PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Automated surgical margin assessment in breast conserving surgery using SFDI with ensembles of self-confident deep convolutional networks

Pardo, Arturo, Gutiérrez-Gutiérrez, José, Streeter, Samuel, Maloney, Benjamin, López-Higuera, José, et al.

Arturo Pardo, José A. Gutiérrez-Gutiérrez, Samuel S. Streeter, Benjamin W. Maloney, José M. López-Higuera, Brian W. Pogue, Olga M. Conde, "Automated surgical margin assessment in breast conserving surgery using SFDI with ensembles of self-confident deep convolutional networks," Proc. SPIE 11362, Clinical Biophotonics, 113620I (1 April 2020); doi: 10.1117/12.2554965

SPIE.

Event: SPIE Photonics Europe, 2020, Online Only

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Automated surgical margin assessment in breast conserving surgery using SFDI with ensembles of self-confident deep convolutional networks

Arturo Pardo[1,2], José A. Gutiérrez-Gutiérrez[1,2], Samuel S. Streeter[3], Benjamin W. Maloney[3], José M. López-Higuera[1,2,4], Brian W. Pogue [3], and Olga M. Conde[1,2,4,*]

[1] *Photonics Engineering Group (GIF), TEISA department, University of Cantabria. Edificio I+D+i Telecomunicación, Avda. Los Castros S/N, 39005 Santander, Cantabria, Spain*
[2] *Instituto de Investigación Sanitaria Valdecilla (IDIVAL), 39011 Santander, Cantabria, Spain*
[3] *Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755*
[4] *Biomedical Research Networking Center – Bioengineering, Biomaterials, and Nanomedicine (CIBER-BBN),Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0 28029 Madrid*

(*) Corresponding author: olga.conde@unican.es

Abstract: With an adequate tissue dataset, supervised classification of tissue optical properties can be achieved in SFDI images of breast cancer lumpectomies with deep convolutional networks. Nevertheless, the use of a black-box classifier in current ex vivo setups provides output diagnostic images that are inevitably bound to show misclassified areas due to inter- and intra-patient variability that could potentially be misinterpreted in a real clinical setting. This work proposes the use of a novel architecture, the self-introspective classifier, where part of the model is dedicated to estimating its own expected classification error. The model can be used to generate metrics of self-confidence for a given classification problem, which can then be employed to show how much the network is familiar with the new incoming data. A heterogenous ensemble of four deep convolutional models with self-confidence, each sensitive to a different spatial scale of features, is tested on a cohort of 70 specimens, achieving a global leave-one-out cross-validation accuracy of up to 81%, while being able to explain where in the output classification image the system is most confident.

OCIS codes: 100.0100, 200.4260, 170.3880, 170.3010, 110.0113, 200.4700, 200.

1. Introduction

This work is the last of a three-part technical series in classification of SFDI data of breast cancer lumpectomies with neural networks [1, 2]. This particular work focuses on positively improving a surgeon's ability to assess surgical margins in gross pathology images of breast-conserving surgery (BCS). The objective is to reduce current re-excision rates of about 22%–30% [3]. In this contribution, we present how real time automated assessment methods may improve these statistics, by combining the enhanced contrast that can be obtained via Spatial Frequency Domain Imaging (SFDI) with a self-explanatory neural network that avoids false positives or negatives.

Currently, analyzing SFDI data with ensembles of convolutional neural networks can differentiate between various tissue types, namely adipose tissue, connective tissue, benign cysts and malignant tumors, with up to 85% accuracy [1]. However, there are several issues that may impede the use of these systems in a clinical setting. First, the presence of collagen and elastin fibers in both normal and cancer tissues [4] will produce confounding results in a supervised classifier that must provide a binary output representing tissue margins. Second, it is well known that neural networks can respond erroneously when provided data from a previously unseen distribution. This may occur, for instance, when the signal-to-noise ratio of a given image is reduced for a given sample [5], or when shown adversarial examples [6]. Such situations must never happen in real-world scenarios; thus, we require a framework where unfamiliar data does not evoke a response.

We seek to produce a classifying network that is capable of reasoning whether or not the input data belongs to a known or familiar distribution. The problem is divided into three networks: (1) a classifier, which uses supervised data to attempt to provide a diagnosis; (2) an autoencoder, which will observe all the internal states of the classifier network, storing and memorizing specific signatures of the input data; (3) a PDF estimator, which will infer a likelihood metric of the current data for each of the output categories of interest. This method, termed *self-introspection* or

Clinical Biophotonics, edited by Daniel S. Elson, Sylvain Gioux, Brian W. Pogue,
Proc. of SPIE Vol. 11362, 113620I · © 2020 SPIE · CCC code: 0277-786X/20/$21
doi: 10.1117/12.2554965

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

self-reflection [5], allows the end-user of the classification system to include an additional degree of freedom to the classifier's response: a measure of *familiarity* or likelihood that an incoming datapoint belongs to a specific distribution.

2. Materials and methods

We shall briefly describe the methodology, which is extensively explained in previous work [5], indicate the architecture of the networks used for these experiments, as well as the simulations that were prepared. The main objective is to produce a composite system capable of inferring how much new incoming data is similar to the data used during training.

2.1. Self-introspection as a tool for diagnosis

As explained in the Introduction section, a self-introspective model consists of three individual parts:

- **Classifier.** The classifier will be a network $x \overset{f_{\text{cls}}}{\rightarrow} \hat{y}$ that provides an output prediction $\hat{y}$ for a given input x. The prediction $\hat{y}$ should be identical to the actual label, y. In practice, this function is constructed from a series of units or nodes, h, which respond differently to input stimuli x. Thus, we may rewrite this function as $x \rightarrow h \rightarrow \hat{y}$ (Fig. 1.(a)). It is expected that, since the neural network is a deterministic function, it must respond differently to different stimuli, and thus its hidden states must also be different for different input data. In other words, different inputs should evoke different responses within the network.

- **Autoencoder.** An unsupervised model, $h \overset{f_{\text{enc}}}{\rightarrow} z \overset{f_{\text{dec}}}{\rightarrow} \hat{h}$, studies the full range of internal responses within the classifier network and produces a low-dimensional representation of them (Fig. 1.(b)). Functions f_{enc} and f_{dec} represent the encoder and decoder functions, respectively. If output categories are separable, then the internal states of the classifier network will be separable as well. Otherwise, the input data does not provide sufficient information for adequate classification to take place. Misclassification occurs when the internal states of the network do not match with responses learned during training, thus resulting in an abnormal output response (Figure 1.(c)).

- **PDF estimator.** We use the autoencoder's dimensionality reduction capabilities to escape the *curse of dimensionality* and observe the map of all possible classifier responses to data in the two-dimensional space provided by the autoencoder's bottleneck, z. In this space, we can define probability density functions (PDFs) for each of the initial tissue class hypotheses, namely $f_Z(z|H_k)$, where $Z|H_k$ is assumed to be a random variable for a given tissue category hypothesis, H_k. This estimate can be achieved via traditional non-parametric methods, or neural networks.

