BLIND SOURCE SEPARATION IN DYNAMIC CELL IMAGING USING NON-NEGATIVE MATRIX FACTORIZATION APPLIED TO BREAST CANCER BIOPSIES

D. Mandache[a,b], E. Benoit à la Guillaume[b], J.-C. Olivo-Marin[a], V. Meas-Yedid[a]

[a] BioImage Analysis Unit, CNRS UMR 3691, Institut Pasteur, 25 rue du docteur Roux, 75015, Paris, France

[b] LLTech, 58 rue du dessous des berges, Bâtiment A, 75013, Paris, France

ABSTRACT

We propose a method to fully exploit the dynamic signal produced by a recently developed non-invasive imaging modality: Dynamic Cell Imaging based on Full Field Optical Coherence Tomography, towards fast extemporaneous tissue assessment. The non-negative matrix factorisation method is used in an interpretable and quantifiable fashion to extract the signals coming from different structures of breast tissue in order to characterize cancerous tissue.

***Index Terms*—** dynamic full field optical coherence tomography, data mining, non-negative matrix factorization

1. INTRODUCTION

Breast cancer is the most frequent cancer in women worldwide, representing almost 25% of all cancers in women, it is also the second most deadly (15.4% of deaths) after lung cancer. Standard treatment involves the surgical removal of the tumor, with partial or total breast ablation. Even after heavy surgery, the risk of recurrence after 5 years is above 10%, suggesting that an imperfect removal of the tumor was performed during surgery. Hence, there is a crucial need to improve real-time intraoperative characterization of the tumor margins, in order to reduce the ablation of healthy tissue, the surgery time, and the risk of additional surgery and cancer resurgence.

The gold standard procedure for margin assessment is performed only after surgery and it is based on fixed, cut and Hematoxylin & Eosin (H&E) stained samples, this can take up to a couple of days. As for the preferred intraoperative procedure, it is also histology (frozen section or imprint cytology), despite being labour-intensive and time-consuming ($\sim$30 min) so still difficult to employ during surgery. Other techniques include X-ray or Ultrasound examination of the specimen, but the insufficient resolution does not ensure a sensitive diagnosis. Since this is a field of major interest, novel techniques are emerging, they are mostly founded on physical properties of the tissue that could discriminate between normal and cancerous cells [1]. The existing non-invasive margin assessment techniques are either too laborious or offer only a coarse resolution, some being just sensing devices with no interpretable feedback to the clinician. We propose the use of an imaging solution offering histology-like structure appearance without requiring any tissue preparation. The technique is Dynamic Cell Imaging (DCI) [2], a time-resolved variant of Full Field Optical Coherence Tomography (FFOCT) [3] and offers a intracellular resolution of 1 µm, compatible with the need of clinicians.

FFOCT and DCI images are however still difficult to interpret by surgeons because the obtained contrasts are significantly different from standard histology (See Fig. 1). Furthermore, since DCI signal origin is not yet fully characterized, there is variability in the signal acquisition and display, hindering reproducible cancer tissue analysis.

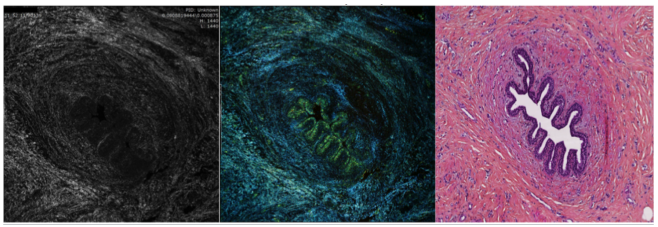

Fig. 1: Breast duct in FFOCT, DCI and H&E (left to right).

In [2] it was showed that the DCI signal could be revealing glycolysis-related movements, a metabolic pathway through which the cell produces energy. Knowing that the metabolism of cancer cells is inherently different from that of normal cells [4], we believe these are clear motivations to explore the dynamic signal provided by DCI.

In this study we present an exploratory approach towards discriminating between cancerous and normal tissue based on the dynamic profiles extracted from the DCI acquisition, while remaining in the scope of interpretability. The approach is performed on 33 patients after mastectomy.

