BASAL CELL CARCINOMA DETECTION IN FULL FIELD OCT IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

D. Mandache[a], *E. Dalimier*[b], *JR. Durkin*[c], *C. Boccara*[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, Pépinière Paris Santé Cochin, 29 rue du faubourg Saint-Jacques, 75014, Paris, France
[c] Department of Dermatology, Drexel University College of Medicine, Philadelphia, PA, USA

ABSTRACT

In this paper we introduce a new application that exploits the emerging imaging modality of full field optical coherence tomography (FFOCT) as a means of optical biopsy. The objective is to build a computer-aided diagnosis (CAD) tool that can speed up the detection of tumoral areas in skin excisions resulting from Mohs surgery. Since there is little prior knowledge about the appearance of cancer cell morphology in this type of imagery, deep learning techniques are applied. Using convolutional neural networks (CNN), we train a feature extractor able to find representative characteristics for FFOCT data and a classifier that learns a generalized distribution of the data. With a dataset of 40 high-resolution images, we obtained a classification accuracy of 95.93%.

***Index Terms*—** convolutional neural networks, non-melanoma skin cancer, digital pathology, full field optical coherence tomography

1. INTRODUCTION

Skin cancer is the most common human malignancy, predominantly represented by non-melanoma types with 5.4 million cases per year, 80% of which are Basal Cell Carcinomas (BCC) with the majority the remaining being Squamous Cell Carcinomas (SCC) [1]. The gold standard procedure for treating non-melanoma skin cancer in high risk areas is Mohs Surgery [2]. The technique involves the consecutive removal of thin layers of skin, followed by histological preparation and microscopical examination for tumor clearance. This process can take up to an hour and guides further tissue extraction. We investigate the feasibility of using a non-invasive optical slicing modality, together with an automated diagnosis of the cancerous areas, which would lead to speeding up the procedure, consequently, improving patient comfort and physician throughput.

Histology slides are $3\,\mu\text{m}$ thick, performed at any desired depth in the resection, and observed with high resolution microscopes. So only thin optical slicing systems which ensure cellular-level resolution and enough penetration depth can compete with this standard. The technologies generally employed in this domain are optical coherence tomography (OCT) and confocal microscopy, while the first favors penetration depth ($1\,\text{mm}$) with a low axial resolution ($10\,\mu\text{m}$), the other has a better resolution ($0.8\,\mu\text{m}$) and insufficient penetration ($100\,\mu\text{m}$). FFOCT relies on the same principle as classical OCT, light interferometry, but produces ”en face” slices, instead of cross-sections. This allows for an intra-cellular resolution ($1\,\mu\text{m}$) and sub-surface penetration ($200\,\mu\text{m}$-$1\,\text{mm}$, depending on the numerical aperture and optical properties of the tissue). Given its specifications and resemblance with transverse histological slices, FFOCT proves to be a powerful technique for optical pathology [3].

There is a growing tendency for medical diagnosis applications to rely on deep learning [4], especially convolutional neural networks (CNN) which are most suitable for image inputs. Their efficiency is proven by the fact that CNN architectures are winning the grand challenges in the domain: TUPAC16 (MICCAI), CAMELYON16-17 (ISBI), however, all their datasets consist of bright-field whole-slide images. In OCT imaging, deep learning is mostly used for segmentation of the retinal layers [5]. Recently, neural networks conquered the field of dermatology, with [6], which classifies cancerous lesions from macro images of the skin surface. Still, to our knowledge, there is almost no research in automatic diagnosing for the FFOCT modality, including using deep learning methods. In this work we propose exploiting FFOCT images, as exposed in Section 2, which will be used to train a CNN, as detailed in Section 3. The results obtained are presented in Section 4 and the concluding remarks in Section 5.

