

PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Scatter signatures in SFDI data enable breast surgical margin delineation via ensemble learning

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

Arturo Pardo, Samuel S. Streeter, Benjamin W. Maloney, José M. López-Higuera, Brian W. Pogue, Olga M. Conde, "Scatter signatures in SFDI data enable breast surgical margin delineation via ensemble learning," Proc. SPIE 11253, Biomedical Applications of Light Scattering X, 112530K (21 February 2020); doi: 10.1117/12.2546945

SPIE.

Event: SPIE BiOS, 2020, San Francisco, California, United States

Scatter signatures in SFDI data enable breast surgical margin delineation via ensemble learning

Arturo Pardo^{1,2}, Samuel S. Streecher³, Benjamin W. Maloney³, José M. López-Higuera^{1,2,4}, Brian W. Pogue³, 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: Margin assessment in gross pathology is becoming feasible as various explanatory deep learning-powered methods are able to obtain models for macroscopic textural information, tissue microstructure, and local surface optical properties. Unfortunately, each different method seems to lack enough diagnostic power to perform an adequate classification on its own. This work proposes using several separately trained deep convolutional networks, and averaging their responses, in order to achieve a better margin assessment. Qualitative leave-one-out cross-validation results are discussed for a cohort of 70 samples.

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

1. Introduction

As of 2019, breast-conserving surgery (BCS) followed by radiotherapy remains as one of the preferred procedures for breast cancer treatment, while still exhibiting re-excision rates of about 22%–30% [1]. This may be due, among other factors, to the lack of automated methods that can assess lumpectomy margins in real time with sufficient accuracy. As a consequence, for example, surgery outcomes for BCS have been shown to be dependent on surgeon case volume or hospital specialization [2]. Spatially-modulated backscattered reflectance provides different responses for the various types of tissues that can be found in gross pathology samples of BCS [3] but, as can be seen in the following pages, there exist a series of caveats that must be overcome in order to use modulated imaging as a viable, robust tool for margin assessment.

Analyzing and classifying tissue subtypes with SFDI poses an additional challenge. While pixel-wise classification provides tumor probability maps with higher resolution, it is more sensitive to inter- and intra-sample variability, in the sense that different tissue classes may respond similarly to light and thus would be easily misclassified, as their optical signatures describe similar morphological and molecular properties. This is further exacerbated if the number of acquisitions (i.e. wavelengths and/or spatial frequencies captured per sample) is minimized, which seems to be the case whenever acquisition speed is a fundamental requirement. On the other hand, patch-wise textural analysis in the sub-diffuse domain [4] shows a much more defined accuracy, precision and recall, at the expense of losing spatial resolution. Both methods show strengths and weaknesses that, ideally, should be combined to obtain a model with robust accuracy and sufficient resolution, which would be essential for surgical margin assessment.

To this end, the following contribution is an attempt at combining these models using state-of-the-art deep learning models. Convolutional networks are flexible systems that aim to provide the best possible classification for a given dataset. In practice, the input of a deep neural net may vary in size, while keeping the classification objective constant. This allows us to produce an ensemble of classifiers. Each of the models observes a given specimen at a different spatial scale, providing maps of varied spatial resolutions and (expectedly) dissimilar accuracies. In this work, we use an ensemble of modified convolutional DenseNets [5], and study the performance of each separate model versus the complete ensemble. Two ensemble learning models are employed: (a) ensemble averaging and (b) stacking with a meta-network. Performance is evaluated via leave-one-out cross-validation on a cohort of $n = 70$ lumpectomy samples, using the best possible accuracy for each specimen separately.

2. Materials and methods

Imaging system and image acquisition The SFDI instrument is a Perkin-Elmer IVIS SpectrumCT system with a retrofitted Spatial Frequency Domain Imaging (SFDI) device [6]. The FOV is about 10×10 cm. Lumpectomy samples are cut into slices of 5 mm in thickness, following a specific clear-cut protocol [4]. One of the cuts is chosen for imaging; such cut 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.

