Depósito de Investigación de la Universidad de Sevilla

https://idus.us.es/

This is an Accepted Manuscript of an article published by Elsevier in Journal of Investigative Dermatology (JID) , Vol. 1291 - 1299.e2 , on May 2022, available at: https://doi.org/10.1016/j.jid.2021.09.029

Elsevier Copyright © 2021 The Authors. Published by Elsevier Inc. All rights reserved . En idUS Licencia Creative Commons CC BY-NC-ND

Article Type: Original article

Article Title: CLASSIFICATION OF BASAL CELL CARCINOMA IN EX VIVO CONFOCAL MICROSCOPY IMAGES FROM FRESHLY EXCISED TISSUES USING A DEEP LEARNING ALGORITHM

Short Title: Deep learning for confocal ex vivo images

Mercedes Sendín-Martín[1] (ORCID ID# 0000-0003-0156-381X), Manuel Lara-Caro[1] (ORCID ID# 0000-0001-7424-2000), Ucalene Harris[2] (ORCID ID# 0000-0002-7845-7460), Matthew Moronta[2] (ORCID ID# 0000-0002-7845-7460), Anthony Rossi[2] (ORCID ID# 0000-0003-2295-1934), Erica Lee[2] (ORCID ID# 0000-0001-9670-3242), Chih-Shan Jason Chen[2] (ORCID ID# orcid.org/0000-0002-0977-2707), Kishwer Nehal[2] (ORCID ID# 0000-0001-5791-7184), Julián Conejo-Mir Sánchez[1,3] (ORCID ID# 0000-0001-9108-9538), José-Juan Pereyra-Rodríguez*[1,4] (ORCID ID# 0000-0001-6843-5877), Manu Jain*[2,4] (ORCID ID# 0000-0002-6035-0825).

1. Hospital Universitario Virgen del Rocío, Dermatology Department, Sevilla (Spain).
2. Memorial Sloan Kettering Cancer Center, Dermatology Service, Department of Medicine, New York (USA).
3. School of Medicine, University of Seville (Spain).
4. Weill Cornell Medicine, Dermatology Service, Department of Medicine, New York (USA).

* Both authors contributed equally.

Corresponding author: Manu Jain

Memorial Sloan Kettering Cancer Centre (MSKCC)

530 E74th St NY, NY 10021

Email: jainm@mskcc.org

Funding sources: This research is funded by a grant from the National Cancer Institute /National Institutes of Health (P30-CA008748) made to the Memorial Sloan Kettering Cancer Center.

Manuscript word count: 3500.

Figures: 4; **Tables**: 2; **Supplementary figures**: 2; **Supplementary video**: 1; **Supplementary tables**: 1.

Keywords: *ex vivo* confocal microscopy, digitally stained purple and pink images, basal cell carcinoma, convolutional neural network, deep learning

Abbreviations Used: BCC is Basal Cell carcinoma, MMS is Mohs Micrographic Surgery, AI is Artificial Intelligence, EVCM is *Ex Vivo* Confocal Microscopy, ROC is Receiver Operator Characteristic, AUC is Area Under the Curve, CNN is Convolutional Neural Network.

ABSTRACT

Ex vivo confocal microscopy (EVCM) generates digitally colored purple-pink images similar to H&E, without time-consuming tissue processing. It can be used during Mohs surgery for rapid detection of basal cell carcinoma (BCC); however, reading EVCM images requires specialized training. An automated approach using a Deep Learning algorithm to BCC detection in EVCM images can aid in diagnosis. 40 BCCs and 28 negative ("not-BCC") samples were collected at Memorial Sloan Kettering Cancer Center to create three training datasets: 1) EVCM image dataset (663 images), 2) H&E image dataset (516 images), and 3) a combination of the two datasets. 7 BCCs and 4 negative samples were collected to create a EVCM test dataset (107 images). The model trained with the EVCM dataset achieved 92% diagnostic accuracy, similar to the H&E model (93%). The area under ROC was 0.94, 0.95, and 0.94 for EVCM, H&E, and combination trained models, respectively**.** We developed an algorithm for automatic BCC detection in EVCM images (comparable accuracy to dermatologists). This approach could be used to assist with BCC detection during Mohs surgery. Furthermore, we found that a model trained with only H&E images (which are more available than EVCM images) can accurately detect BCC in EVCM images.

INTRODUCTION

Basal cell carcinoma (BCC) is the most common skin cancer accounting for ~2 millions of cases annually in the United States alone [Rogers et al. 2015]. Biopsy, followed by histopathology, is the gold standard for diagnosis and subtyping of BCCs (as aggressive or non-aggressive) for appropriate management — surgical treatment for aggressive tumors versus topical treatment (non-surgical) for less aggressive BCCs. For aggressive or recurrent tumors, especially those located in cosmetically sensitive sites such as the face, Mohs micrographic surgery (MMS) is the treatment of choice with high cure rates [Van Loo et al. 2014]. However, MMS is a tedious and time-consuming procedure as it involves careful removal of skin cancer layer-by-layer. Each excised layer undergoes frozen sectioning and microscopic evaluation, requiring up to 20-45 minutes. Often times multiple layers are removed to achieve complete tumor clearance; thus, the entire surgical procedure may last for several hours [Keena and Que 2016], increasing patient's waiting time, complications, and cost of the procedure. Additionally, frozen sectioning can cause tissue destruction and create artefacts that may hinder the final diagnosis.

To expedite the surgical procedure, various *ex vivo* optical imaging devices have been developed [Bennàssar et al. 2013; Dalimier and Salomon 2012; Gareau et al. 2009; Karen et al. 2009]. These devices can rapidly image freshly excised tissues at "near-histopathological" resolution, obviating the need for destructive and time-consuming tissue processing.

Ex vivo confocal microscopy (EVCM) is an emerging imaging technique that can evaluate freshly (un-processed) excised whole-tissue samples without the need for tissue processing (frozen sectioning). As there is no tissue processing involved, EVCM can image tissues rapidly (less than a minute for a tissue measuring up to 2 cm), reducing the time for the Mohs surgery and enabling real-time imaging in the surgical suite [Keena and Que 2016]. Furthermore, EVCM creates digitally colored purple and pink images (similar to H&E images) by converting

fluorescence signal originating from the nucleus into purple color and reflectance signal from the cytoplasm into pink color. These digitally colored images can be read by Mohs surgeons trained in pathology [Mu et al. 2016]. Although EVCM has demonstrated an overall high sensitivity and specificity (~ 90%) for detection of BCC [Gareau et al. 2009; Karen et al. 2009] during MMS [Bennàssar et al. 2013], it is only been utilized in very few academic centers. We believe that the integration of an automatic algorithm for the detection of BCC in EVCM images could immensely aid Mohs surgeons, increasing adoption of this technology. Moreover, EVCM technology may be useful for surgical pathologists and dermatologists to obtain faster results from standard excisions [Bennàssar A et al. 2012; Debarbieux et al. 2015] or even for biopsies of inflammatory skin lesions [Bağcı et al. 2019; Bağcı et al. 2021; Bertoni et al. 2018]. This approach can achieve the goal of a real "bedside" pathology, similar to the ongoing integration of this technique for assessment of non-dermatology specimens in surgical pathology [Panarello et al. 2020].

Artificial intelligence (AI) is currently transforming healthcare [Hinton 2018]. A popular AI technique for image classification is Convolutional Neural Networks (CNNs), a deep-learning approach inspired from the human brain. In CNNs, inputs such as images go through several layers of artificial "neurons" before an output is finally rendered such as the diagnoses of those images. CNN algorithms are being used in radiology and pathology [Topol 2019] to classify images as neoplastic or non-neoplastic and have shown a proficiency at par or even exceeding human performance [Campanella et al. 2019]. Likewise, in dermatology, CNNs have reported comparable performance to an expert dermatologist in skin cancer diagnosis using clinical [Esteva et al. 2017; Fujisawa et al. 2019, Han et al. 2018] and dermoscopy images [Brinker et al. 2019; Codella et al. 2016; Haenssle et al. 2018; Haenssle et al. 2020]. Recently, CNN has also been applied successfully to reflectance confocal microscopy (RCM) images to classify skin lesions [Kose et al. 2020, Wodzkinski et al. 2019; Wodzkinski et al. 2020, Campanella et

al. 2021]. However, to the best of our knowledge, CNNs have not been developed and tested to diagnose skin cancers in digitally colored EVCM images.