2. DATA

2.1. Data Acquisition

Our data set consists of 40 FFOCT images of tissue excisions obtained from Mohs surgery, biopsies and conventional excisions, which were then imaged using the Light-CT$^{\text{TM}}$ scanner developed by LLTech, France. The samples did not undergo any preparation. The scanner has a resolution of $1\,\mu\text{m}^3$ and a penetration depth of 200 μm, which means reaching the level of the dermis. Each image is a 2D transverse slice of a unique tissue sample imaged at 20 μm below the surface. Samples measure between 2-2.5cm^2 which gives high-resolution images of around 200 Megapixels. The speed of acquisition is 1 cm^2 per minute.

2.2. Data Annotation

As shown in Fig. 1 the images were manually segmented and diagnosed by a dermatopathologist with experience with this modality and who could access the gold standard H&E frozen sections of the specimen for validation. The data was gathered and annotated using the Cytomine [7] platform. 26% of the total imaged area was segmented, the rest (unlabeled areas) being background or abnormal tissue (it sometimes surrounds the tumors but its appearance is not relevant for either class, it should be treated separately). The images are preponderantly annotated with the *normal* label and only 10 images present some cancerous areas, annotated as *BCC*, more precisely, only 9.5% of the annotated data is pathological. Therefore, as it is the case of most of the medical applications, we face the class imbalance problem, which we try to solve by oversampling the minority class.

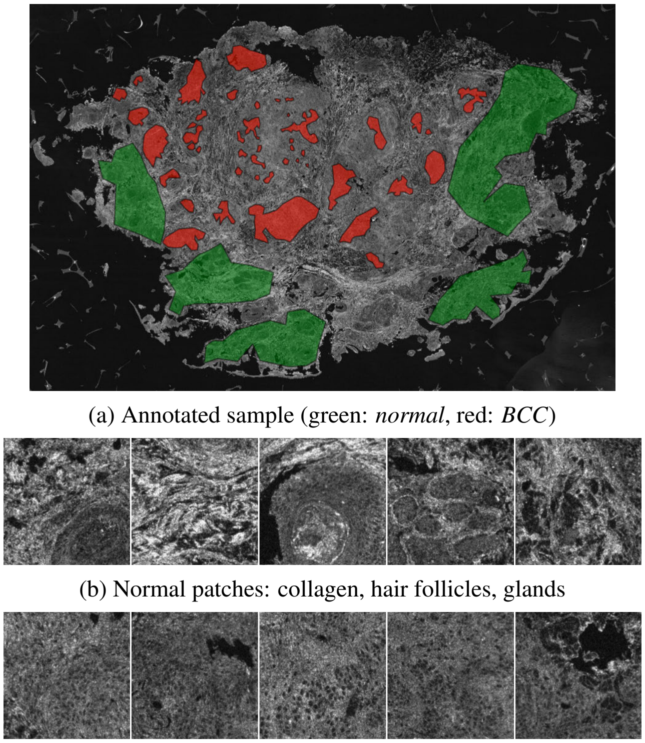
2.3. Data Sampling & Preprocessing

The scanner produces 16-bit DICOM images, but only 10 to 12 bits are actually used; they were converted to 8-bit JPEG for convenience, so they can be tested with out of the box pre-trained architectures that only accept this depth.

Speckle noise is strongly present in FFOCT images, but proper denoising algorithms are too computationally costly (of the order of hours) therefore, since one of the requirements of our application is speed, we applied a 3 × 3 Gaussian filter to provide some smoothing while preserving the structures (e.g. cancerous cell nuclei appear as dark blobs with 10 pixels in diameter).

A constraint imposed by the computational power needed to train artificial neural networks is its number of parameters. This is a function of the depth (number of filters, layers) and width (input size of the layers) of the network. To satisfy this constraint while capturing enough context to discern normal skin structures from the cancerous cell organization, we split the images in patches of 256 × 256 pixels. With the aim of augmenting and also balancing the data set, we oversampled the patches with different step values for the two classes: 170px for the normal class, while BCC patches overlap more, with a stride of 40px. This produces 108 082 patches: 59 112 normal and 48 970 BCC; 80% of which serve as a training set and the rest is used for measuring the performance.