A general outline of the system is summarized in Figure 1. For a given incoming datapoint x_i, the classifier network provides a category prediction $\hat{y}_i$, consequence of its internal states h_i. These states are observed by the autoencoder, which provides a position in activation space z_i. Finally, the PDF estimator outputs the likelihood that z_i belongs to the distribution of each of the known categories, $f_Z(z_i|H_0), \ldots, f_Z(z_i|H_N)$, with N the number of classes.

Thus, a given output response $\hat{y}_i$ can be accompanied by the likelihood that the classifier network is responding with a familiar signature, $f_Z(z_i|H_0), \ldots, f_Z(z_i|H_N)$. The output response is a vector $\hat{y}_i \in \mathbb{R}^N$, and thus each unit is associated with the familiarity of the network's response to that specific signature. We thus define $c_i = f_Z(z_i|H_i) / \sum_{k=1}^{N} f_Z(z_k|H_k)$ as the *confidence* or *familiarity* metric for a given input datapoint x_i, which will be based on information from z-space. The original sequence h_i is referred to as the *activation sequence* for input x_i. The output response $\hat{y}_i$ can then be weighed by how typical the classifier's response is, $\tilde{y}_i = \hat{y}_i \cdot c_i$. If the network is fed with an abnormal example, its response will be different from the set of well-known activation sequences (learned by the autoencoder) and its representation at the bottleneck will fall in an area outside of the range of known responses, which will result in a low likelihood by the PDF estimator, and thus in an overall inhibition of the network. In summary, the system will only provide a diagnosis if both (1) the input is identified by the classifier as belonging to a specific category and (2) if the classifier's response is similar to responses observed during training.

As a consequence of this process, we expect that the network's response will be modulated by how similar incoming data is to the training set. If data is notably different from the training set, the internal activation sequence inside the classifier will be abnormal, and then the network will be inhibited by the likelihood that the activation belongs to any of the training classes. We test this empirically, by weighing the classic accuracy metric by the aforementioned PDFs, such that we can verify that the network is more proficient when it is familiar with incoming data than when it is not.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Expectedly, if intra-patient variability is greater than the classifier's discriminative capacity, it will be misclassified either way, thus requiring either a different approach, or perhaps more data for training.

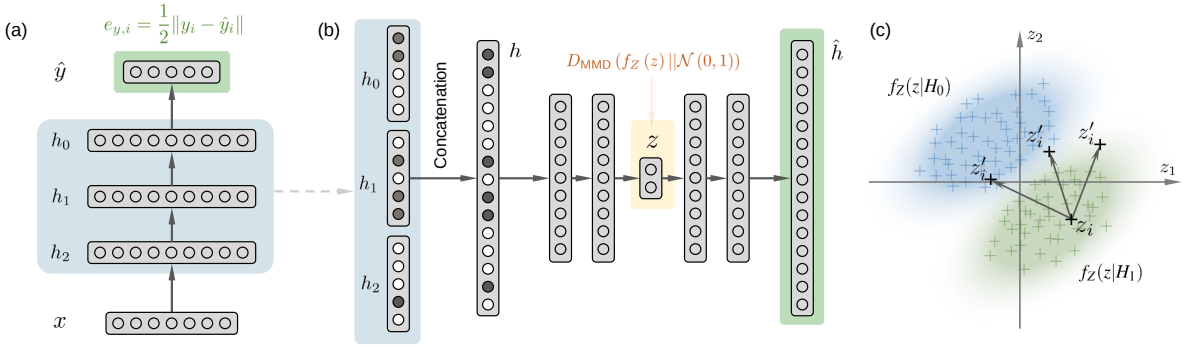

Fig. 1. General outline of the *self-reflection* algorithm. A classifier (a) is trained on a given dataset of input-output pairs (x_i, y_i) with a specific classification error/loss. If well trained, the network should respond differently for different input-output pairs, and thus its hidden units respond with a specific activation sequence h_0, h_1, h_2 that should be different as well. The full range of possible internal states for a training set can be then learned by an autoencoder (b) that allows us to produce a 2D representation of the behavior of the classification network, z (c). This low-dimensional representation also simplifies calculating certain metrics, such as the probability density function of the classifier's hidden activation sequence for a specific type of input data $f(z|H_k)$. In this way, it is possible to detect when the network strays away from its archetypal behavior learned in training.

2.2. Dataset, calibration, and other properties

Imaging system and image acquisition The SFDI instrument is a Perkin-Elmer IVIS SpectrumCT system with a Spatial Frequency Domain Imaging (SFDI) device [7] inside. Field of view is about 10×10 cm, with a spatial resolution of 1024×1024 pixels. Lumpectomy samples are cut into five-millimeter-thick *bread-loafed* slices [8]. One of the cuts is chosen for imaging; the selected section is placed between two optically transparent acrylic plates, held together by elastic bands. Each specimen is imaged at four spatial frequencies (0.0, 0.1488, 0.6053, 1.3736 mm^{-1}) and four wavelengths (490.0, 550.0, 600.0, and 700.0 nm). The sample is then studied by a board-certified breast pathologist via standard H&E staining, providing a high-resolution ground truth diagnostic map. Conservative Regions of Interest (ROIs) are manually generated for the SFDI images, by visually cross-referencing the ground truth H&E stain images with the modulated imaging data. These manually-crafted ROIs only cover areas where only one particular tissue type is present. A complete summary of all tissue categories, samples and ROIs is left in Table 1. Examples for the original ROIs are shown in the Results section.

Initial considerations and category superclasses In a similar way to the previous contribution, we have selected a total of four tissue superclasses (shown in Table 2): (1) Adipose, (2) Connective, (3) Benign, (4) Fibrocystic Disease, and (5) Malignant. As shown in [1], we consider Fibrocystic Disease, despite the fact that typically presents as stromal fibrosis (i.e. formations and/or bundles of connective tissue cells) and benign cyst formation [9], so that this section of the dataset can be used to help in the classification of Connective tissue.

Patch dataset generation Model training was possible via dataset augmentation. A balanced dataset of 40,000 31×31 patches was generated via random population subsampling. Patches were successively extracted from random samples, at random locations, within existing Regions of Interest. The only requirement for this random subsampling protocol is that each tissue superclass must have the same number of patches, thus avoiding over-representation of more frequent classes. This results in a balanced dataset that makes each classifier consider every tissue superclass as equally relevant. About 400-500 patches are extracted per sample.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 1. Number of samples with ROIs of each different tissue types.

Tissue subtype	Samples with tissue subtype	*n*
Adipose Tissue	2, 3, 4, 5, 8, 9, 11, 12, 14, 15, 18, 22, 24, 28, 30, 34, 37, 39, 40, 41, 42, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 58, 59, 60, 63, 64, 67, 69, 70.	39
Connective Tissue	6, 8, 12, 13, 22, 23, 33, 35, 40, 43, 47, 48, 49, 50, 51, 53, 55, 58, 60, 63, 64, 65.	22
Myofibroblastic	57	1
Phyllodes	38	1
Normal Treated	20	1
Fibroadenoma	1, 19, 27, 32	4
Fibrocystic Disease	4, 5, 7,9, 11, 32, 36	7
IDC (Low Grade)	12, 18, 28, 40, 47, 50, 52, 54, 61, 65, 68	11
IDC (Intermediate Grade)	6, 8, 13, 29, 30, 33, 35, 36, 37, 41, 42, 53, 58, 63, 66, 67, 70	17
IDC (High Grade)	5, 15, 23, 25, 26, 39, 46, 56, 59, 64	10
ILC	2, 4, 21, 22, 24, 45, 48, 49, 51, 55	10
DCIS	3, 17, 31, 34, 37, 44, 62	7
Mucinous	10, 14, 43, 60	4
Tubular	11, 69	2
Metaplastic	15	1
Total	**70 samples**	**137 ROIs**

Table 2. Tissue superclasses and total number of ROIs.

