folded retinas where several layers overlap at the same depth

(Panels B1, B2, B3 at different depths from retinal surface - 20, 70, and 80 µm respectively). A second strategy using CNN is to propagate the gradients from the last convolutional layer using the GradCAM algorithm (Panel C) in order to retrieve the spatial information from the CNN. From a ROI showing healthy breast lobule surrounded by isolated infiltrating cancerous cells, correctly predicted as cancerous with 97 % confidence by the CNN (Panel D1), the tumor positive attention map (panel D2) focuses on the regions with the infiltrating cells, while the the tumor negative attention map (panel D3) shows healthy regions, including the healthy breast lobule. Scale bars are 25 µm in A, 50 µm in B and 200 µm in D.

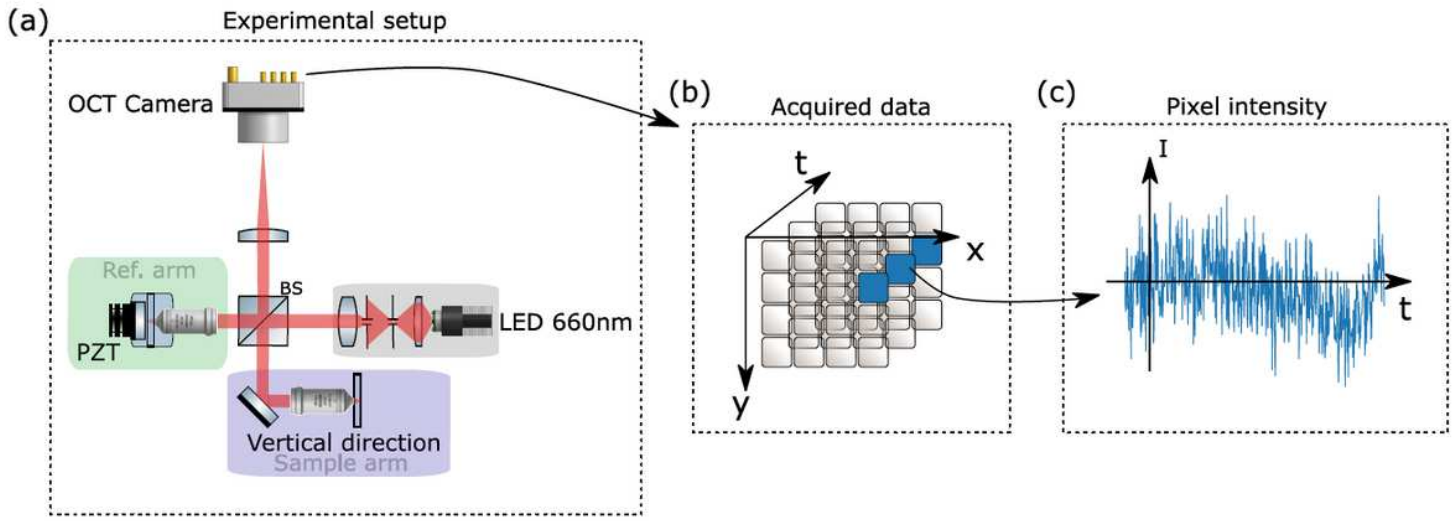

Figure 5

PZT: piezoelectric translation - BS: Beam splitter. Experimental D-FFOCT setup in an inverted configuration optimized for tissue imaging. 512 images are acquired by the CMOS camera (a). The resulting (1440, 1440, 512) 3D tensor (b) is then processed independently for each pixel (c).