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

Tissue subtype	Samples with tissue subtype	<i>n</i>
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	68 samples	137 ROIs

Initial considerations and category superclasses Two minor issues must be dealt with before approaching the classification problem. First, and foremost, the reduced number of samples ($n = 70$) forces us to only study local tissue properties. This is easily overcome by producing a patch dataset, as described in the next paragraph. Additionally, the fact that ROIs are conservative (i.e. the binary masks do not encompass the entirety of each sample) makes using models such as the U- and W-Nets infeasible, since those would require complete ROIs, labeling all pixels in the image. Second, there are considerable differences in terms of the number of samples per tissue subtype; rarer categories are only represented by a single specimen, while more frequent types are represented by ten or more. Considering that this approach is an initial study of the influence of spatial scale in SFDI tissue characterization, we have chosen to produce a total of five tissue superclasses (shown in Table 2): (1) Adipose, (2) Connective, (3) Benign, (4) Fibrocystic Disease (FD), and (5) Malignant. Adipose, FD and Connective tissue were the only tissue subtypes that are represented as a separate category for this study. Particularly, FD usually presents as stromal fibrosis (i.e. formations and/or bundles of connective tissue cells) and benign cyst formation [7], and thus cannot be considered as belonging to either the Benign or Connective subtypes. Further work may (and should) study more specific categories, and can support itself on these models (i.e. stacked classification).

Patch dataset generation Model training was possible via dataset augmentation. A balanced dataset of 40000 31×31 patches was generated via random population subsampling. Patches were successively extracted from random

Table 2. Tissue superclass and total number of ROIs.

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

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 600 patches are extracted per sample.

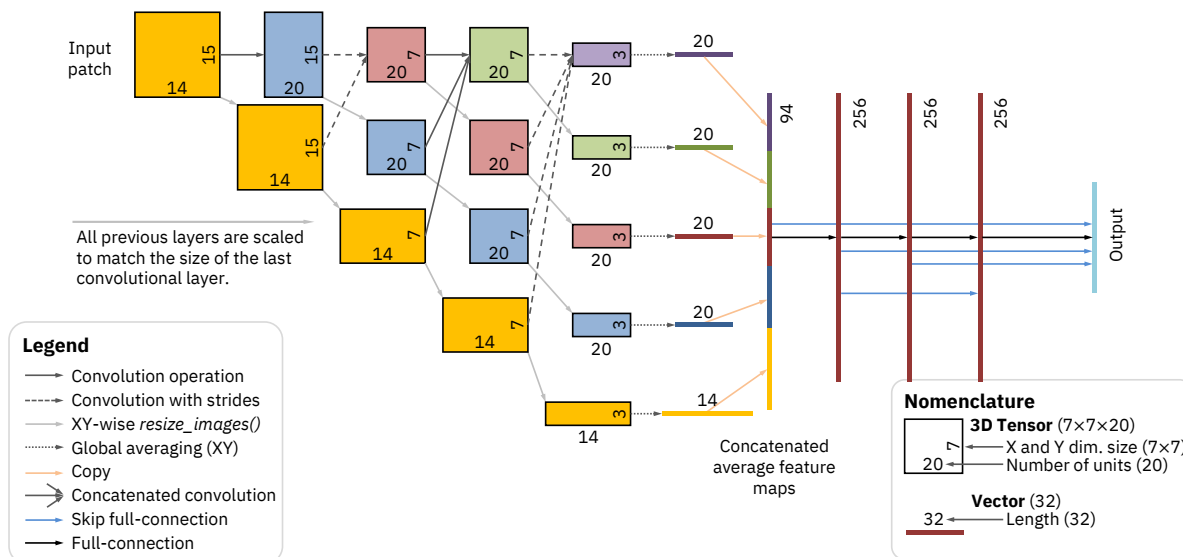


Fig. 1. Shape of the all-convolutional DenseNet classifier structure: lateral view. 3D tensors are represented sideways, as in U-Net and W-Net descriptions. This is a form of skip-connection deep neural network, where each convolutional layer can observe the activations of all previous layers. Fully connected layers prior to the output layer are all connected to the output, as well as to all previous fully connected layers.

Compute infrastructure and timing Two computers are used to train separate leave-one-out cross-validation simulations: (1) an Intel Core i3-7100 CPU, with 32 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. Each model reaches its optimal validation value after 20-30 minutes of training.

Individual networks The selected neural network for all scales was an all-convolutional modified DenseNet [5]. For this first approach, and with the specified hardware limitations, we have selected a network size and dimensions that can train in less than 30 minutes on an *nVidia RTX 2080 Super*, obtaining a typical VRAM consumption of 3GB and GPU utilization circa 75-80% maximum. An example schematic and table for this neural architecture is shown in Figure 1. The complete tabular description for the actual network is provided in Table 3. All hidden layers are ELUs to avoid issues related to *ReLU death* and gradient vanishing [8]. Output classification layers are sigmoidal to allow for more than one potential diagnosis.

Table 3. Individual classifier structures. As an illustrative reference, the first network is the one depicted in Figure 1. 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 ($= 3, 5, 11, 21, 31$) and $n_{\text{cat}} = 5$ is the number of output categories.