Tissue group	Tissue subtypes	*n*
Adipose	Adipose	38
Connective	Connective Tissue	22
Benign	Fibroadenoma, Myofibroblastic, Benign Phyllodes	6
Fibrocystic Disease	Fibrocystic Disease	7
Malignant	IDC (Low Grade), IDC (Interm. Grade), IDC (High Grade), ILC, DCIS, Mucinous, Tubular, Metaplastic	62
Total	**68 samples**	**135 ROIs**

Synthetic miscalibration Height variability may produce two unexpected, minor artifacts in the specimen images. First, the effective frequency that is projected onto local regions of a lump may vary or be shifted due to the pattern projector's throw ratio. Secondly, and for the same reason, these changes in height may vary the effective incident fluence per unit area, thus modifying the effective received backscattered reflectance. To improve generalization robustness and train the network beyond associating certain heights and intensities with specific pathologies, reflectance intensity values and effective spatial frequencies are shifted by $\pm 10\%$ randomly. The same shift is applied to all pixels in an extracted patch.

Measuring confidence At a given location in z-space, confidence in a specific diagnosis could be calculated as the likelihood ratio

$$c_k(z) = \frac{f_Z(z|H_k)}{\sum_{j=0}^{N_{\text{classes}}} f_Z(z|H_j)}, \tag{1}$$

where N_{cls} is the number of classes, and $f_Z(z|H_k)$ represents the probability density function of tissue subclass k. This metric is equal to 1 only when $f_Z(z|H_j) = 0, \forall j \neq k$, and is equal to zero only when $f_Z(z|H_k) = 0$.

2.3. Neural network architectures

Classifier network The selected neural network for all scales was a convolutional modified DenseNet [10]. For this first approach, and with the specified hardware limitations, we have selected a network size and dimensions that can

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

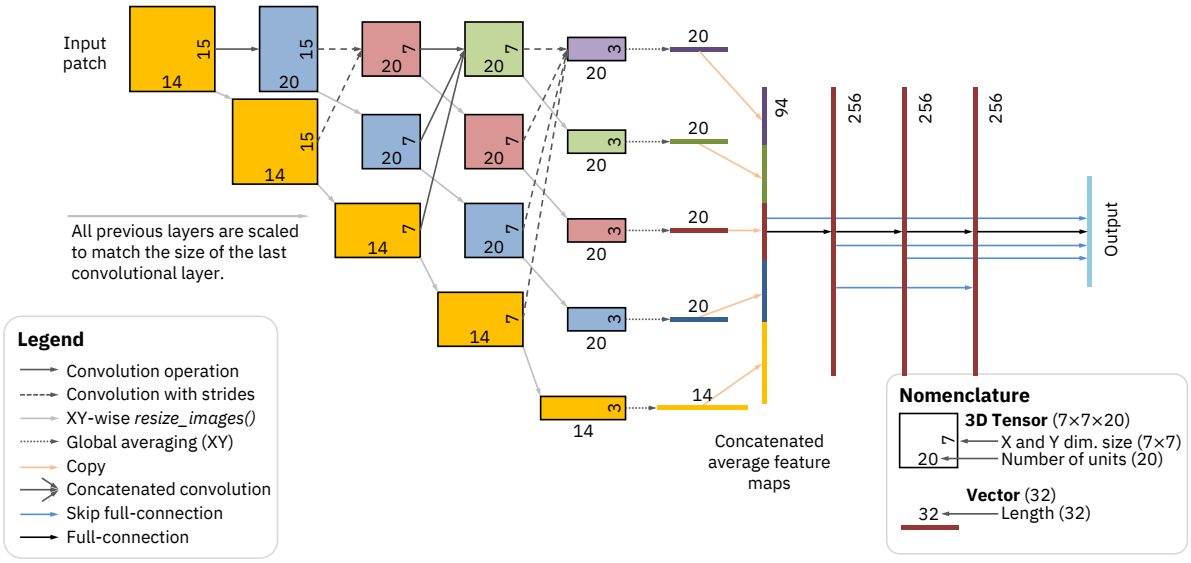

Fig. 2. Lateral view of the convolutional DenseNet classifier structure. 3D tensors are represented sideways, as in U-Net and W-Net descriptions. This is a form of skip-connection deep neural network, where each convolutional layer can observe the activations of all previous layers. Fully connected layers prior to the output layer are all connected to the output, as well as to all previous fully connected layers.

train in less than 30 minutes on an *nVidia RTX 2080 Super*, obtaining a typical VRAM consumption of 4-5GB and GPU utilization circa 75-80% at worst. An example schematic and table for this neural architecture is shown in Figure 2. The complete tabular description for the actual network is provided in Table 3. All hidden layers are hyperbolic tangent (*tanh*) layers, since their dynamic range is bounded to the interval $[-1, 1]$.

Autoencoder network The autoencoder is a skip-connection multi-layer perceptron InfoVAE [11, 12]. The encoder and decoder networks are multi-layer perceptrons composed of 4×500 *tanh* layers each. Hidden layers have skip-connections within the encoder and decoder. The encoder observes the last fully connected layers of the convolutional network, namely the last 4×400 skip-connection *tanh* layers. The bottleneck of the autoencoder is a two-dimensional vector. No skip connections can traverse the bottleneck layer, ensuring that no information reaches the decoder via any other path than z.

PDF estimation PDF estimation in 2D space is achieved via a 6×300 skip-connection multilayer perceptron network, by means of a well-known procedure for calculating a PDF within a bounded domain [13].

Compute infrastructure and timing Two computers are used to train separate 10-fold cross-validation simulations: (1) an Intel Core i9-9700K CPU, with 64 GB of RAM and a dedicated SSD swap space of 100 GB, and an nVidia RTX 2080Ti GPU; (2) an AMD Ryzen 5 3600, 32 GB of RAM, 100 GB dedicated SSD swap space, and an nVidia RTX 2080 Super GPU; (3) a series of Docker deployments operated remotely via a JupyterLab interface, namely consisting of a total of eight nVidia 1080Ti and three Tesla V100 GPUs (11 and 32 GB of VRAM, respectively), up to 384 GB of RAM, and a 64-core Intel Xeon Gold 6230 CPU. Classifiers are trained for about 30 minutes while autoencoders require more time, i.e. about an hour and a half. Each PDF estimator network takes about 20 additional minutes to train.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 3. Individual classifier structures. As an illustrative reference, the first network is the one depicted in Figure 2. The actual network used is the one in the second column. Each layer includes the dimensions of its respective input and output tensors. n_{patch} indicates patch width ($= 3, 11, 21, 31$) and $n_{\text{cat}} = 5$ is the number of output categories.