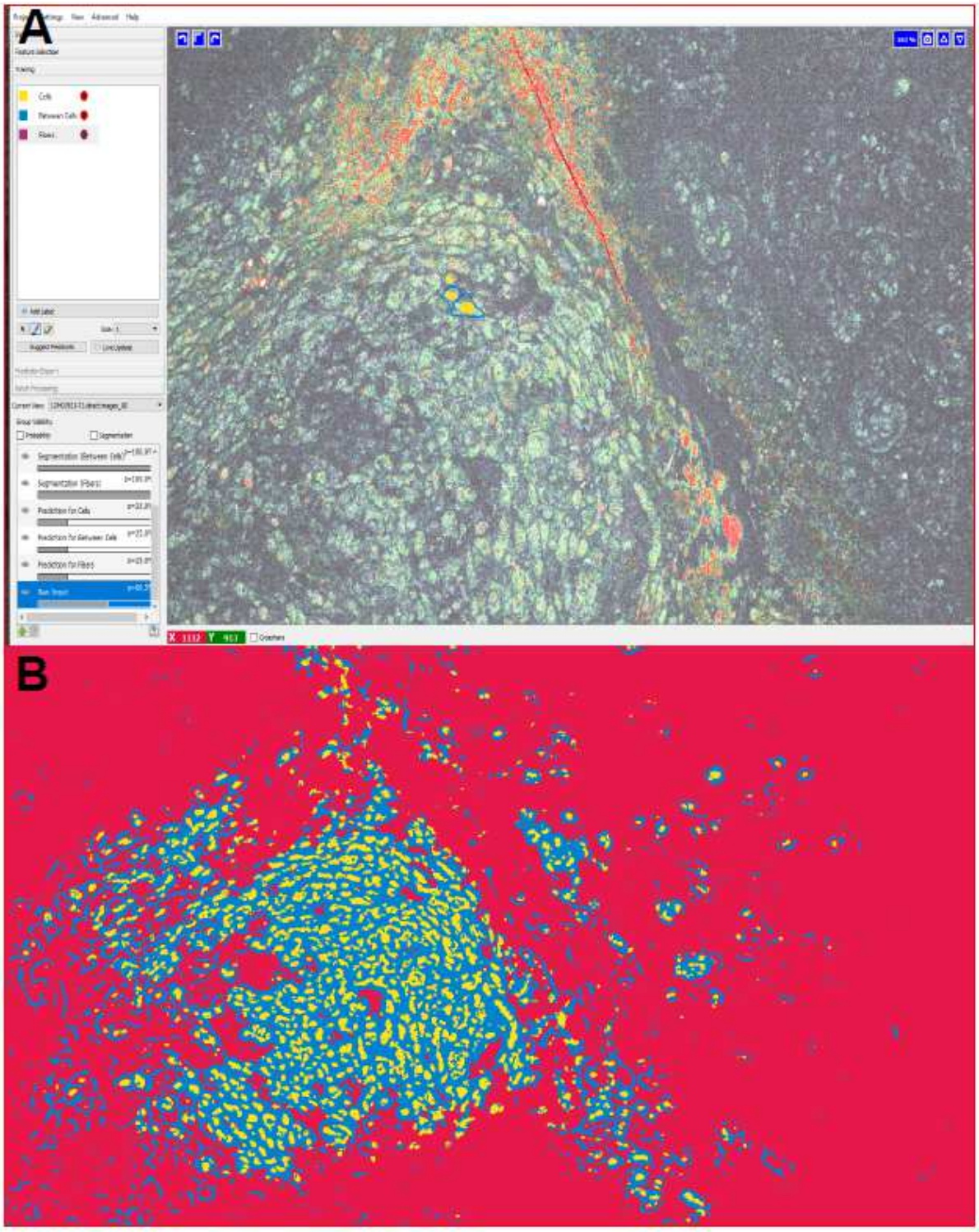

Figure 6

Machine learning based segmentation with iLastik on D-FF-OCT images. Panel A shows the user interface of iLastik, and illustrates the learning procedure. Pixels corresponding to each class of interest are manually drawn (opaque pixels). To start with initial prediction, a small number of pixels in the first image were drawn, and then the prediction is refined step by step by correcting misclassified pixels. Panel B shows the segmentation result after the learning process. For DCI images, only the cell segmentation is

used (class 1 - in yellow here), but the others two classes were used to segment the cells with higher precision.