Example from Fig. 1	Tensor shapes (I/O)	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 ELU, f.s. 3×3	$(? \times 15 \times 15 \times 14)$ in $(? \times 15 \times 15 \times 24)$ out	1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 16)$ in $(? \times n_{\text{patch}} \times n_{\text{patch}} \times 66)$ out
1×20 ELU, f.s. 3×3 , stride = 2	$(? \times 15 \times 15 \times 24)$ in $(? \times 7 \times 7 \times 54)$ out	1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 66)$ in $(? \times n_{\text{patch}} \times n_{\text{patch}} \times 116)$ out
1×20 ELU, f.s. 3×3	$(? \times 7 \times 7 \times 54)$ in $(? \times 7 \times 7 \times 74)$ out	1×50 ELU, f.s. 3×3 , stride = 2	$(? \times n_{\text{patch}} \times n_{\text{patch}} \times 116)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 166)$ out
1×20 ELU, f.s. 3×3 , stride = 2	$(? \times 7 \times 7 \times 74)$ in $(? \times 3 \times 3 \times 94)$ out	1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 166)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 216)$ out
Global avg.	$(? \times 3 \times 3 \times 94)$ in $(? \times 94)$ features	1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 216)$ in $(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 266)$ out
3×256 skip-connection ELU	$(? \times 256)$ per layer	1×50 ELU, f.s. 3×3 , stride = 2	$(? \times n_{\text{patch}}/2 \times n_{\text{patch}}/2 \times 266)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 316)$ out
n_{cat} Sigmoid output	$(? \times n_{\text{cat}})$ Output	1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 316)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 366)$ out
		1×50 ELU, f.s. 3×3	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 366)$ in $(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 416)$ out
		1×50 ELU, f.s. 3×3 , stride = 2	$(? \times n_{\text{patch}}/4 \times n_{\text{patch}}/4 \times 416)$ in $(? \times n_{\text{patch}}/8 \times n_{\text{patch}}/8 \times 466)$ out
		Global avg.	$(? \times n_{\text{patch}}/8 \times n_{\text{patch}}/8 \times 466)$ in $(? \times 466)$ features
		3×512 skip-connection ELU	$(? \times 512)$ 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.

Training schedule Training for each model was allowed for about 60 epochs (120,000 iterations, minibatch size set to 16). Adam was the optimizer used for training ($\epsilon = 1 \times 10^{-5}$). Dropout regularization was used to avoid overfitting, with a drop probability of $p_{\text{drop}} = 0.1$. For leave-one-out cross-validation, patches belonging to the sample under test was used as the validation set. The model with the lowest MSE error for the validation set was saved as the optimal classifier for each specimen. To minimize overhead due to model performance evaluations, train and validation errors were calculated every 1000 iterations (~ 1 epoch). Early stopping was implemented with a patience of 50,000 iterations.

Generating category maps The method for obtaining a diagnostic map with a neural network trained with patches is sketched out in Figure 2. A given classifier is considered as a function with an input receptive field ($m \times n$) and a

given diagnostic output \hat{y} . To produce a complete probability map for a given classifier, each model is swept across the complete image, producing a diagnostic output at every filter position. All pixels within the receptive field are diagnosed with the output category. A diagnostic map for a given classifier is obtained by finding the most frequent diagnosis for each pixel.

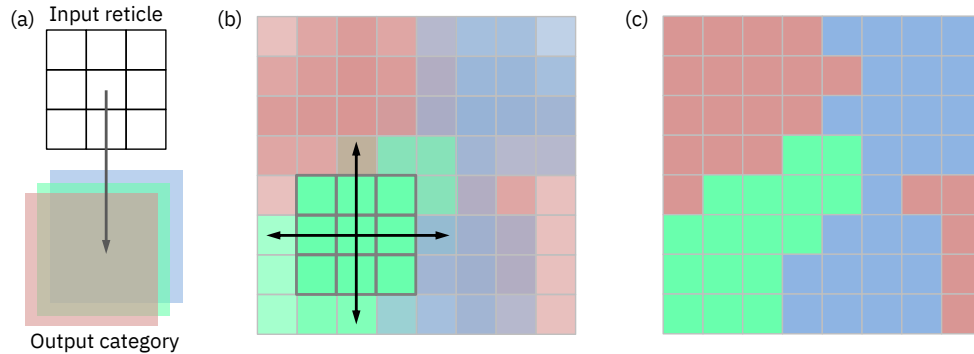


Fig. 2. Obtaining pathology maps with patch classifiers. The patch classifier is treated as a filter (a) that transforms data into diagnostic information. To obtain a complete diagnostic image, the filter is swept across a given sample of interest (b). The final diagnostic map is obtained by weighing all contributions and finding the most agreed upon output category at each pixel (c).