(a) Example from Fig. 2	Tensor shapes (I/O)	**(b) Network used**	Tensor shapes (I/O)
$15 \times 15 \times 14$ Input	$(? \times 15 \times 15 \times 14)$ in	$n_{\text{patch}} \times n_{\text{patch}} \times 16$ Input	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 16)$ in
1×20 tanh, f.s. 3×3	$(? \times 15 \times 15 \times 14)$ in $(? \times 15 \times 15 \times 24)$ out	1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 16)$ in $(? \times n_{\text{patch}} \times n_{\text{patch}} \times 166)$ out
1×20 tanh, f.s. 3×3, stride $= 2$	$(? \times 15 \times 15 \times 24)$ in $(? \times 7 \times 7 \times 54)$ out	1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 166)$ in $(? \times n_{\text{patch}} \times n_{\text{patch}} \times 316)$ out
1×20 tanh, f.s. 3×3	$(? \times 7 \times 7 \times 54)$ in $(? \times 7 \times 7 \times 74)$ out	1×150 tanh, f.s. 3×3, stride $= 2$	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 316)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 466)$ out
1×20 tanh, f.s. 3×3, stride $= 2$	$(? \times 7 \times 7 \times 74)$ in $(? \times 3 \times 3 \times 94)$ out	1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 466)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 616)$ out
Global avg.	$(? \times 3 \times 3 \times 94)$ in $(? \times 94)$ features	1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 616)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 766)$ out
3×256 skip-connection tanh	$(? \times 256)$ per layer	1×150 tanh, f.s. 3×3, stride $= 2$	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 766)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 916)$ out
n_{cat} Sigmoid output	$(? \times n_{\text{cat}})$ Output	1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 916)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 1066)$ out
		1×150 tanh, f.s. 3×3	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 1066)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 1216)$ out
		1×150 tanh, f.s. 3×3, stride $= 2$	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 1216)$ in $(? \times n_{\text{patch}}/8 \times n_{\text{patch}}/8 \times 1366)$ out
		Global avg.	$(? \times n_{\text{patch}}/8 \times n_{\text{patch}}/8 \times 1366)$ in $(? \times 1366)$ features
		4×400 skip-connection tanh	$(? \times 400)$ per layer
		n_{cat} Sigmoid output	$(? \times n_{\text{cat}})$ out neurons

Notation: f.s. stands for *filter size* (per unit). Layer notation is $c \times n$, $c :=$ number of concatenated layers, $n :=$ number of hidden neurons in each layer. Finally, *stride* indicates stride steps, if there are any. Tensor shape notation as in TensorFlow: `(batch, height, width, channels)`. The token '`?`' denotes an unknown input batch size.

2.4. Training simulations

Models were tested via leave-one-out cross-validation. No early stopping method was used. Validation errors were monitored and verified *a posteriori* to ensure that the networks under evaluation were not overfitting and thus returning suboptimal classification results.

Background segmentation We consider background segmentation to be a task that is secondary with respect to tissue diagnosis. To this end, a single neural network with a (21×21) input receptive field and the structure presented in Table 3.(b) is used for classifying all images. No cross-validation is employed for this specific instance, and thus the same network is used for all samples. This allows most compute time to focus on tissue classification alone.

Classifier schedule Classifier networks were trained with Adam, for a total of 60,000 iterations, minibatch size 64 (circa 60 epochs), and an initial learning rate of $l_r = 0.0001$. Dropout was set to $p = 0.1$. Accuracy was registered at the end of each epoch. Cost function was set to the standard *categorical cross-entropy loss*. Gradient values were clipped by norm to one.

Autoencoder schedule AEs were trained via Cyclic Learning Rates (CLR) [14] combined with AdInitial learning rate was se to $l_r = 0.0001$. Cycle period was $T = 10,000$ iterations. Dropout was $p = 0.1$. Networks were run for a total of $N = 25$ cycles each. Minibatch size was 64. Cost function was given by the network's structure: Maxmimum

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Mean Discrepancy at the bottleneck [11] and relative Mean Squared Error at the reconstruction, i.e.

$$\mathscr{L}_{\text{rMSE}}\left(h,\hat{h}\right)=\frac{1}{2n}\frac{\|h-\hat{h}\|^2}{\|h\|^2+\varepsilon} \tag{2}$$

with $\varepsilon=0.1$ as a regularization parameter that avoids NaN resulting from the term $1/\|h\|^2$. Gradient clipping was set to have a maximum L2-norm of 1.0. Concatenated classifier activation sequences are column vectors, i.e. $h\in\mathbb{R}^n$.

PDF estimator schedule The PDF estimator were trained for 100,000 iterations each, with a minibatch size of 128. Learning rate was fixed to $l_r=0.00001$. Adam was used as optimizer. A total of 5 estimators per classifier/autoencoder pair were trained, one for each output category.

2.5. Ensemble learning

Four different input receptive field sizes are used in patch classification: (3×3), (11×11), (21×21), and (31×31)-pixel windows are used. Dataset patch dimensions are adjusted by clipping the edges of the original 31×31-pixel patches while keeping the center of the patch at the center of the input receptive field. The total number of color channels is 16, corresponding to the 4 demodulated spatial frequencies and 4 wavelengths per spatial frequency in every patch.

For every sample and receptive field size, a classifier and autoencoder are trained with the specified protocol and all training data except those coming from the current sample under evaluation. The networks are stored and then used to classify the aforementioned image. A confusion matrix is created for each input receptive field size. The final ensemble average is also evaluated with a last confusion matrix and corresponding statistics [1].

Output layer reassignment If one of the categories cannot be adequately classified, the softmax output layer $\hat{y}\in\mathbb{R}^n$ can then be globally edited by supressing or disabling the affected output unit (that is, multiplying the k-th unit by zero, $\hat{y}_k=0$) and then dividing by the norm again, so that the output vector can still represent exclusive probabilities. This allows the network to learn from data that belong to difficult categories and would be discarded otherwise, so that at least it can try to learn that such data does not belong to any of the other categories.

2.6. Weighted accuracy

The standard procedure for classifier characterization involves generating a confusion matrix where the predicted and actual categories are compared, and well-known metrics (sensitivity, specificity, accuracy) are calculated from it. To verify if likelihoods can directly improve performance, we may evaluate the results in two different ways. First, we verify that the networks behave correctly initially and therefore have been adequately trained. Second, we multiply the output maps from the classifier networks by the output confidence maps provided by the autoencoder and PDF estimators, and then choose the category that scores highest. A network ensemble can be thus weighted by confidence, by calculating the final output diagnosis as $\hat{y}_{\text{en}}=\frac{1}{N}\left(\alpha_1\hat{y}_1+\cdots+\alpha_N\hat{y}_N\right)$, where N is the size of the ensemble and $\alpha_1,\ldots,\alpha_N$ are the respective confidences of each of the members of the ensemble. This is expected to either benefit performance, or keep it constant.