Among the popular practices in deep learning is data standardization (zero centering + normalization) which translates into imposing the data to follow a normal distribution. This influences the robustness of the algorithm to variations in the images caused by the acquisition conditions, for example, and it also ensures a better convergence of the learning process. Data standardization is done by subtracting the mean intensity value over the training set and dividing by their standard deviation. Note that the same preprocessing has to be applied on the test data for consistency. Furthermore, some basic data augmentation was performed, which led to doubling the training set, by adding synthetically generated images obtained through horizontal and vertical flipping, slight rotations and shifts.



(a) Annotated sample (green: *normal*, red: *BCC*)

(b) Normal patches: collagen, hair follicles, glands

(c) BCC patches: aggregates of cancer cells, retraction artifacts

Fig. 1. An example of image (11808 × 8352) annotated in Cytomine and some patches (256 × 256) extracted from it.

3. METHOD

Using some popular architectures like VGG-16 [8] or InceptionV3 [9], pre-trained on ImageNet database, we obtained an accuracy of 89.30% and 90.79%, respectively. Moreover, the overfitting phenomenon (*i.e.* ”memorizing” training data, rather than learning to generalize) was quite important and was fast to appear. We inferred that those architectures were too deep and complex for our data distribution, therefore, data over specification caused overfitting. However further investigation is intended. Moreover, using pre-trained networks or even fine-tuning (*i.e.* continuing training) them seems inappropriate for our problem, since we are dealing with a new

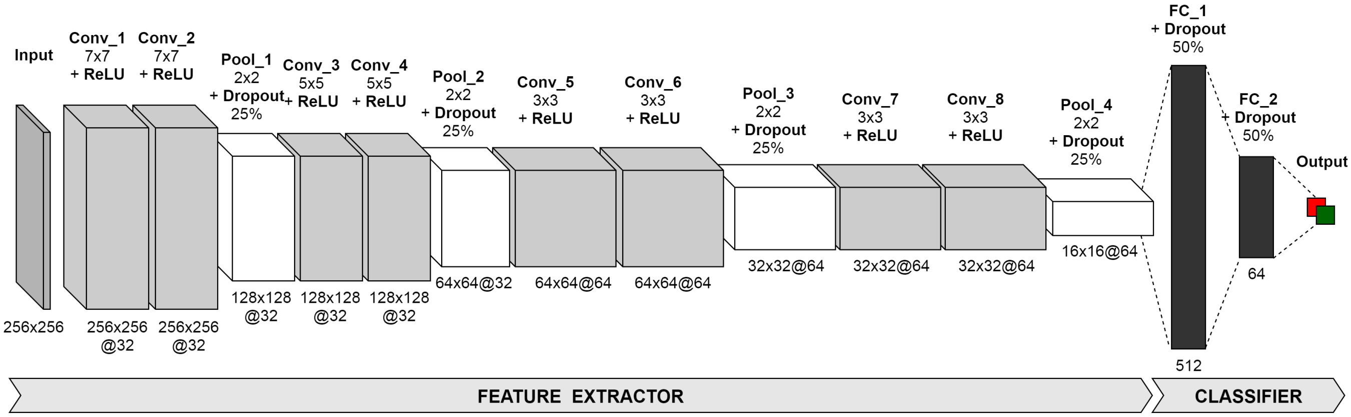


Fig. 2. Proposed CNN architecture.

modality and one of our objectives is to discover relevant features that characterize FFOCT images. We build and train from scratch, a CNN that is able to learn a generalized distribution of our data, with respect to our two classes, *normal* and *BCC*.

The proposed architecture follows the classical construction of a multi-layer CNN while having a smaller number of parameters than state of the art architectures. Nevertheless, it takes advantage of the ideas employed by VGG, like: 1) convolutional blocks: consecutive convolutional layers to capture larger input with a spare of parameters; 2) dropout layer: a fraction of neurons is randomly removed to avoid overfitting; 3) rectified linear unit (ReLU): activation function used to speed up the computations.