2. OCT SIGNALS: NON INVASIVE IMAGING

FFOCT [3] provides a fast non-invasive intracellular level tissue examination tool suitable for extemporaneous analysis as

This study was performed in line with the principles of the Declaration of Helsinki. Informed consent was obtained from all individual participants involved in the study.

The authors acknowledge financial support of the ANRT (grant CIFRE no 2018/0139) and the Inception program (Investissement d′Avenir grant ANR-16-CONV-0005) for providing the computing resources.

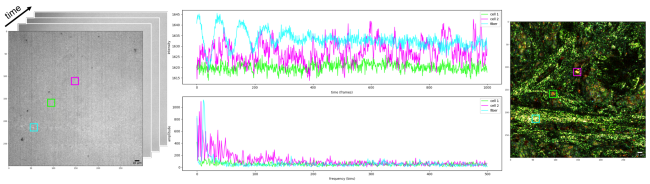

Fig. 2: Example of DCI signal for three pixels (two inside cells and one fiber): raw interferometric signal and FFT.

the sample does not require any preparation, i.e. in fresh tissue. It is based on the same optical principle as classical OCT, however, the latter performs a single-point raster scan (producing a cross-section image), while FFOCT produces *en face* images (i.e. full field front images) thanks to its array detector. The obtained image captures the optical properties of the tissue, such as the absorption coefficients of the structures imaged or their differences in refractive indexes which lead to scattering variations. The result is a gray scale image where highly backscattering elements, mostly fibrous structures (e.g. collagen), appear white while weakly backscattering content like cells appears dark gray or black. The technique reveals structures at a resolution of $1\,\mu\mathrm{m}$ up to a depth of $200\,\mu\mathrm{m}$ inside the sample.

DCI [2] relies on the same optical set-up as the classical static FFOCT, except that there are multiple acquisitions (here 1000), resulting in a time stack of interferometric frames. This captures the movement of scatterers in the sample, therefore it is only suitable for studying fresh tissue. The resulting data cube for a $1.3\,\mu\mathrm{m} \times 1.3\,\mu\mathrm{m}$ field of view (FOV) is quite significant: $1440 \times 1440 \times 1000$ pixels ($\sim$4 GB). For visualization purpose the 3D data is transformed to an RGB image according to the image formation algorithm [2], which consists in performing a Fast Fourier Transform (FFT) and averaging the amplitudes over 3 sets of frequency bands, resulting in 3 channels coded in RGB (Fig. 2). However, the noise from the environment (i.e. vibrations) induces a contamination of the frequency maps resulting in a poor separation of the channels (Fig. 3). Hence, in the present work the full frequency spectrum is considered, as detailed thereafter.

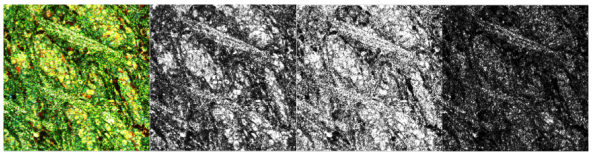

Fig. 3: DCI crop processed in RGB and the individual channels, showing poor signal separation.

In order to extract the pertinent metabolic information and remove the incoherent part of the signal the raw interferometric (time) domain is transformed to the frequency domain. Starting with 1000 frames acquired at $150\,\mathrm{Hz}$ the next steps are performed per FOV: (1) normalize frame to constant energy to remove frame-to-frame inconsistencies introduced by the acquisition; (2) average the frames by groups of 2 to attenuate noise, obtaining 500 frames pseudo-acquired at $75\,\mathrm{Hz}$; (3) pass to frequency domain with pixel-wise FFT obtaining 250 frequency maps with a step of $0.15\,\mathrm{Hz}$ (with respect to the Nyquist limit); (4) normalize FFT by its norm L_1; (5) pass to logarithmic scale to compensate the skewness of the amplitude towards low frequencies. This results in a $1440 \times 1440 \times 250$ frequency stack holding both spatial and dynamical information.