The goal of this study was to develop and test the performance of CNN algorithms for detecting BCC in digitally colored EVCM images obtained from freshly excised tissues from Mohs surgery. For this study, 40 BCCs and 28 negative ("not-BCC") skin tissue samples were collected from 42 patients to create three different image datasets to train CNN models: 1) an EVCM image dataset with 663 images, 2) an H&E image dataset with 516 images, and 3) a combination of the two datasets (EVCM and H&E image datasets) with 1179 images. The performance of these 3 trained models were evaluated and compared on a separate test set (not used in training), which comprised of 97 EVCM images created using 7 BCCs and 4 negative ("not-BCC") skin tissue samples were from 11 patients.

RESULTS

Patient demographics and lesion characteristics:

A total of 53 patients were enrolled in the study. Mean age was 61 years (± 13, range 36-95 years); 64% (34/53) were males and 36% (19/53) females. Majority of the lesions, 63.3% (50/79) were located on the head and neck. A total of 47 BCCs were imaged including 18 nodular BCCs (nBCCs), 11 superficial BCCs (sBCCs), 10 infiltrative BCCs (iBCCs), 8 mix-subtype BCCs and 32 skin samples did not present BCC. Patient demographics and lesion characteristics are detailed in the Supplementary Table S1.

Model performance:

The main outcome measures were sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) (Table 1 and Supplementary Figure S1).

The model trained with the EVCM dataset (EVCM model) achieved 92% diagnostic accuracy, similar to the H&E model (93%). Compared to the H&E model, the EVCM model had a higher sensitivity (96% vs. 93%) but lower specificity (89% vs. 92%). The combined model had the lowest diagnostic accuracy (86%) with a high specificity (92%; similar to H&E model) but the lowest sensitivity (78%). Area under the curve (AUC) of receiver operating characteristics (ROC) for diagnostic dichotomous classification was calculated for each the three training datasets (Figure 1). The AUC was 0.94, 0.95, and 0.94 for EVCM, H&E and combination trained models, respectively.

Gradient maps:

Gradient maps created with Grad-CAM [Selvaraju RR, 2016] highlighted important regions in red color in the images for predicting the presence/absence of tumor after all the three trainings. Figure 2 shows gradient map examples of a true positive (TP), a true negative (TN), a false positive (FP), and a false negative (FN) tissue sample. For the TP example, the algorithm identified even small BCC nodules (asterisks) as important areas for BCC prediction, while not taking the hair follicle (arrow) as an important region. Likewise, for the TN example, sebaceous gland was correctly identified as an important region within the image for the negative prediction for tumor. We had a very few images with false positive and false negative results. For FP example, sebaceous glands (arrowhead) and eccrine ducts (star) were detected as important for BCC detection. On the other hand, a FN example where BCC nodules (asterisks) were not considered important and the algorithm prediction was no-BCC.

DISCUSSION

BCC is the most common skin cancer worldwide [Leiter et al. 2014]. Although, BCC has a low metastatic potential it can be locally invasive causing extensive tissue damage and loss of

regional function [Nehal and Bichakjian 2018]. MMS is a specialized surgical procedure capable of achieving complete clearance of BCC, while maximizing normal tissue preservation, making it a preferred treatment for recurrent BCCs and BCCs located on cosmetically sensitive and functionally challenging sites, such as the face [Jain et al. 2017]. However, MMS is not only a time-consuming surgery (due to frozen section analysis) but it is also an expensive procedure that requires an extensive laboratory set-up and specialized surgeons and technicians. *Ex vivo* confocal microscopy (EVCM) is an emerging imaging technique that generates digitally colored purple and pink images similar to H&E, without any time-consuming tissue processing [Mu et al. 2016; Schüürmann et al. 2019]. It can be used during Mohs surgery (MMS) for rapid detection of residual basal cell carcinoma (BCC); however, reading EVCM images requires specialized training. An automated approach to BCC detection in EVCM images can aid in diagnosis.

Currently, AI is being implemented extensively in the field of dermatology and pathology for the automated diagnosis of skin cancers and non-neoplastic lesions (psoriasis, atopic dermatitis and onychomycosis) [Han et al. 2018] in clinical and dermoscopy images using CNN, a deep-learning algorithm [Esteva et al. 2017, Schüürmann et al. 2019]. In our study we used CNN for detection of BCC in EVCM images. CNN was first used in skin cancer detection by Nasr-Esfahani et al. (2016) for the diagnosis of melanoma. They trained the algorithm with a small dataset of 170 clinical images from melanocytic lesions. Similar to our study, the authors used augmentation methods such as random rotation, and resizing of the images to increase the number of images (from 170 original images to 6120 images) in the training dataset, which yielded a sensitivity of 81% and a specificity of 80% in the diagnosis of melanoma.

CNNs has been applied to non-invasive *in vivo* imaging technique such as reflectance confocal microscopy (RCM). Wodzinski et al. (2019) reported an accuracy of 91% in the diagnosis of BCC in RCM images. Recently, Campanella et al. (2021) developed a deep learning model to

automatically detect BCC in RCM images acquired from lesions clinically equivocal for BCC and compared the results with the RCM expert readers. The proposed model achieved an area under the curve (AUC) for the receiver operator characteristic (ROC) curve of 89.7%, which was on par with the expert readers. We achieved similar results with 92% diagnostic accuracy for detection of BCC in EVCM images using EVCM training dataset. The use of H&E images to train deep-learning algorithms has been extensively used in pathology. Towards this end, Campanella et al. (2019) reported an accuracy above 98% in the diagnosis of BCC in conventional H&E stained images. On the contrary, we achieved a diagnostic accuracy of 93% with H&E-trained model for the detection of BCC in EVCM images. This difference in diagnostic accuracy could be attributed to the differences in the type of images used for the training and testing datasets in our study i.e. H&E images trained model was tested on EVCM image dataset compared to the use of only H&E images in the training and testing datasets in their study. Another reason could be the use of a relatively larger number of images (9,962) used to train their model.

Our study demonstrates CNN's high-level performance in classifying BCC in EVCM images. Even with the use of freshly discarded tumor margins in this study, which typically has less tumor burden than the central tumor de-bulk tissue, we achieved a high sensitivity and specificity in the diagnosis of BCC in EVCM images, which is at par with dermatologists' reported level in the literature [Gareau et al. 2009; Karen et al. 2009; Mu et al. 2016]. The highest sensitivity value of 96% was obtained when the CNN was trained with EVCM images, whereas the specificity decreased to 89%. Our results showed the best accuracy with H&E-trained model, which could be attributed to the better and sharper image quality of H&E images compared to digitally-colored EVCM images. Because we demonstrated that H&E images can successfully train CNNs to diagnose BCC in EVCM images, one can collect a large dataset comprising of only H&E images to train such a model (as H&E images are more readily

available than EVCM images). Due to large number of images used in the H&E training dataset, less time-consuming weakly supervised CNN training could be used [Campanella et al. 2019] instead of the completely supervised training used in our study.

Although the combined (H&E and EVCM training dataset) model improved the specificity of BCC detection in EVCM images, it had the lowest diagnostic accuracy and sensitivity compared to the other two models (EVCM and H&E models). It is possible that the model trained with the combined dataset may not have had sufficient training time (i.e., number of epochs), which can be explored in future work.