We trained a 10 layer CNN (see Fig. 2) including the feature extraction part, composed of 4 convolutional blocks (with two convolutional layers each) followed by max-pooling and 25% dropout and a classifier consisting of two fully-connected layers of 512 and 64 neurons, respectively, each followed by 50% dropout, lastly, there is one output neuron whose firing signals the classification of the input patch as *BCC* or *normal* otherwise. The layers from the first blocks have 32 filters and the rest have 64 filters each, the receptive fields of the convolutions vary from 7×7 and 5×5 to 3×3 as we go deeper into the network.

The network has in total 8 654 369 parameters to train, most of them corresponding to the classifier part (*i.e.* fully connected layers), while 232 417 represent the filters encoding the features, meaning 60× less than VGG-16. The weights are initialized using the Glorot method [10] which is based on the idea that the gradients of each layer should follow more or less the same distribution at the beginning of training and it is proven to converge faster and towards a ”better” minimum. The training process consists in minimizing the binary cross entropy loss. When computing it, class weighting applies a higher penalization for misclassifying cancerous class with respect to the under representation of the minority class: $1 \div 1.2$ (there is 1 normal sample for 1.2 cancerous samples). Learning is possible using a gradient decent optimization algorithm, for our application Adaptive Moment Estimation (Adam) [11] worked best. Adam is one of the adaptive methods of gradient descent whose particularity is that they adapt the learning rate (*i.e.* step of the descent) to the parameters, performing larger updates for infrequent parameters (*i.e.* the ones which were rarely updated) and smaller updates for frequent ones. Adam also multiplies the learning rate by the momentum (*i.e.* average of the previous gradients) providing accelerated optimization. The mini-batch gradient descent approach is a trade-off between computational accuracy and convergence time, so between batch (entire dataset) and stochastic (one example at a time) gradient descent. We chose a mini-batch of 40 samples as it was the biggest size that respected the memory constraints.

The proposed network was implemented using Keras [12] with Tensorflow [13] backend. Training time was about a day (25 hours and 17 minutes) on 4 Nvidia Tesla P100 GPUs for 2000 epochs (45 seconds per epoch); an epoch represents one forward and one backward pass over all the training examples. However, testing a full-size image patch by patch takes up to a few seconds and is possible on any kind of basic system configuration.

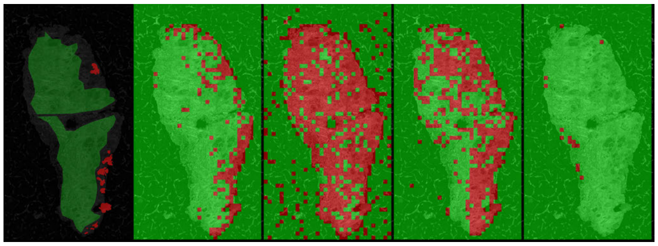


Fig. 3. Ground truth labeling (left) and predictions (from left to right): **proposed** method, **InceptionV3** pre-trained, **VGG16** pre-trained, **VGG16** fine-tuned (green: *normal*, red: *BCC*).

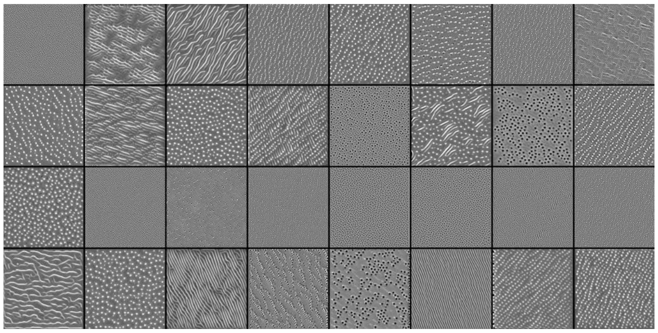


Fig. 4. Textures maximizing activations for Conv_3 layer.