3. BLIND SOURCE SEPARATION

The motivation towards isolating different structures in the dynamic stack came from the prior intuition that there were multiple behaviours present: combined signals from the sample and perturbations (e.g. vibrations of the setup), multiple types of scatterers in the tissue (e.g. mitochondria and collagen), multiple sources of signal in one pixel (e.g. superposition of fiber and cell) or even at a lower scale, given the resolution of $1\,\mu\mathrm{m}$, different biological phenomena firing inside the cells at organelle level. Therefore, a blind source separation approach is appropriate for tackling this problem. Suitably, we employed the Non-Negative Matrix Factorization (NMF) method for its highly interpretable results by virtue of its positivity constraint leading to part-based decomposition.

Introduced by Paatero *et al.* [5], and popularised by Lee *et al.* [6], NMF is successfully used in many domains [7]: hyperspectral imaging, audio source separation, topic modeling, face recognition, furthermore, biomedical domain where it gives excellent results in stain separation [8] and is used to segment cells in calcium imaging [9]. NMF formulates a feasible model for learning object parts, relevant to perception mechanism [10].

The purpose of NMF is factorizing a data matrix $X \in R^{n \times d}$ into two low-rank positive matrices $H \in R^{k \times d}$ and $W \in R^{n \times k}$ representing the extracted feature basis and its corresponding activation, respectively : $X \approx WH$, where n is the number of data points, d the dimension of each data point and k the number of chosen components to split into. Finding the two composing matrices is achieved by minimizing the error (e.g. squared Frobenius norm - sum of squares) between the original data matrix and the result of the factorization: $\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2$. To solve this optimization problem the algorithm of multiplicative update [6] is used; it updates alternatively and iteratively for W and H in the direction of the gradient until convergence.

4. APPLICATION

In the scope of this work, we employ NMF decomposition as a feature extraction technique towards classifying cancerous and normal tissue. in order to probe the importance of the metabolic signal revealed by DCI imaging, we will only take into account the H components i.e. the dynamical profiles found in each FOV. We then apply multiple tree-based models on 382 DCI FOVs from 47 samples coming from a cohort of 33 patients after mastectomy.

4.1. Feature Extraction

The NMF algorithm was applied individually on the flattened frequency cube of each DCI FOV, passing from $1440 \times 1440 \times 250$ to $2\,073\,600 \times 250$, so the spectrum of each pixel in the cube is treated as an individual data point, disregarding the spatial configuration. One drawback of NMF is the empirical choice of the rank of factorization k; it can be set using some prior knowledge about the data together with trial and error experiments. Given the lack of a validation metric, the optimal heuristic choice of rank $k = 5$ was based on qualitative assessment of activation maps and energy of frequency components. Accordingly, there were obtained frequency signatures $H \in R^{5\times250}$ and their corresponding spatial activations $W \in R^{1440\times1440\times5}$ (see Fig. 4 for the feature extraction pipeline). The revealed components correspond to: baseline signal, fibers, noise, cells, motion artifact and they offer proper signal separation (see Fig. 5 for a representative example of the factorization result). However, further validation through biological experimentation needs to be conducted.

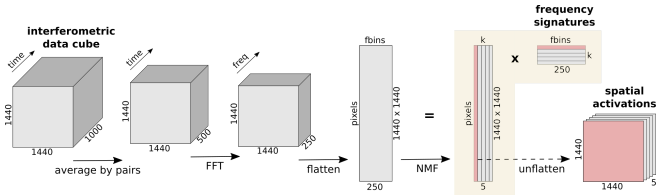

Fig. 4: Overview of pre-processing and decomposition algorithm.

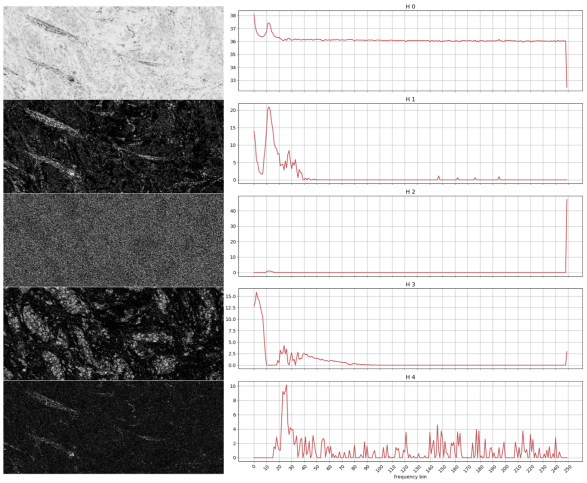

Fig. 5: NMF factorization results for $k = 5$: activations W (left) and signatures H (right) showing : baseline, fibers, noise, cells, motion artifact (top to bottom).