Furthermore, the gradient map created in this study could be combined with the model prediction to aid surgeons in the real-time diagnosis and also as a teaching-training tool for novices [Campanella et al. 2021].

Our study had some limitations. First, while our dataset covered all the common subtypes of BCCs, it had an overall small size samples of each subtype, which could not represent all the morphological appearance of BCCs encountered in clinical practice. Also, this study did not include analysis on pigmented BCCs, which may be important to be tested in future studies. Secondly, the algorithm has yet to be tested in actual clinical practice. Also, for imaging with the EVCM device, we used freshly discarded surgical margins from Mohs surgery, which often has less tumor burden than the central de-bulk tissue. Even with the smaller tumor burden in these samples our algorithm achieved high diagnostic values; thus, we anticipate better results using de-bulk specimens with higher tumor burden. Lastly, all the images were acquired at a single institution, which does not account for the variability in staining protocols and tissue processing. Thus, our results should be further validated on EVCM images acquired from various centers (multi-centric study).

In conclusion, we present the results of a deep-learning algorithm in classifying BCC in EVCM images. The various models developed could diagnose BCC in digitally colored purple and

pink EVCM images which was at par with reported dermatologists' accuracy in the literature. Furthermore, we found that a model trained with only H&E images (which are more available than EVCM images) can accurately detect BCC in EVCM images. Training deep-learning technology with H&E images to diagnose EVCM images expands the possibility our approach to be generalized to diagnose a variety of skin lesions (neoplastic and non-neoplastic) in excised tissues. Ultimately, deep-learning models could be integrated in existing EVCM devices to aid Mohs surgeons in identifying BCCs automatically. Prospective and larger scale studies are needed to validate this technology in real clinical practice.

MATERIAL AND METHODS:

Tissue sample collection and image acquisition:

All the tissues used for creating the training and test datasets were collected at Memorial Sloan Kettering Cancer Center (NY, USA) under institutional review protocols (IRBs #17-078 and # 08-066) approved by Memorial Sloan Kettering Cancer Center Ethics Committee and after written, informed consent from patients.

a) Tissue sample collection and image acquisition for EVCM images:

Freshly discarded whole tissues samples (excised *en face* tumor margins) were collected consecutively from BCC cases undergoing Mohs surgery under a prospective IRB protocol (# 08-066). These samples were dipped in 0.6 mM acridine orange dye (a fluorescent dye) for 20 seconds and immediately placed on a commercial EVCM device (Vivascope 2500; Caliber ID, Rochester, NY, USA) for imaging [Jain et al. 2017]. The tissue was similarly oriented as for the frozen sectioning, which enabled an *en face* view of the entire tissue. Digitally-colored purple and pink images were acquired by combining signals from the fluorescent and reflectance channels, respectively (Supplementary Figure S2). First, we acquired an overview image, covering the entire tissue (measuring up to 2 cm in maximum diameter) section. Then

we acquired multiple smaller sized zoomed-in images (within the overview image) of BCC tumors and normal surrounding skin structures such as sebaceous glands, hair follicles, epidermis, and eccrine ducts.

b) Tissue sample collection and image acquisition for H&E images:

Under another IRB approved protocol (#17-078), a Dataline search was performed to identify lesions with a histopathology confirmed diagnosis of BCC. We retrieved routine histopathology H&E stained slides from these lesions from the Department of Dermatopathology. The glass slides were then digitized using Aperio AT2 slide Scanner (Leica Biosystems, Nussloch, Germany) in the Dermatology research lab. Similar to EVCM image acquisition, we acquired multiple H&E images at various magnifications (2x, 8x,10x) from BCC tumors and normal surrounding skin structures (sebaceous glands, hair follicles, epidermis, and eccrine ducts). On average, we acquired 24 images (ranging from 1 to 49 images) of varied sizes measuring ~1200x600 pixels to ~12000x12000 pixels per tissue sample using EVCM and H&E images.

Image Labelling:

The H&E and EVCM images were analyzed for the presence or absence of BCC by a dermatologist (MSM) and a pathologist specialized in optical imaging techniques (MJ). Each image was labelled as "BCC" and "not-BCC" and used to create training and test datasets (see below).

Dataset Creation:

Training Datasets:

1. **EVCM training dataset:** A total of 663 digitally stained purple and pink EVCM images (190 "BCC" images and 473 "not-BCC" images) were obtained from 14 fresh

BCC tissues (5 nodular BCC, 4 infiltrative BCC, 3 superficial BCC, and 2 infiltrative-nodular BCCs) and 15 negative ("not-BCC") controls normal skin tissues.

2. **H&E training dataset**: A total of 516 H&E images (170 "BCC" images and 346 "not-BCC") were obtained from 26 H&E stained slides of BCC (11 nodular BCC, 4 infiltrative BCC, 6 superficial BCC, 3 infiltrative-nodular and 2 superficial-nodular BCC), and 13 negative ("not-BCC") controls normal skin tissues.
3. **Combined EVCM and H&E training dataset:** A total of 1,179 images, which was created by combining all 516 images from the EVCM database and all 663 images from the H&E dataset.

Test Dataset: The CNN models built using the above three training datasets were tested on a new set of 107 EVCM images (45 BCC and 62 normal images) obtained from 7 BCCs and 4 normal skin tissue samples that were not previously shown to the algorithm. Images artifacts were introduced for both training and testing sets to simulate real conditions. Composition of the training and test datasets are detailed in Table 2. Example images from the training and test datasets can be seen in Figure 3.

CNN Architecture:

In this study, we trained and evaluated ResNet50 (Residual Neural Network) [He et al. 2016], a type of CNN, using both EVCM and H&E images. Unlike a standard CNN architecture, a ResNet architecture can handle greater number of hidden layers (higher model complexity), allowing for the extraction of more complex patterns and features. We used 181 hidden layers in our ResNet. We integrated Transfer Learning in our CNN to improve the efficacy of feature extraction. Transfer Learning stores knowledge gained from another problem and applies it to a different problem [Pan and Yang, 2010]. In this case, we used a CNN model pre-trained with the popular ImageNet dataset (which includes images from a large number of categories,

including animal, plant, and objects but not medical images) and applied its knowledge to train our BCC-detection models.

The distribution of images labelled as "BCC" and "not-BCC" were imbalanced because we used different number of images per class within each training dataset (Table 2). Images in each class were augmented, increasing the total number of images to 10000 per class. To perform data-augmentation, synthetic copies were created by applying image transformation methods such as random rotation, shifting, and resizing [Wodzinski et al. 2020]. To speed up model training time and reduce model complexity, all input images were resized to 300x300 pixels [Wodzinski et al. 2019]. A total of 20 epochs (number of complete passes through a training dataset) were used to train each of the models. A schematic of the methodology is shown in Figure 4.

The dataset was divided in three different subsets (training, validation and testing) in order to avoid the overfitting. The training and validation subsets were used to train the CNN and the testing subset was only used to see how well the model performs on unseen data. Besides, we also used Dropout Regularization [Liang et al. 2021] to reduce overfitting. Dropout works by probabilistically removing, or "dropping out," inputs to a layer, which may be input variables in the data sample or activations from a previous layer. It has the effect of simulating a large number of networks with very different network structure and, in turn, making nodes in the network generally more robust to the inputs.

Gradient maps:

Gradient maps were created with Grad-CAM [Selvaraju RR et al. 2016]. Grad-CAM is a technique for visual explanation of CNNs that highlights the regions of the input that are "important" for the prediction in a particular CNN model. Grad-CAM determines the weight for each of CNN feature maps to compute the weighted sum of the activations and then up

sampling the result to the image size to plot the original image with the heatmap, highlighting the important regions (red color) for the model prediction. A subset of gradient maps is shown in Figure 2.