4. RESULTS

We obtained a classification accuracy of 95.93%, corresponding to a sensitivity of 95.2% and 96.54% in specificity. Fig. 3 shows a comparison between the ground truth labeling and the patches classified with our method. We notice that the cancerous regions are coarsely detected and, interestingly, the abnormal tissue that was unlabeled (so unknown to the network during training) is classified as BCC. Note that background removal was not performed when testing, however, the method doesn't detect any abnormality outside the sample. There are a few misclassified solitary patches which inspired us to impose some neighborhood relationship constraints in the future work. The results are still raw but they clearly represents a promising step towards an automated segmentation of FFOCT images.

However, caution should be taken with the statistical results which can be misleading in interpreting the overall efficiency of the method and its behavior with different data. Since artificial neural networks are black-box models, gaining an intuition about the reasoning performed by the network is not straightforward. To do that we visualize what the network is learning. This is possible by viewing the weights of the neurons which correspond to the convolutional filters, but since they are very small, the textures encoded are not easily deductible. Still, to get the texture that a filter is responsive to, we can visualize the simulated input that would maximize the activation of its corresponding neurons. This is achieved by performing gradient ascent in the input space with respect to the filter activation loss. In Fig. 4 are plotted the patterns learned by the 3rd convolutional filter. Without any clinical feedback from a pathologist, we deduce that they could encode different distributions of cells and orientations of collagen fibers, but this matter is still open for discussion.

5. CONCLUSION

In this work we trained a convolutional neural network in the purpose of discriminating basal cell carcinoma from normal skin. We show preliminary results that open a promising research direction, which is analyzing FFOCT images with the powerful methods of deep learning. Developing computer-aided diagnosis tools could ease the integration of this novel optical biopsy technology in the clinical environment by assisting pathologists in their familiarization with the new modality and, ultimately, it could reduce the costs and duration of certain medical procedures, like Mohs surgery.

To improve our results we will firstly need a more consistent data set and also a better understanding of the decision flow of the pathologists in diagnosing the samples. This would allow us to translate the knowledge of an expert to artificial neural networks. Additionally, we will include a 3rd class for the unlabeled areas. For future work we also intend to adopt a multi-scale approach, inspired by MIMO-Net [14], for capturing a larger context and extracting specific information at different levels of zooming.

Another idea is that, in order to accurately assess the efficiency of such a model, we need to understand the reasoning learned by the machine. Therefore, we are ambitiously aiming towards demystifying artificial neural networks in the hope of also gaining knowledge about the data set.

6. REFERENCES

[1] Amercian Cancer Society, "Cancer facts & figures 2016," website, last accessed in October 2017.

[2] S. Z. Aasi, D. J. Leffell, and R. Z. Lazova, *Atlas of Practical Mohs Histopathology*, Springer, 2013.

[3] E. Dalimier and D. Salomon, "Full-field optical coherence tomography: a new technology for 3d high-resolution skin imaging," *Dermatology*, vol. 224, pp. 84–92, 2012.

[4] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[5] F. G. Venhuizen et al., "Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks," *Biomedical Optics Express*, vol. 8, no. 7, pp. 3292–3316, 2017.

[6] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature Letter*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] R. Marée et al., "Collaborative analysis of multi-gigapixel imaging data using cytomine," *Bioinformatics*, vol. 32, no. 9, pp. 1395–1401, 2016.

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

[9] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2818–2826.

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics (PMLR)*, 2010, vol. 9, pp. 249–256.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[12] F. Chollet, "Keras," 2015, Available from github.com/fchollet/keras.

[13] M. Abadi, A. Agarwal, et al., "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015, Available from tensorflow.org.

[14] S. E. A. Raza et al., "Mimo-net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images," in *Proc. ISBI*. IEEE, 2017, pp. 337–340.