To construct a unified feature vector for each FOV, the H components are ordered by their energy (area under curve) and the ones with the minimum and maximum energy are removed, since they correspond to the noise and baseline component, respectively. Then the 3 remaining components are concatenated to form a single feature vector that will represent each FOV. Note that ordering the components by their energy also ensures some consistency of the feature vector between FOVs.

4.2. Classification

In terms of classification models, we focused our attention on tree-based classifiers [11] whose constituting element is the Decision Tree: a flowchart-like graph where each internal node represents a rule-based test on an attribute (i.e. feature), each branch representing the outcome of the test, and each leaf node a decision (class label). The paths from root to leaf represent classification rules. We prefer this approach mainly by virtue of their interetability, generalization capabilities and efficiency on small sample size. They are robust to outliers because the node splits are determined based on the sample proportions in each child node and not on their absolute values. Therefore, this approach also leads to finding the combination of features which can best classify the data i.e. the most important features.

Training a single decision tree can be limiting in the sense that simple trees will have a large bias (oversimplification of the model - underfitting) while complex trees will display a large variance (lack of generalization - overfitting). The bias-variance trade-off is improved by ensemble methods, so combining multiple decision trees (weak classifier) towards building a stronger classifies. There are two main approaches: bagging [12], which consists in independently training multiple trees on random sub-samples of data points and/or features and then aggregating their predictions by a voting mechanism and boosting [13] which incrementally trains trees on samples previously misclassified. In the proposed work, multiple tree-based classifiers were tested, from the simplest (single Decision Tree) to the more complex ensemble methods i.e bagged trees (Random Forest, Extra Trees) or boosted (Adaptive Boosting, Gradient Boosting, XGBoost [14]).

The splitting of the dataset into trainset and testset was done in a stratified manner, meaning that the class proportionality of the whole dataset was kept. Also, to tackle the class imbalance (thus, avoiding learning a biased model and also having clear interpretable performance metrics) an oversampling of the minority class (Cf. normal) was performed: for the train set we applied the SMOTE [15] algorithm which generates synthetic samples from interpolation and for the validation set we only applied random oversampling to avoid introducing any ambiguity in the performance metrics.

5. RESULTS

We trained multiple tree-based models using 4-fold cross validation: 75% of the samples for training (286 samples: 174 cancerous, 112 normal) and 25% for validation (96 samples: 58 cancerous, 38 normal). Only the lower half of spectrum (up to $f_{bin} = 120$) was considered since there was observed that the higher part of the spectrum has low SNR, hence overfitting on noise is avoided. Take note of some of the most important hyperparameters chosen: maximum tree depth = 10 (for bagging ensembles and simple decision tree) or 1 (for

boosting models), number of trees in ensembles = 100. Table 1 presents the classification metrics obtained. The model with the best generalization power proves to be AdaBoost, this can be deduced by the fact that it obtains the best accuracy, while also keeping consistency between train and test metrics. We also notice a lower specificity compared to the sensitivity which is due to the under-representation of the normal class in our dataset. The results are promising, being comparable with the other state-of-the-art non-invasive margin assessment techniques [1].

Classifier	Train Acc.	Test Acc.	Test Sn.	Test Sp.
AdaBoost	90.95 ±1.39	**70.91** ±6.38	77.59 ±7.21	**64.22** ±10.37
XGBoost	91.38 ±5.66	70.69 ±5.21	82.33 ±5.5	59.05 ±18.48
RandomForest	98.13 ±0.32	65.73 ±7.93	**83.62** ±4.95	47.84 ±15.73
ExtraTrees	96.77 ±1.81	65.52 ±4.00	78.45 ±5.52	52.59 ±10.73
GradientBoosting	98.42 ±0.72	65.09 ±4.66	76.72 ±4.64	53.45 ±10.20
DecisionTree	99.93 ±0.12	57.54 ±3.41	64.66 ±3.11	50.43 ±4.46

Table 1: Classification performance metrics: accuracy. sensitivity, specificity (mean percentage ± standard deviation for 4-fold cross validation).