Background segmentation A separate network is used to perform background segmentation. The same learning protocol and dataset management rules are applied here but, considering background segmentation as an issue with little to no clinical relevance, we do not perform leave-one-out and simply let the network learn with the complete patch dataset. To save time, the same network is used for background segmentation in all images. This segmentation is not used in the performance metrics of Section 3.1.

Methods for ensemble classification Two typical ensemble learning methods are employed to combine each of the expert systems: (a) ensemble averaging and (b) stacking.

- **Ensemble averaging.** Each of the networks provides a diagnostic prediction $\hat{y}_1, \dots, \hat{y}_n$. The average of the predictions is considered the final diagnostic prediction: $\hat{y} = \frac{1}{N} (\hat{y}_1, \dots, \hat{y}_n)$. This method should provide an average sensitivity and specificity to all pathologies and patch sizes, and an increased reliability due to redundancy.
- **Stacking.** Each network provides a diagnostic prediction $\hat{y}_1, \dots, \hat{y}_n$. An additional neural network, called the meta-learner or meta-classifier $\hat{y} = g(\hat{y}_1, \dots, \hat{y}_n)$ is used to produce a final diagnostic decision. The objective of the network will be maximizing the overall accuracy of the result.

3. Results

Two main results are shown in this contribution. First, the five classifier networks are compared against each other, as well as against the ensemble average and the stacked ensemble model. The practicality of applying ensemble learning to this specific problem will depend on whether or not its performance shows improvements with respect to using the networks separately.

3.1. Separate and ensemble accuracy

Figures 3 and 4 show the best possible performances of each of the models, as well as the performance of the average and stacked ensemble. Note that these calculations are obtained with leave-one-out cross-validation, and thus they express the most optimistic scenario possible. The global accuracy of each of the models is more or less similar to already existing results, within 75-80% [3]. Some of the results, however, are particularly noteworthy.

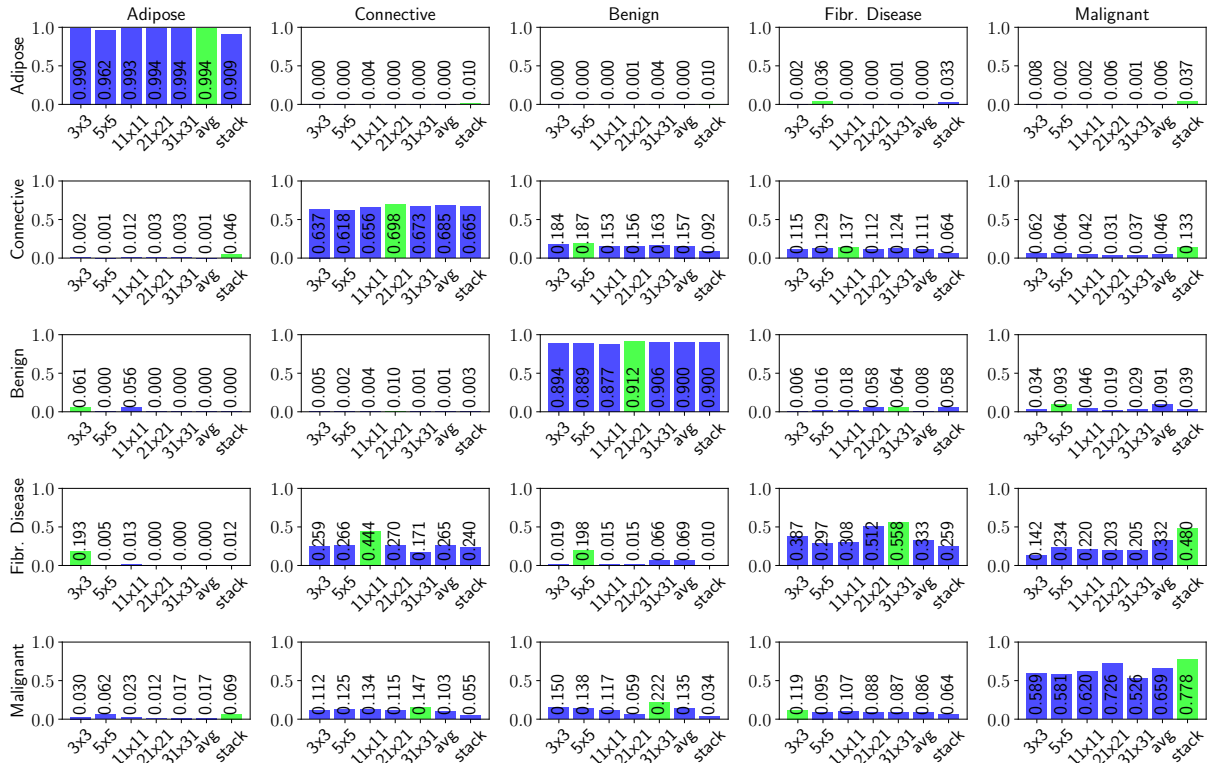


Fig. 3. Confusion matrix for each of the networks and tissue supercategories. Rows are actual categories, and columns represent predicted categories. The highest numbers are shown in green.