3. Results

The following sections study the potential advantages of using self-reflecting (or self-introspective) models instead of standard classifiers. We shall present the properties of activation maps, compare the responses of a softmax output versus a confidence-weighted output, and provide some proof of the method's influence on overall accuracy.

3.1. Activation maps

Figure 3 depicts a total of three samples in leave-one-out cross-validation for each of the four patch sizes under analysis. In this figure, rows represent patch sizes, while columns show the same sample under test for each patch size. Each subplot displays the 2D map (or constellation) of all possible internal activation sequences in the classifier, as studied by the autoencoder. Each of the networks produces a different response constellation due to the low number of samples ($n=70$). In each subplot of Fig. 3, three different things are represented: (1) training set activations as dots, (2) test set activations as crosses, and (3) the probability density functions for each tissue subtype estimated via KDE and the training set.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

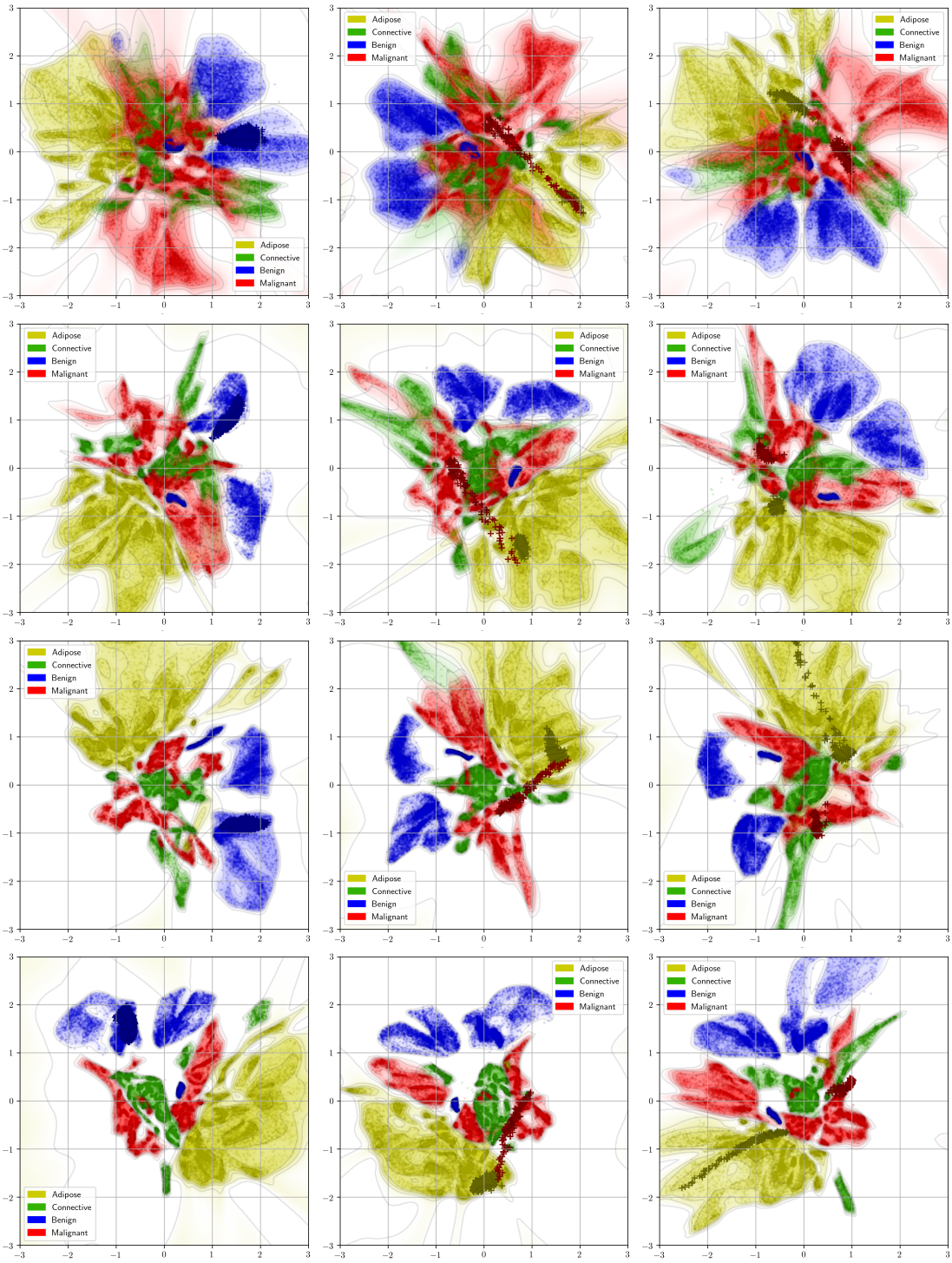

Fig. 3. Classifier response constellations for Samples 1, 15 and 16 (columns from left to right, respectively) and for each of the four patch sizes under study (top to bottom, 3, 11, 21, and 31, respectively). Points represent activations elicited by the training set, while crosses indicate the network's responses to a fraction of the sample under test's ROIs. The PDF estimates are superimposed on top, in log scale, normalized by their maximum value. Best viewed in color.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

A series of initial conclusions may be made with this figure. First, it is possible to observe how textural information is crucial in order to separate between malignant and connective tissue. In other words, microstructural properties and its influence in the resulting spatial frequency reflectance data may not be sufficient in separating connective tissue and malignant tissue subtypes, at least for this dataset. This may be due, among other things, to the presence of elastin and collagen in the latter. Consequently, textural information is fundamental when classifying benign, striated structures, such as those presenting in connective tissue. These results are comparable to those obtained via more traditional machine learning methods [8]. Given the selected network structure, it is also possible that larger input sizes allow the network to better exploit its available capacity to detect textural information in its receptive field.

Secondly, networks with larger receptive fields appear to be more robust to incoming data. Robustness can be intuitively defined by comparing the spread of the test set constellation against the training set constellation. If both of them spread similarly, then the network is similarly robust to training and test data, suggesting good generalization. If test set activations stray away from training set activations, they end up in areas with low likelihood $f(z|H_k)$ for their true label. The cause of this difference in robustness may be suboptimal training (i.e. overfitting), as well as inter- and intra-patient variability, among other factors. Due to the low number of samples, we are inclined to consider that it is much more likely that the latter has a stronger influence on the network's responses than the former.

Further research is needed to properly assess these constellations; for now, we will evaluate how much the test set constellation relates to the training set constellation via an estimate of the PDF of the training set constellation for each of the labels of interest. In this way, activations that stray away from known behavior will score a lower likelihood than those resting within the boundaries of the training set constellation. However, if data elicits such a singular response in the classifier that an activation sequence lands in a section of the constellation with high likelihood for a different tissue subtype, this method will not be able to detect the error. Such a result is analogous to standard digital communications systems (i.e 64-QAM constellation maps).

3.2. Confidence-weighted diagnostic maps

As is observed in the previous section, some of the incoming samples elicit responses that are too different from those observed during training. These events can be overlaid on top of the image in the form of confidence-weighted diagnostic maps. In image segmentation, it is typical to show binary classification maps as well as the raw response of the classifier in all the positions within the image. Figure 4 shows the behavior of various receptive field sizes to Sample 1, a benign Fibroadenoma, as well as where the data under test falls in the activation map (middle plots for all rows). Figure 5 shows unweighted and weighted ensemble diagnostic maps for three samples, as well as their original ROIs.