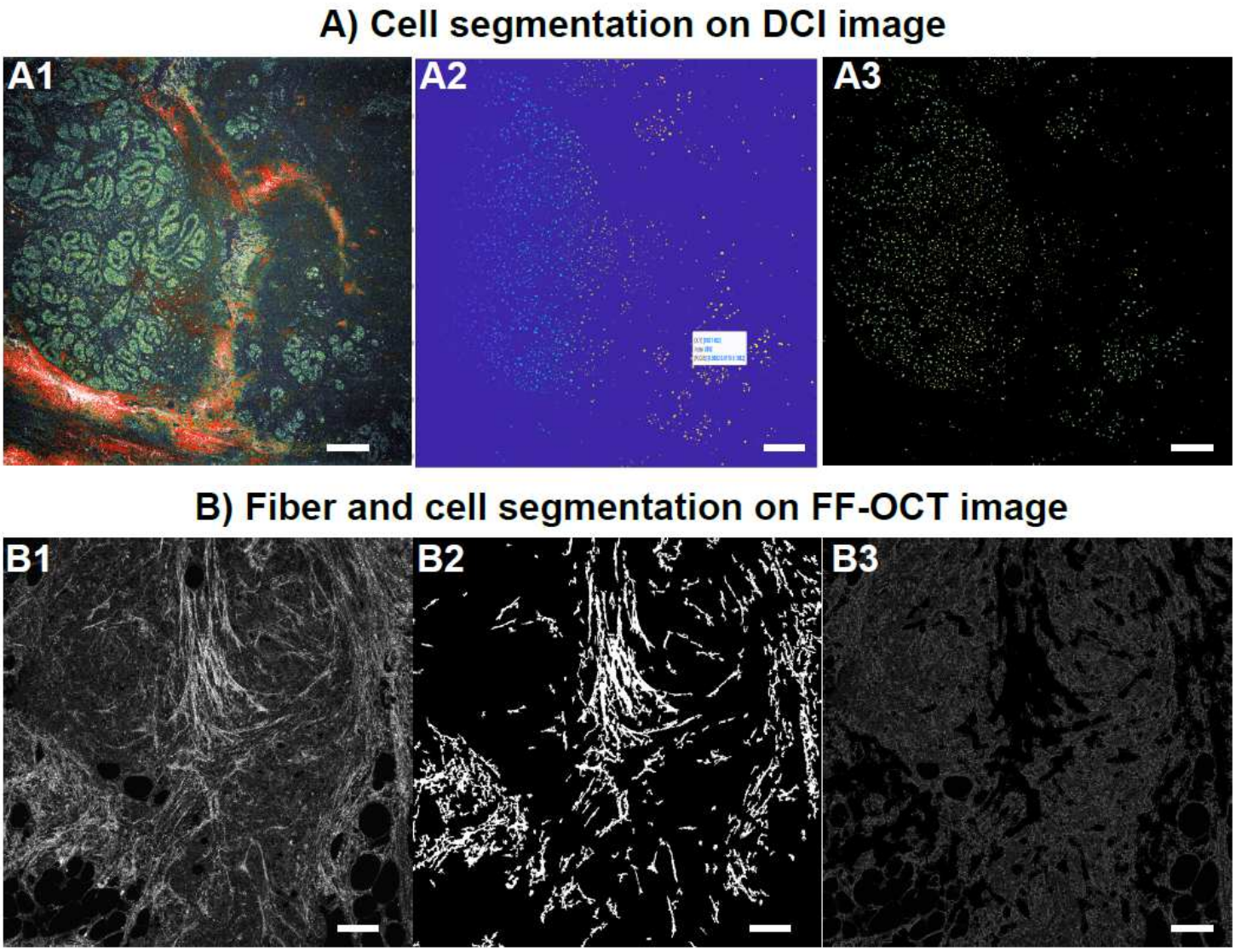

Figure 7

Cell and fiber segmentation output. Panel A shows the cell segmentation using iLastik and Matlab. From original DCI image (A1), after iLastik segmentation, the first class (cells) is selected, and separated in independent regions with Matlab (A2). Here, the increasing color from blue to red gives the number of the identified region (from 1 to 3951 here).Each region of interest (ROI) can be analyzed independently, and only the regions of area above 20 pixels are kept and considered as cells. The obtained mask image is multiplied by the D-FF-OCT image to obtain panel A3. Panel B shows the fiber and cell segmentation results on FF-OCT image. The original image (B1) is processed by iLastik and classified between fibers, pixels between fibers, and cells. The first and third classes are extracted and filtered using Matlab to obtain fibers (B2) and cells (B3)