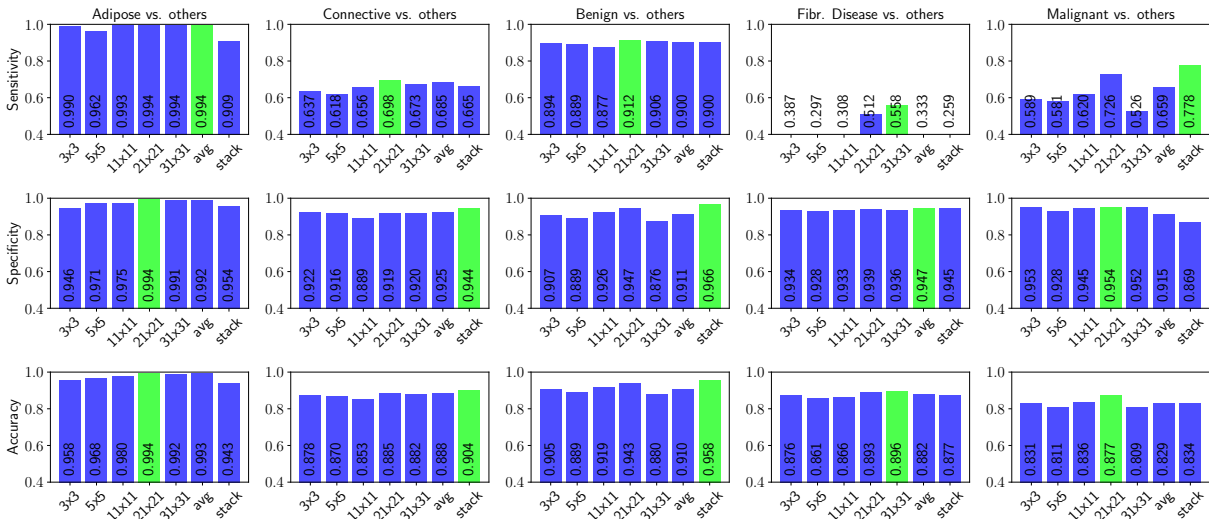


Fig. 4. One-vs-others study. Each row represents the Sensitivity, Specificity, and Accuracy for each of the tissue superclasses against the other four. The bar with the highest score is shown in green.

As can be observed in the confusion matrix of Figure 3, Fibrocystic Disease, which would usually presents as connective tissue and/or presence of benign growths [7], is the most misdiagnosed category. This (typically benign) subtype is often misdiagnosed as connective tissue and/or malignant growths, and is only detected half the time by the 31×31 network (i.e. the neural network that can best observe textural information within the dataset). The best individual network for detecting malignancy is the 21×21 network, reaching about 79% global accuracy, mostly due