Statistical analysis:

All statistical analyses was performed in R (v4.03) [R Core Team, 2020]. The ability to discriminate the classes ("BCC" or "not-BCC") inferred by the model was used to generate ROC curves with 95% intervals, using the package "pROC" (version 1.17.0.1) [Robin et al. 2011]. ROC curves and 95% confidence intervals for sensitivity, specificity, positive predictive and negative predictive values, and accuracy measures were generated.

DATA AVAILABILITY

The deep learning model that was trained to generate the presented results and the code is available at https://github.com/boxyware/confocal-ex-vivo.

CONFLICT OF INTERESTS

The authors declare none.

ACKNOWLEDGMENTS

We would like to thank Dr. Kivanc. Kose, PhD at Memorial Sloan Kettering Center for his critical review of the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization: MSM, MJ. Data curation: MSM, MLC, MM, UH, MJ. Resources: AR, EL,

CJC, NK. Software: MLC. Formal Analysis: MLC. Results analysis: MSM, MJ, MLC, JJPR. Writing – original draft: MSM. Writing – review & editing: MSM, MLC, JCS, JPR, MJ.

ORCID IDs

Mercedes Sendín-Martín[1] (ORCID ID# 0000-0003-0156-381X), Manuel Lara-Caro[1] (ORCID ID# 0000-0001-7424-2000), Ucalene Harris[2] (ORCID ID# 0000-0002-7845-7460), Matthew Moronta[2] (ORCID ID# 0000-0002-7845-7460), Anthony Rossi[2] (ORCID ID# 0000-0003-2295-1934), Erica Lee[2] (ORCID ID# 0000-0001-9670-3242), Chih-Shan Jason Chen[2] (ORCID ID# orcid.org/0000-0002-0977-2707), Kishwer Nehal[2] (ORCID ID# 0000-0001-5791-7184), Julián Conejo-Mir Sánchez[1,3] (ORCID ID# 0000-0001-9108-9538), José-Juan Pereyra-Rodríguez*[1,4] (ORCID ID# 0000-0001-6843-5877), Manu Jain*[2,4] (ORCID ID# 0000-0002-6035-0825).

REFERENCES

Bağcı IS, Aoki R, Krammer S, Ruzicka T, Sárdy M, French LE, Hartmann D. Ex vivo confocal laser scanning microscopy for bullous pemphigoid diagnostics: new era in direct immunofluorescence? J Eur Acad Dermatol Venereol. 2019 Nov;33(11):2123-2130. doi: 10.1111/jdv.15767.

Bağcı IS, Aoki R, Vladimirova G, Ergün E, Ruzicka T, Sárdy M, et al. New-generation diagnostics in inflammatory skin diseases: Immunofluorescence and histopathological assessment using ex vivo confocal laser scanning microscopy in cutaneous lupus erythematosus. Exp Dermatol. 2021 May;30(5):684-690. doi: 10.1111/exd.14265.

Bennàssar A, Vilalta A, Carrera C, Puig S, Malvehy J. Rapid diagnosis of two facial papules using ex vivo fluorescence confocal microscopy: toward a rapid bedside pathology. Dermatol Surg. 2012 Sep;38(9):1548-51. doi: 10.1111/j.1524-4725.2012.02467.x.

Bennàssar A, Carrera C, Puig S, Vilalta A, Malvehy J. Fast evaluation of 69 basal cell carcinomas with ex vivo fluorescence confocal microscopy: criteria description, histopathological correlation, and interobserver agreement. JAMA Dermatol. 2013 Jul;149(7):839-47. doi: 10.1001/jamadermatol.2013.459.

Bertoni L, Azzoni P, Reggiani C, Pisciotta A, Carnevale G, Chester J, et al. Ex vivo fluorescence confocal microscopy for intraoperative, real-time diagnosis of cutaneous inflammatory diseases: A preliminary study. Exp Dermatol 2018;27:1152–9. https://doi.org/10.1111/exd.13754.

Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019;111:30–7. https://doi.org/10.1016/j.ejca.2018.12.016.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9. https://doi.org/10.1038/s41591-019-0508-1.

Campanella G, Navarrete-Dechent C, Liopyris K, Monnier J, Aleissa S, Minas B, et al. Deep Learning for Basal Cell Carcinoma Detection for Reflectance Confocal Microscopy. J Invest Dermatol. 2021 Jul 12:S0022-202X(21)01437-8. doi: 10.1016/j.jid.2021.06.015. Epub ahead of print. PMID: 34265329.

Codella N, Nguyen Q-B, Pankanti S, Gutman D, Helba B, Halpern A, et al. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. IBM J Res Dev 2016;61.

Dalimier E, Salomon D. Full-field optical coherence tomography: A new technology for 3D high-resolution skin imaging. Dermatology 2012;224:84–92. https://doi.org/10.1159/000337423.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8. https://doi.org/10.1038/nature21056.

Debarbieux S, Gaspar R, Depaepe L, Dalle S, Balme B, Thomas L. Intraoperative diagnosis of nonpigmented nail tumours with ex vivo fluorescence confocal microscopy: 10 cases. Br J Dermatol. 2015 Apr;172(4):1037-44. doi: 10.1111/bjd.13384.

Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019;180:373–81. https://doi.org/10.1111/bjd.16924.

Gareau DS, Karen JK, Dusza SW, Tudisco M, Nehal KS, Rajadhyaksha M. Sensitivity and specificity for detecting basal cell carcinomas in Mohs excisions with confocal fluorescence mosaicing microscopy. J Biomed Opt 2009;14:034012. https://doi.org/10.1117/1.3130331.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836–42. https://doi.org/10.1093/annonc/mdy166.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836–42. https://doi.org/10.1093/annonc/mdy166.

Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less

456 artificial conditions. Ann Oncol 2020;31:137–43.
457 https://doi.org/10.1016/j.annonc.2019.10.013.
458
459 Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images
460 for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. J Invest
461 Dermatol 2018;138:1529–38. https://doi.org/10.1016/j.jid.2018.01.028.
462
463 Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an
464 equivalent and often superior performance to dermatologists in onychomycosis diagnosis:
465 Automatic construction of onychomycosis datasets by region-based convolutional deep neural
466 network. PLoS One 2018;13. https://doi.org/10.1371/journal.pone.0191493.
467
468 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc. IEEE
469 Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016- December, IEEE Computer
470 Society; 2016, p. 770–8. https://doi.org/10.1109/CVPR.2016.90.
471
472 Hinton G. Deep learning-a technology with the potential to transform health care. JAMA - J
473 Am Med Assoc 2018;320:1101–2. https://doi.org/10.1001/jama.2018.11100.
474
475 Jain M, Rajadhyaksha M, Nehal K. Implementation of fluorescence confocal mosaicking
476 microscopy by “early adopter” Mohs surgeons and dermatologists: recent progress. J Biomed
477 Opt 2017;22:24002. https://doi.org/10.1117/1.JBO.22.2.024002.
478

Karen JK, Gareau DS, Dusza SW, Tudisco M, Rajadhyaksha M, Nehal KS. Detection of basal cell carcinomas in Mohs excisions with fluorescence confocal mosaicing microscopy. Br J Dermatol 2009;160:1242–50. https://doi.org/10.1111/j.1365-2133.2009.09141.x.

Keena S, Que T. Research Techniques Made Simple: Noninvasive Imaging Technologies for the Delineation of Basal Cell Carcinomas. J Invest Dermatol 2016;136:e33–8. https://doi.org/10.1016/j.jid.2016.02.012.

Kose K, Bozkurt A, Alessi-Fox C, Brooks DH, Dy JG, Rajadhyaksha M, et al. Utilizing Machine Learning for Image Quality Assessment for Reflectance Confocal Microscopy. J Invest Dermatol 2020;140:1214–22. https://doi.org/10.1016/j.jid.2019.10.018.