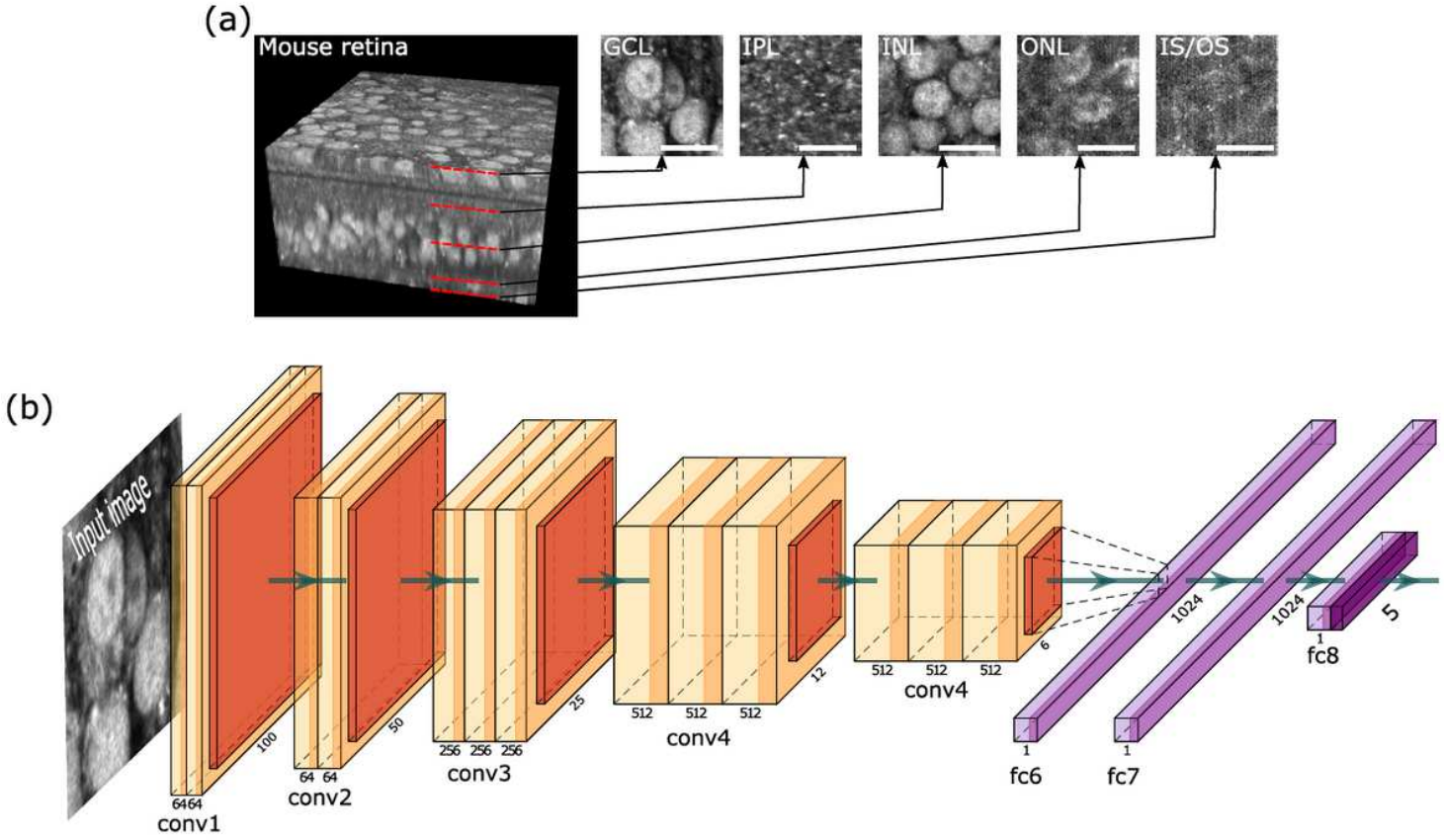

Figure 8

(a) 3D stack of mouse retina on 80 µm depth. 697 thumbnail images of 100 × 100 pixel (corresponding to 27 × 27 µm) were randomly selected and manually classified on a single retina in 5 different classes: GCL, IPL, INL, ONL and IS/OS. (b) Neural network architecture based on VGG16 used for classifying thumbnail images in 5 different classes. Only the lasts fully connected layers (depicted in purple) were trained. Scale bar: (a) 25 µm.