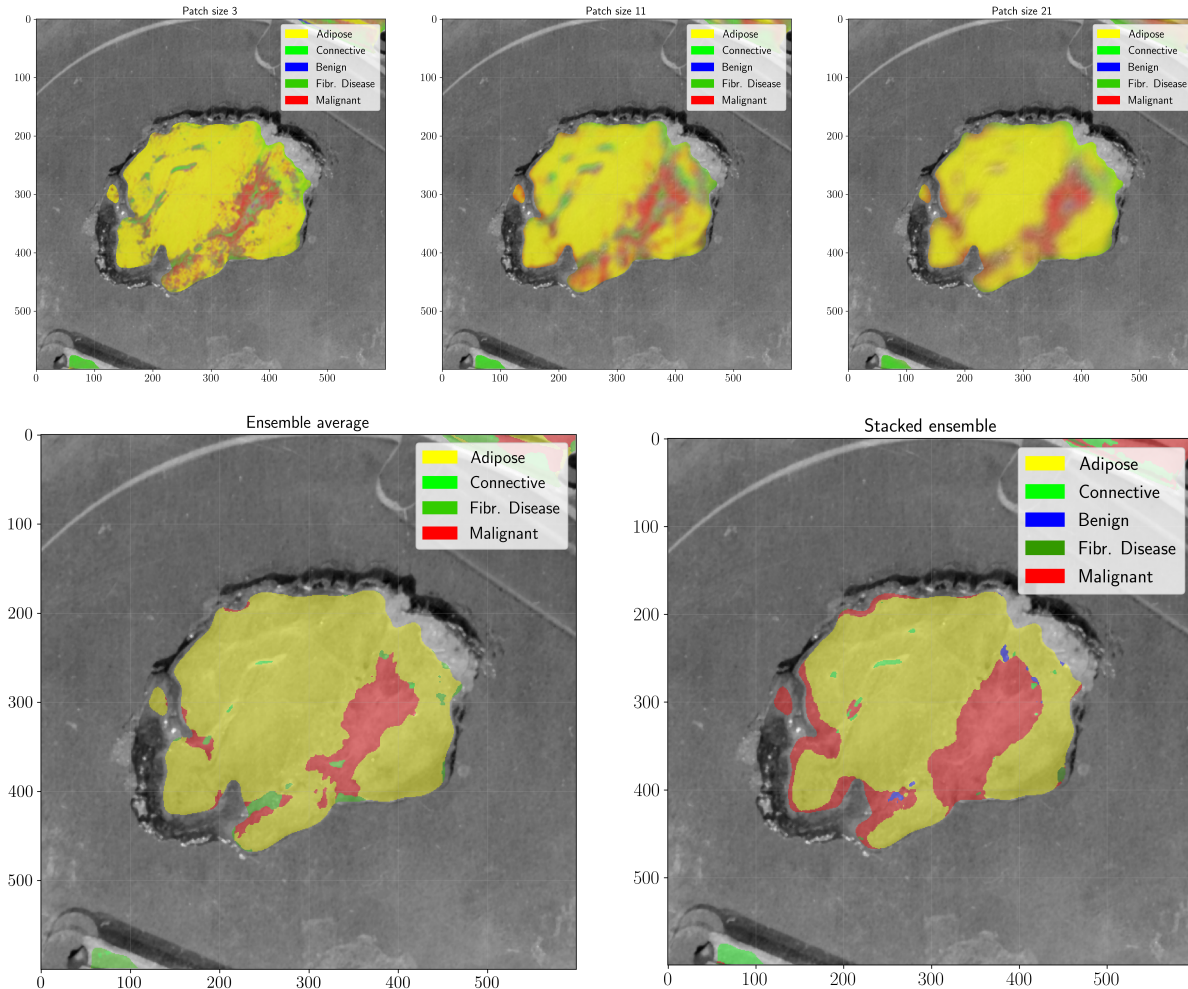


Fig. 5. Diagnostic maps for Sample 16. Top row: output (sigmoid layer) probability maps for three networks, namely those with reticle sizes 3×3 , 11×11 , and 21×21 each, respectively. Bottom row: diagnostic maps using ensemble averaging (left) and ensemble stacking (right).

to an improvement in detecting malignant tissues of over 10% with respect to the rest of the ensemble (Figure 3, bottom right subplot). This may suggest either that the 31×31 network lacks capacity to classify the large patches, or that a particular textural depth is specifically significant for detecting malignancy.

The two ensemble methods show different results, which can be interpreted in various ways. Expectedly, ensemble averaging obtains an average performance with respect to each of the individual classifiers. The method (as can be seen in Figures 5 and 6) provides much better diagnostic resolution than the large-input networks, and on-par specificity for malignant subtypes, as well as Fibrocystic Disease. Ensemble stacking shows a 15% boost in sensitivity with respect to averaging, perhaps due to a better exploitation of the performance of the 21×21 network. However, as can be seen in Figure 4, the stacked ensemble exhibits a much worse specificity for malignant tissue types. This is notably visible in the qualitative cases: while the ensemble average tends to equalize all the diagnostic maps, the stacked ensemble output draws a much more drastic boundary between malignant and non-malignant regions.

3.2. Qualitative leave-one-out results

Once the quantitative behavior of the model has been assessed, we show some case results for tumor delineation. Figures 5 and 6 show the classification results for two particular samples. The top row of plots shows the output probability maps of the first four classifiers, while the bottom two maps exhibit the behavior of the ensemble average