Liang X, Wu L, Li J, Wang Y, Meng Q, Qin T, et al. R-Drop: Regularized Dropout for Neural Networks 2021. arXiv:2106.14448

Mu EW, Lewin JM, Stevenson ML, Meehan SA, Carucci JA, Gareau DS. Use of digitally stained multimodal confocal mosaic images to screen for nonmelanoma skin cancer. JAMA Dermatology 2016;152:1335–41. https://doi.org/10.1001/jamadermatol.2016.2997.

Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, et al. Melanoma detection by analysis of clinical images using convolutional neural network. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2016- October, Institute of Electrical and Electronics Engineers Inc.; 2016, p. 1373–6. https://doi.org/10.1109/EMBC.2016.7590963.

503 Nehal KS, Bichakjian CK. Update on keratinocyte carcinomas. N Engl J Med 2018;379:363–
504 74. https://doi.org/10.1056/NEJMra1708701.
505
506 Leiter U, Eigentler T, Garbe C. Epidemiology of Skin Cancer. Sunlight, Vitam. D Ski. Cancer,
507 vol. 810, New York, NY: Springer New York; 2014, p. 120–40. https://doi.org/10.1007/978-1-
508 4939-0437-2_7.
509
510 Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–
511 59. https://doi.org/10.1109/TKDE.2009.191.
512
513 Panarello D, Compérat E, Seyde O, Colau A, Terrone C, Guillonneau B. Atlas of Ex Vivo
514 Prostate Tissue and Cancer Images Using Confocal Laser Endomicroscopy: A Project for
515 Intraoperative Positive Surgical Margin Detection During Radical Prostatectomy. Eur Urol
516 Focus. 2020 Sep 15;6(5):941-958. doi: 10.1016/j.euf.2019.01.004
517
518 R Core Team (2020). A language and environment for statistical computing. R Foundation for
519 Statistical Computing, Vienna, Austria. n.d.
520
521 Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source
522 package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12.
523 https://doi.org/10.1186/1471-2105-12-77.
524
525 Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence estimate of nonmelanoma
526 skin cancer (keratinocyte carcinomas) in the us population, 2012. JAMA Dermatology
527 2015;151:1081–6. https://doi.org/10.1001/jamadermatol.2015.1187.

528 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual
529 Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis
530 2016;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.
531
532 Schüürmann* M, Stecher MM, Paasch U, Simon JC, Grunewald S. Evaluation of digital
533 staining for ex vivo confocal laser scanning microscopy. J Eur Acad Dermatology Venereol
534 2019:jdv.16085. https://doi.org/10.1111/jdv.16085.
535
536 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence.
537 Nat Med 2019;25:44–56. https://doi.org/10.1038/s41591-018-0300-7.
538
539 Van Loo E, Mosterd K, Krekels GAM, Roozeboom MH, Ostertag JU, Dirksen CD, et al.
540 Surgical excision versus Mohs' micrographic surgery for basal cell carcinoma of the face: A
541 randomised clinical trial with 10 year follow-up. Eur J Cancer 2014;50:3011–20.
542 https://doi.org/10.1016/j.ejca.2014.08.018
543
544 Wodzinski M, Skalski A, Witkowski A, Pellacani G, Ludzik J. Convolutional Neural Network
545 Approach to Classify Skin Lesions Using Reflectance Confocal Microscopy. Proc. Annu. Int.
546 Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2019, Institute of Electrical and Electronics
547 Engineers Inc.; 2019, p. 4754–7. https://doi.org/10.1109/EMBC.2019.8856731.
548
549 Wodzinski M, Pajak M, Skalski A, Witkowski A, Pellacani G, Ludzik J. Automatic Quality
550 Assessment of Reflectance Confocal Microscopy Mosaics using Attention-Based Deep Neural
551 Network. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2020- July, Institute

of Electrical and Electronics Engineers Inc.; 2020, p. 1824–7. https://doi.org/10.1109/EMBC44109.2020.9176557.

TABLES

Table 1. Comparison of the performance of trained models for diagnosis of BCC in the EVCM images

Training Dataset	**Results of BCC detection in EVCM images- Metrics (CI 95%)**				
	Sensitivity	**Specificity**	**PPV**	**NPV**	**Accuracy**
EVCM TRAINING	0.96 (0.88-1)	0.89 (0.80-0.97)	0.86 (0.75-0.97)	0.96 (0.91-1)	0.92 (0.86-0.98)
H&E TRAINING	0.93 (0.85-1)	0.92 (0.84-1)	0.89 (0.8-0.99)	0.95 (0.87-1)	**0.93** (0.87-0.98)
EVCM + H&E TRAINING	0.78 (0.65-0.91)	0.92 (0.84-1)	0.87 (0.76-0.99)	0.85 (0.76-0.94)	0.86 (0.79-0.93)

Abbreviations: CI= confidence interval, PPV= positive predictive value, NPV =negative predictive value. Bold fonts indicate highest accuracy results.

Table 2. Samples and images used for creating training and test datasets.

<table>
<tr><th>Training datasets</th><th colspan="2">Nº of Tissue Samples and their diagnosis</th><th>Total Nº of images/dataset</th><th>Nº of BCC positive and negative images</th></tr>
<tr><td rowspan="2">1. EVCM TRAINING DATASET</td><td>14 BCCs</td><td>5 nBCC (35.7%)
4 iBCC (28.6%)
3 sBCC (21.4%)
2 inBCC (14.3%)</td><td rowspan="2">663</td><td>“BCC”
190 (28,7%)</td></tr>
<tr><td colspan="2">15 normal (“not-BCC”) skin samples</td><td>“not-BCC”
473 (71,3%)</td></tr>
<tr><td rowspan="2">2. H&E TRAINING DATASET</td><td>26 BCCs</td><td>11 nBC (42.3%)
4 iBCC (15.4%)
6 sBCC (23.1%)
3 inBCC(11.5%)
2 snBCC (7.7%)</td><td rowspan="2">516</td><td>“BCC”
170 (32,9%)</td></tr>
<tr><td colspan="2">13 normal (“not-BCC”) skin samples</td><td>“not BCC”
346 (67,1%)</td></tr>
<tr><td>3. COMBINED EVCM + H&E TRAINING DATASET (Datasets 1 and 2)</td><td colspan="4">40 BCC and 28 normal (“not-BCC”) skin samples
1,179 images: 360 “BCC” images and 820 “not-BCC” images</td></tr>
<tr><td>Test-set</td><td colspan="2">Nº of Tissue Samples and their diagnosis</td><td>Nº of images</td><td>Nº of BCC positive and negative images</td></tr>
<tr><td rowspan="2">EVCM TEST SET</td><td>7 BCCs</td><td>2 sBCCs (28.6%)
2 nBCCs(28.6%)
2 iBCCs (28.6%)
1 snBCC(14.3%)</td><td rowspan="2">107</td><td>45 “BCC”
(44,4%)</td></tr>
<tr><td colspan="2">4 normal (“not-BCC”) skin samples</td><td>62 “not-BCC”
(53,6%)</td></tr>
</table>

Abbreviations: BCC= basal cell carcinoma, nBCC=nodular BCC, sBCC= superficial BCC, iBCC= infiltrative BCC, inBCC= infundibular BCC, snBCC= superficial and nodular BCC.

608 **FIGURES**

609 **Figure 1. Receiver Operating Characteristic (ROC) Curves obtained using**: (a) EVCM-
610 trained model, (b) H&E-trained model, and (c) combined (EVCM+H&E) trained model. The
611 95% CI bounds of the ROC curve were calculated via bootstrapping. The proposed algorithm
612 achieves an AUC of 94.4%, 95.7% and 93.8%, respectively. The shaded ellipse represents the
613 95% CI area for the estimate of the sensitivity and specificity of the algorithm calculated via
614 bootstrapping.