3.3. Accuracy

The initial results of leave-one-out cross-validation are provided in Figure 6. As was observed in previous work [1], Fibrocystic Disease (FD) cannot be accurately diagnosed with this training set, imaging setup and classification framework. However, by training the network with FD and then suppressing the output neuron corresponding to this diagnosis, we can obtain an overall accuracy of over 82% in all four major categories (Adipose, Connective, Benign, Malignant). Classifier sensitivity to connective tissue is still a challenge (currently at 52% sensitivity and 95% specificity), suggesting that the presence of connective-like structures that are detected in multiphoton microscopy [4] can also be observed in macroscopic SFDI imaging.

The effects of confidence weighing are also provided. There is approximately $\leq 1\%$ increase in global accuracy for all receptive field sizes by using confidence weighing. One-vs-others statistics for all four main pathology groups are provided in Figure 7. Thus, overall accuracy is not hindered by the use of self-reflective networks.

4. Summary

This work showcases a first step towards explainable artificial intelligence for margin assessment in SFDI images of breast cancer lumpectomies. The models are capable of correctly diagnosing up to 80% of the dataset ROI data in leave-one-out cross-validation, despite potential overfitting and PDF estimate errors. The use of weighted confidence provides diagnostic maps that are inhibited in regions corresponding to abnormal internal behavior in the classifier.

However, there are many challenges and limitations that must be systematically solved before clinical applicability can be considered. For instance, inter-patient variability seems to be particularly damaging in the final accuracy benchmarks. Some samples still remain in the sub-50% accuracy range, and cannot be correctly classified by any supervised means. Other samples score low in confidence and thus result in no image contrast. This implies that there is still

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

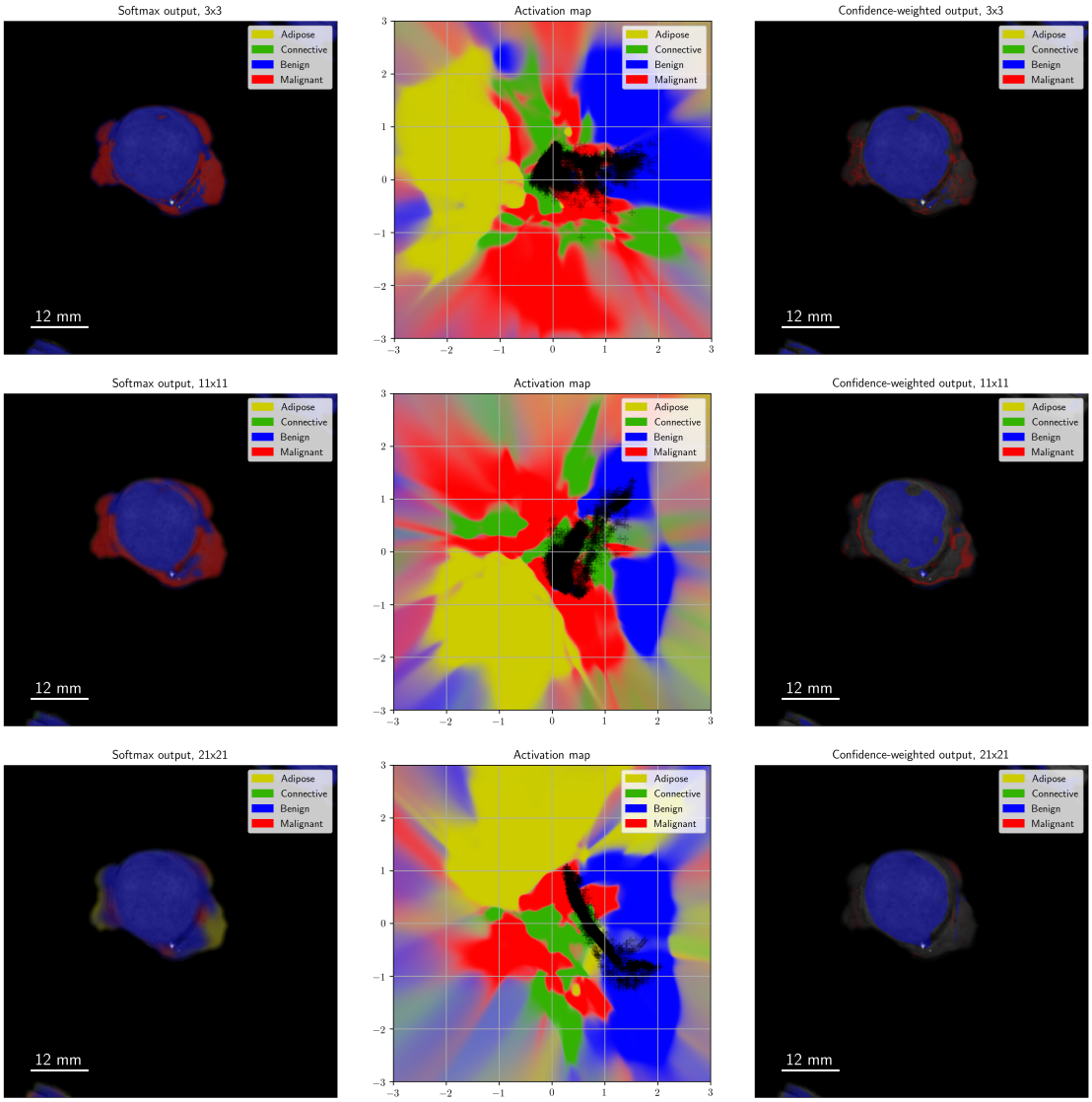

Fig. 4. Network responses to Sample 1 at three different receptive field sizes (rows correspond to 3×3, 1×11, and 21×21 receptive fields, respectively). At each row: original diagnostic map (left image); corresponding internal activation maps (in color) with 1/50 of specimen spectra plotted as black crosses; confidence-weighted output (right image). Tissue flaps surrounding this benign Fibroadenoma are not as strongly classified as malignant due to height variations after confidence-weighing is applied.

much to learn about inter- and intra-specimen variability, suggesting the need for further data acquisition campaigns. Additionally, employing conservative ROIs pushes the networks to operate in a limited domain, where only a fraction of the available data is used. Further research shall focus in the use of unsupervised methods that parse the complete dataset and may provide human-readable, automated labeling of complete datasets, ultimately finding which specific pathologies can be distinguishable from one another.