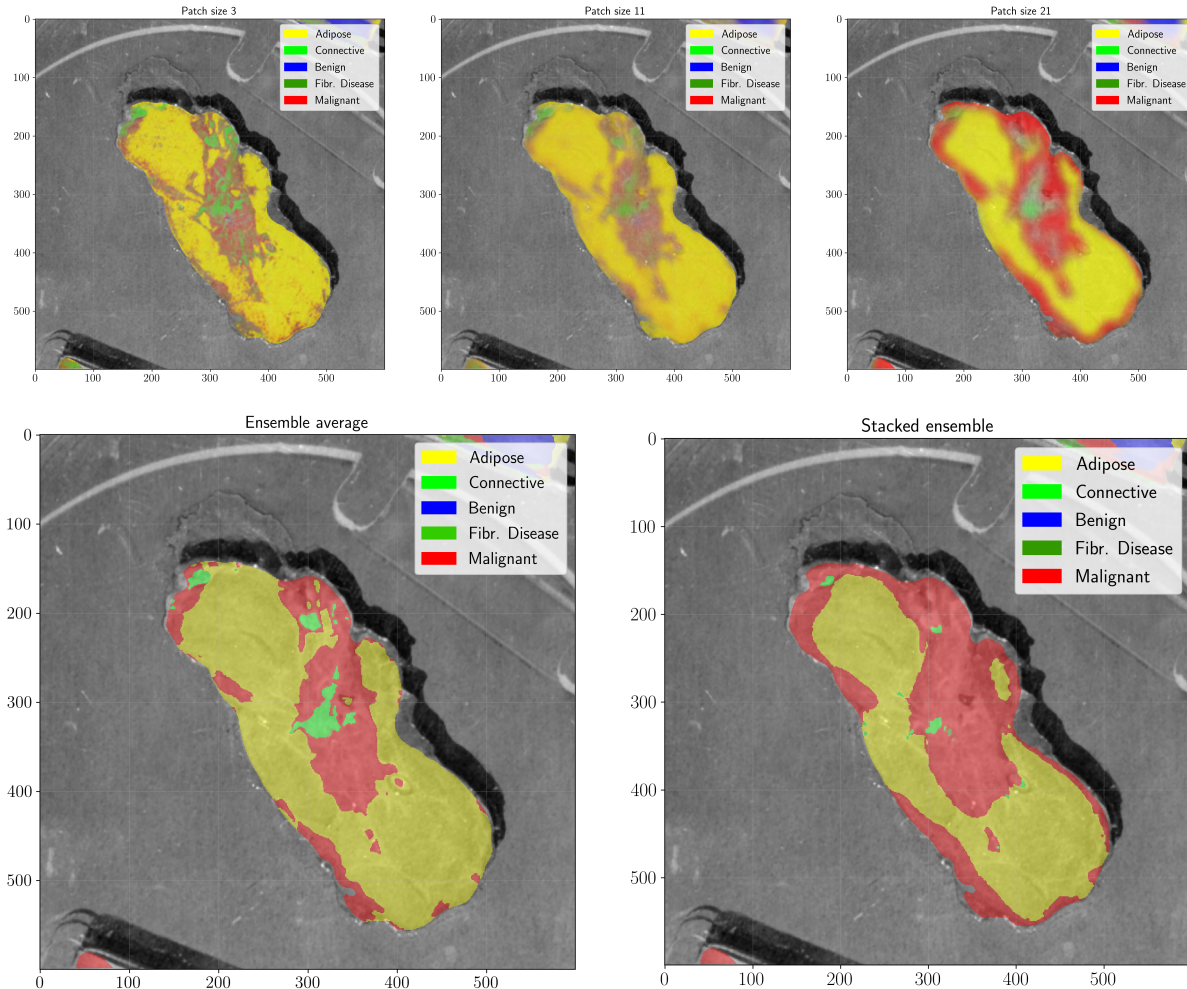


Fig. 6. Diagnostic maps for Sample 42. Top row: output (sigmoid layer) probability maps for three networks, namely those with reticle sizes 3×3 , 11×11 , and 21×21 each, respectively. Bottom row: diagnostic maps using ensemble averaging (left) and ensemble stacking (right).

(left) and the stacked ensemble (center). The original Regions of Interest for the two samples are also provided, for reference, in Figure 7.

In these output maps, many of the notions described in the quantification stage can be observed. Stacking tends to over-simplify pathology maps at the expense of maximizing overall accuracy, increasing its sensitivity to malignant scatter signatures and producing anomalous artifacts. In Figure 5, the edges of the lumpectomy are not significantly classified as malignant by any individual network. The meta-classifier, however, responds with a malignant diagnosis to maximize the accuracy within the malignant ROI. This implies that the ensemble average may be preferable in terms of textural and spatial resolution as well as fidelity to the responses of the ensemble.

4. Conclusions

This work shows a pre-emptive step in applying multi-scale analysis of SFDI data to tissue characterization and diagnosis. As a general rule, large-scale textural analysis results in more reliable diagnostic maps with lower spatial resolution, while pixel-wise classification provides higher spatial fidelity, at the expense of reducing diagnostic accuracy. The suggested architecture provides tumor probability maps that are as balanced as possible in terms of both spatial resolution and diagnostic performance.