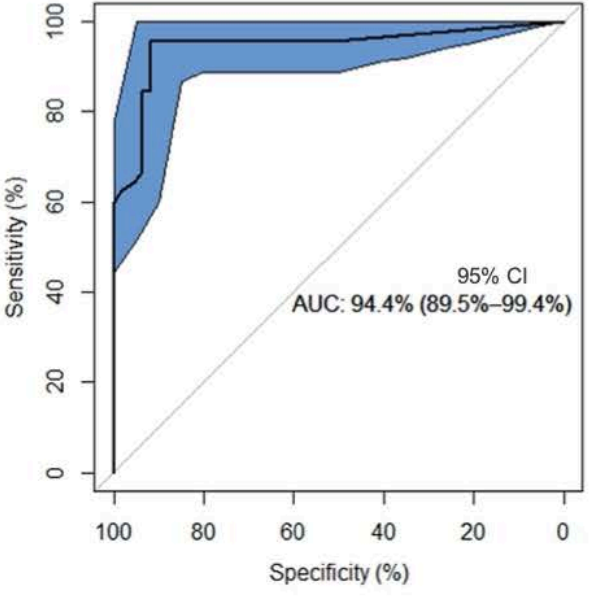
615 **Figure 2. Gradient Map: Left: EVCM images and Right: Gradient maps.** (a) True positive
616 (TP) example of a BCC with small tumor nodules (asterisks). High prediction attributes over
617 the tumor nodules in the Gradient map. Note, that a hair follicle (arrow) in the same field was
618 not considered "important" for the prediction by the algorithm. (b) True negative example of
619 an area with no-BCC, where a sebaceous gland (arrowhead) was identified "important" region
620 for negative prediction. (c) False positive (FP) example where eccrine glands (star) and
621 sebaceous gland (arrowhead) were detected as important for the prediction of tumor. (d) False
622 Negative (FN) example, where BCC tumor nodules (asterisks) were not considered important
623 by the algorithm for tumor detection. Note, this BCC tumor nodule has an extensive cystic
624 degeneration in the center (dashed arrow), which could have resulted in the false prediction in
625 this case. Color scalebar: Red color, high attribution and blue color low attribution for a given
626 prediction by the algorithm. Scale bars: a, b, c, d) 200 μm.

627 **Figure 3. Images used to create training and test datasets:** EVCM images (upper and lower
628 panel: Training and Test datasets): Purple and pink digitally colored EVCM images of: (a, i)
629 nodular BCCs, (b, j) superficial BCC, (c, k) infiltrative BCC, and (d, l) normal ("not-BCC")
630 skin tissue with epidermis (arrow) and pilosebaceous gland (arrowhead). Conventional H&E-
631 stained images (middle panel) of a: (e) a nodular BCC, (f) a superficial BCC, (g) an infiltrative
632 BCC, and (h) normal ("not-BCC") skin tissue with epidermis (arrow) and hair follicles

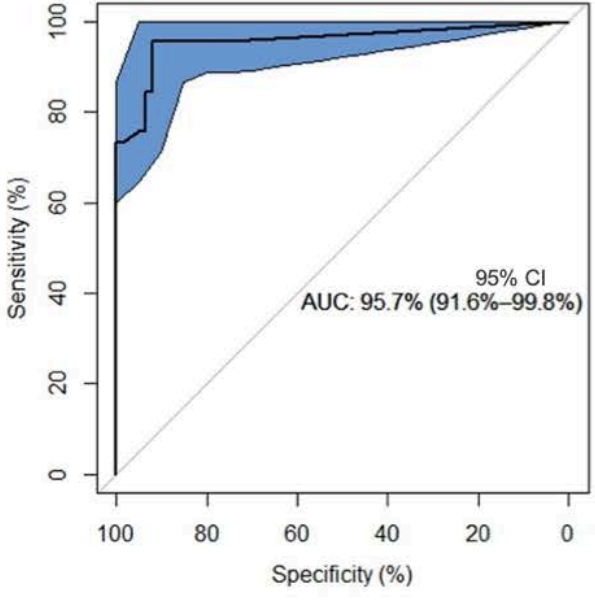
(arrowheads). BCCs are shown with an asterisk. Scale bars: a, b, c, e, i, j) 100 µm and d, f, g, h, k) 200 µm.

Figure 4. Experimental workflow: a) Tissue acquisition, imaging and datasets generation, and b) CNN model used in this study. A ResNet50 of 181 hidden layers and pretrained with ImageNet was used on "BCC" and "not-BCC" labelled images. Abbreviations: BCC= Basal Cell Carcinoma.