5.1. Feature Importance

As one of the main motivations for choosing this type of models was their semantic interpetability, we are looking at the most discriminating features as established by the best-performing algorithm (AdaBoost). They are highlighted in Fig. 6, plotted over the average components of the whole dataset. Feature importance is calculated for each attribute in a decision tree as the amount by which the split points over the considered attribute improve the performance measure. Then, for each feature, its importance is averaged over all the trees of the ensemble. In other words, for the given situation, feature importance is the ability of an attribute (here frequency bin in a NMF component) to discriminate towards normal or cancerous class.

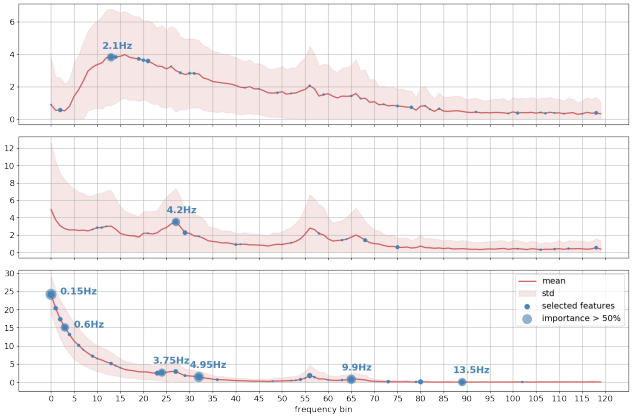

Fig. 6: Avg. and std. of the features (3 NMF components) over the training set (orange) and the f_{bins} selected by AdaBoost (blue).

We observe the following frequencies appearing: the peak at f=2.1 Hz T=0.5 s corresponding to the cell component, as well as the lower part of the spectrum for the more static fiber components T=2 s to 6 s. However, for the peaks in the vicinity of f=4.2 Hz and 9 Hz we intend to further investigate their corresponding spatial maps to characterize their nature, but based on our preliminary observations we believe they correspond to vibration artifacts due to external nuisance.

6. DISCUSSION AND CONCLUSION

In the present work we demonstrate the feasibility of employing a blind source separation technique, namely NMF, to better extract the signal coming from different types of moving scatters in breast tissue imaged with the non-invasive DCI technique in an interpretable and quantifiable way that can overcome the noise and motion artifacts. For now we used NMF decomposition to classify between cancerous and normal FOVs with 70.91% accuracy and we revealed some salient frequencies.

For future work we are planning to also include the morphological information i.e. the spatial activations associated with the dynamical signatures towards achieving improved classification results. Moreover, we aim to apply a scaled-up NMF algorithm on more FOVs whereby we could achieve separation between cell types (cancerous *vs* immune cell).

7. REFERENCES

[1] B. W. Maloney et al., "Review of methods for intraoperative margin detection for breast conserving surgery," *Journal of Biomedical Optics*, vol. 23, no. 10, pp. 1, oct 2018.

[2] C. Apelian et al., "Dynamic full field optical coherence tomography: subcellular metabolic contrast revealed in tissues by interferometric signals temporal analysis," *Biomedical Optics Express*, vol. 7, no. 4, pp. 1511–24, 2016.

[3] A. Dubois et al., "High-resolution full-field optical coherence tomography with a linnik microscope," *Applied Optics*, vol. 41, no. 4, pp. 805–12, 2002.

[4] R. J DeBerardinis and N. S. Chandel., "Fundamentals of cancer metabolism.," *Science Advances*, vol. 2, no. 5, 2016.

[5] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values," *Environmetrics*, 1994.

[6] D. Lee and H. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] A. Cichocki et al., *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*, John Wiley and Sons, oct 2009.

[8] A. Vahadane et al., "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[9] R. Maruyama et al., "Detecting cells using non-negative matrix factorization on calcium imaging data," *Neural Networks*, vol. 55, pp. 11–19, 2014.

[10] S. Ullman, *High-level vision : object recognition and visual cognition*, MIT Press, 1996.

[11] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[12] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 421, pp. 123–140, 1996.

[13] Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," Tech. Rep. 5, 1999.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[15] N V Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," Tech. Rep., 2002.