Acknowledgments The authors would like to thank Lara Lloret from the Advanced Computing Group at Instituto de Física de Cantabria (IFCA) for their technical support and assistance in using the *DEEP Hybrid DataCloud* infrastructure.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

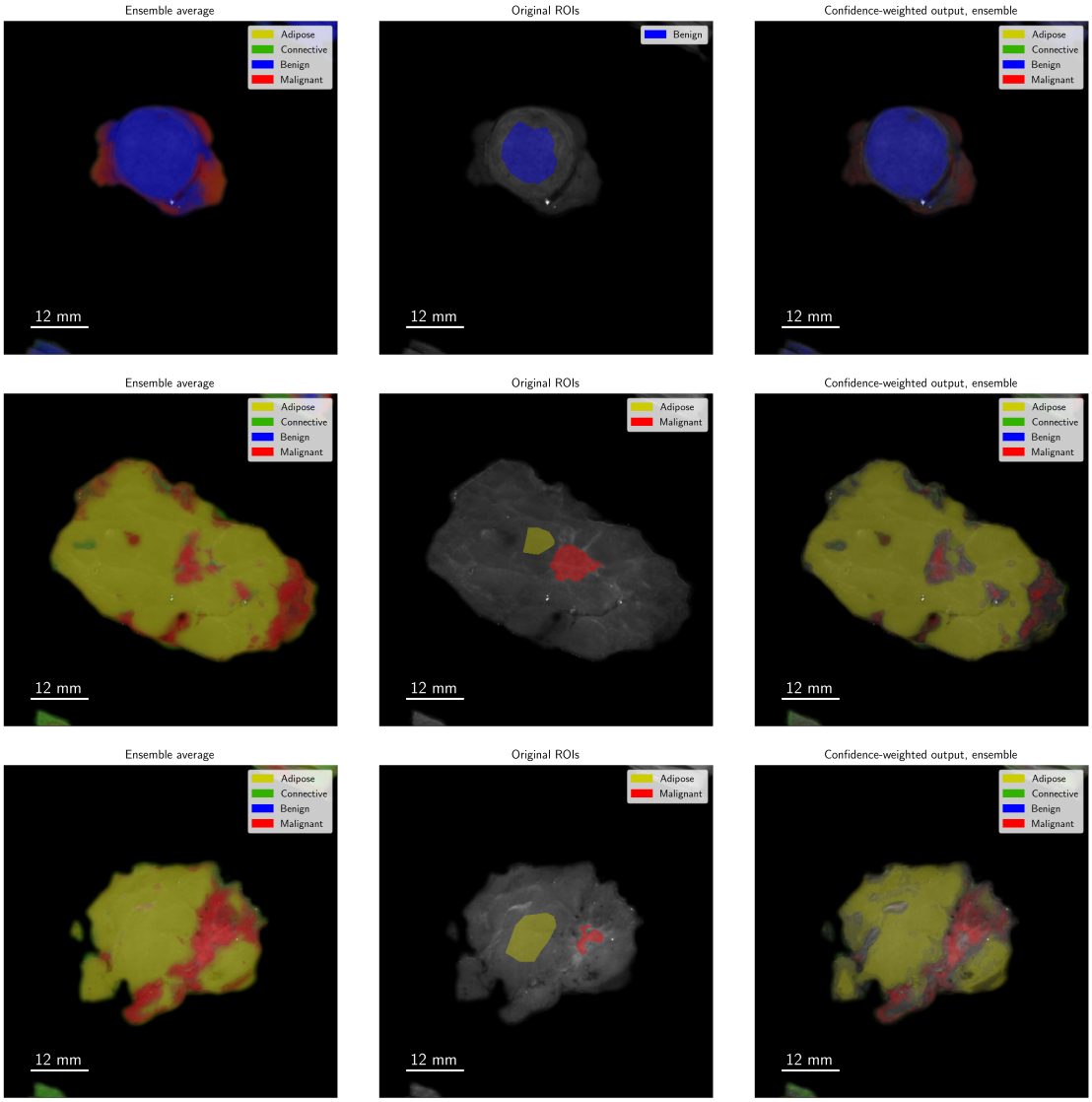

Fig. 5. Ensemble average (left column) and original Regions of Interest (center column) of Samples 1, 16 and 17 (as rows). Confidence-weighted ensemble averaging (right column) attempts to avoid highlighting areas that elicit unprecedented responses in the classifier.

Funding Spanish Ministry of Science, Innovation and Universities (FIS2010-19860, TEC2016-76021-C2-2-R), Spanish Ministry of Economy, Industry and Competitiveness and Instituto de Salud Carlos III (DTS17-00055, DTS15-00238), Instituto de Investigación Valdecilla (INNVAL16/02, INNVAL18/23), Spanish Ministry of Education, Culture, and Sports (FPU16/05705).

References

1. A. Pardo, S. S. Streeter, B. W. Maloney, J. M. López-Higuera, B. W. Pogue, and O. M. Conde. Scatter signatures in SFDI data enable breast surgical margin delineation via ensemble learning. In Adam Wax and Vadim Backman, editors, *Biomedical Applications of Light Scattering X*, volume 11253, pages 1 – 10. International Society for Optics and Photonics, SPIE, 2020.
2. A. Pardo, J. A. Gutiérrez-Gutiérrez, J. M. López-Higuera, B. W. Pogue, and O. M. Conde. Affinity-based color enhancement methods for contrast enhancement in hyperspectral and multimodal imaging. In Sylvain