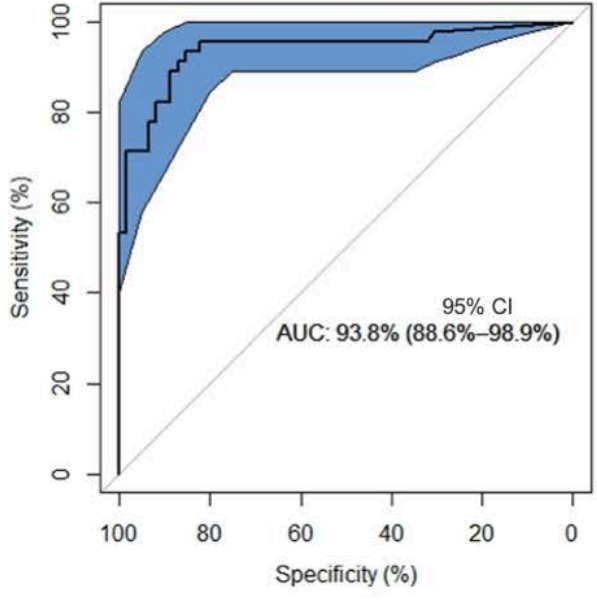
a. EVCM TRAINING

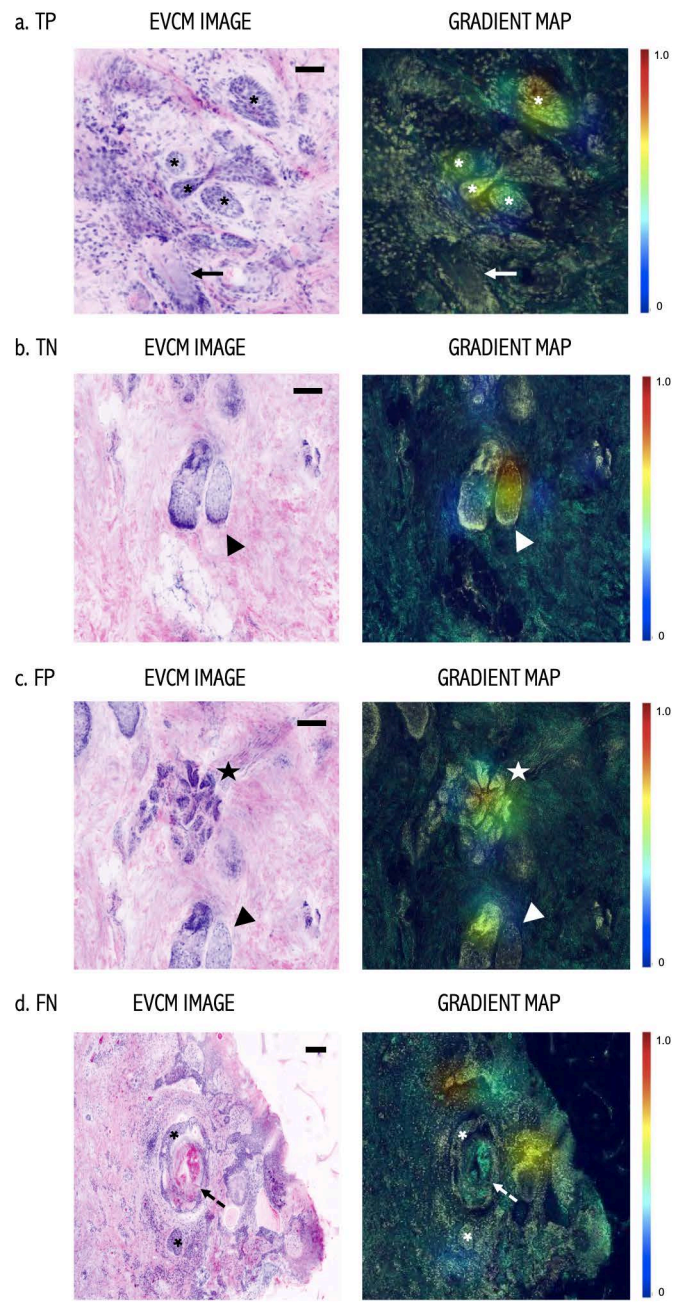


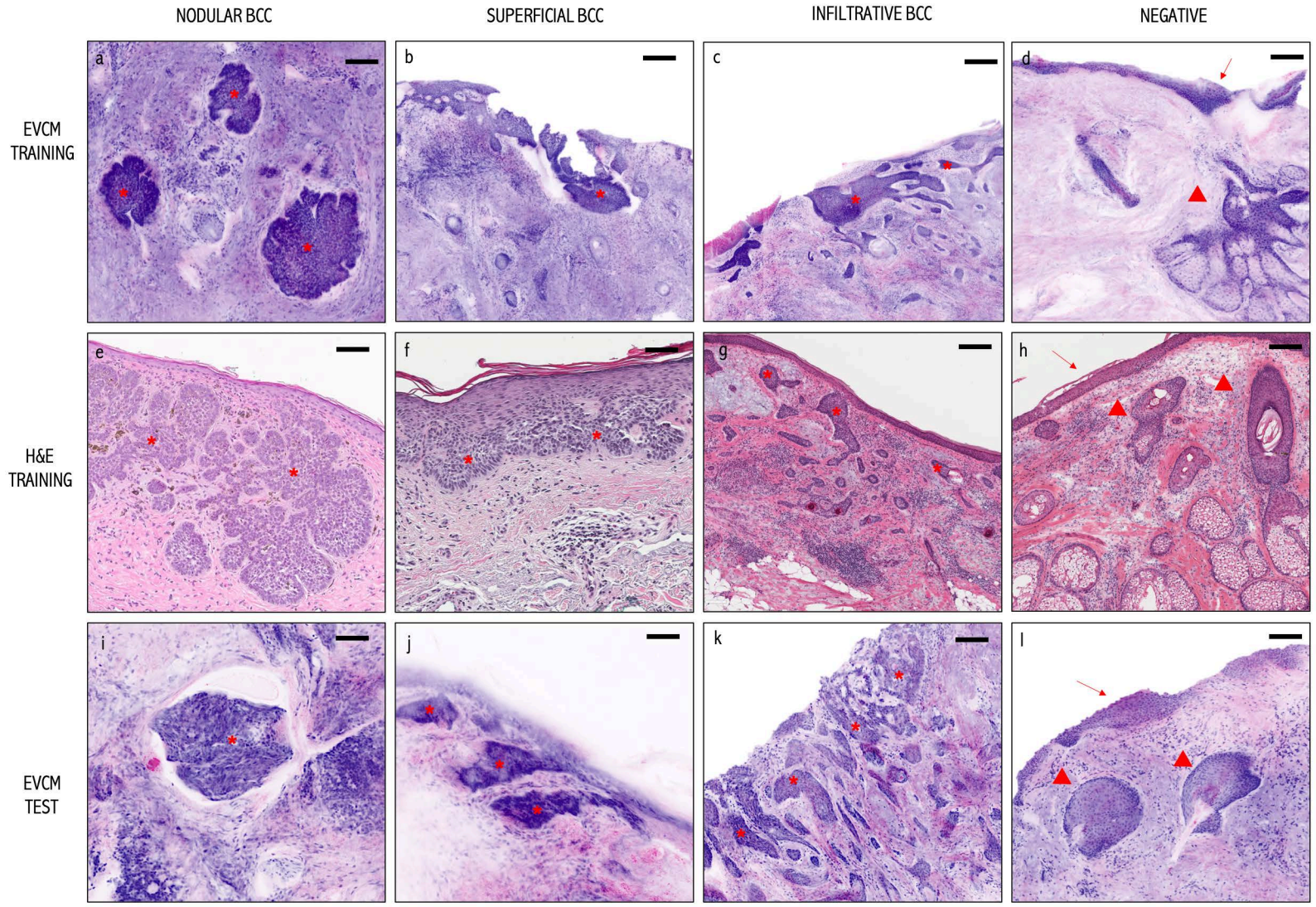
b. H&E TRAINING

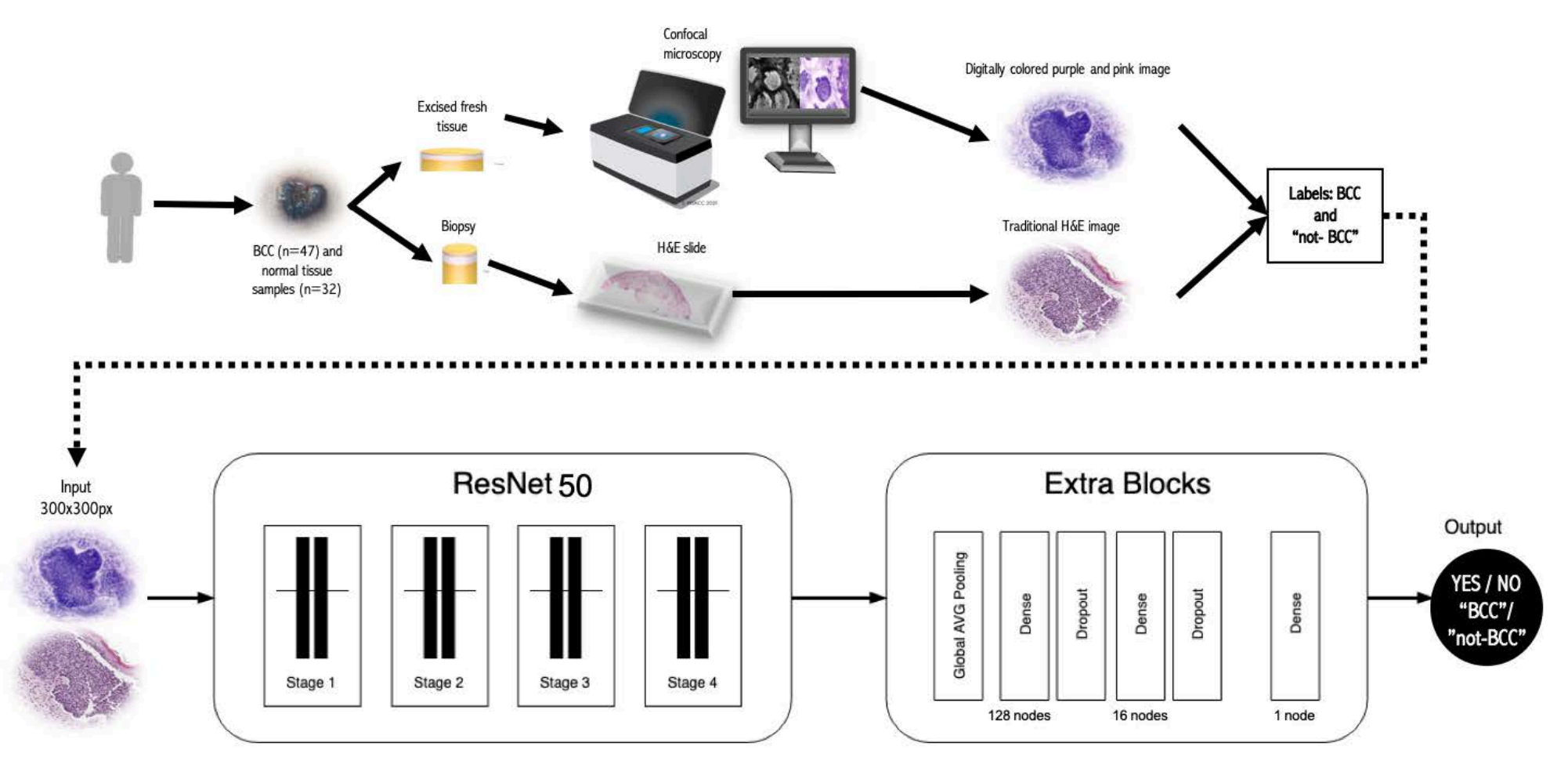


c. EVCM + H&E TRAINING




a. TP
EVCM IMAGE
GRADIENT MAP
1.0
0
b. TN
EVCM IMAGE
GRADIENT MAP
1.0
0
c. FP
EVCM IMAGE
GRADIENT MAP
1.0
0
d. FN
EVCM IMAGE
GRADIENT MAP
1.0
0