Nevertheless, much more research is needed to fully assess the effectiveness of ensemble learning in real-time mar-

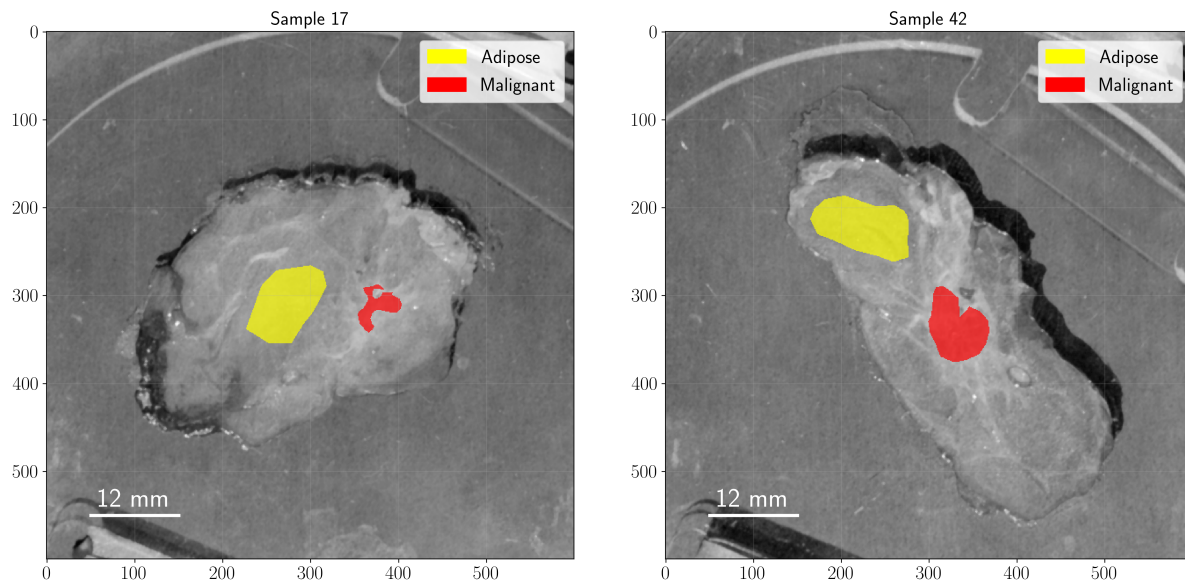


Fig. 7. Original ROIs for Sample 17 (left) and Sample 42 (right).

gin assessment of breast cancer lumpectomies. For example, some of the leave-one-out simulations failed to converge to an adequate output, resulting in misdiagnosis. This suggests that performance could significantly improve if a larger dataset was procured, minimizing the effect of inter-sample variability and obtaining a better model with greater generalization capabilities. Additionally, further improvement in acquisition technology could allow for the capture of more spatial frequencies and wavelengths, providing redundancy and perhaps improving model response further.

Future lines of work include: (a) Including pre-biopsy information as causal input data to the classifier, (b) evaluating height maps for spatial frequency calibration (height variations modify the effective projected frequency on the images), (c) exploring explanatory AI models that can provide some form of diagnostic certainty.

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 (INIVAL16/02, INIVAL18/23), Spanish Ministry of Education, Culture, and Sports (FPU16/05705).

References

1. K. Kaczmarek, P. Wang, R. Gilmore, H. N. Overton, D. M. Euhus, L. K. Jacobs, M. Habibi, M. Camp, M. J. Weiss, and M. A. Makary. Surgeon Re-Excision Rates after Breast-Conserving Surgery: A Measure of Low-Value Care. *J. Am. Coll. Surg.*, 228(4):504–512, 2019.
2. P. J. Lovrics, S. D. Cornacchi, F. Farrokyar, A. Garnett, V. Chen, S. Franic, and M. Simunovic. Technical factors, surgeon case volume and positive margin rates after breast conservation surgery for early-stage breast cancer. *Can J. Surg.*, 53(5):305–312, 2010.
3. David "Bo" McClatchy III, Elizabeth Rizzo, Wendy Wells, Candice Black, Keith Paulsen, Stephen Kanick, and Brian Pogue. Light scattering measured with spatial frequency domain imaging can predict stromal versus epithelial proportions in surgically resected breast tissue. *Journal of Biomedical Optics*, 24:1, 09 2018.
4. 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.
5. Gao Huang, Shichen Liu, Laurens van der Maaten, and Kilian Weinberger. Condensenet: An efficient densenet using learned group convolutions. pages 2752–2761, 06 2018.

6. 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.
7. Susan L Norwood. Fibrocystic breast disease: An update and review. *Journal of Obstetric, Gynecologic and Neonatal Nursing*, 19(2):116–121, 1990.
8. Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015.