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

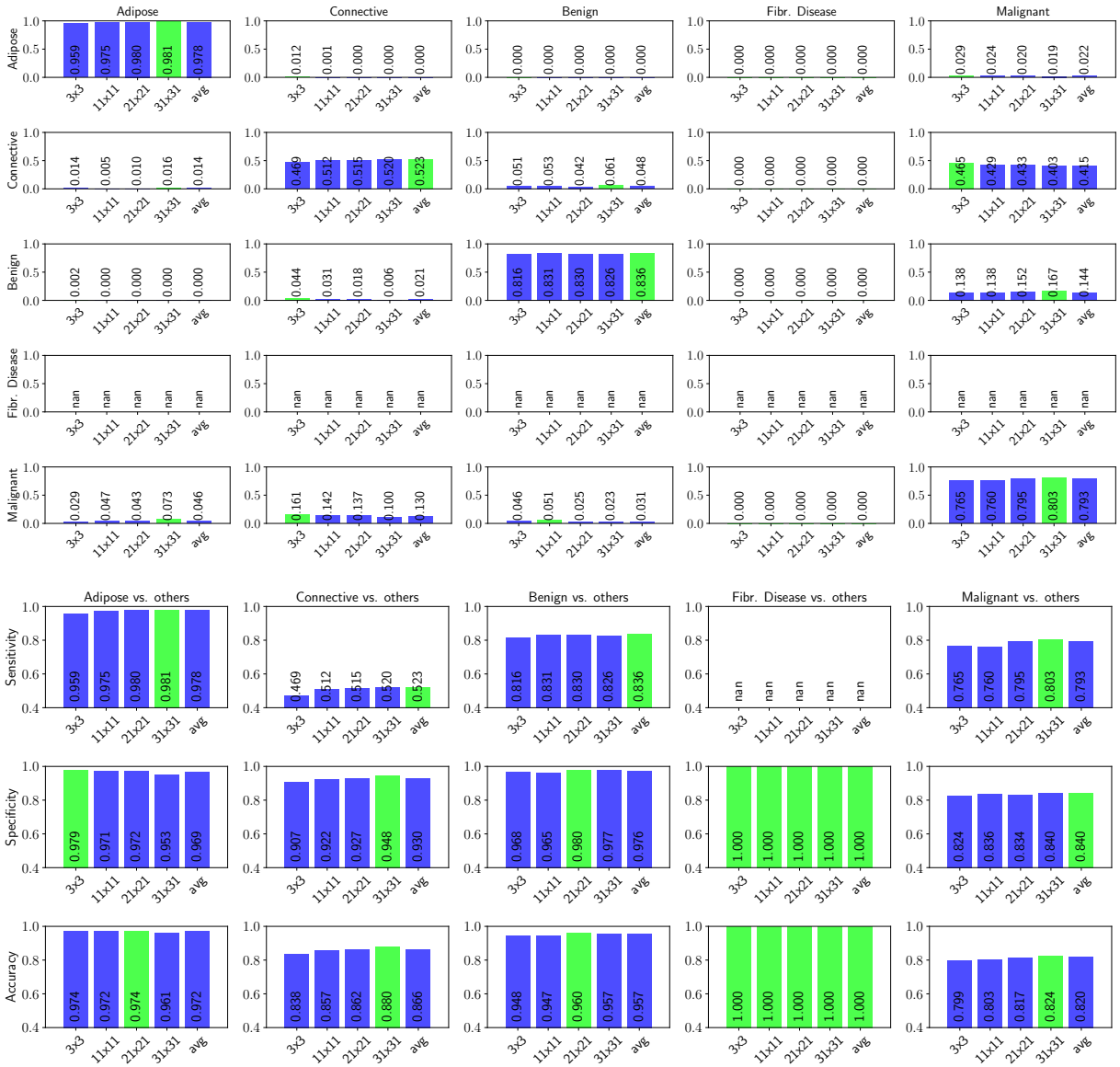

Fig. 6. Confusion matrix (top subplot) and one-vs-others analysis (bottom subplot) for the network ensemble with the Fibrocystic Disease output neuron disabled. FD samples are discarded here and in all accuracy calculations.

Gioux, Summer L. Gibbs, and Brian W. Pogue, editors, *Molecular-Guided Surgery: Molecules, Devices, and Applications VI*, volume 11222, pages 7 – 11. International Society for Optics and Photonics, SPIE, 2020.

3. K. Kaczmarski, P Wang, R Gilmore, H. N. Overton, D. M. Euhus, L. K. Jacobs, M. Habibi, M. Camp, M. J. Weiss, and M. A. Makary. Surgeron Re-Excision Rates after Breast-Conserving Surgery: A Measure of Low-Value Care. *J. Am. Coll. Surg.*, 228(4):504–512, 2019.
4. S. You, Y. Sun, L. Yang, J. park, H. Tu, M. Marjanovic, S. Sinha, and S. A. Boppart. Real-time intraoperative diagnosis by neural network driven multiphoton virtual histology. *Nature Precision Oncology*, 3(33), 2019.
5. A. Pardo, J. A. Gutiérrez-Gutiérrez, J. M. López-Higuera, B. W. Pogue, and O. M. Conde. Coloring the black box: Visualizing neural network behavior with a self-introspective model, 2019.
6. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

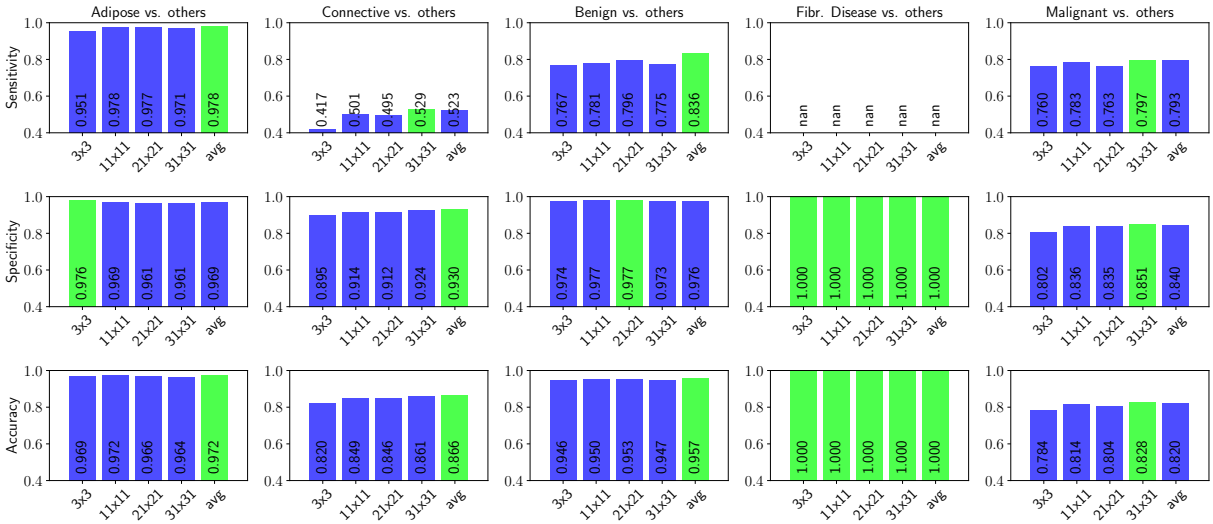

Fig. 7. One-vs-others statistics of confidence-weighted ensemble averaging. FD cases are discarded and the corresponding output neuron remains disabled.

7. D. M. McClatchy, E. J. Rizzo, J. Meganck, J. Kempner, J. Vicory, W. A. Wells, K. D. Paulsen, and B. W. Pogue. Calibration and analysis of a multimodal micro-CT and structured imaging system for the evaluation of excised breast tissue. *Phys. Med. Biol.*, 62(23):8983–9000, 2017.
8. Samuel S. Streeter, Benjamin W. Maloney, David M. McClatchy, Michael Jermyn, Brian W. Pogue, Elizabeth J. Rizzo, Wendy A. Wells, and Keith D. Paulsen. Structured light imaging for breast-conserving surgery, part II: texture analysis and classification. *Journal of Biomedical Optics*, 24(9):1 – 12, 2019.
9. Susan L Norwood. Fibrocystic breast disease: An update and review. *Journal of Obstetric, Gynecologic and Neonatal Nursing*, 19(2):116–121, 1990.
10. Gao Huang, Shichen Liu, Laurens van der Maaten, and Kilian Weinberger. CondenseNet: An Efficient DenseNet Using Learned Group Convolutions. pages 2752–2761, 06 2018.
11. Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders. *arXiv e-prints*, page arXiv:1706.02262, Jun 2017.
12. Huangjie Zheng, Jiangchao Yao, Ya Zhang, and Ivor Wai-Hung Tsang. Degeneration in vae: in the light of fisher information loss. *ArXiv*, abs/1802.06677, 2018.
13. Leonardo Reyneri, Valentina Colla, and Marco Vannucci. Estimate of a probability density function through neural networks. In Joan Cabestany, Ignacio Rojas, and Gonzalo Joya, editors, *Advances in Computational Intelligence*, pages 57–64, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
14. Leslie Smith. Cyclical Learning Rates for Training Neural Networks. pages 464–472, 03 2017.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 04 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use