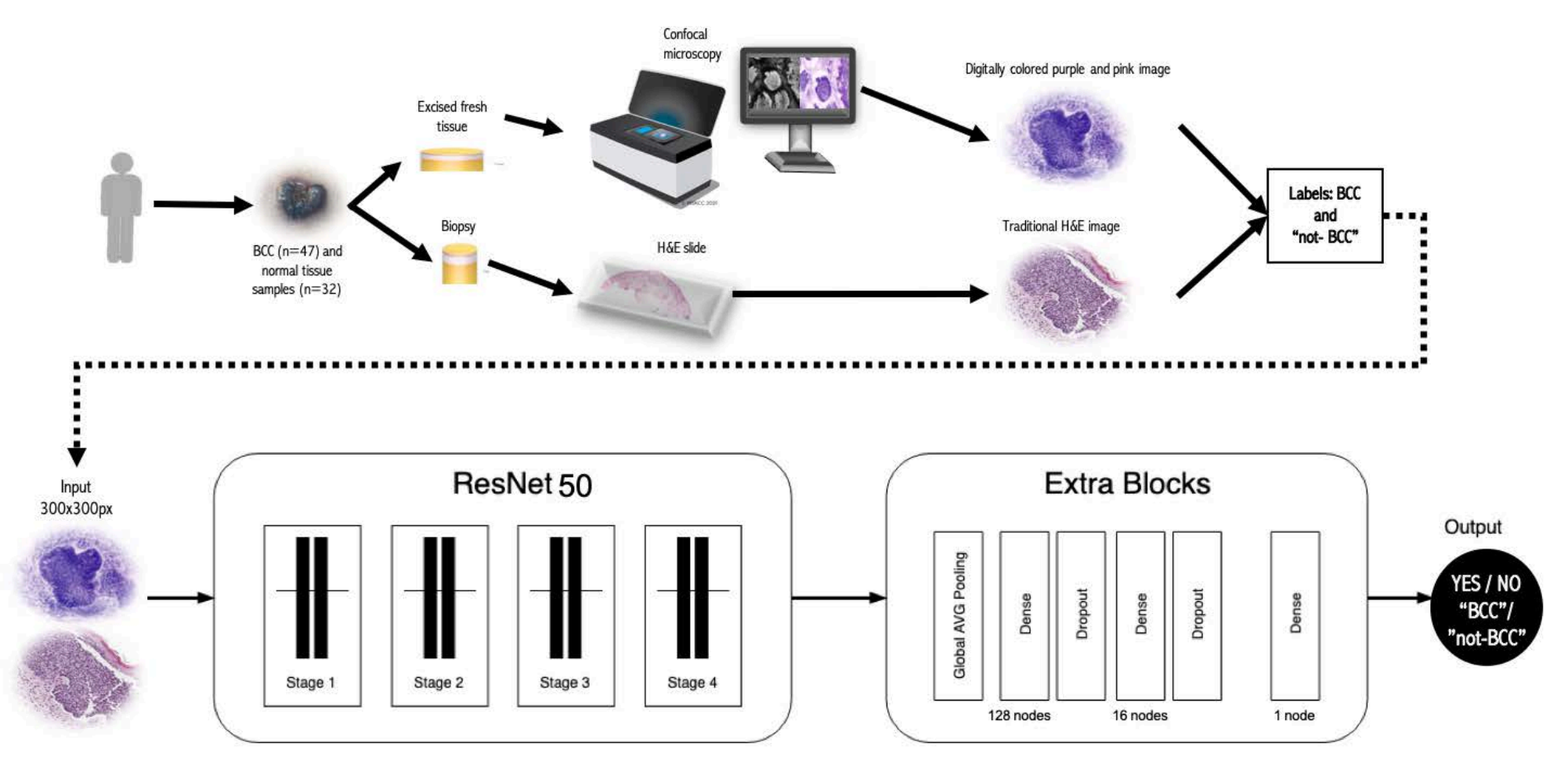

NODULAR BCC
SUPERFICIAL BCC
INFILTRATIVE BCC
NEGATIVE
EVCM TRAINING
H&E TRAINING
EVCM TEST
a
b
c
d
e
f
g
h
i
j
k
l


Confocal microscopy
Excised fresh tissue
Digitally colored purple and pink image
BCC (n=47) and normal tissue samples (n=32)
Biopsy
H&E slide
Traditional H&E image
Labels: BCC and "not- BCC"
Input 300x300px
ResNet 50
Stage 1
Stage 2
Stage 3
Stage 4
Extra Blocks
Global AVG Pooling
Dense
Dropout
Dense
Dropout
Dense
128 nodes
16 nodes
1 node
Output
YES / NO "BCC"/ "not-BCC"

Supplementary Table 1. Summary of patient´s demographic data and lesion location

Variable	Total (53 patients/79 lesions)
Age, mean (SD), y*	61 (13)
Sex, %(n)	
Male	64% (34/53)
Female	36% (19/53)
Diagnosis, n(%)	79 lesions
BCC	47 (59.5%)
Nodular	18 (22.8%)
Superficial	11 (13.9%)
Infiltrative	10 (12.6%)
Mix-subtype	8 (10.1%)
"Not-BCC"	32 (40.5%)
Location, n(%)	79 lesions
Head and neck	50 (63.3%)
Upper extremities	8 (10.1%)
Trunk	15 (19%)
Lower extremities	6 (7.6%)

Abbreviation: SD, standard deviation


Confocal microscopy
Excised fresh tissue
Digitally colored purple and pink image
BCC (n=47) and normal tissue samples (n=32)
Biopsy
H&E slide
Traditional H&E image
Labels: BCC and "not- BCC"
Input 300x300px
ResNet 50
Stage 1
Stage 2
Stage 3
Stage 4
Extra Blocks
Global AVG Pooling
Dense
Dropout
Dense
Dropout
Dense
128 nodes
16 nodes
1 node
Output
YES / NO "BCC"/ "not-BCC"

SUPPLEMENTARY MATERIAL

Supplementary Figure 1. Algorithm´s performance in each separate category with three different trainings.

Supplementary Figure 2. EVCM image acquisition: a) Purple and pink digitally colored EVCM image of a nodular BCC (yellow boxed area and arrows) acquired from a freshly excised whole-tissue obtained from Mohs surgery. EVCM images tissues simultaneously in fluorescence and reflectance mode, where fluorescence signal from nucleus is converted to purple color (b) and reflectance signal from cytoplasm and collagen is converted to pink color (c) to generate a combined purple and pink image (d), resembling conventionally H&E- stained image. Scale bars: a) 400 μm and b, c, d) 200 μm.

Supplementary Table 1. Summary of patient´s demographic data and lesion location

Supplementary Video 1. Steps involved in EVCM image acquisition.


a. EVCM TRAINING
b. H&E TRAINING
c. EVCM + H&E TRAINING
No
Yes
true label
55
7
2
43
57
5
3
42
57
5
10
35
No
Yes
predicted label